ANDERSON CRISTIANO NEISSE

SÍNDROME DA FADIGA CRÔNICA E ABSENTEÍSMO: ESTUDO DE TRABALHADORES EM TURNOS COMPARANDO *STEPWISE* E *ELASTIC-NET*

Dissertação apresentada à Universidade Federal de Viçosa como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para a obtenção do título de *Magister Scientiae*.

Orientador: Fernando Luiz P. de Oliveira

VIÇOSA - MINAS GERAIS
2020

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
N416s
2020

Neisse, Anderson Cristiano, 1993-
Síndrome da fadiga crônica e absenteísmo : estudo de trabalhadores em turnos comparando *stepwise* e *elastic-net* / Anderson Cristiano Neisse. – Viçosa, MG, 2020.
56 f. : il. (algumas color.) ; 29 cm.

Inclui anexo.
Orientador: Fernando Luiz Pereira de Oliveira.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Síndrome da Fadiga Crônica. 2. Biometria. 3. Análise de regressão logística. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.536

ANDERSON CRISTIANO NEISSE

SÍNDROME DA FADIGA CRÔNICA E ABSENTEÍSMO: ESTUDO DE TRABALHADORES EM TURNOS COMPARANDO *STEPWISE* E *ELASTIC-NET*

Dissertação apresentada à Universidade Federal de Viçosa como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para a obtenção do título de *Magister Scientiae*.

APROVADA: 28 de fevereiro de 2020.

Assentimento:

Anderson Cristiano Neisse

(Autor)

Fernando Luiz Pereira de Oliveira

(Orientador)

Dedico este trabalho à minha namorada, Caren Batista Alves, e à minha família.

Agradecimentos

Agradeço aos meus pais, Pedro e Luzenha, e meus irmãos, Mirtes, Mirna, Mirian e Alexandre pelo amor, carinho e por toda a dedicação e apoio.

À minha irmã, Mirtes e meu cunhado Duljon por todo o incentivo e apoio incondicionais sem os quais eu não teria trilhado este caminho do qual me orgulho.

À minha namorada, Caren, pela paciência, dedicação, amor, apoio nos momentos difíceis e fé em mim.

À Universidade Federal de Viçosa, em especial aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria, pelos momentos de descontração e em especial pelos ensinamentos que contribuíram para a minha formação acadêmica.

Ao secretário da pós-graduação Júnior Pires, pelo empenho e dedicação em nos atender sempre prontamente e com pró-atividade.

Aos professores Fernando Luiz Pereira de Oliveira, Anderson Castro Soares de Oliveira e Frederico Rodrigues Borges da Cruz, pela parceria e confiança durante o mestrado. Agradeço ainda pela orientação, atenção, momentos de descontração e especialmente por todo o aprendizado.

Aos professores Frederico Rodrigues Borges da Cruz, Graziela Dutra Rocha Gouvêa e Raimundo Marques do Nascimento Neto, por terem aceitado o convite para participar da banca.

À CAPES pelo apoio financeiro para o desenvolvimento deste trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"A dúvida é o princípio da sabedoria."

Aristóteles

Resumo

NEISSE, Anderson Cristiano, M.Sc., Universidade Federal de Viçosa, fevereiro de 2020. **Síndrome da Fadiga Crônica e Absenteísmo: Estudo de trabalhadores em turnos comparando *Stepwise* e *Elastic-Net***. Orientador: Fernando Luiz Pereira de Oliveira.

Caracterizada por fadiga persistente, dor muscular, dificuldades cognitivas e de sono, a Sídrome da Fadiga Crônica (CFS) tem se tornado comum nas práticas clínicas nas últimas décadas desde sua recente definição, em 1988. Estudos resultantes da contínua busca por fatores relacionados à CFS citam, dentre outros: sono irregular/insatisfatório, estresse psicológico, disfunção hormonal, deficiência de nutrientes, disfunção imunológica e infecções. Em condições de trabalho de risco o desenvolvimento da CFS pode aumentar a chance de acidentes fatais, tal como o trabalho em turnos na área de mineração que naturalmente já possui fatores evidentemente relacionados à CFS. Estudos indicam que indivíduos com má qualidade de sono e ciclos circadianos irregulares têm risco elevado de CFS, neuroticismo e absenteísmo. Uma vez que modelagem preditiva pode se mostrar efetiva tanto na prevenção da fadiga quanto na detecção de fatores, este estudo tem o objetivo de utilizar de regressão logística ajustada por meio de dois métodos de seleção/regularização (*Stepwise* e *Elastic-Net*) para procurar modelo que descreva a relação entre variáveis bioquímicas e antropométricas com o absenteísmo. Desta forma, por meio do absenteísmo e utilizando de efeitos encontrados na bibliografia, o objetivo é procurar evidência de relação entre a CFS e absenteísmo. Os resultados obtidos mostram indícios de relação do colesterol total, HDL, LDL e Triglicerídeos com o risco de absenteísmo, relação também presente para as variáveis de sódio e potássio. Com exceção ao potássio, todas as variáveis também possuem relação similar com a CFS, de acordo com a literatura.

PALAVRAS-CHAVE:

Síndrome da Fadiga Crônica. Biometria. Regressão Logística. Elastic-Net.

Abstract

NEISSE, Anderson Cristiano, M.Sc., Universidade Federal de Viçosa, February 2020. **Chronic Fatigue Syndrome and Absenteeism: A study of shift workers comparing Stepwise and Elastic-Net**. Adviser: Fernando Luiz Pereira de Oliveira.

Characterized by persistent fatigue, muscle pain, cognitive impairment and sleep difficulties, Chronic Fatigue Syndrome (CFS) has become common in clinical practices in recent decades due to its recent definition, in 1988. Studies resulting from the continuous search for related factors to CFS mention, among others: irregular/unsatisfactory sleep, psychological stress, hormonal dysfunction, nutrient deficiency, immunological dysfunction and infections. In risky working conditions, the development of CFS might increase the chance of fatal accidents, such as shift work in mines, which presents naturally some factors that are related to CFS. Studies indicate that individuals with poor sleep quality and irregular circadian cycles are at high risk for CFS, neuroticism and absenteeism. Since predictive modeling can be effective both in preventing fatigue and detecting its factors, this study aims to use logistic regression using two selection/regularization methods (Stepwise and Elastic-Net) to look for a model that describes the relationship between biochemical and anthropometric variables with absenteeism. Thus, through absenteeism and using effects found in the bibliography, the objective is to look for evidence of a relationship between CFS and absenteeism. Results show evidence of relationship between total cholesterol, HDL, LDL and triglycerides with the risk of absenteeism, relation which is also present for the sodium and potassium variables. With the exception of potassium, all variables had similar relationship with CFS, according to the literature.

KEYWORDS:

Chronic Fatigue Syndrome. Biometrics. Logistic Regression. Elastic-Net.

Lista de figuras

Lista de tabelas

Sumário

Parte I

Revisão bibliográfica e metodologia

1 Introdução

A Síndrome da Fadiga Crônica (CFS), ou Encefalomielite miálgica, é uma síndrome que têm sido comum nas práticas clínicas das últimas décadas (Afari e Buchwald, 2003). Estudo epidemiológico conduzido por Jason et al. (1999) estimou a prevalência de CFS de 410 por 100.000 indivíduos, o que sugere uma prevalência de 1 milhão de indivíduos somente nos Estados Unidos. Outro estudo de Jason e Njoku (2005) estimou que por volta de 850 mil americanos sofram da síndrome. A CFS foi definida pela primeira vez em 1988 pelo *United States Center for Disease Control and Prevention (CDC)* de forma a padronizar sua caracterização para estudos epidemiológicos (Holmes et al, 1988). Houveram diferentes revisões da definição de caso da CFS desde então (Williams et al, 2014). De acordo com Fukuda et al. (1994) um caso de CFS deve apresentar fadiga crônica persistente ou relapsante por mais de 6 meses acompanhada da ocorrência de 4 ou mais dos sintomas: Problemas de concentração ou memória, dor de garganta, gânglios linfáticos sensíveis, dor muscular, dor articular, dor de cabeça, sono inefetivo e mal-estar pós-esforço. Apesar de seu uso pela maioria da comunidade acadêmica, há críticas ao fato de o critério não demandar presença de sintomas de importância no diagnóstico como o mal-estar pós-esforço, requerindo quaisquer 4 dos 8 sintomas listados (Williams et al, 2014). A definição de caso mais utilizada atualmente, proposta pelo *United States Center for Disease Control and Prevention (CDC)* em 2015, requer fadiga profunda persistente por mais de seis meses necessariamente acompanhada de mal estar pós-esforço, sono não reparador e dificuldade cognitiva.

Os fatores que contribuem para o desenvolvimento de CFS ainda são estudados pela comunidade acadêmica, dentre eles estão: sono irregular/insatisfatório, estresse psicológico, disfunções hormonais, deficiência de nutrientes, disfuções imunológicas e infecções. Em pesquisa realizada com enfermeiras, Samaha et al. (2007) identificou qualidade de

sono como um fator que contribui significativamente para a CFS. Em um estudo de espectometria de massa realizado por Naviaux et al. (2018), 45 pacientes com CFS apresentaram níveis alterados de fosfolipídios, colesterol, aminoácidos ramificados, vitaminas e metabólitos mitocondriais. Segundo Litleskare et al. (2018) a prevalência de CFS após 10 anos de ocorrência da infecção Giardíase foi de 2,22 a 4,08 vezes maior do que no grupo de controle. Estudo caso-controle realizado por Nagy-Szakal et al. (2018) revelou níveis baixos de betaína e lipídios complexos assim como níveis elevados de triglicerídeos e fenilacetilglutamina em indivíduos com CFS quando comparados com os de controle. Existe evidência de relação entre CFS e a síndrome metabólica, que inclui hipertensão arterial, níveis elevados de açúcar no sangue, gordura em torno da cintura e níveis anormais de colesterol (Maloney et al., 2010). Há indícios de baixos níveis de nutrientes em indivíduos com CFS, como vitamina C, vitamina B, sódio, magnésio, ácido folico, ácidos graxos que também parecem ter importância na severidade da CFS (Bjorklund et al., 2019). Segundo Bou-Holaigah et al. (1995), existe relação significativa entre a CFS e hipotensão mediada neuralmente, seu tratamento que inclui consumo moderado de sódio se mostrou eficiente em reduzir sintomas da CFS em uma parcela de indivíduos. Os esforços para definir marcadores eficientes para a CFS claramente estão em andamento, isso se deve em por sua importância ter sido reconhecida recentemente. Segundo Kennedy et al. (2010) a qualidade de vida de crianças com CFS se compara à de crianças com diabetes mellitus tipo 1 ou asma.

Trabalhadores em turnos são naturalmente mais sucetíveis ao desenvolvimento da CFS devido aos hábitos inusuais de sono e descanso. De acordo com Costa (2010) o trabalho em turnos noturnos é uma das condições mais estudadas, pois perturba o ciclo de sono e modifica o padrão de descanso, resultando em estresse significativo na regulagem dos ritmos circadianos biológicos de humanos, seres naturalmente diurnos. Estudo realizado por Shen et al. (2005) detectou correlação positiva com frequência de trabalho em turnos e intensidade da fadiga sentida pelos indivíduos. A alteração dos ritmos circadianos de funções corporais é responsável pela síndrome de "*shift-lag*", caracterizada por sentimentos de fadiga, sonolência, insônia, dificuldades digestivas, irritabilidade, menor agilidade mental e performance reduzida (Costa, 2010). Em condições de trabalho de risco o desenvolvimento da CFS pode aumentar a chance de acidentes fatais. Resultados de modelagem de equações estruturais realizada por Useche et al. (2017) em dados de

motoristas encontrou relação significativa entre presença de fadiga e comportamentos de risco na condução de ônibus. O trabalho em turnos alternados da indústria de mineração, além de se enquadrar neste cenário também apresenta naturalmente os fatores de sono irregular/insatisfatório e estresse psicológico.

Tendo em vista a severidade de sintomas da CFS e suas implicações na vida social e profissional, esforços para sua efetiva prevenção têm sido feitos, além de sua caracterização. De acordo com Murphy et al. (2011), modelagem preditiva pode ser uma ferramenta efetiva na prevenção da CFS. Em um estudo realizado por Huang et al. (2007) três métodos de classificação foram comparados na predição de CFS utilizando de dados genéticos, onde o *nayve bayes* atingiu 0.7 de área abaixo da curva ROC (AUC). Buscando fatores de relevância para fadiga em portadoras de câncer de mama, Servaes et a (2002) utilizou de regressão linear para examinar a contribuição de fatores físicos, psicológicos, sociais e cognitivos na severidade da fadiga das pacientes. Com árvores de decisão, Bronikowski et al. (2011) obteve 71.88% de acurácia predizendo CFS com base em respostas a um questionário médico aplicado a uma comunidade em um estudo sobre fadiga crônica.

Trabalhadores em turnos frequentemente reclamam de irritabilidade, ansiedade e condições de trabalho estressante. Déficit de sono e alteração persistente do ritmo cirdadiano pode levar a CFS, neuroticismo, ansiedade crônica e/ou depressão, aumentando risco de absenteísmo e necessidade de medicamentos psicotrópicos (Cole et al., 1990; Colquhoun et al., 1996; Nakata et al., 2004). No intuito de buscar evidências que permitam auxiliar na prevenção da CFS e procurar fatores relacionados, este trabalho propõe o estudo de seu risco relacionando variáveis bioquímicas e antropométricas com o absenteísmo de trabalhadores em turnos da indústria de mineração.

1.1 Objetivos

Este trabalho visa contribuir para o corpo de evidências de potenciais fatores e obter modelo de predição do risco de CFS por meio do absenteísmo. Desta maneira contribuindo para o esforço da comunidade acadêmica caracterização da CFS e seus fatores causadores, em busca de maneiras eficientes de sua prevenção e tratamento. Dentre os objetivos

específicos, estão:

- Comparar diferentes modelos para predição de absenteísmo;
- Obter modelo de predição do absenteísmo com melhor desempenho;
- Explorar fatores relacionados ao absenteísmo e relacioná-los à CFS.

1.2 Organização

O restante desta dissertação está organizada conforme se segue.

Para auxílio do leitor, a próxima seção apresenta a revisão bibliográfica do referencial teórico que será útil no entendimento das análises feitas. A seção três apesenta uma breve introdução dos dados utilizados e, tendo feito a revisão metodológica, da abordagem de análise utilizada. Por fim, na segunda parte, a seção 4 traz o artigo redigido com os resultados e discussões da análise, assim como considerações, forças e limitações.

2 Referencial teórico

2.1 Dados faltantes

Dados faltantes têm sido um problema para pesquisadores desde o início da pesquisa em campo, maior parte deste problema se deve ao fato de que os procedimentos analíticos utilizados, muitos dos quais foram desenvolvidos no começo do século 20, foram desenvolvidos para dados completos (Graham, 2009). É um problema comum na maioria da pesquisa científica, dentre elas biologia e medicina, podendo surgir a partir de amostras mal manuseadas, erros de medidas, não-resposta ou valores aberrantes que necessitam ser deletados (Schmitt et al., 2015). Não é incomum pesquisadores removerem casos com dados faltantes da análise, o que é chamado de método ou análise de casos completos, mas este tipo de análise pode introduzir viés de acordo com a natureza do fenômeno que resultou nos dados faltantes (Sterne et al., 2009). De acordo com Rubin (1976) os mecanismos de dados faltantes podem ser classificados em três perfis:

- **MAR**: Padrão dos dados faltantes é dito somente aleatório, ou *missing at random* (MAR). Neste caso o fato de os dados estarem faltando pode ter relação com os dados observados, mas não pode ter nenhuma dependência em dados não observados;
- **MCAR**: Dados que estejam faltando com padrão completamente aleatório, ou *missing completely at random* (MCAR). Neste caso o fato de haverem dados faltantes não pode ter dependência com os dados, sejam eles observados ou não;
- **MNAR**: O padrão não é aleatório ou *missing not at random* (MNAR). Este mecanismo implica que os dados dependem pelo menos de dados não observados.

Como Graham (2009) enfatiza, "mecanismo" não está relacionado necessariamente

com mecanismos causais, no sentido estatístico procura-se mais por descrever o comportamento dos dados faltantes. Em dados faltando com padrão MCAR, pode-se pensar nos casos com dados faltantes como uma amostrra aleatória do banco de dados. Já a palavra *random* no termo MAR quer dizer que, uma vez que os dados faltantes sejam condicionados a todas as variáveis observadas seu padrão será aleatório. Por fim, o MNAR não obtêm padrão aleatório mesmo que o condicionamento aos dados observados seja feito, mas este diagnóstico deve ser baseado no procdimento da coelta de dados uma vez que essencialmente não é detectável.

O método de análise de casos completos pode ser uma opção plausível para dados MCAR, entretanto para dados MAR esta abordagem introduz viés e uma técnica de imputação deve ser aplicada (Sterne et al., 2009). Muitos métodos de imputação foram propostos, um subgrupo deles se baseia na média dos dados observados, como: k vizinhos mais próximos, componentes principais bayesianos (bPCA) e imputações múltiplas por equações em cadeia (Schmitt et al., 2015). Segundo Graham (2009), dados MNAR introduzem viés de estimativa nas análises e seria preferível a análise de casos completos. Apesar de dados MCAR proporcionarem perda de poder estatístico, as análises resultam em estimativas não-viesadas, o que também ocorre com dados MAR quando as variáveis observadas são consideradas na imputação.

2.2 Imputação e o método *missForest*

Muitos métodos de imputação de dados faltantes já foram propostos, sendo que uma parcela considerável baseia-se nos dados observados (não-faltantes), como os métodos de K vizinhos mais próximos (KNN), análise de componentes principais bayesiana (bPCA), florestas aleatórias (missForest) e imputações múltiplas por equações em cadeia (MICE) (Schmitt et al., 2015). Um estudo de comparação de métodos de imputação, por Cihan (2018), analisou a média, KNN, decomposição de valores singulares (SVD), bPCA e missForest. Os resultados mostram melhor performance do método missForest em todos os conjuntos de dados quando são considerados o Erro Quadrático Médio (EQM) e a acurácia de classificação. Comparação realizada por Stekhoven and Bühlmann (2012) em 10 conjuntos de dados biológicos e médicos também mostra melhor performance do método

missForest em comparação com os métodos MICE e KNN, especialmente na presença de relações complexas entre as variáveis. Em dados laboratoriais a análise de Waljee et al. (2013) mostra o método missForest com performance melhor ou equiparável à performance dos métodos KNN e MICe em todos os conjuntos de dados, tais resultados permaneceram consistentes com 10%, 20% e 30% de dados faltantes usando tanto regressão logística quanto florestas aleatórias como modelos de predição.

O método missForest foi proposto por Stekhoven e Buhlmann (2012), como um método não paramétrico que lida com diferentes tipos de variáveis simultaneamente, usando o modelo de florestas aleatórias para prever valores ausentes. Em resumo, uma floresta aleatória é ajustada aos dados observados e usada para estimar os casos ausentes. O método ordena variáveis de acordo com sua porcentagem de dados faltantes e começa com a variável com a menor proporção, definindo-a como a variável dependente y. Em seguida, divide o conjunto de dados em dois: as linhas para as quais o y atual está faltante (X_{mis}, y_{mis}) e as que apresentam casos completos para y (X_{obs}, y_{obs}). Todos os casos ausentes de X são estimados, a princípio, usando a média ou qualquer outro método de imputação. Então, uma floresta aleatória é ajustada a (X_{obs}, y_{obs}) e usada para prever y_{mis} com base em X_{mis}. Outra variável é então definida como variável de dependente e o processo é repetido, prosseguindo iterativamente até que um critério de parada seja atendido, sendo esse critério geralmente o aumento do erro de imputação estimado e, finalmente, a última iteração é considerada o resultado final. Mais detalhes sobre o método e o algoritmo são descritos por Stekhoven e Bühlmann (2012).

2.3 Regressão Logística

Na regressão logística, a variável binária Y_i segue uma distribuição Bernoulli com $P(Y_i = 1) = \pi_i$ variando de acordo com uma função logística inversa do vetor de observações $\mathbf{x}_i j (j = 1, \ldots, k)$ incluindo uma constante e $k - 1$ variáveis explanatórias:

$$Y_i \sim Bernoulli(Y_i|\pi_i)$$
$$\pi_i = \frac{1}{1 + \exp -\mathbf{x}_i\beta}$$

em que o parâmetro desconhecido $\beta = (\beta_0, \beta_1)$ é um vetor de tamaho k com β_0 sendo escalar constante e β_1 o vetor de parâmetros correspondentes às variáveis explicativas (King e Zeng, 2001).

O modelo logístico é um caso particular de modelos lineares generalizados (GLM), modelos que estendem a regressão linear (Distribuição Normal) para outras distribuições da família exponencial. Neste caso a Bernoulli (Binomial com $n = 1$) é a distribuição que resulta no modelo logístico (Witten et al., 2016). Uma forma alternativa de expressar o modelo é imaginar uma variável contínua não observada Y_i^* com distribuição logística com média μ_i. Então o modelo estaria próximo de uma regressão linear se Y_i^* fosse observado:

$$Y_i^* \sim Logistica(Y_i^* | \mu_i)$$

$$\mu_i = \mathbf{x}_i \beta$$

$$P(Y_i^*) = \frac{\exp -(Y_i^* - \mu_i)}{(1 + \exp -(Y_i^* - \mu_i))^2}$$

em que μ_i varia de acordo com as observações como uma função linear de $\mathbf{x}_i$ (King e Zeng, 2001). A estimativa dos parâmetros é feita por máxima verossimilhança considerando observaçoes independentes e identicamente distribuídas. A aplicação de regressão logística é comum na área da saúde para criação de escores de risco (Kulothungan et al., 2014).

2.4 Regularização e seleção de variáveis

Técnicas de regularização são utilizadas para reduzir o erro ajustando uma função apropriadamente aos dados e prevenir super-ajuste. A proposta da regularização é evitar problemas de super-ajuste (*overfitting*) por meio do equilíbrio viés-variância. A regularização em regressão é obtida geralmente por meio de restrição no processo de ajuste dos modelos. Alguns tipos de restrição também combinam seleção de variáveis automática no ajuste do modelo.

2.4.1 Método Stepwise

A seleção de variáveis *Stepwise* é um método de ajuste de modelo que realiza um algoritmo para seleção de variáveis automaticamente com base em um critério de performance

(Hocking, 1976). É uma alternativa à técnica de *best-subset*, que é computacionalmente custosa por testar todos os possíveis subconjuntos de variáveis. As técnicas de *Stepwise* são variações de duas abordagens para a seleção:

- **Seleção *Forward***: Inicia-se com um modelo sem nenhuma variável. Em cada passo do algoritmo é gerado um modelo para cada variável candidata, incluindo ela no modelo obtido do passo anterior. Então é gerada uma medida de desempenho ou qualidade de ajuste e a variável com melhor resultado será incluída no modelo a ser utilizado no passo seguinte.
- **Eliminação *Backward***: O modelo inicial contém todas as variáveis, então é obtido um modelo para cada variável por meio de sua retirada. O modelo que tiver o melhor desempenho terá a variável removida no passo e será utilizado no próximo passo.

Variações do *Stepwise* envolvem a escolha da métrica e abordagem utilizadas (Hocking, 1976). Dentre as métricas mais utilizadas estão a soma de quadrados dos resíduos (SQR), erro quadrático médio (EQM) e coeficiente de determinação (R^2). Quanto às variações de abordagem além das duas apresentadas, a mais utilizada é a bidirecional que consiste em realizar a seleção *forward* e em cada passo considerar a retirada de variáveis utilizando a eliminação *backward* (Efroymson, 1960).

2.4.2 Métodos Elatic-Net

O método Elastic-Net é utilizado para seleção e regularização de variáveis em um banco de dados que combina os métodos de Ridge e Lasso, sendo considerado como uma generalização deles. Proposto por Hoerl and Kennard (1988), o método de regularização Ridge minimiza a Soma de Quadrado dos Resíduos (SQR) condicionada a uma penalização do tipo L2:

$$SQR_{Ridge} = SQR + \lambda \sum_j \beta_j^2$$

onde o hiperparâmetro λ é escolhido de forma a otimizar uma medida de qualidade do modelo. Esta regularização permite melhor performance do modelo mas não resulta em modelos parsimoniosos. Isto se deve ao fato de que a regularização encolhe os coeficientes

mas teoricamante nunca os define como zero (Zou e Hastie, 2005). A regressão Lasso, proposta por Tibshirani (1995), resolve o quesito de parsimônia propondo uma regularização do tipo L1:

$$SQR_{Ridge} = SQR + \lambda \sum_j |\beta_j|$$

que resulta no encolhimento contínuo dos coeficientes, assim como a Ridge, mas também faz seleção de variáveis pelo fato de os coeficientes poderem ser encolhidos para zero. Apesar de a Lasso ser mais promissora pelo fato de também efetuar seleção de variáveis em tempos de grandes bancos de dados, ela não domina a Ridge em termos de performance (Tibshirani, 1995). Apesar de performar bem em muitas situações, algumas limitações foram identificadas no método Lasso, conforme indicadas por Zou e Hastie (2005):

- Em casos de o número de variáveis p ser muito maior do que o de observações n, a Lasso seleciona no máximo n variáveis;
- Se existirem grupos de variáveis com correlação alta, a Lasso escolherá somente uma das variáveis, não seguindo um critério na escolha;
- Em casos que $n > p$ com alta correlação entre as variáveis, a Ridge domina a Lasso em termos de performance.

Tendo em mente estas limitações, Zou e Hastie (2005) propuseram a Elastic-Net. Este novo método de regularização combina as regularizações L1 (Lasso) e L2 (Ridge) utilizando de um segundo hiperparâmetro α:

$$SQR_{Ridge} = SQR + \lambda \left[(1-\alpha) \sum_j \beta_j^2 + \alpha \sum_j |\beta_j| \right]$$

que é selecionado, em conjunto com λ, de forma a otimizar uma medida de performance do modelo. Em aplicações de dados de genes, o Elastic-Net se mostrou eficiente em agrupar genes com correlações altas ao invés de selecionar somente um deles. Zou e Hastie (2005) descrevem o Elastic-Net como uma generalização do Lasso, que performa bem nos casos em que o Lasso tinha limitações.

2.5 Desempenho de modelos

Em modelos de predição se faz necessário o uso de medidas de performance, pois modelos podem se ajustar bem aos dados mas performar mal, ou ter sua performance alterada no decorrer do tempo com a possível alteração do comportamento nos dados. Medidas comumente utilizadas para classificadores binários (como a regressão logística) são a acurácia, sensitividade e especificidade (Fawcett, 2005), que se baseiam em quatro quantidades: Verdadeiros Positivos (VP), Falsos Positivos (FP), Verdadeiros Negativos (VN) e Falsos Negativos (FN). Nas nomenclaturas a palavra "Falso"diz respeito a classificação errônea enquanto que "Verdadeiro"diz respeito a classificação correta na predição do evento de interesse. Comumente estas quatro quantidades são mostradas em uma amtriz, chamada matriz de confusão.

A medida de acurácia do classificador é a proporção de classificações corretas feitas por ele em relação ao total de classificações realizadas, calculada por:

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN}$$

não levando em conta cada classe, somente se a classificação foi correta ou não. Por não discernir performance entre classes, a acurácia pode levar a conclusões errôneas em bancos de dados com proporção de classes desbalanceadas, quando considerada sozinha (Provost et al., 1998). Por esse motivo geralmente é acompanhada das medidas de sensibilidade (taxa de verdadeiros positivos ou de detecção) e especificidade (taxa de verdadeiros negativos) que são obtidas da forma:

$$Sensibilidade = \frac{VP}{VP + FN}$$
$$Especificidade = \frac{VN}{VN + FP}$$

permitindo melhor mensuração de performance do modelo para tomada de decisão (Altman e Bland, 1994). Pode-se ter interesse em um modelo com especificidade reduzida em troca de alta sensibilidade, por exemplo, quando se quer garantir a detecção da maior taxa de verdadeiros positivos possíveis (Provost et al., 1998).

Quando se trata de escolher o melhor classificador com base nas medidas de sensibilidade e especificidade, a curva ROC é o método mais utilizado para visualizar o balanço entre taxas de classificações corretas e de alarmes falsos (Fawcett, 2005). Basicamente consiste em um gráfico 2D com a taxa de cerdadeiros positivos (Sensitividade) no eixo Y e a taxa de falsos negativos (1-Especificidade) no eixo X. A Figura 1 exemplifica o gráfico com a performance de cinco classificadores.

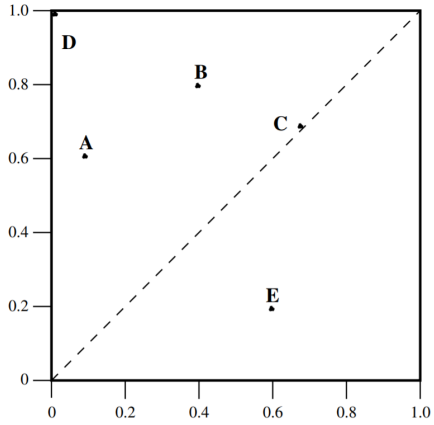


Figura 1: Gráfico ROC básico para cinco classificadores
Fonte: Fawcett et al. (2005)

O classificador perfeito ($Sensibilidade = 1$ e $Especificidade = 1$) seria o ponto D da Figura 1, a proximidade de classificadores deste ponto permite a comparação baseada nas duas medidas em conjunto. A linha pontilhada representa a performance de um classificador aleatório, que escolha completamente ao acaso.

Quando o classificador é obtido a partir de um modelo que retorna uma probabilidade contínua, a sua performance está condicionada ao limite de decisão escolhido, neste caso pode-se obter uma curva ROC com os classificadores utilizando múltiplos valores dentro do intervalo de possíveis limites de decisão de classe (Fawcett, 2005). Uma métrica que se torna útil para classificadores que gerem valores contínuos é a área abaixo da curva ROC (AUC), por estar diretamente relacionada com a proximidade da curva ao ponto de classificação perfeita, como mostrado na Figura 2, em que o classificador B mostra maior dicriminância entre as classes.

É importante perceber que, realisticamente, nenhum classificador deveria obter AUC abaixo de 0,5, uma vez que a linha diagonal $x = y$ representa um classificador que faz

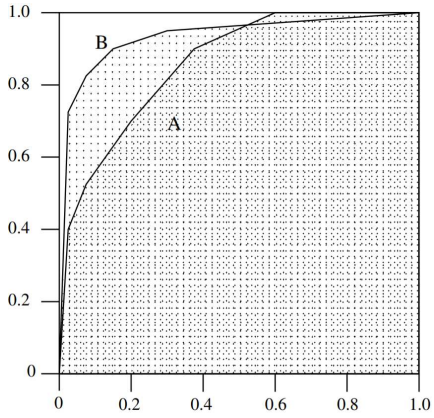


Figura 2: Curva ROC comparando dois classificadores
Fonte: Fawcett et al. (2005)

escolhas aleatóriamente (Fawcett, 2005). Uma importante propriedade da AUC é sua equivalência à probabilidade de que um classificador ranqueie um valor escolhido positivo acima de um valor escolhido negativo, o que é equivalente ao teste de ranque de Wilcoxon (Hanley and McNeil, 1982). Também existe relação com o coeficiente de Gini ($Gini + 1 = 2 \times AUC$), como mostrado por Hand and Till (2001).

2.6 Validação Cruzada

Validação cruzada (*cross-validation*) é um método de seleção de modelos que, a partir de uma medida de performance de escolha, estima performance em dados não vistos anteriormente. Considerando um banco de dados com $X_1, X_2, \ldots, X_n$ observações, os métodos básicos de validação cruzada são:

- *leave-one-out*: Consiste em remover a i-ésima observação dos dados e ajustar o modelo nas observações restantes $(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ e utilizar o modelo obtido para obter a medida de performance na observação X_i. Este processo é repetido para cada observação do banco de dados, por fim obtendo n estimativas de performance do modelo. Desta forma é possível estimar a distribuição da medida de performance obtendo uma performance média estimada e sua medida de incerteza;

- *k-fold*: Este método segue a mesma idéia do *leave-one-out*, mas separa as n observações em k subgrupos mutuamente exclusivos (os $folds$) de tamanho aproximadamente igual. Como resultado teremos k performances estimadas com base na medida

de escolha. A partir das k medidas é possível então obter estimativa de performance em novos dados e também uma medida de incerteza.

Segundo Vanwinckelen e Blockeel (2012), sob o argumento da alta variabilidade dos métodos *leave-one-out* e *k-fold*, uma variação frequentemente utilizada é o *repeated k-fold* em que a validação cruzada é feita repetidas vezes, variando a separação dos k grupos de forma a obter mais estimativas. Como foi mostrado por Vanwinckelen e Blockeel (2012), a performance do modelo é subestimada quando a validação cruzada é utilizada, por estimar a performance de modelos treinados em dados com tamanho $(k-1)/k \times 100\%$ do total. O método *repeated 10-fold* de validação cruzada se mostra mais acurado para seleção de modelos (Arlot and Celisse, 2010), mas devido a viés ele não é sugerido para estimativa de performance do modelo (Kohavi, 1995).

2.6.1 Método não-paramétrico de *Bootstrap*

A técnica não-paramétrica de bootstrap, proposta por Efron (1979), é utilizada para estimar a variabilidade de uma medida de interesse que seja obtida a partir de dados. Considerando conjunto de dados $X_1, X_2, \ldots, X_n$, uma reamostragem de bootstrap consiste em obter uma nova amostra $Y_1, Y_2, \ldots, Y_n$ retirada dos dados originais, com reposição. A técnica envolve obter k reamostragens de bootstrap e suas respectivas medições $m_i = f(Y_i1, Y_i2, \ldots, Y_in) \quad \forall i = 1, \ldots, k$, o que resultará na estimativa de bootstrap:

$$Est_{Bootstrap} = \sum_{i=1}^{k} \frac{m_i}{k}$$

assim como medidas de incerteza também podem ser obtidas a partir de $m_1, \ldots, m_k$.

A partir das medidas de variabilidade obtidas por bootstrap pode-se obter intervalos de confiança, o método mais utilizado é o de quantis, que obtém intervalo de confiança $(1-\alpha) \times 100\%$ por meio dos quantis $\alpha/2$ e $(1-\alpha)/2$ das k medidas m_i de bootstrap. Em estudo buscando quantidade ideal k de reamostragens , Efron (1987) mostra que o coeficiente de variação é considerável para 200 reamostragens (9%) e reduz para 4% em 1.000, consequentemente o autor indica $k = 1000$ como um valor apropriado. Como mostrado por (Kohavi, 1995) em análise de simulação, quando aplicado em grandes amostras,

estimativas de bootstrap tendem a apresentar menor variabilidade do que as estimadas por *k-fold* ($k = 10$ e $k = 20$), enquanto nenhum dos métodos domina quanto ao viés relativo, variando dentre diferentes bancos de dados.

2.6.2 *Bootstrap 0.632*

Ao considerar o desempenho das regras de previsão treinadas nos dados, Efron e Tibshirani (1997) propuseram o bootstrap 632+ para corrigir o viés inerente ao bootstrap clássico e melhorar o processo de validação cruzada. Ele reduz a variabilidade em relação ao *k-fold* e reduz o viés, transformando o bootstrap 632+ em um método mais atraente para a estimativa de desempenho do modelo. Conforme descrito por Witten (2016), em qualquer amostra de bootstrap, à medida que n cresce, a proporção de casos não selecionados tenderá a $(1-1/n)^n \approx e^{-}1 = 0,368$, que é a chance de uma observação em particular não ser escolhida. O desempenho do modelo ajustado à reamostragem fornecerá estimativas tendenciosamente otimizadas se avaliado na reamostragem em si, uma vez que estimou os coeficientes a partir do mesmo conjunto de dados. O conjunto de treinamento (nova amostra) possui apenas 0,632x100% dos casos originais; portanto, o modelo originado resultará em uma estimativa de desempenho pessimista, enviesada, quando avaliada nos 0,368 casos restantes (teste), apesar de ter tamanho n. A idéia principal do bootstrap 632+ é avaliar as medidas de desempenho de treinamento ($\varepsilon_{training}$) e de teste (ε_{test}) e obter a estimativa de desempenho ponderada

$$\varepsilon_{632} = 0,632 \times \varepsilon_{test} + 0,368 \times \varepsilon_{training}$$

combinando o desempenho pessismista do teste com o desempenho otimista do treinamento.

3 Material e métodos

3.1 Dados

Um estudo de coorte foi realizado no ano de 2012 utilizando como indivíduos de interesse trabalhadores de uma mina localizada nos Inconfidentes Mineiros. O estudo, entitulado Sindrome Metabólica em Trabalhadores da Mineração do Estado de Minas Gerais, foi aprovado pelo comitê de ética da Universidade Federal de Ouro Preto (CAAE: 0018.0.238.000-11). Os indivíduos de interesse trabalham em turnos de 6 horas operando caminhões *offroad* de mineração, turnos estes seguidos por um período de 12 horas de descanso. Os dados coletados consistem de 22 variáveis divididas entre 8 antropométricas, 11 bioquímicas, sexo, idade e a variável "Absenteísmo", que é o indicador binário de absenteísmo emqualquer dia do ano de 2012. A variável "bsenteísmo"é a variavel de interese do estudo, ou seja, a variável dependente.

3.2 Método de Análise

Após da análise descritiva/exploratória, todas as variáveis foram padronizadas parao restante da análise. O primeiro passo é classificar o padrão de dados faltantes e então realizar a imputação de todos os valores com o método *missForest*. Modelos foram ajustados aos dados imputados utilizando tanto Regressão Logística quanto *Elastic-Net*. No caso da regressão logística foi utilizado stepwise para seleção de variáveis enquanto que a *Elastic-Net* realizar regularização/seleção naturalmente. O método do *Bootstrap 0.632* é, então, utilizado para comparar o poder preditivo dos modelos utilizando as métricas de acurácia, sensitividade e especificidade. Por fim, os efeitos significativos dos modelos são ponderados com efeitos encontrados na literatura para a CFS e as variáveis do estudo.

REFERÊNCIAS

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

Altman, D. G., and Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal, 308(6943), 1552.*

Arlot, S., and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys, 4, 40-79.*

Bjorklund, G., Dadar, M., Pen, J. J., Chirumbolo, S., and Aaseth, J. (2019). Chronic fatigue syndrome (CFS): Suggestions for a nutritional treatment in the therapeutic approach. *Biomedicine and Pharmacotherapy, 109, 1000-1007.*

Bou-Holaigah, I., Rowe, P. C., Kan, J., and Calkins, H. (1995). The relationship between neurally mediated hypotension and the chronic fatigue syndrome. *Jama, 274(12), 961-967.*

Bronikowski, C. M., Weng, A., Furst, J. D., and Raicu, D. S. (2011). Prediction of chronic fatigue syndrome using decision tree-based ensemble methods. In Proceedings on the International Conference on Artificial Intelligence (ICAI) (p. 1). *The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).*

Cihan, P. (2018). A Comparison of Five Methods for Missing Value Imputation in Data Sets. *International Scientific and Vocational Studies Journal. 2(2), 80-85.*

Cole, R. J.; Loving, R. T.; Kripke, D. F. (1990). Psychiatric aspects of shiftwork. *Occupational medicine (Philadelphia, Pa.), 5(2), 301-314.*

Colquhoun, W. P., Costa, G., Folkard, S., and Knauth, P. (1996). *Shiftwork. Problems and solutions.*

Costa, G. (2010). Shift work and health: current problems and preventive actions. *Safety and health at Work, 1(2), 112-123.*

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics.*

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association, 82(397), 171-185.*

Efron, B., and Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association, 92(438), 548-560.*

Efroymson, M. A. (1960). Multiple regression analysis. *Mathematical methods for digital computers, 191-203.*

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27(8), 861-874.*

Fukuda, K., Straus, S. E., Hickie, I., Sharpe, M. C., Dobbins, J. G., and Komaroff, A. (1994). The chronic fatigue syndrome: a comprehensive approach to its definition and study. *Annals of internal medicine, 121(12), 953-959.*

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual*

review of psychology, 60, 549-576.

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143(1), 29-36.*

Hand, D. J., and Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning, 45(2), 171-186.*

Hocking, R. R. (1976). A Biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics, 32(1), 1-49.*

Hoerl, A., and Kennard, R. (1988). Ridge regression, in *Encyclopedia of Statistical Sciences, Vol. 8.*

Holmes, G. P., Kaplan, J. E., Gantz, N. M., Komaroff, A. L., Schonberger, L. B., Straus, S. E., ... and Tosato, G. (1988). Chronic fatigue syndrome: a working case definition. *Annals of internal medicine, 108(3), 387-389.*

Huang, L. C., Hsu, S. Y., and Lin, E. (2009). A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data. *Journal of Translational Medicine, 7(1), 81.*

Huber-Carol, C., Balakrishnan, N., Nikulin, M., and Mesbah, M. (Eds.). (2012). *Goodness-of-fit tests and model validity.* Springer Science and Business Media.

Jason, L. A. T. H. S., and Njoku, M. G. C. (2006). The face of CFS in the US. *CFIDS Chronicle, 16-21.*

Jason, L. A., Richman, J. A., Rademaker, A. W., Jordan, K. M., Plioplys, A. V., Taylor,

R. R., ... and Plioplys, S. (1999). A community-based study of chronic fatigue syndrome. *Archives of internal medicine, 159(18), 2129-2137.*

Kennedy, G., Underwood, C., and Belch, J. J. F. (2010). Physical and functional impact of chronic fatigue syndrome/myalgic encephalomyelitis in childhood. *Pediatrics, 125(6), e1324-e1330.*

King, G., and Zeng, L. (2001). Logistic regression in rare events data. *Political analysis, 9(2), 137-163.*

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. *In Ijcai (Vol. 14, No. 2, pp. 1137-1145).*

Kulothungan, V., Ramakrishnan, R., Subbiah, M., and Raman, R. (2014). Risk Score Estimation of Diabetic Retinopathy: Statistical Alte rnatives using Multiple Logistic Regression. *Journal of Biometrics and Biostatistics, 5(5), 1.*

Litleskare, S., Rortveit, G., Eide, G. E., Hanevik, K., Langeland, N., and Wensaas, K. A. (2018). Prevalence of irritable bowel syndrome and chronic fatigue 10 years after Giardia infection. *Clinical Gastroenterology and Hepatology, 16(7), 1064-1072.*

Maloney, E. M., Boneva, R. S., Lin, J. M. S., and Reeves, W. C. (2010). Chronic fatigue syndrome is associated with metabolic syndrome: results from a case-control study in Georgia. *Metabolism, 59(9), 1351-1357.*

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

Nagy-Szakal, D., Barupal, D. K., Lee, B., Che, X., Williams, B. L., Kahn, E. J., ... and Levine, S. (2018). Insights into myalgic encephalomyelitis/chronic fatigue syndrome phenotypes through comprehensive metabolomics. *Scientific reports, 8(1), 10056.*

Nakata, A., Haratani, T., Takahashi, M., Kawakami, N., Arito, H., Kobayashi, F., ... and Araki, S. (2004). Association of sickness absence with poor sleep and depressive symptoms in shift workers. *Chronobiology International, 21(6), 899-912.*

Naviaux, R. K., Naviaux, J. C., Li, K., Bright, A. T., Alaynick, W. A., Wang, L., ... and Gordon, E. (2016). Metabolic features of chronic fatigue syndrome. *Proceedings of the National Academy of Sciences, 113(37), E5472-E5480.*

Provost, F. J., Fawcett, T., and Kohavi, R. (1998, July). The case against accuracy estimation for comparing induction algorithms. *In ICML (Vol. 98, pp. 445-453).*

R CORE TEAM (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* URL https://www.R-project.org/.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63(3), 581-592.*

Samaha, E., Lal, S., Samaha, N., and Wyndham, J. (2007). Psychological, lifestyle and coping contributors to chronic fatigue in shift?worker nurses. *Journal of advanced nursing, 59(3), 221-232.*

Schmitt, P., Mandel, J., and Guedj, M. (2015). A comparison of six methods for missing data imputation. *Journal of Biometrics and Biostatistics, 6(1), 1.*

Servaes, P., Verhagen, S., and Bleijenberg, G. (2002). Determinants of chronic fatigue in disease-free breast cancer patients: a cross-sectional study. *Annals of Oncology, 13(4),*

589-598.

Shen, J., Botly, L. C., Chung, S. A., Gibbs, A. L., Sabanadzovic, S., and Shapiro, C. M. (2006). Fatigue and shift work. *Journal of sleep research, 15(1), 1-5.*

Stekhoven, D. J., and Bühlmann, P. (2011). MissForest: non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28(1), 112-118.*

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj, 338, b2393.*

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.*

Useche, S., Cendales, B., and Gómez, V. (2017). Work stress, fatigue and Risk Behaviors at the Wheel: Data to assess the association between psychosocial work factors and risky driving on Bus Rapid Transit drivers. *Data in brief, 15, 335-339.*

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

Waljee, A. K., Mukherjee, A., Singal, A. G., ... and Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open, 3(8), e002847.*

Williams, Y. J., Jantke, R. L., and Jason, L. A. (2014). Chronic fatigue syndrome: case definitions and diagnostic assessment. *New York State psychologist, 26(4), 41.*

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

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.*

Parte II

Artigo

4 Artigo: *Chronic fatigue syndrome and its relation with absenteeism: Elastic-net and stepwise applied to biochemical and anthropometric clinical measurements*

Redigido conforme as normas da revista PLOS ONE - versão preliminar.

Chronic fatigue syndrome and its relation with absenteeism: Elastic-net and stepwise applied to biochemical and anthropometric clinical measurements

* Corresponding author: a.neisse@gmail.com

Abstract

Characterized by persistent fatigue, pain, cognitive impairment and sleep difficulties, the Chronic Fatigue Syndrome (CFS) has been common in clinical practice in the last decades. Studies indicate multiple factors that contribute to CFS development: poor sleep, dehydration, psychological stress, hormonal dysfunction, infections, nutrient deficiencies, among others. In risk work conditions, like the shift work of mines, development of CFS increases significantly the chance of fatal accidents. The work environment of mines suggests the presence of factors that increase the risk of developing CFS. Considering the severity of CFS's symptoms and its implications on the social and professional lives as well as on the economy, efforts are targeting not only its characterization but also effective 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 imputing missing data, logistic regression was fitted using Stepwise selection as well as Lasso and Elastic-Net regularization. The three variations were also applied to the complete-cases data set, 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 significative 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.

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 mean 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 the majority of the academic community, there is criticism to this definition case due to the fact that 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 Center 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-extertional 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., 2018). 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) suggests 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. (2005), 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. (2007) 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.

Materials and methods

A cross-sectional study was performed in 2012 on shift workers of a mine localized 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. 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.

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), missing mechanisms have three classifications: Missing Completely at Random

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

Variable	Description	Method	Type	Mean	SD	% Missing
Skipped	Was there abseteeism?	Mining Company Files	Binary (Yes = 1)	0.2238	-	0.00
Sex	Sex of the individual	Medical file	Binary (Male = 1)	0.9646	-	0.00
Age	Individual's age	Medical file	Discrete	36.7504	7.1336	0.00
Height	Individual's height (m)	Platform stadiometer	Continuous	174.1032	7.3123	5.25
Weight	Individual's weight (kg)	0.1 kg precision scale	Continuous	80.3670	12.8352	2.48
BMI	Body Mass Index	Weight/Height^2	Continuous	26.4756	3.6749	2.48
WHRatio	Avg. Waist-to-Hip Ratio (n=3)	Simple tape measure	Continuous	0.8566	0.1220	2.64
TBF	Total Body Fat (%)	Bioimpedance	Continuous	24.3923	8.1730	3.50
ViscFat	Visceral Fat (%)	Bioimpedance	Continuous	7.6712	3.4312	3.33
AvDBP	Avg. Dias. Blood Press. (n = 3)	Semi-autom. Monitor	Continuous	82.7661	9.3659	0.16
AvSBP	Avg. Sist. Blood Press. (n = 3)	Semi-autom. Monitor	Continuous	131.5532	13.7510	0.16
HDL	HDL Cholesterol	Enzyme-colorimetric	Continuous	55.4063	16.3538	5.79
LDL	LDL Cholesterol	Chol - (HDL + VLDL)	Continuous	109.4432	36.2181	7.63
Trig	Triglycerides	Enzyme-colorimetric	Continuous	151.5527	83.6749	5.61
Chol	Total Cholesterol	Enzyme-colorimetric	Continuous	194.5765	45.4948	5.61
Calcium	Calcium	Ion selective electrode	Continuous	9.5704	1.2965	5.61
Phosphorus	Phosphorus	Kinetic U.V. test	Continuous	3.5085	0.5965	42.43
VitD	Vitamin D	Electrochemiluminescence	Continuous	25.4991	7.7247	38.62
PTH	PTH Hormone	Chemiluminescence	Continuous	30.8635	10.8913	25.96
Glucose	Fasting Glucose	Enzyme-colorimetric	Continuous	86.9475	14.5506	5.08
Sodium	Sodium	Ion selective electrode	Continuous	144.2196	8.1133	12.70
Potassium	Potassium	Ion selective electrode	Continuous	4.7249	0.7105	9.72

(MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). As Graham (2009) emphasizes, the term "mechanism" is not necessarily related to causation, in the statistical sense the interest resides in the description of the missing data patterns themselves. In MCAR the pattern of missing data is completely independent and the cases with missing data can be considered a random sample from the whole dataset. As for MAR, the missingness only achieves randomness when the missing pattern is conditioned on the observed data. However, when data is missing not at random (MNAR), it is not random even when conditioned on observed data because the missing pattern depends on unobserved data. Since the reason that the data is missing in this is 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). A study comparing imputation methods performed by Cihan (2018) compared mean, KNN, singular values decomposition (SVD), bPCA and missForest. The results show that, when considering residual mean squared error (RMSE) and classification accuracy, the missForest method outperformed the other methods in all the datasets. A comparison performed by Stekhoven and Bühlmann (2012) in 10 medical and biological datasets showed that missForest outperforms KNN and MICE approaches to imputation, especially when there are complex relations between variables. In a study comparing imputation methods for laboratory data, performed by Waljee et al. (2013), the missForest method either outperformed or remained competitive with mean imputation, KNN and MICE. These results were consistent in 10%, 20% and 30% of missing data using both Logistic Regression and Random Forest models for prediction.

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 method orders variables according to their missing percentage and starts with the variable with the least missing data defining it as the target variable y. Then it splits the dataset in two: the rows for which the current y is missing (X_{mis}, y_{mis}) and the ones with complete cases for y (X_{obs}, y_{obs}). Then any missing cases of X are guessed by using mean imputation or any other imputation method. Finally a Random Forest is fitted to (X_{obs}, y_{obs}) and used to predict y_{mis} based on X_{mis}. Another variable is then defined as the target variable and the process is repeated, proceeding iteratively until a stop criterion is satisfied, that criteria usually being the increase of the estimated imputation error, then the last iteration is considered as the final result. 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 the study at the beginning while the remaining cases were imputed with the missForest method.

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. In the Logistic Regression framework the dependent variable follows a Bernoulli distribution with $P(Y_i = 1) = \pi_i$. It is usual to consider it as an unobserved continuous Y_i^* variable that follows a Logistic distribution ($P(Y_i^* = 1) \sim Logistic(Y_i|\pi_i)$) with mean μ_i that varies over the observations as a linear function of $\mathbf{x}_i$ ($\mu_i = \mathbf{x}_i\beta$). Therefore the model with logistic density $P(Y_i^*) = e^{-(Y_i^* - \mu_i)}/(1 + e^{-(Y_i^* - \mu_i)})^2$ would be very close to a linear regression if Y_i^* was observed but despite only Y_i being observed the model remains the same, details are described by King and Zeng (2001). This framework 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 on step one evaluating the p possible models with one variable in terms of the chosen metric in order to choose the variable to be included. Then, on step two, $p - 1$ models are evaluated to choose the variable to be included on the one-variable model from the previous step. At the end each p selected models are compared in terms of the performance metric and the one with the best performance is chosen to be the final model. 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 this approach 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 (1995) 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 $\lambda \left[(1-\alpha) \sum_j \beta_j^2 + \alpha \sum_j |\beta_j| \right]$ 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 had 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 will be fitted by using Stepwise, Lasso and Elastic-Net for comparison of predictive performance. All these methods need to be used in conjunction with a model selection procedure. The most commonly 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 and respectively.

Model performance measurement

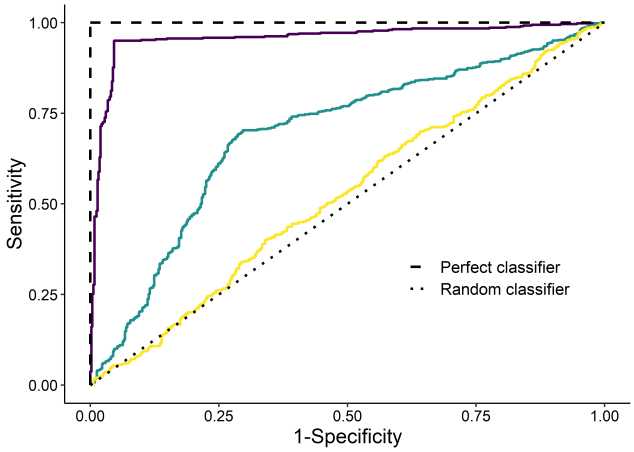
Measuring a predictive model's performance is useful both for model selection and validation, it allows to verify whether it over-fitted the data and/or result in poor prediction in new data. According to Fawcett (2005) the most common measures for binary classifiers are Accuracy (ACC), Sensitivity (SNS), Specificity (SPC) and the Area Under the ROC curve (AUC), all being functions of True Positives (TP), False

Positives (FP), True Negatives (TN) and False Negatives (FN):

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad SNS = \frac{TP}{TP + FN} \quad SPC = \frac{TN}{TN + FP}$$

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 (Fawcett, 2005). It basically consists of a graph with the Sensitivity in the y-axis and 1-Specificity in the x-axis for every possible cut-off probability forming a line, as shown by Figure 1 in an example for 3 different classifiers and the theoretically perfect and random classifiers. The perfect classifier ($Sensitivity = 1$ and $Specificity = 1$) showed in Figure 1 would be the cut-off corresponding to the upper left point in the figure, the proximity of this point is an indicator of the model's discrimination level. Both the Random and the perfect classifiers establish a baseline for comparison of any obtained classifier, the ones closer to the upper left corner of Figure 1 (Perfect classifier) discriminate better between the classes.



Source: Author

Fig 1. Theoretical curves for perfect and random classifiers (in black) and three examples of curves for classifiers of different discrimination powers.

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 ($Gini + 1 = 2 \times AUC$) as shown by Hand and Till (2001).

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 method is the k-fold in which, considering a data set with n observations, the basic idea is to divide them into k groups and evaluate the model obtained from the k-1 groups' observations in the remaining group that has a $1/k$ proportion of the original data set. This is performed for each of the k groups resulting in k measurements of performance, which then are averaged and used as an estimate of the performance in new data. The method called leave-one-out consists of considering $k = n$ and therefore generating n measurements of performance. According to Vanwinckelen and Blockeel (2012) due to the argument of the high variability of these methods a frequently used variation is the repeated k-fold, where each repetition m varies the choice of the k subgroups in order to obtain more estimates. In the classification context, it is also common to perform the stratified cross-validation, which guarantees class balance similar to the original data's 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, showed by Arlot and Celisse (2010), it is still biased and therefore 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 evaluate 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 (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 0.632x100% 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 allows 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.

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, 2019) and the `RStudio IDE v1.3.125`. The packages used were: `caret` for model fitting and selection (Kuhn, 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. Source code to reproduce the results is available as Supporting Information on the journal's web page (http://onlinelibrary.wiley.com/doi/xxx/suppinfo). The script `Analysis_reproduction.R` should be executed first, which sources `auxiliary_functions.R` on the run. The script `Figure-Table_reproduction.R` contains code to reproduce the table and figures presented in this paper.

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.

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.

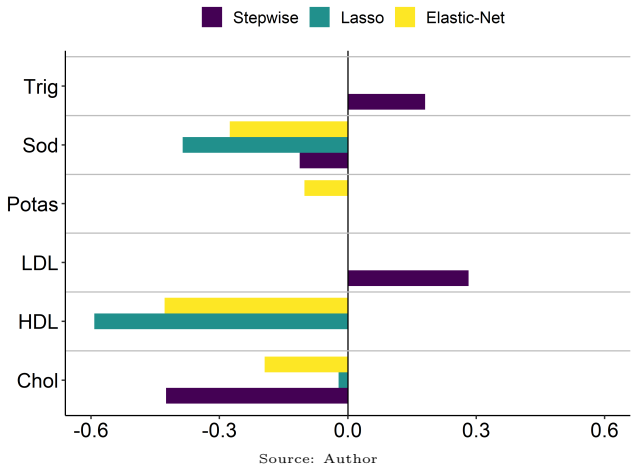


Fig 2. Relative effects of coefficients for each model. Each model had its coefficients divided by the sum of absolute values of all their coefficients.

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 regularize models might be explained by their complementary

property as decompositions of the Total Cholesterol. The coefficients are presented in Table 2 with the bootstrap estimations and confidence intervals.

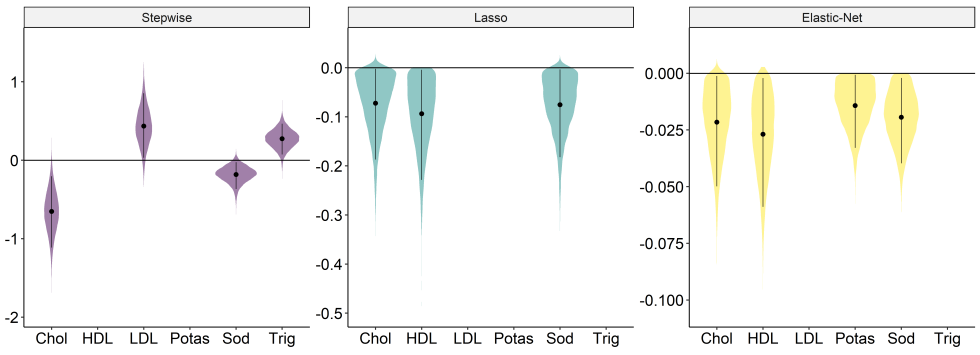
Table 2. Coefficient estimates and bootstrapped means with 90% confidence intervals for each model. Coefficients scale is in log-odds and models were fitted to standardized variables.

	Stepwise		Lasso		Elastic-Net	
Names	**Estimate**	**Bootstrap [90% C.I.]**	**Estimate**	**Bootstrap [90% C.I.]**	**Estimate**	**Bootstrap [90% C.I.]**
(Intercept)	-1.2412**	-1.2564 [-1.4277; -1.0925]	-1.2364 (80.58)	-1.2280 [-1.3867; -1.0720]	-1.2407 (80.86)	-1.2288 [-1.3969; -1.0713]
HDL	-	-	-0.0641 (89.65)	-0.0912 [-0.2165; -0.0053]	-0.0173 (24.16)	-0.0269 [-0.0595; -0.0021]
LDL	0.4288.	0.4423 [0.0446; 0.8491]	-	-	-	-
Trig	0.2747*	0.2718 [0.0666; 0.4651]	-	-	-	-
Chol	-0.6458*	-0.6610 [-1.1060; -0.2245]	-0.0023 (0.26)	-0.0711 [-0.1856; -0.0020]	-0.0078 (0.88)	-0.0212 [-0.0489; -0.0012]
Sodium	-0.1711.	-0.1826 [-0.3658; -0.0217]	-0.0418 (34.9)	-0.0754 [-0.1882; -0.0032]	-0.0111 (9.3)	-0.0197 [-0.0403; -0.0025]
Potassium	-	-	-	-	-0.0041 (6.5)	-0.0142 [-0.0332; -0.0007]

"**", "*" and "." represent significative results at 0.01, 0.05 and 0.1 respectively.

In parenthesis is the percentage represented by the coefficient when compared to the full model's coefficient.

Despite the confidence intervals for Lasso and Elastic-Net having limits relatively close to zero, all confidence intervals indicated significative 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 significative effects for LDL and Triglycerides, the effect of HDL is significative 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 significative 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.



Source: Author

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

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

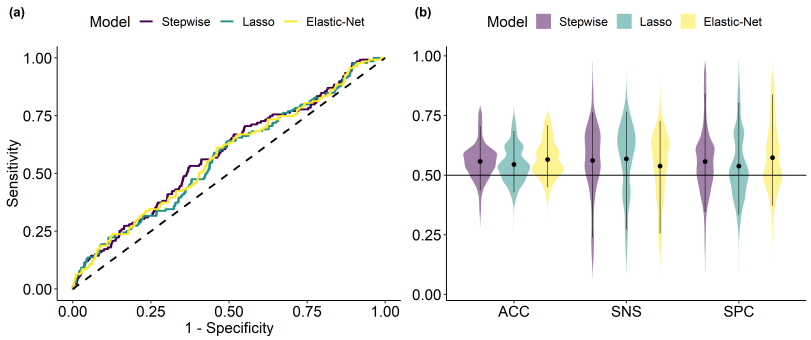
al. (2018). One should note the difference between regularized estimates for Total Cholesterol and their bootstrap counterparts in Table 2 meaning that, according to the bootstrap estimates, the regularization techniques (Lasso and Elastic-Net) estimated a coefficient far from the average of its empirical 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 doesn't 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. (2010) 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.

Performance evaluation

All models showed significative effects for variables relating them to the risk of absenteeism and also, 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 of 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 it's 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.



Source: Author

Fig 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 significative 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 due to the fact that Accuracy is nothing more than a weighted average of the latter two.

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.

Conclusion

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 on 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 significative effects for several variables. However, the detected effects of 5 out of 6 significative 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 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 also use of longitudinal frameworks might reveal stronger effects of the selected variables as well as actually significant discrimination power.

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

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

2. Altman, D. G., and Bland, J. M. Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal* (1994) **308(6943)**, 1552.

3. Arlot, S., and Celisse, A. A survey of cross-validation procedures for model selection. *Statistics surveys* (2010) **4**, 40–79.

4. Bjørklund, G., Dadar, M., Pen, J. J., Chirumbolo, S., and Aaseth, J. Chronic fatigue syndrome (CFS): Suggestions for a nutritional treatment in the therapeutic approach. *Biomedicine and Pharmacotherapy* (2019) **109**, 1000–1007.

5. Bou-Holaigah, I., Rowe, P. C., Kan, J., and Calkins, H. The relationship between neurally mediated hypotension and the chronic fatigue syndrome. *Jama* (1995) **274(12)**, 961–967.

6. Bronikowski, C. M., Weng, A., Furst, J. D., and Raicu, D. S. Prediction of chronic fatigue syndrome using decision tree-based ensemble methods. *WorldComp* (2011) textbf1-6.

7. Cihan, P. A Comparison of Five Methods for Missing Value Imputation in Data Sets. *International Scientific and Vocational Studies Journal* (2018) **2(2)**, 80–85.

8. Colquhoun, D., and Senn, S. Is NADH effective in the treatment of chronic fatigue syndrome?. *Annals of allergy, asthma and immunology: official publication of the American College of Allergy, Asthma, and Immunology* (2000) **84(6)**, 639–640.

9. 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* (2018) **114**, 45–48.

10. Costa, G. Shift work and health: current problems and preventive actions. *Safety and health at Work* (2010) **1(2)**, 112–123.

11. Dechene, L. Chronic fatigue syndrome: influence of histamine, hormones and electrolytes. *Medical hypotheses* (1993) **40(1)**, 55–60.

12. De Lorenzo, F., Xiao, H., Mukherjee, M., Harcup, J., Suleiman, S., Kadziola, Z., and Kakkar, V. V. Chronic fatigue syndrome: physical and cardiovascular deconditioning. *QJM: monthly journal of the Association of Physicians* (1998) **91(7)**, 475–481.

13. Efron, B. Computers and the theory of statistics: thinking the unthinkable. *SIAM review* (1979) **21(4)**, 460–480.

14. Efron, B. Better bootstrap confidence intervals. *Journal of the American statistical Association* (1987) **82(397)**, 171–185.

15. Efron, B., and Tibshirani, R. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association* (1997) **92(438)**, 548–560.

16. Efroymson, M. A. Multiple regression analysis. *Mathematical methods for digital computers* (1960) **1**, 191–203.

17. Fawcett, T. An introduction to ROC analysis. *Pattern recognition letters* (2006) **27(8)**, 861–874.

18. Fukuda, K., Straus, S. E., Hickie, I., Sharpe, M. C., Dobbins, J. G., and Komaroff, A. International Chronic Fatigue Syndrome Study Group: The chronic fatigue syndrome: a comprehensive approach to its definition and study. *Ann Intern Med* (1994) **121(12)**, 953–959.

19. Graham, J. W. Missing data analysis: Making it work in the real world. *Annual review of psychology* (2009) **60**, 549–576.

20. Hanley, J. A., and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* (1982) **143(1)**, 29–36.

21. Heinze, G., Wallisch, C., and Dunkler, D. Variable selection â€“ a review and recommendations for the practicing statistician. *Biometrical Journal* (2018) **60(3)**, 431–449.

22. Henry, L. and Wickham, H. purrr: Functional Programming Tools *R package version 0.3.2.* (2019) **https://CRAN.R-project.org/package=purrr**

23. Hocking, R. R. The analysis and selection of variables in linear regression. *Biometrics* (1976) **32**, 1–49.

24. Hoerl and Kennard Ridge regression *Encyclopedia of Statistical Sciences* (1988) **Vol. 8**

25. Holmes, G. P., Kaplan, J. E., Gantz, N. M., ... and Brus, I. Chronic fatigue syndrome: a working case definition. *Ann Intern Med* (1988) **108(3)**, 387–389.

26. Huang, L. C., Hsu, S. Y., and Lin, E. A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data. *Journal of Translational Medicine* (2009) **7(1)**, 81–88.

27. Jason, L. A., Richman, J. A., Rademaker, A. W., ... and Plioplys, S. Testing equivalence simultaneously for location and dispersion of two normally distributed populations. *Biometrical Journal* (1999) **36**, 643–660.

28. Jason, L. A. T. H. S., and Njoku, M. G. C. The face of CFS in the US. *CFIDS Chronicle* (2006) **1**, 16–21.

29. Kennedy, G., Underwood, C., and Belch, J. J. F. Physical and functional impact of chronic fatigue syndrome/myalgic encephalomyelitis in childhood. *Pediatrics* (2010) **125(6)**, 1324–1330.

30. King, G., and Zeng, L. Logistic regression in rare events data. *Political analysis* (2001) **9(2)**, 137–163.

31. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai* (1995) **14(2)**, 1137–1145.

32. Kuhn, M., Wing J., Weston, S., ... and Hunt, T. caret: Classification and Regression Training. *R package version 6.0-84.* (2019) **https://CRAN.R-project.org/package=caret**

33. Kulothungan, V., Ramakrishnan, R., Subbiah, M., and Raman, R. Risk Score Estimation of Diabetic Retinopathy: Statistical Alternatives using Multiple Logistic Regression. *Journal of Biometrics and Biostatistics* (2014) **5(5)**, 1–6.

34. Lerner, A. M., Goldstein, J., Chang, C. H., ... and 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* (1997) **6**, 327–333.

35. Litleskare, S., Rortveit, G., Eide, G. E., Hanevik, K., Langeland, N., and Wensaas, K. A. Prevalence of irritable bowel syndrome and chronic fatigue 10 years after Giardia infection. *Clinical Gastroenterology and Hepatology* (2018) **16(7)**, 1064–1072.

36. Maloney, E. M., Boneva, R. S., Lin, J. M. S., and Reeves, W. C. Chronic fatigue syndrome is associated with metabolic syndrome: results from a case-control study in Georgia. *Metabolism* (2010) **59(9)**, 1351–1357.

37. Müller, K. and Wickham, H. tibble: Simple Data Frames. *R package version 2.1.1.* (2019) **https://CRAN.R-project.org/package=tibble**

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

39. Nagy-Szakal, D., Barupal, D. K., Lee, B., ... and Levine, S. Insights into myalgic encephalomyelitis/chronic fatigue syndrome phenotypes through comprehensive metabolomics. *Scientific reports* (1994) **8(1)**, 10056–10067.

40. Nakata, A., Haratani, T., Takahashi, M., ... and Araki, S. Association of sickness absence with poor sleep and depressive symptoms in shift workers. *Chronobiology International* (1994) **21(6)**, 899–912.

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

42. Nijs, J., De Becker, P., Demanet, C., McGregor, N. R., Englebienne, P., Verhas, M., and De Meirleir, K. Monitoring a hypothetical channelopathy in chronic fatigue syndrome: preliminary observations. *Journal of Chronic Fatigue Syndrome* (2003) **11(1)**, 117–133.

43. Pavlou, M., Ambler, G., Seaman, S. R., Guttmann, O., Elliott, P., King, M., and Omar, R. Z. How to develop a more accurate risk prediction model when there are few events. *Bmj* (2015) **351**, h3868.

44. Provost, F. J., Fawcett, T., and Kohavi, R. The case against accuracy estimation for comparing induction algorithms. *ICML* (1998) **98**, 445-453.

45. Rowe, P., and Calkins, H. Neurally mediated hypotension and chronic fatigue syndrome. *The American journal of medicine* (1998) **105(3)**, 15–21.

46. Rubin, D. B. Inference and missing data. *Biometrika* (1976) **63(3)**, 581–592.

47. R Core Team R: A language and environment for statistical computing *R Foundation, Viena, Austria* (2019) **https://www.R-project.org/**

48. Samaha, E., Lal, S., Samaha, N., and Wyndham, J. Psychological, lifestyle and coping contributors to chronic fatigue in shift-worker nurses. *Journal of advanced nursing* (2007) **59(3)**, 221–232.

49. Schmitt, P., Mandel, J., and Guedj, M. A comparison of six methods for missing data imputation. *Journal of Biometrics and Biostatistics* (2015) **6(1)**, 1–6.

50. Servaes, P., Verhagen, S., and Bleijenberg, G. Determinants of chronic fatigue in disease-free breast cancer patients: a cross-sectional study. *Annals of oncology* (2002) **13(4)**, 589–598.

51. Shen, J., Botly, L. C., Chung, S. A., Gibbs, A. L., Sabanadzovic, S., and Shapiro, C. M. TFatigue and shift work. *Journal of sleep research* (2006) **15(1)**, 1–5.

52. Sterne, J. A., White, I. R., Carlin, J. B., ... and Carpenter, J. R. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj* (2009) **338**, b2393.

53. Stekhoven, D. J., and Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* (2011) **28(1)**, 112–118.

54. Stekhoven, D. J. missForest: Nonparametric Missing Value Imputation using Random Forest *R package version 1.4.* (2013) **https://CRAN.R-project.org/package=missForest**

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

56. Tomic, S., Brkic, S., Maric, D., and Mikic, A. N. Lipid and protein oxidation in female patients with chronic fatigue syndrome. *Archives of medical science: AMS* (2012) **8(5)**, 886–892.

57. Useche, S. A., Ortiz, V. G., and Cendales, B. E. Stress-related psychosocial factors at work, fatigue, and risky driving behavior in bus rapid transport (BRT) drivers. *Accident Analysis and Prevention* (2017) **104**, 106–114.

58. Vanwinckelen, G., and Blockeel, H. On estimating model accuracy with repeated cross-validation. *21st Belgian-Dutch Conference on Machine Learning* (2012) **39–44**

59. Vaughan, D. and Matt Dancho, M. furrr: Apply Mapping Functions in Parallel using Futures. *R package version 0.1.0* (2018) **https://CRAN.R-project.org/package=furrr**

60. Waljee, A. K., Mukherjee, A., Singal, A. G., ... and Higgins, P. D. Comparison of imputation methods for missing laboratory data in medicine. *BMJ open* (2013) **3(8)**, e002847.

61. Wickham, H. ggplot2: Elegant Graphics for Data Analysis *Springer-Verlag New York* (2016)

62. Wickham, H. and Henry, L. tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions *R package version 0.8.3* (2019) **https://CRAN.R-project.org/package=tidyr**

63. Williams, Y. J., Jantke, R. L., and Jason, L. A. Chronic fatigue syndrome: case Definitions and diagnostic Assessment. *New York State psychologist* (2014) **26(4)**, 41–49.

64. Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. Data Mining: Practical machine learning tools and techniques *Morgan Kaufmann* (2016)

65. Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* (2005) **67(2)**, 301–320.

ANEXO A

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
COMITÊ DE ÉTICA EM PESQUISA
Campus Universitário - Morro do Cruzeiro - ICEB-II, Sala 29
35400-000 - Ouro Preto - MG - Brasil
Fone (31) 3559-1368 Fax (31) 3559-1370
Email: [illegible]

OFÍCIO CEP Nº. 074/2011, de 17 de outubro de 2011.

Ilmo. Sr.
Prof. Dr. Raimundo Marques do Nascimento Neto
DECME/EF/UFOP

Senhor Pesquisador,

É com prazer que comunicamos a **Aprovação**, pelo Comitê de Ética em Pesquisa da Universidade Federal de Ouro Preto, de seu projeto intitulado "*Síndrome Metabólica em Trabalhadores da Mineração do Estado de Minas Gerais*" (CAAE: 0018.0.238.000-11).

Atenciosamente,

Prof. Dr. André Talvani Pedrosa
Vice-Coordenador do Comitê de Ética em Pesquisa/UFOP