Universidade de Aveiro
2023

GUILHERME SILVA RODRIGUES

DECODING THE GENETIC ARCHITECTURE OF FUNCTIONAL STATUS IN COPD: LASSO AND ITS DERIVATIVES FOR FEATURE SELECTION

DESCODIFICANDO A ARQUITETURA GENÉTICA DO ESTADO FUNCIONAL NA DPOC: LASSO E OS SEUS DERIVADOS PARA SELEÇÃO DE VARIÁVEIS

Universidade de Aveiro
2023

GUILHERME SILVA RODRIGUES

DECODING THE GENETIC ARCHITECTURE OF FUNCTIONAL STATUS IN COPD: LASSO AND ITS DERIVATIVES FOR FEATURE SELECTION

DESCODIFICANDO A ARQUITETURA GENÉTICA DO ESTADO FUNCIONAL NA DPOC: LASSO E OS SEUS DERIVADOS PARA SELEÇÃO DE VARIÁVEIS

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Estatística Médica, realizada sob a orientação científica da Doutora Vera Afreixo, Professora Associada com Agregação do Departamento de Matemática da Universidade de Aveiro e coorientação científica da Doutora Alda Marques, Professora Coordenadora com Agregação da Escola Superior de Saúde da Universidade de Aveiro.

Esta dissertação integra o projeto “CENTR(AR): pulmões em andamento”, financiado pelo Programa de Parcerias para o Impacto, Portugal Inovação Social, mediante o Progarama Operacional Inclusão Social e Emprego (POISE-03-4639-FSE-000597), a Fundação para a Ciência e Tecnologia (SFRH/BD/148738/2019) e o Programa Operacional e Competitividade e Internacionalização (COMPETE 2020 – POCI-01-0145-FEDER-007628; UIDB/04501/2020).

LAB Laboratório de Investigação e Reabilitação Respiratória
essua universidade de aveiro escola superior de saúde
universidade de aveiro ibimed instituto de biomedicina
estga
AVEIRO CÂMARA MUNICIPAL
ESTARREJA MUNICÍPIO
Oliveira do Bairro NO CORAÇÃO DA BAIRRADA
montemor-o-velho MUNICÍPIO
Cofinanciado por:
FCT Fundação para a Ciência e a Tecnologia
PO ISE PROGRAMA OPERACIONAL INCLUSÃO SOCIAL E EMPREGO
Portugal INOVAÇÃO SOCIAL
PORTUGAL 2020
LabEx DRIIHM
ohm Estarreja
UNIÃO EUROPEIA Fundo Social Europeu

O júri

Presidente	Prof. Doutor Pedro Miguel Ferreira de Sá Couto Professor Auxiliar da Universidade de Aveiro
Vogais	
Arguente principal	Doutor João Firmino Domingues Barbosa Machado Professor Associado Convidado em Regime Laboral da Universidade de Aveiro
Orientador	Prof. Doutora Vera Mónica Almeida Afreixo Professora Auxiliar com Agregação da Universidade de Aveiro

Agradecimentos

O primeiro, e o maior de todos, vai para a minha mãe, pelo amor, dedicação e sacrifício.

O segundo, para a minha avó, que é mãe duas vezes; obrigado pelo teu ouvido amigo, pelos conselhos de quem já viveu e sabe mais do que eu, e pelo abraço que só nós sabemos.

Obrigado, professora Vera Afreixo, por descomplicar o complicado, por estar sempre disposta e disponível, e, acima de tudo, pelo voto de confiança.

À professora Alda Marques, expresso o meu mais sincero e profundo agradecimento por ter me acolhido desde os meus primeiros passos, por todas as portas que abriu para mim, por acreditar no meu potencial e pela sua incansável paciência.

Obrigado às duas pela vossa compreensão e por me ajudarem a superar todas as dificuldades, mas acima de tudo, por aquilo que me ensinaram.

Agradeço ao Rui por todas as vezes em que me ajudou sem esperar nada em troca.

À Joana, pela amizade e companheirismo diários, a qualquer hora.

À Patrícia, por todas as vezes em que ela me apoiou e segurou as pontas por mim, além da sua paciência em lidar com minha desorganização.

A todas as pessoas do Lab3R, sem exceção, pelo trabalho que fazem com e pelas pessoas com doença respiratória crónica, e por tudo o que aprendi convosco.

Aos doentes, que são o verdadeiro motivo por trás de todo este trabalho.

Um agradecimento especial à professora Gabriela Moura, por me ter recebido de braços abertos, e ao Miguel Pinheiro por toda a ajuda técnica.

palavras-chave

Doença pulmonar obstrutiva crónica; Força muscular; Capacidade funcional; Estudo de associação genética; Métodos de seleção de variáveis; Least absolute shrinkage and selection

resumo

Introdução: A doença pulmonar obstrutiva crónica (DPOC) é um problema de saúde pública que causa incapacidade e mortalidade significativa. Pessoas com DPOC sofrem frequentemente com fraqueza muscular periférica e redução da capacidade funcional, o que afeta o seu bem-estar e aumenta a sua dependência de terceiros. A genética pode desempenhar um papel nestas manifestações, mas a sua análise é ainda desafiante. Métodos de regressão penalizada, como Lasso e suas derivações, oferecem uma abordagem alternativa tanto para a seleção de variáveis quanto para a estimativa de parâmetros em dados genómicos (de grande dimensão).
Objetivo: Este estudo teve como objetivo investigar a possível associação entre polimorfismos genéticos (SNPs) e estado funcional em indivíduos com DPOC. Além disso, o estudo abordou o desafio da seleção de variáveis em dados de grande dimensão.
Métodos: O teste de sentar e levantar de um minuto e o teste de marcha de seis minutos foram utilizados para avaliar a capacidade funcional. A força de preensão manual e a contração voluntária máxima do quadricípite foram medidas para determinar a força muscular periférica. Os indivíduos foram classificados utilizando análise de componentes principais e análise de cluster hierárquico. O resultado da classificação obtida por meio do cluster hierárquico foi considerado como fenótipo, assumindo um modelo genético aditivo. Foi realizado um estudo de associação genética (GWAS) baseado em regressão logística não ajustada (univariada). Foram aplicados e comparados quatro modelos de regressão penalizada: regressão Lasso logística, bem como duas versões ponderadas do Lasso, conhecidas como Lasso relaxado e Lasso adaptativo, e um modelo elastic net. Métricas de pseudo-R^2 foram usadas para avaliar o desempenho dos modelos, permitindo a comparação do ajuste do modelo. Todas as análises estatísticas foram realizadas utilizando os softwares PLINK 1.9 e R (versão 4.3.0).
Resultados: Um total de 211 pessoas com DPOC foram incluídos na análise, sendo que dados de genotipagem estavam disponíveis para 167 deles. O Cluster A era composto principalmente por indivíduos mais jovens e do sexo masculino, com menos sintomas e maior incidência de obesidade. Em contraste, o Cluster B era composto principalmente por indivíduos mais velhos, incluindo uma proporção maior de mulheres, que referiram maior severidade dos sintomas, menor qualidade de vida relacionada à saúde e apresentaram pontuações mais baixas de força muscular e capacidade funcional em comparação com o Cluster A. Nenhum polimorfismo alcançou o nível de significância na regressão logística GWAS. Os estimadores Lasso e Lasso relaxado exibiram resultados idênticos, identificando 8 variáveis (incluindo a constante do modelo) com coeficientes diferentes de zero. Em contraste, o modelo elastic net resultou num conjunto maior de 52 variáveis com coeficientes diferentes de zero. Por fim, a abordagem Lasso adaptativo selecionou um total de 99 variáveis com coeficientes diferentes de zero.
Conclusão: Este estudo destaca a presença de 99 polimorfismos genéticos associados à deterioração funcional na DPOC. O conjunto de covariáveis selecionadas constitui agora um bom ponto de partida para futuras investigações científicas, incluindo validação externa e estudos funcionais, para validar os resultados e elucidar os mecanismos biológicos subjacentes.

keywords

Chronic obstructive pulmonary disease; Muscle strength; Functional capacity; Genome-wide association study; Feature selection; Least absolute shrinkage and selection

abstract

Introduction: Chronic obstructive pulmonary disease (COPD) is a public health problem that causes significant disability and mortality. People with COPD often suffer from peripheral muscle weakness and reduced functional capacity, which affects their own well-being and increases their dependence on others. It is possible that genetics play a role in these manifestations, but analysis remains difficult. Penalised regression methods, such as Lasso and its derivatives, offer a promising approach for both feature selection and parameter estimation for analysing high-dimensional data.

Aim: The aim of this study was to investigate the potential association between single nucleotide polymorphisms (SNPs) and functional status in individuals with COPD. In addition, the study addressed the challenge of feature selection in high-dimensional data.

Methods: Functional capacity was assessed using the one-minute sit-stand test and the six-minute walk test. Peripheral muscle strength was measured using handgrip strength and quadriceps maximum voluntary contraction. Patients were classified using principal component analysis and hierarchical cluster analysis. An unadjusted (univariate) logistic regression-based genome-wide association study (GWAS) was performed. Cluster membership was considered as the phenotype, assuming an additive genetic model. In addition, four penalised regression models were applied and compared: the (ordinary) logistic Lasso regression as well as two weighted versions of Lasso, namely relaxed Lasso and adaptive Lasso, and finally an elastic net model. Pseudo-R^2 metrics were used to evaluate the performance of the models, allowing comparison of model fit. All statistical analyses were performed using PLINK 1.9 and R statistical software (version 4.3.0).

Results: A total of 211 patients with COPD were included in the analysis, with genotyping data available for 167 of them. Cluster A consisted mainly of younger, male patients who had fewer symptoms and a higher incidence of obesity. Cluster B consisted primarily of older individuals, including a higher proportion of women, who reported higher symptom severity, lower health-related quality of life, and exhibited lower muscle strength and functional capacity scores compared to Cluster A. No SNP reached genome-wide significance in the logistic regression GWAS. The Lasso and relaxed Lasso estimators showed identical results, identifying 8 variables (including the model intercept) with non-zero coefficients. In contrast, the elastic net model yielded a larger set of 52 variables with non-zero coefficients. Finally, the adaptive Lasso approach selected a total of 99 variables with non-zero coefficients.

Conclusion: This study highlights the presence of 99 genetic polymorphisms associated with functional impairment in COPD. These selected covariates provide a starting point for further scientific investigation, including external validation and laboratory-based functional studies, to validate the findings and understand the underlying biological pathways.

Table of Contents

List of Abbreviations

CAT – COPD assessment test

COPD – Chronic obstructive pulmonary disease

CV – Cross-validation

FEV1 – Forced expiratory volume in the first second

FVC – Forced vital capacity

GOLD – Global initiative for obstructive lung disease

GWAS – Genome-wide association studies

LASSO – Least absolute shrinkage and selection

mMRC – Medical research council modified dyspnoea score

OLS – Ordinary least squares

PCA – Principal component analysis

RSS – Residual sum of squares

SGRQ – Saint george’s respiratory questionnaire

SNPs – Single-nucleotide polymorphisms

STROBE – Strengthening the reporting of observational studies in epidemiology statement

List of Figures

List of Tables

Introduction

Chronic obstructive pulmonary disease

Chronic obstructive pulmonary disease (COPD) is a heterogeneous lung disease characterised by chronic respiratory symptoms (dyspnoea, cough, sputum) due to airway and/or alveolar abnormalities that result in persistent, often progressive, airflow obstruction[1]. COPD is the third leading cause of death worldwide and represents a significant individual, social and economic burden[1-4]. In 2019, approximately 212.3 million people worldwide had COPD, resulting in 3.3 million deaths and making it the fourth and third leading cause of disability-adjusted life years among adults aged 50-74 years and 75 years and older, respectively[4, 5]. Cigarette smoking, air pollution, occupational exposure to (in)organic dusts, chemicals and fumes, and household air pollution are important risk factors for the development of COPD[1, 6]. The disease starts in the respiratory system, but those affected often suffer debilitating symptoms such as peripheral muscle weakness and impaired functional capacity, which affect their quality of life[7, 8], prevent them from participating in activities of daily living[9, 10] and lead to dependence on informal caregivers[11, 12]. Muscle weakness and reduced functional capacity are present early in disease development[13, 14], have been shown to be associated with poorer health outcomes and independently predict healthcare utilisation and mortality in COPD[15-20]. The annual cost of COPD has been found to be higher in patients with COPD and muscle weakness than in patients without this weakness[20]. The aetiology is multifactorial and includes extrinsic factors such as smoking, physical inactivity/deconditioning, malnutrition and systemic corticosteroid use as well as intrinsic factors such as hypoxia, hypercapnia, inflammation and oxidative/nitrosative stress[21]. Recently, there is evidence that single-nucleotide polymorphisms (SNPs), i.e., base pair variations in the genome that are common in the population, may also play a role in the pathological processes underlying functional impairment in COPD[21, 22], although the available literature remains limited.

Genome-wide association studies

Genome-wide association studies (GWAS) are commonly used to detect associations between genetic variants and common diseases or traits in a population[23, 24]. In a GWAS, the phenotype and genotype of n subjects are measured, where $y = (y_1, \dots, y_n)$ is the phenotype of interest, such as height, blood pressure, or disease status, which can be either quantitative or dichotomous[23, 24]. Let x_{ij} denote the number of minor alleles the ith subject has at the jth SNP. Suppose two alleles of a SNP are A and a. A dominant model for A translates the genotypes (AA, Aa, aa) into (1, 1, 0) - the presence of the A allele increases the risk of disease for the genotypes AA and Aa to the same extent, compared to the baseline risk for aa; an additive or codominant model encodes (AA, Aa, aa) as (2, 1, 0) - one additional copy of the A allele increases disease risk; a recessive genetic model for A encodes (AA, Aa, aa) as (1, 0, 0) - two copies of the A allele are required to express the phenotypic traits associated with that allele. The results of a GWAS are usually summarised in a Manhattan plot of all individual P-values, and a SNP is considered significant if its P-value is less than or equal to a predefined significance level α[23, 24]. A fixed P-value threshold of 5×10^{-8} is the standard for reporting genome-wide association at a minor allele frequency $\geq 5\%$ for populations of European ancestry; this threshold was derived based on the number of independent tests taking into account the linkage disequilibrium of the genome[25-27]. This approach is simple and easy to use, and software such as PLINK[28] can process and analyse SNP array data from across the genome in a computationally efficient manner[29]. Nonetheless, due to their univariate nature, GWAS suffer from information losses and are often statistically underpowered due to heavy multiple testing (to control for type I errors)[24]. This will be exacerbated in the future due to the increasing density of genotype arrays, along with the advancement of imputation methods[30] and the expansion of imputation reference panels[31]. Multiple regression, on the other hand, includes all relevant predictors and can thus provide a more accurate description of the influence of the covariates on the outcome, resulting in less residual variance. However, these high-dimensional data pose a challenge for multivariate statistical analyses, including conventional regression models, or make them practically unfeasible because there are more independent variables than observations/patient samples[32].

Dimensionality reduction

Currently, the two most commonly used dimensionality reduction techniques are feature extraction and feature selection[33]. Feature extraction maps the high-dimensional data to a low-dimensional subspace by converting the raw data into numerical features that can be processed while retaining as much information as possible from the original dataset. Feature selection reduces dimensionality by selecting a subset of the original features according to a specific criterion. This process eliminates irrelevant and redundant data, resulting in improved model performance and reduced computation time. Unlike feature extraction, where the original data is altered by mathematical transformations, feature selection preserves the original meaning of the features in a dataset.

Feature selection: penalised regression

A general model of multiple linear regression is written as follows:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \tag{1}$$

where $y_i \in \mathbb{R}$ is the dependent variable, $x_i = (x_{i1}, \dots, x_{ip})$ is a p-dimensional vector of independent variables, $\beta = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$ is a vector of regression coefficients, $\beta_0 \in \mathbb{R}$ is the constant or model intercept, and ε_i is the random (unobservable) error term. The linear model can be fitted using the ordinary least squares (OLS) method and the parameters (β_0, β) are estimated by minimising the residual sum of squares (RSS):

$$\mathrm{RSS} = \left[\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\right]. \tag{1.1}$$

Logistic regression is another commonly used model, but this time to estimate the probability of a binary response based on the value of multiple predictor variables. The logistic regression model represents the class-conditional probabilities by a linear function of the predictors:

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum_{j=1}^{p} x_{ij}\,\beta_j = \beta_0 + \beta^T x. \tag{2}$$

Logistic models (2) are fitted by maximising the likelihood (i.e., maximum likelihood estimation) or by minimising the negative log likelihood, which can be written as follows:

$$\mathcal{L} = -\ln(L(\beta)) = -\sum_{i=1}^{n}\big(y_i\ln(p_i) + (1-y_i)\ln(1-p_i)\big). \tag{2.1}$$

In the high-dimensional setting, where the number of features p is larger than the sample size, these models cannot be used without modification; when $p > n$, any linear model is overparameterised and regularisation is required to achieve a stable fit[34]. Let us consider the standard linear regression framework. It is known that the OLS solution for estimating the coefficient vector corresponds to (3).

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \cdot \frac{1}{n}\sum_{i=1}^{n} x_i y_i = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'y \tag{3}$$

The OLS estimator has some desirable properties, e.g., it is an unbiased and consistent estimator (minimum variance). However, if the predictor variables are highly correlated, the OLS method often leads to unsatisfactory or even incorrect estimates, which are often inflated (in absolute values) and in extreme cases even lead to sign reversals[35]. Moreover, the sample covariance matrix is singular (and therefore cannot be inverted) if $p > n$, but a valid covariance matrix must be positive-definite[36]. Art Hoerl and Bob Kennard introduced ridge regression[37] in 1970, which was later extended to the high-dimensional setting and can be written as follows:

$$\hat{\beta}^R = \operatorname{argmin}\left[\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right]. \tag{4}$$

In equation (4), the first term represents the RSS, while the second term corresponds to a ℓ_2 penalty on the regression coefficients, and $\lambda > 0$ serves as a tuning parameter that controls the amount of shrinkage. Ridge regression goes back to earlier work, in particular that of Tikhonov, a Soviet mathematician and geophysicist; Tikhonov, and later Hoerl and Kennard, found that adding a single positive constant to the diagonals makes the matrix behave more like an orthogonal system[37]. Put simply, adding the regularisation parameter lambda (λ) to singular or nearly singular matrices leads to a matrix for which an inverse exists[37, 38]. Unfortunately, similar to least squares, ridge regression cannot perform variable selection, which means that all coefficients are assigned non-zero estimates[37].

When dealing with high-dimensional data, one assumes sparsity of the coefficient vector, which refers to the phenomenon that an underlying data structure can usually be explained by a few features out of many[34]. Least absolute shrinkage and selection (Lasso)[39], introduced by Robert Tibshirani in 1996, is an example of an embedded feature selection method where feature selection and parameter estimation (shrinkage) are performed simultaneously to improve the predictive accuracy and interpretability of the regression model. The Lasso for linear regression can be written as follows:

$$\hat{\beta}^L = \operatorname{argmin}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right]. \tag{5}$$

In equation (5), the first term represents the RSS, while the second term corresponds to a ℓ_1 penalty on the regression coefficients, and $\lambda > 0$ serves as a tuning parameter that controls both the number of variables selected and the extent to which their estimated coefficients are shrunk to zero[38, 39]. The set of predictor variables selected by the Lasso estimator is denoted by $M_\lambda \ = \{1 \leq k \leq p | \beta_k^L \neq 0\}$; some coefficients are set to zero and thus excluded from the selected model, while all variables in the selected model are shrunk towards zero compared to the least squares solution.

Hyperparameter tuning cross-validation

High-dimensional data contains many noisy and redundant features that not only significantly increase the learning and inference time of the algorithm, but also greatly reduce the performance of the model. An appropriate selection of λ is required to achieve a good balance between simplicity and selection accuracy[40]. Small values of λ allow the model to fine-tune to the noise in each data set, resulting in a large variance. Conversely, a large value of lambda causes the weighting parameters to approach zero, resulting in a large bias (bias-variance trade-off). Lambda is data dependent and can be calculated using a cross-validation method (CV) or a generalised information criterion (i.e., Akaike and/or Bayesian information criteria). Often, the optimal hyperparameter λ is determined to minimise the mean square error with 5 or 10-fold CV in the training data set[34]. In K-fold CV, the data are first divided into K equally sized (or nearly equally sized) segments or folds. Then K iterations of training and validation are performed, such that in each iteration a different fold of the data is held out for validation, while the remaining $K-1$ folds are used for learning (i.e., we apply the Lasso to the training data using different lambda values, using each fitted model to predict the outcome in the test set and record the mean square error). We can also achieve smaller, simpler models with comparable predictive performance by applying a simple rule called the "one standard error rule"[34]. Instead of choosing the value of the tuning parameter that minimises the CV error curve, the one-standard-error rule chooses the value of the tuning hyperparameter corresponding to the simplest model whose CV error is within one standard error of the minimum.

Ordinary lasso: drawbacks and inconveniences

The Lasso also has some disadvantages that need to be addressed. Lasso shrinkage reduces estimates of large coefficients, leading to an asymptotic bias that could negatively affect risk estimates[41]. A two-step procedure is usually used to correct for bias: Lasso regression is used to select variables, and then a least squares estimator is obtained over the selected variables. Other options are weighted versions of the Lasso method based on iterative schemes[42]. Another pitfall of Lasso relates to how it handles correlated data: if there is a group of variables between which the pairwise correlations are very high, Lasso tends to (randomly) select only one variable from the group, which can lead to confounding and loss of information[41, 43, 44].

From now on we will focus on the relaxed Lasso, the adaptive Lasso and the elastic net and briefly describe each model. For simplicity, the models will be presented as in linear regression. However, to extend it to generalised linear models, e.g., logistic regression, the RSS term in the objective function must be replaced by a negative log-likelihood term (2.1).

Lasso derivatives

Relaxed Lasso

Relaxed Lasso[45] performs model selection and shrinkage estimation with the two hyperparameters λ and ϕ:

$$\hat{\beta}^{RL} = \operatorname{argmin}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\{\beta_j \cdot 1_M\})^2 + \phi\lambda\sum_{j=1}^{p}|\beta_j|\right] \quad (5.1)$$

where $\lambda > 0$, $\phi \in (0, 1]$, M_λ is the set of predictor variables selected by the Lasso estimator and 1_M is the indicator function on the set of variables $\beta_j \cdot 1_M = \begin{cases} 0, & j \notin M_\lambda \\ \beta_j, & j \in M_\lambda \end{cases}$, for $j \in \{1, \dots, p\}$.

The hyperparameter λ controls the number of predictors with non-zero coefficients in the model, while the hyperparameter ϕ determines the degree of shrinkage of the selected predictors. If $\phi = 1$, the Lasso and relaxed Lasso estimators are identical. For $\phi < 1$ the shrinkage of the coefficients in the selected model is reduced compared to the ordinary Lasso estimator. Relaxed Lasso produces a sparse model that avoids excessive shrinkage for non-zero coefficients and outperforms Lasso when the number of predictors is large relative to the sample size[45]. Moreover, the number of selected coefficients in relaxed Lasso is generally much smaller than in ordinary Lasso without affecting the prediction accuracy, since in Lasso the number of selected variables is often large and contains many noise variables, especially when the signal-to-noise ratio is high (i.e., given estimated regression coefficients and residual variance, the signal-to-noise ratio is defined as the ratio of estimated signal variance to estimated noise variance)[45].

Adaptive Lasso

In the case of the ordinary Lasso and the relaxed Lasso, the shrinkage is constant, irrespective of the size of the parameter to be estimated, which leads to undesirable properties regarding the prediction of the resulting estimator. In adaptive Lasso, each covariate is weighted differently when penalised and readjusted at each iteration step until convergence[42]. First, a weight vector $\hat{w}$ is estimated from the ridge regression, given by $\hat{w} = 1/|\hat{\beta}^R|^\gamma$, $\gamma > 0$. Second, for this weight vector $w = (w_1, \dots, w_p)^T$ the adaptive Lasso is formulated:

$$\hat{\beta}^{AL} = \operatorname{argmin}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\hat{w}_j|\beta_j|\right] \quad (5.2)$$

Elastic net

The elastic net[44] was introduced by Zou and Hastie in 2005 to extend the ordinary Lasso and improve some of its limitations. It minimises the RSS and a regularisation term that is a mixture of ℓ_1 and ℓ_2 penalties:

$$\hat{\beta}^{EN} = \operatorname{argmin}\left[\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{i=1}^{p}\left(\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha\left|\beta_j\right|\right)\right] \qquad (5.3)$$

where $\lambda > 0$ is a tuning parameter and $\alpha \in [0,1]$ is a higher level hyperparameter. The elastic net is particularly useful when $p > n$ or in a situation where there are many correlated predictor variables[44, 46]. When α increases from 0 to 1, for a given λ, the sparsity of the solution of the above formula increases monotonically from 0 to the Lasso solution.

The aim of this work was, therefore, to find out whether there are polymorphisms associated with functional limitations in COPD, while addressing the problem of feature selection in high-dimensional data.

Methodology

A cross-sectional secondary analysis was conducted on data collected in “GENetic and clinIcAL markers in COPD trajectory” – GENIAL (PTDC/DTP-PIC/2284/2014), “Pulmonary Rehabilitation Innovation and Microbiota in Exacerbations of COPD” – PRIME (PTDC/SAU-SER/28806/2017; POCI-01-0145-FEDER-028806), “Revitalizing Respiratory Rehabilitation” – 3R (SAICT-POL/23926/2016; POCI-01-0145-FEDER-016701), and CENTR(AR) (POISE-03-4639-FSE-000597). Ethical approval was obtained from the ethics committees of the Administração Regional de Saúde do Centro, I.P. (3NOV'2016:64/2016), Centro Hospitalar do Baixo Vouga (22MAR'2017:777638), Unidade Local de Saúde de Matosinhos (17FEB'2017:10/CE/JAS), Centro Hospitalar do Médio Ave (27AUG'2018), Hospital Distrital da Figueira da Foz (18JUL'2017) and from the Portuguese Data Protection Authority (8828/2016). Each participant signed a written informed consent form prior to data and sample collection. This work is reported according to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement[47].

Participants

The analysis included individuals aged ≥ 18 years diagnosed with COPD according to the Global Initiative for Obstructive Lung Disease (GOLD) criteria (FEV1/FVC < 0.7 after bronchodilation) and clinically stable (i.e., no exacerbations or change in medication) in the previous month. Those with other respiratory conditions or a clinical condition that may have affected test performance were excluded. Eligible participants were identified by their physicians in hospitals or primary healthcare centres during routine outpatient visits.

Data collection

Data collection was carried out by trained physiotherapists and assessments took place in the Respiratory Research and Rehabilitation Laboratory (Lab3R), School of Health (ESSUA), University of Aveiro, or in partner institutions. At baseline, sociodemographic, anthropometric and clinical data as well as saliva samples were collected from the participants. Lung function was either measured with a portable spirometer (Micro-Lab 3535, CareFusion, Kent, United Kingdom) or gathered from the respective medical records. The quadriceps muscle strength on the dominant side was assessed using a handheld dynamometer (microFET®2 Digital Handheld Dynamometer, Hoggan Scientific LLC, Salt Lake City, Utah); the patient, in a seated position, was instructed to extend their knee against resistance applied to the anterior tibia for a period of 6 seconds[48]. The isometric strength of the lower limbs determined with the handheld dynamometer proved to be valid compared to the Biodex™ dynamometer (ICC: 0.79-0.94; Pearson's r = 0.73-0.90) and showed excellent intra- and inter-rater reliability (ICC: 0.97-0.98; 0.83-0.95)[49]. Handgrip strength was measured with a hydraulic dynamometer on the dominant hand (Baseline® 12-0241 LiTE Hydraulic Hand Dynamometer, Fabrication Enterprises Inc., White Plains, New York) with the patient seated, the elbow flexed at 90°, the forearm in a neutral position, and the wrist between 0 and 30° of dorsiflexion, according to the American Society of Hand Therapists[50, 51]. Three repetitions were performed for both muscle strength tests to ensure less than 10% variability between measurements. The highest value was recorded and used for analysis. Functional capacity was measured using the six-minute walk test and the 1-minute sit to stand test according to the technical standards established by the European Respiratory Society/American Thoracic Society[52, 53]. In the six-minute walk test, the patient is asked to walk as far as possible along a 30-metre, low-traffic, straight, flat corridor over a 6-minute period[54]. Two tests are performed to account for a possible learning effect, and the longer distance covered in metres is recorded and used for further analysis. A recent meta-analysis found a moderate to strong correlation (Pearson's r = 0.65) between 6MWT distance and peak oxygen uptake in patients with COPD during cardiopulmonary exercise testing[55]. The 1-minute sit-to-stand test is a quick and simple assessment conducted on a standard-height chair (46 cm) without armrests[56]. Participants were asked to stand up and sit down as many times as possible within one minute at their own pace, without using their arms for support. The number of complete repetitions was recorded. It has demonstrated high reliability (ICC: 0.90-0.99) and validity (Pearson's r = 0.716) as a tool for evaluating functional capacity in patients with COPD[57, 58]. Other assessments included the Medical Research Council Dyspnoea Score (mMRC)[59], the COPD Assessment Test (CAT™)[60], and the Saint George's Respiratory Questionnaire (SGRQ)[61]. The SGRQ is an important tool for assessing health-related quality of life in patients with respiratory diseases. It consists of three main parts: a symptom questionnaire, an activity limitation questionnaire and a daily life impact questionnaire. The score ranges from 0 to 100, with higher scores indicating more limitations. The CAT™ is a multidimensional questionnaire with 8 items to assess the impact of COPD on health status. Scores range from 0 to 40, with higher scores indicating greater symptom burden. The GOLD 2023 report[1] suggests a CAT cut-off score of ≥ 10 points for classifying patients as highly symptomatic, but a more recent cut-off score of ≥ 18 points is now recommended[62, 63]. The mMRC dyspnoea score is a simple and reliable measure used to assess the severity of activity-related breathlessness in individuals with chronic lung disease. It ranges from 0 (no breathlessness) to 4 (very severe breathlessness), and a cut-off score of ≥ 2 is used to classify patients into those with low or high symptom burden[1].

Statistical analysis

Hierarchical cluster analysis

Hierarchical cluster analysis (using Ward's method) was performed to classify the patients[64]. A principal component analysis was conducted to reduce the correlation between the independent variables (such as the six-minute walk test, one-minute sit-to-stand test, and muscle strength measurements). This is particularly important for hierarchical clustering using Euclidean distance as the distance measure, as it is sensitive to the level of correlation between variables compared to other distance measures. Principal components were retained until the cumulative percentage of the total variance reached at least 70%[65]. Multivariate outliers were identified by comparing the Mahalanobis distance to a chi-square distribution with a critical value of 0.99. Any multivariate outliers detected were excluded from the analysis. Differences between the clusters were examined using t-tests for independent samples, with the homogeneity of variance assessed using Bartlett's test. If the variances between groups were unequal, the Welch test was employed. The Shapiro-Wilk test was used to assess the normality assumption, and if violated, the Mann-Whitney U test was used instead. For categorical variables, differences in distribution between groups were evaluated using the Chi-square test/Fisher's Exact test or the two-sample proportion Z-test. All statistical analyses were conducted using R statistical software (version 4.3.0), with a significance level set at $P < 0.05$.

Genomic data pre-processing and imputation

Genotyping was performed using the Infinitum Global Screening Array-24 v1.0 and according to the Illumina Infinitum HTS assay protocol. Sample and SNP quality control was performed with GenomeStudio 2.0 and PLINK 1.9 software using standard protocols[28, 66, 67]. Samples with a genotyping rate of less than 95%, sex discrepancies, a divergent ancestry from the study cohort, a higher heterozygosity rate than expected or an unreported relatedness to another study participant were excluded[66, 67]. SNPs with a missing genotype rate of more than 5%, a deviation from Hardy-Weinberg equilibrium or a minor allele frequency below 5% were also excluded[66, 67]. Imputation was performed on the Michigan Imputation Server[68]. After imputation, a quality control of the SNPs was carried out again. A total of 167 samples and 7,035,690 SNPs passed quality control. Despite having high call rates, the genotype data still had 2.9% missing values after undergoing quality control. To address this, the bigstatsr R package[69] was employed to impute the genotype matrix using a simple imputation method (mean). The imputed matrix, which contained numerous zero values, was then converted into a sparse matrix format to optimize storage usage and enhance computational efficiency.

Regression analysis

An unadjusted (univariate) logistic regression-based GWAS was conducted in PLINK 1.9, using cluster membership as the phenotype and assuming an additive genetic model. The Lasso regression was performed using the glmnet R package[46], and the optimal hyperparameter λ was determined by minimising the binomial deviance through a 3-fold CV on the training dataset. It is important to add that the solutions are computed in a descending sequence of values for λ, starting with the smallest value $\lambda_{\max}$ for which the entire vector $\hat{\beta} = 0$; the aim is to select the minimum value $\lambda_{\min} = \epsilon\lambda_{\max} > 0$, and construct a sequence of K values of λ decreasing from $\lambda_{\max}$ to $\lambda_{\min}$ on the logarithmic scale. The default options of the glmnet R package were used, i.e., K = 100 and $\epsilon = 0.001$. For the elastic net model, the hyperparameter α was selected via CV from a range of 0 to 1 with 0.05 increments, choosing the α value that resulted in the smallest binomial deviance. Similarly, the gamma value for the relaxed Lasso was determined by CV using a standard grid with the values 0, 0.25, 0.5, 0.75 and 1, and the model with the optimal pair (λ, γ) was selected. In the adaptive Lasso, weights were constructed using ridge regression and then multiplied by lambda for differential shrinkage. The goodness of fit of the penalised regression models was tested using the McFadden pseudo-R^2 measure[70]. In contrast to the least squares R^2, the log-likelihood-based pseudo-R^2 value does not represent the proportion of variance explained, but the improvement in model likelihood over a null model (intercept only)[70].

Results

The dataset included 1087 records of people who had participated in all of the above projects, of which 946 were deemed eligible according to the projects' eligibility criteria, but only 727 were adults with a primary diagnosis of COPD. Of the 727 people, 293 participants remained in the study after excluding those who did not have complete data for the 6-minute walk test, 1-minute sit-to-stand test, and quadriceps and handgrip muscle strength. Of these, 49 were duplicates of the same individuals, 30 had an FEV1/FVC > 0.7 after bronchodilation, and 3 were considered multivariate outliers and therefore excluded. Ultimately, 211 patients with COPD were included in the analysis, of whom 167 had genotyping data.

Two principal components were extracted, which together accounted for 74.5% of the total variance; the NbClust R package[71] was used to determine the optimal data partition, which was 2 based on a majority rule (Figure 1). Cluster A consisted mainly of younger male patients with fewer symptoms and a higher prevalence of obesity. Cluster B, on the other hand, consisted mainly of older people who reported more severe symptoms and lower health-related quality of life, as well as a higher prevalence of women compared to cluster A. In addition, compared to cluster A, cluster B was characterised by individuals with lower muscle strength and functional capacity scores. Detailed characteristics of the total sample and the subgroups in clusters A and B can be found in Table 1. Of the 167 individuals for whom genotyping data were available, 110 belonged to cluster A and 57 to cluster B.

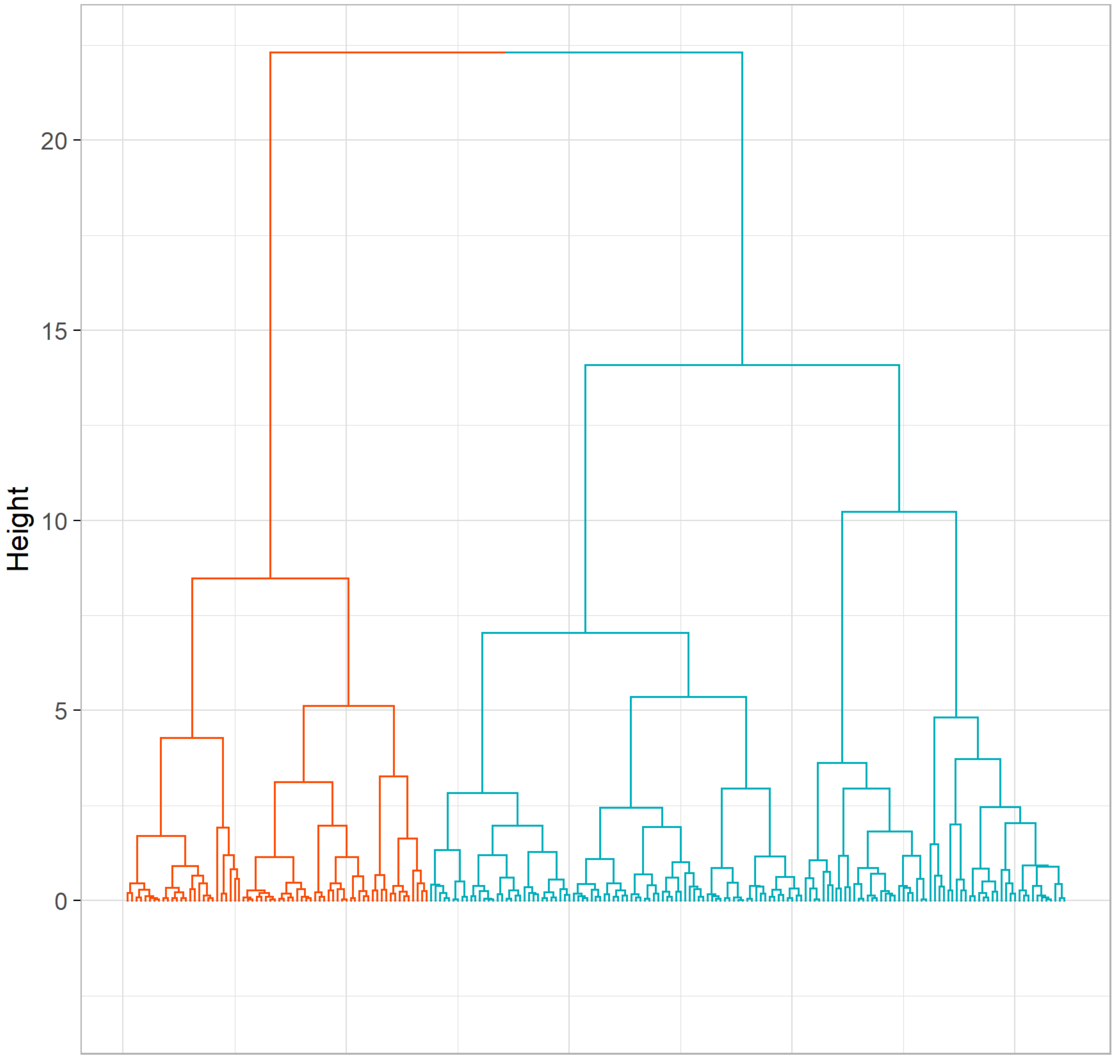

Figure 1. Hierarchical clustering dendrogram of principal component analysis-transformed data based on the 6-minute walk test, the 1-minute sit-to-stand test, and quadriceps and handgrip muscle strength in patients with chronic obstructive pulmonary disease.

Table 1. Descriptive statistics for the total sample and for each cluster of patients with chronic obstructive pulmonary disease.

	Total sample (n = 211)	Cluster A (n = 139)	Cluster B (n = 72)	P-value
Age, years	67.88 (7.98)	66.38 (7.33)	70.78 (8.43)	<0.001
Female, %	44 (20.9)	15 (10.8)	29 (40.3)	<0.001
BMI, kg/m^2	26.26 [23.40 – 30.30]	26.6 [23.7 – 30.8]	26.2 [22.8 – 28.7]	0.1157
BMI > 30 kg/m^2, %	56 (26.5)	44 (31.7)	12 (16.7)	0.0298
BMI < 21.75 kg/m^2, %	28 (13.3)	19 (13.7)	9 (12.5)	0.9814
FEV1, % of predicted	53.0 [40.5 – 67.0]	53.0 [42.0 – 68.0]	53.5 [38.0 – 66.0]	0.3667
FEV1/FVC	0.54 [0.44 – 0.62]	0.54 [0.44 – 0.62]	0.56 [0.43 – 0.62]	0.7464

FVC, % of predicted	80.0 [66.2 – 95.0]	82.03 (22.54)	79.22 (21.86)	0.383
Smoking status				
Current smoker, %	36 (17.1)	22 (15.8)	14 (19.4)	0.2375
Former smoker, %	136 (64.5)	95 (68.3)	41 (56.9)	
Never smoker, %	39 (18.5)	22 (15.8)	17 (23.6)	
Pack-years	43.5 [21.6 – 74.8]	45.0 [27.0 – 80.0]	40.0 [17.1 – 61.5]	0.1392
GOLD grade				
GOLD 1, %	27 (12.8)	21 (15.1)	6 (8.3)	0.5437
GOLD 2, %	91 (43.1)	57 (41.0)	34 (47.2)	
GOLD 3, %	75 (35.5)	49 (35.3)	26 (36.1)	
GOLD 4, %	18 (8.53)	12 (8.6)	6 (8.3)	
CAT, points	13.0 [8.0 – 20.0]	12.0 [7.0 – 17.0]	18.0 [11.0 – 22.2]	<0.001
CAT ≥ 10, %	146 (69.2)	88 (63.3)	58 (80.6)	0.0157
CAT ≥ 18, %	68 (32.2)	31 (22.3)	37 (51.4)	<0.001
mMRC				
mMRC ≥ 2, %	116 (55.0)	63 (45.3)	53 (73.6)	<0.001
SGRQ Activities	60.26 [42.97 – 73.80]	55.10 [41.21 – 67.69]	72.44 [53.62 – 85.87]	<0.001
SGRQ Symptoms	44.57 [31.53 – 62.74]	42.48 [29.50 – 56.73]	53.95 [37.38 – 69.16]	0.0096
SGRQ Impact	29.15 [14.90 – 47.11]	23.97 [12.07 – 42.37]	39.06 [24.48 – 56.63]	<0.001
SGRQ Total	42.83 [27.08 – 57.43]	37.47 [23.83 – 53.27]	52.11 [38.51 – 63.51]	<0.001
6MWT, m	420.0 [344.5 – 492.0]	462.0 [401.5 – 508.0]	313.0 [225.5 – 408.5]	<0.001
1-minute STS, reps	23.0 [19.0 – 28.0]	26.0 [22.0 – 31.0]	18.0 [13.75 – 21.25]	<0.001
QMS, kgF	31.10 [26.0 – 36.2]	33.50 [30.80 – 38.20]	25.0 [20.23 – 28.45]	<0.001
Handgrip muscle strength, kg	34.11 (9.48)	38.28 (7.36)	26.07 (7.83)	<0.001

Notes: Data presented as mean (standard deviation) or median [1st quartile; 3rd quartile] for continuous variables and number (percentage) for categorical variables. **Abbreviations:** 1-minute STS, one-minute sit-to-stand test; 6MWT, six-minute walk test; BMI, body mass index; CAT, COPD Assessment Test; FEV1, forced expiratory volume in the first second; FVC, forced vital capacity; GOLD, Global Initiative for Obstructive Lung Disease; mMRC, Medical Research Council modified dyspnoea score; QMS, quadriceps muscle strength; SGRQ, Saint George's Respiratory Questionnaire.

In the logistic regression GWAS, no single nucleotide polymorphism (SNP) reached the level required for genome-wide significance (Figure 2).

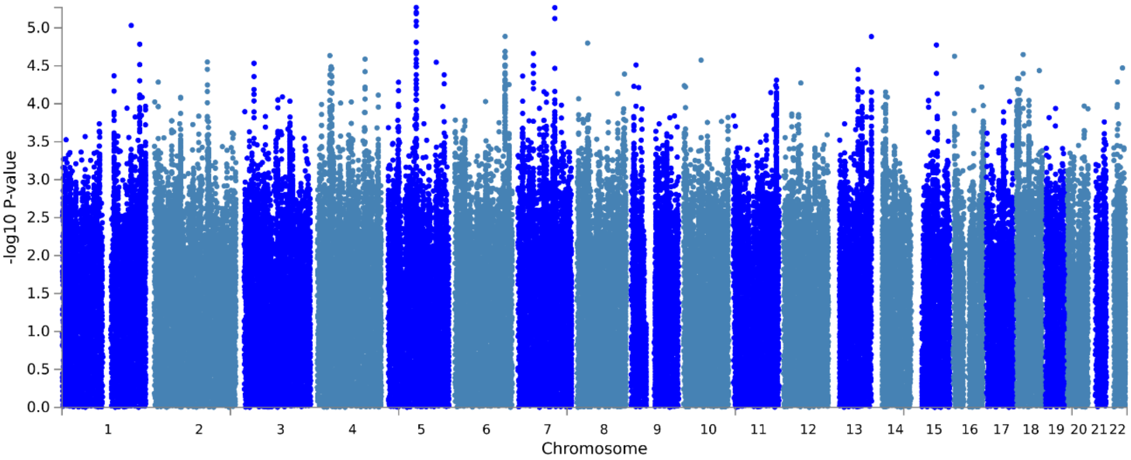

Figure 2. Manhattan plot. The P values were obtained by unadjusted logistic regression analysis assuming an additive genetic model. The x-axis shows the genomic coordinates of the tested SNPs and the y-axis shows the –log10 P value of their association.

For the Lasso, the optimal hyperparameter lambda was 0.1116, resulting in a binomial deviance of 1.29. The model returned 8 variables (including the model intercept) with non-zero coefficients (Figure 3). As with the relaxed Lasso, the optimal pair of hyperparameters that minimised model deviance was gamma = 1 and lambda = 0.1116, yielding identical results to the Lasso estimator (Figure 3).

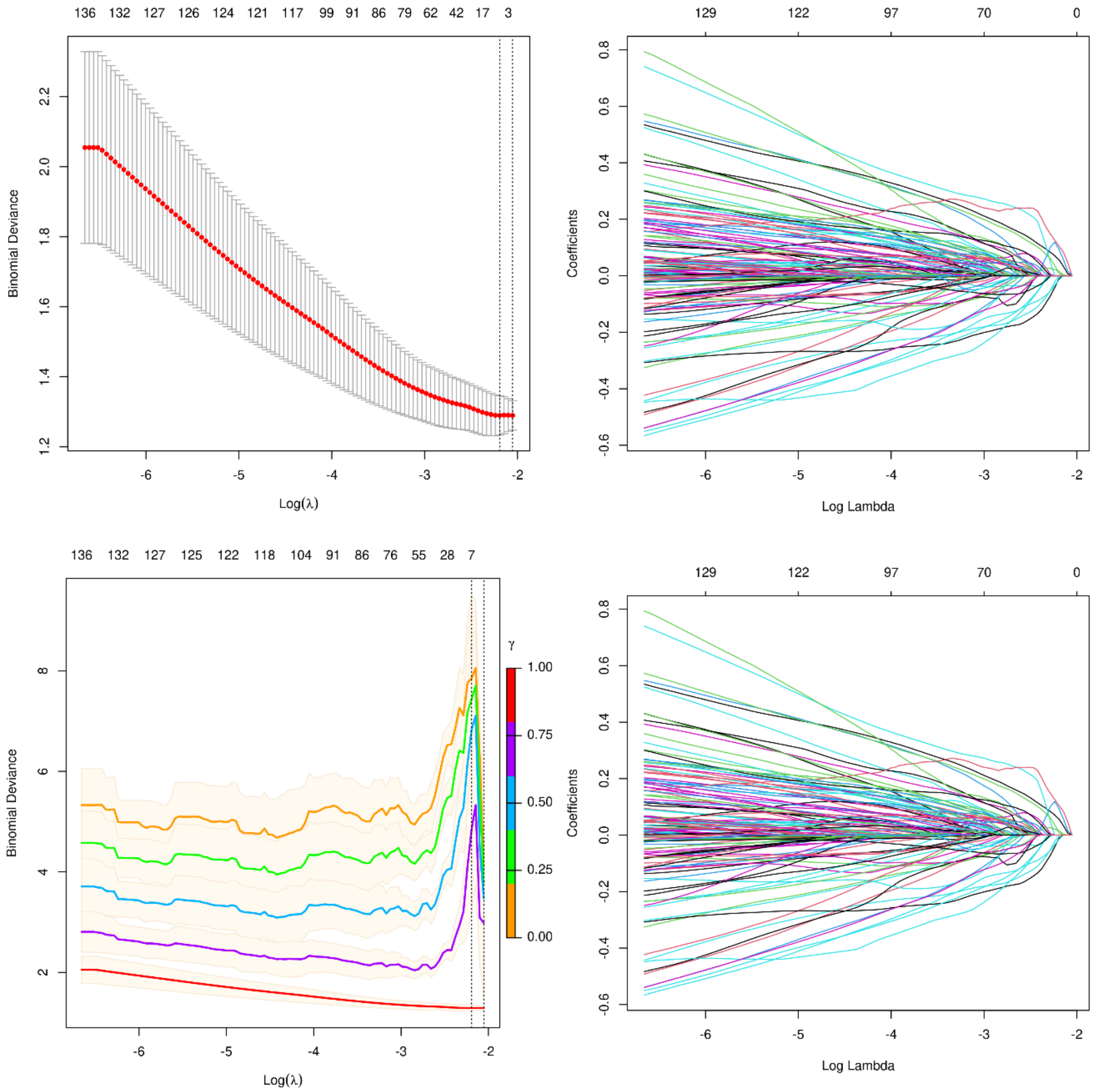

Figure 3. The right panel shows the profiles or regularisation paths of the coefficients for Lasso (top) and relaxed Lasso (bottom) when the tuning parameter lambda varies. The upper left panel shows the cross-validation deviance for each lambda value for Lasso; the dashed vertical lines correspond to the minimum lambda value, and the right line corresponds to the rightmost point of the curve within a standard error of the minimum. The lower left panel shows the cross-validation deviance for the relaxed Lasso for different gamma values. The number of non-zero coefficients is shown at the top of each graph.

For the elastic net model, 0.55 was determined as the alpha value that minimises the deviance (1.26), resulting in a lambda value of 0.1609 and the selection of 52 variables with non-zero coefficients (Figure 4).

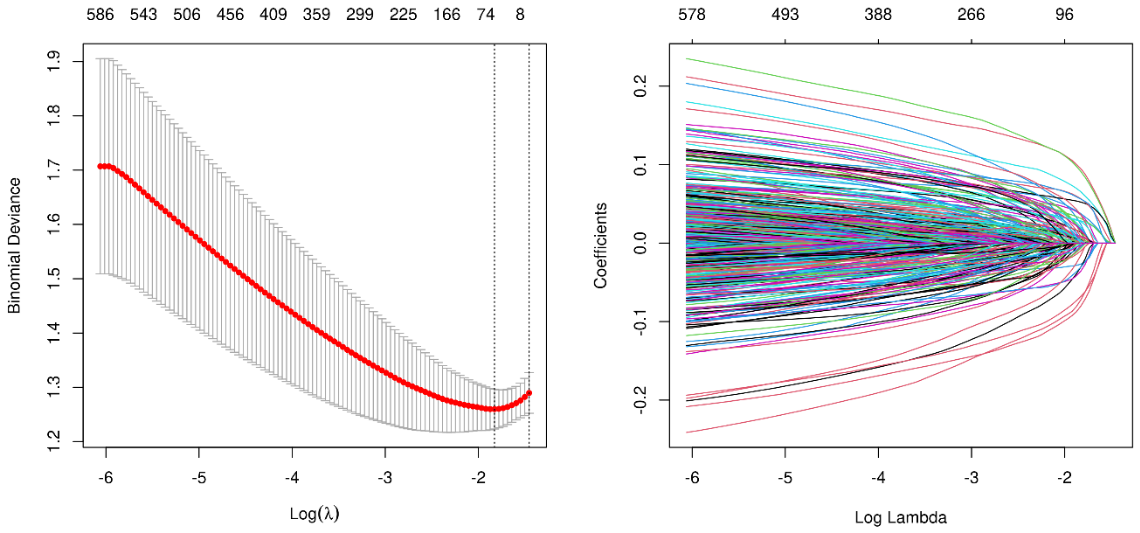

Figure 4. The right panel shows the profiles or regularisation paths of the coefficients for the elastic net when the tuning parameter lambda varies. The left panel shows the cross-validation deviance for each lambda value for the elastic net; the dashed vertical lines correspond to the minimum lambda value, and the right line corresponds to the rightmost point of the curve within a standard error of the minimum. The number of non-zero coefficients is shown at the top of each graph.

A lambda value of 2.326[22] was chosen for the adaptive Lasso, resulting in a model deviance of 0.5675 and the selection of 99 variables with non-zero coefficients (Figure 5). Pseudo-R^2 metrics were calculated for each of the penalised regression models: 0.0558 for Lasso and relaxed Lasso, 0.1969 for elastic net and 0.9920 for adaptive Lasso. These results show that adaptive Lasso has the best model fit among all the models tested.

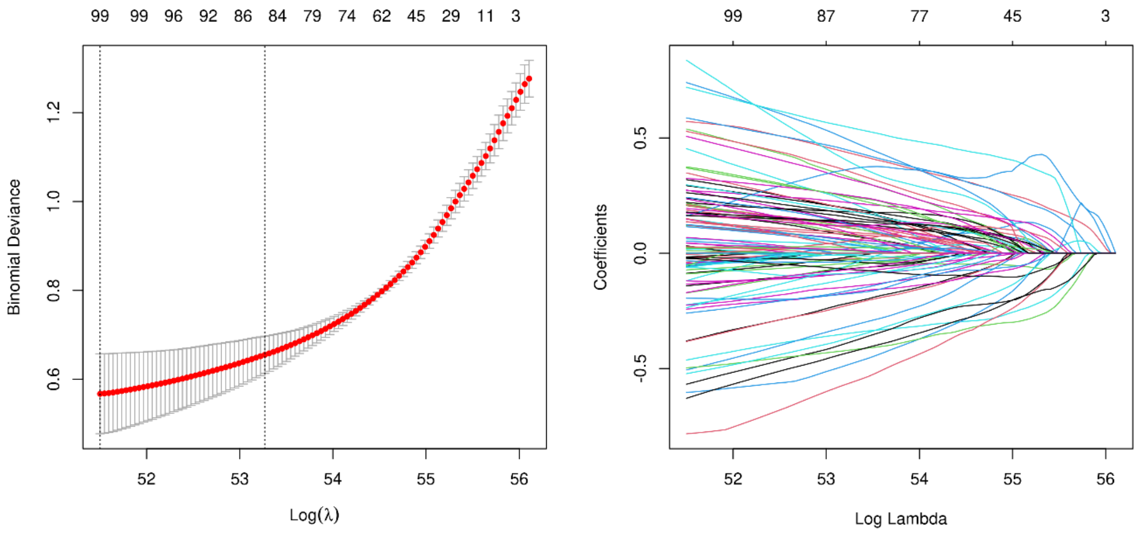

Figure 5. The right panel shows the profiles or regularisation paths of the coefficients for the adaptive Lasso when the tuning parameter lambda varies. The left panel shows the cross-validation deviance for each lambda value for the adaptive Lasso; the dashed vertical lines correspond to the minimum lambda value, and the right

line corresponds to the rightmost point of the curve within a standard error of the minimum. The number of non-zero coefficients is shown at the top of each graph.

Discussion

This study investigated whether common genetic variants are associated with an increased risk of functional impairment, for which 99 SNPs with non-zero coefficients were found. The penalised regression yielded results between 8 and 99 non-zero coefficients compared to the standard approach, univariate GWAS regression, where no SNP was considered genome-wide significant. These results highlight the difficulty of achieving genome-wide significance with GWAS, especially in studies with small sample sizes and due to correction by multiple testing[24, 75]. Alternative, less stringent thresholds have been proposed[72, 73]; however, when the P-value threshold is relaxed, it is observed that the detection rate increases, but this results in an estimated 8-18% of additional loci being incorrectly identified[74]. Moreover, for studies of moderate size, the genome-wide significance threshold is the preferred and recommended criterion[74]. Results obtained from both the Lasso and relaxed Lasso estimators were identical, mainly because the hyperparameter gamma was strongly influenced by the signal-to-noise ratio. In low signal-to-noise ratio scenarios, coefficient shrinkage is preferred and the optimal procedure involves selecting a gamma value close to one to solve noisy problems[45]. These results are consistent with simulation studies evaluating the performance of the relaxed Lasso under various constraints[45]. Both models resulted in the selection of eight variables, but the pseudo-R^2 (0.0558) shows that the improvement of the model compared to the null model is rather small, which can be explained by the size of the estimated parameters, which are close to zero for the chosen lambda value. This underlines one of the well-known disadvantages of the Lasso, namely that it can shrink the coefficients too much[41, 45]. The elastic net model selected a total of 52 variables, including the model intercept, while the ordinary and relaxed Lasso estimators selected only eight variables. This discrepancy can be easily understood by looking at the properties of the respective models. The elastic net allows both the automatic selection of features and the selection of groups of correlated predictors (i.e., strongly correlated predictors are usually included in or removed from the model together)[38, 44]. In contrast, Lasso tends to split such variables and randomly selects only one variable from the group[38, 41]. The adaptive Lasso model selected a total of 99 predictor variables and outperformed all other models tested in terms of the number of variables selected and goodness of fit (pseudo-R^2 = 0.9920). The adaptive Lasso and elastic net are extensions of the Lasso, and both incorporate the ℓ_2 penalty of ridge regression, which is advantageous in the presence of multicollinearity[76, 77]. Unlike the ordinary Lasso and relaxed Lasso models, which apply a constant shrinkage regardless of the size of the parameters, the adaptive lasso assigns different weights to each covariate in the penalty term (derived from ridge regression), thus penalising smaller coefficients more severely, while larger coefficients receive a smaller penalty[42]. While the debiasing process of the adaptive Lasso has the potential to improve the prediction error of the model, it also introduces a higher risk of overfitting the data (i.e., a model performs very well on training data but poorly on new data; the model fits sample-specific random variations in the data rather than the true underlying relationships between variables). Some limitations of this study must be acknowledged. First, the generalisability of the results is limited by the small sample size and the lack of external validation. In addition, there was a sex imbalance, which may affect conclusions. Another drawback of this study is that the genotype matrix was imputed using a simple imputation method. More robust methods such as XGBoost (eXtreme Gradient Boosting) are now available for R[69], but due to the size of the genotype matrix this would require an enormous amount of computation until the matrix is fully imputed. Despite these limitations, this study successfully employs a robust methodology to overcome the challenges of high-dimensional data and the limitations of univariate analysis in GWAS. It also provides a promising starting point for future scientific research, contributing to a body of work that suggests that some individuals with COPD have a genetic predisposition to functional impairment[21].

Conclusion

This study found SNP associations with functional impairment in COPD, but variable selection varied across models. The adaptive Lasso performed best, selecting 99 variables. Future research, including external validation and functional studies (variant-to-function), is needed to confirm the results and elucidate the underlying biological pathways.

References

1. Global Initiative for Chronic Obstructive Lung Disease (GOLD), *Global Strategy for the Diagnosis, Management, and Prevenetion of Chronic Obstructive Pulmonary Disease*. 2023.
2. Adeloye, D., P. Song, Y. Zhu, H. Campbell, A. Sheikh, and I. Rudan, *Global, regional, and national prevalence of, and risk factors for, chronic obstructive pulmonary disease (COPD) in 2019: a systematic review and modelling analysis.* Lancet Respir Med, 2022. **10**(5): p. 447-58.
3. *Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017.* Lancet Respir Med, 2020. **8**(6): p. 585-96.
4. Momtazmanesh, S., S.S. Moghaddam, S.-H. Ghamari, E.M. Rad, N. Rezaei, P. Shobeiri, et al., *Global burden of chronic respiratory diseases and risk factors, 1990-2019: an update from the Global Burden of Disease Study 2019.* eClinicalMedicine.
5. *Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.* Lancet, 2020. **396**(10258): p. 1204-22.
6. Safiri, S., K. Carson-Chahhoud, M. Noori, S.A. Nejadghaderi, M.J.M. Sullman, J. Ahmadian Heris, et al., *Burden of chronic obstructive pulmonary disease and its attributable risk factors in 204 countries and territories, 1990-2019: results from the Global Burden of Disease Study 2019.* BMJ, 2022. **378**: p. e069679.
7. Sepúlveda-Loyola, W., C. Osadnik, S. Phu, A.A. Morita, G. Duque, and V.S. Probst, *Diagnosis, prevalence, and clinical impact of sarcopenia in COPD: a systematic review and meta-analysis.* J Cachexia Sarcopenia Muscle, 2020. **11**(5): p. 1164-76.
8. Holden, M., M. Fyfe, C. Poulin, B. Bethune, C. Church, P. Hepburn, et al., *Handgrip Strength in People With Chronic Obstructive Pulmonary Disease: A Systematic Review and Meta-Analysis.* Phys Ther, 2021. **101**(6).
9. Wong, S.S.L., N. Abdullah, A. Abdullah, S.-M. Liew, S.-M. Ching, E.-M. Khoo, et al., *Unmet needs of patients with chronic obstructive pulmonary disease (COPD): a qualitative study on patients and doctors.* BMC Family Practice, 2014. **15**(1): p. 67.
10. Machado, A., S. Almeida, C. Burtin, and A. Marques, *Giving Voice to People - Experiences During Mild to Moderate Acute Exacerbations of COPD.* Chronic Obstr Pulm Dis, 2022. **9**(3): p. 336-48.
11. Gautun, H., A. Werner, and H. LurÅs, *Care challenges for informal caregivers of chronically ill lung patients: Results from a questionnaire survey.* Scandinavian Journal of Public Health, 2012. **40**(1): p. 18-24.
12. Granados-Santiago, M., R. Romero-Fernández, A. Calvache-Mateo, A. Heredia-Ciuro, J. Martin-Nuñez, L. López-López, et al., *Relationship between patient functionality impairment and caregiver burden: is there a cut off point for the severe COPD patient?* Expert Rev Respir Med, 2023. **17**(3): p. 247-53.
13. Kharbanda, S., A. Ramakrishna, and S. Krishnan, *Prevalence of quadriceps muscle weakness in patients with COPD and its association with disease severity.* Int J Chron Obstruct Pulmon Dis, 2015. **10**: p. 1727-35.
14. Zou, R.H., S.M. Nouraie, H.B. Rossiter, M.L. McDonald, D.L. DeMeo, S. Mason, et al., *Associations Between Muscle Weakness and Clinical Outcomes in Current and Former Smokers.* Chronic Obstr Pulm Dis, 2023. **10**(1): p. 112-21.
15. Swallow, E.B., D. Reyes, N.S. Hopkinson, W.D. Man, R. Porcher, E.J. Cetti, et al., *Quadriceps strength predicts mortality in patients with moderate to severe chronic obstructive pulmonary disease.* Thorax, 2007. **62**(2): p. 115-20.
16. Burtin, C., G. Ter Riet, M.A. Puhan, B. Waschki, J. Garcia-Aymerich, V. Pinto-Plata, et al., *Handgrip weakness and mortality risk in COPD: a multicentre analysis.* Thorax, 2016. **71**(1): p. 86-7.
17. Vaes, A.W., M.A. Spruit, E.H. Koolen, J.C. Antons, M. de Man, R.S. Djamin, et al., *"Can Do, Do Do" Quadrants and 6-Year All-Cause Mortality in Patients With COPD.* Chest, 2022. **161**(6): p. 1494-504.
18. Puhan, M.A., L. Siebeling, M. Zoller, P. Muggensturm, and G. ter Riet, *Simple functional performance tests and mortality in COPD.* Eur Respir J, 2013. **42**(4): p. 956-63.

19. Höglund, J., C. Boström, and J. Sundh, *Six-Minute Walking Test and 30 Seconds Chair-Stand-Test as Predictors of Mortality in COPD - A Cohort Study.* Int J Chron Obstruct Pulmon Dis, 2022. **17**: p. 2461-9.
20. Trantham, L., M.V. Sikirica, S.D. Candrilli, V.S. Benson, D. Mohan, D. Neil, et al., *Healthcare costs and utilization associated with muscle weakness diagnosis codes in patients with chronic obstructive pulmonary disease: a United States claims analysis.* Journal of Medical Economics, 2019. **22**(4): p. 319-27.
21. Henrot, P., I. Dupin, P. Schilfarth, P. Esteves, L. Blervaque, M. Zysman, et al., *Main Pathogenic Mechanisms and Recent Advances in COPD Peripheral Skeletal Muscle Wasting.* Int J Mol Sci, 2023. **24**(7).
22. Yuan, C., D. Chang, G. Lu, and X. Deng, *Genetic polymorphism and chronic obstructive pulmonary disease.* Int J Chron Obstruct Pulmon Dis, 2017. **12**: p. 1385-93.
23. Uffelmann, E., Q.Q. Huang, N.S. Munung, J. de Vries, Y. Okada, A.R. Martin, et al., *Genome-wide association studies.* Nature Reviews Methods Primers, 2021. **1**(1): p. 59.
24. Tam, V., N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre, *Benefits and limitations of genome-wide association studies.* Nature Reviews Genetics, 2019. **20**(8): p. 467-84.
25. Pe'er, I., R. Yelensky, D. Altshuler, and M.J. Daly, *Estimation of the multiple testing burden for genomewide association studies of nearly all common variants.* Genetic Epidemiology, 2008. **32**(4): p. 381-5.
26. Altshuler, D., P. Donnelly, and C. The International HapMap, *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
27. Fadista, J., A.K. Manning, J.C. Florez, and L. Groop, *The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants.* European Journal of Human Genetics, 2016. **24**(8): p. 1202-5.
28. Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* American journal of human genetics, 2007. **81**(3): p. 559-75.
29. Chang, C.C., C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, and J.J. Lee, *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**: p. 7.
30. Li, Y., C. Willer, S. Sanna, and G. Abecasis, *Genotype Imputation.* Annual Review of Genomics and Human Genetics, 2009. **10**(1): p. 387-406.
31. McCarthy, S., S. Das, W. Kretzschmar, O. Delaneau, A.R. Wood, A. Teumer, et al., *A reference panel of 64,976 haplotypes for genotype imputation.* Nature Genetics, 2016. **48**(10): p. 1279-83.
32. Shakeel, N. and T. Mehmood, *Inverse Matrix Problem in Regression for High-Dimensional Data Sets.* Mathematical Problems in Engineering, 2023. **2023**: p. 2308541.
33. Jia, W., M. Sun, J. Lian, and S. Hou, *Feature dimensionality reduction: a review.* Complex & Intelligent Systems, 2022. **8**(3): p. 2663-93.
34. Hastie, T., R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*. 2015: Chapman & Hall/CRC.
35. Hoerl, R.W., *Ridge Regression: A Historical Context.* Technometrics, 2020. **62**(4): p. 420-5.
36. Wang, C., J. Du, and X. Fan, *High-dimensional correlation matrix estimation for general continuous data with Bagging technique.* Machine Learning, 2022. **111**(8): p. 2905-27.
37. Hoerl, A.E. and R.W. Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems.* Technometrics, 1970. **12**(1): p. 55-67.
38. Emmert-Streib, F. and M. Dehmer, *High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection.* Machine Learning and Knowledge Extraction, 2019. **1**(1): p. 359-83.
39. Tibshirani, R., *Regression Shrinkage and Selection Via the Lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 1996. **58**(1): p. 267-88.
40. Fan, Y. and C.Y. Tang, *Tuning parameter selection in high dimensional penalized likelihood.* Journal of the Royal Statistical Society. Series B (Statistical Methodology), 2013. **75**(3): p. 531-52.
41. Freijeiro-González, L., M. Febrero-Bande, and W. González-Manteiga, *A Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates.* International Statistical Review, 2022. **90**(1): p. 118-45.

42. Zou, H., *The Adaptive Lasso and Its Oracle Properties.* Journal of the American Statistical Association, 2006. **101**(476): p. 1418-29.
43. Zou, H. and T. Hastie, *Regularization and Variable Selection via the Elastic Net.* Journal of the Royal Statistical Society. Series B (Statistical Methodology), 2005. **67**(2): p. 301-20.
44. Zou, H. and T. Hastie, *Regularization and Variable Selection Via the Elastic Net.* Journal of the Royal Statistical Society Series B: Statistical Methodology, 2005. **67**(2): p. 301-20.
45. Meinshausen, N., *Relaxed Lasso.* Computational Statistics & Data Analysis, 2007. **52**(1): p. 374-93.
46. Friedman, J.H., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 2010. **33**(1): p. 1 - 22.
47. von Elm, E., D.G. Altman, M. Egger, S.J. Pocock, P.C. Gøtzsche, and J.P. Vandenbroucke, *The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies.* Journal of clinical epidemiology, 2008. **61**(4): p. 344-9.
48. Bohannon, R.W., *Reference values for extremity muscle strength obtained by hand-held dynamometry from adults aged 20 to 79 years.* Arch Phys Med Rehabil, 1997. **78**(1): p. 26-32.
49. Grootswagers, P., A.M.M. Vaes, R. Hangelbroek, M. Tieland, L.J.C. van Loon, and L.C.P.G.M. de Groot, *Relative Validity and Reliability of Isometric Lower Extremity Strength Assessment in Older Adults by Using a Handheld Dynamometer.* Sports Health, 2022. **14**(6): p. 899-905.
50. Roberts, H.C., H.J. Denison, H.J. Martin, H.P. Patel, H. Syddall, C. Cooper, et al., *A review of the measurement of grip strength in clinical and epidemiological studies: towards a standardised approach.* Age and Ageing, 2011. **40**(4): p. 423-9.
51. MacDermid, J., G. Solomon, and K. Valdes, *Clinical Assessment Recommendations*. 3rd ed. 2015: Mount Laurel, NJ: American Society of Hand Therapists.
52. Holland, A.E., M.A. Spruit, T. Troosters, M.A. Puhan, V. Pepin, D. Saey, et al., *An official European Respiratory Society/American Thoracic Society technical standard: field walking tests in chronic respiratory disease.* Eur Respir J, 2014. **44**(6): p. 1428-46.
53. Singh, S.J., M.A. Puhan, V. Andrianopoulos, N.A. Hernandes, K.E. Mitchell, C.J. Hill, et al., *An official systematic review of the European Respiratory Society/American Thoracic Society: Measurement properties of field walking tests in chronic respiratory disease.* European Respiratory Journal, 2014. **44**(6): p. 1447-78.
54. Agarwala, P. and S.H. Salzman, *Six-Minute Walk Test: Clinical Role, Technique, Coding, and Reimbursement.* Chest, 2020. **157**(3): p. 603-11.
55. Chae, G., E.J. Ko, S.W. Lee, H.J. Kim, S.G. Kwak, D. Park, et al., *Stronger correlation of peak oxygen uptake with distance of incremental shuttle walk test than 6-min walk test in patients with COPD: a systematic review and meta-analysis.* BMC Pulmonary Medicine, 2022. **22**(1): p. 102.
56. Ozalevli, S., A. Ozden, O. Itil, and A. Akkoclu, *Comparison of the Sit-to-Stand Test with 6 min walk test in patients with chronic obstructive pulmonary disease.* Respir Med, 2007. **101**(2): p. 286-93.
57. Spence, J.G., J. Brincks, A. Løkke, L. Neustrup, and E.B. Østergaard, *One-minute sit-to-stand test as a quick functional test for people with COPD in general practice.* NPJ Prim Care Respir Med, 2023. **33**(1): p. 11.
58. Reychler, G., E. Boucard, L. Peran, R. Pichon, C. Le Ber-Moy, H. Ouksel, et al., *One minute sit-to-stand test is an alternative to 6MWT to measure functional exercise performance in COPD patients.* Clin Respir J, 2018. **12**(3): p. 1247-56.
59. Bestall, J.C., E.A. Paul, R. Garrod, R. Garnham, P.W. Jones, and J.A. Wedzicha, *Usefulness of the Medical Research Council (MRC) dyspnoea scale as a measure of disability in patients with chronic obstructive pulmonary disease.* Thorax, 1999. **54**(7): p. 581-6.
60. Jones, P.W., G. Harding, P. Berry, I. Wiklund, W.H. Chen, and N. Kline Leidy, *Development and first validation of the COPD Assessment Test.* European Respiratory Journal, 2009. **34**(3): p. 648-54.
61. Jones, P.W., F.H. Quirk, C.M. Baveystock, and P. Littlejohns, *A self-complete measure of health status for chronic airflow limitation. The St. George's Respiratory Questionnaire.* Am Rev Respir Dis, 1992. **145**(6): p. 1321-7.

62. Smid, D.E., F.M.E. Franssen, M. Gonik, M. Miravitlles, C. Casanova, B.G. Cosio, et al., *Redefining Cut-Points for High Symptom Burden of the Global Initiative for Chronic Obstructive Lung Disease Classification in 18,577 Patients With Chronic Obstructive Pulmonary Disease.* J Am Med Dir Assoc, 2017. **18**(12): p. 1097.e11-.e24.

63. Karloh, M., A. Fleig Mayer, R. Maurici, M.M.M. Pizzichini, P.W. Jones, and E. Pizzichini, *The COPD Assessment Test: What Do We Know So Far?: A Systematic Review and Meta-Analysis About Clinical Outcomes Prediction and Classification of Patients Into GOLD Stages.* Chest, 2016. **149**(2): p. 413-25.

64. Ward, J.H., *Hierarchical Grouping to Optimize an Objective Function.* Journal of the American Statistical Association, 1963. **58**(301): p. 236-44.

65. Jollife, I.T. and J. Cadima, *Principal component analysis: a review and recent developments.* Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 2016. **374**(2065).

66. Marees, A.T., H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire, et al., *A tutorial on conducting genome-wide association studies: Quality control and statistical analysis.* International Journal of Methods in Psychiatric Research, 2018. **27**(2): p. e1608-e.

67. Anderson, C.A., F.H. Pettersson, G.M. Clarke, L.R. Cardon, A.P. Morris, and K.T. Zondervan, *Data quality control in genetic case-control association studies.* Nature protocols, 2010. **5**(9): p. 1564-73.

68. Das, S., L. Forer, S. Schönherr, C. Sidore, A.E. Locke, A. Kwong, et al., *Next-generation genotype imputation service and methods.* Nature genetics, 2016. **48**(10): p. 1284-.

69. Privé, F., H. Aschard, A. Ziyatdinov, and M.G.B. Blum, *Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr.* Bioinformatics, 2018. **34**(16): p. 2781-7.

70. Hemmert, G.A.J., L.M. Schons, J. Wieseke, and H. Schimmelpfennig, *Log-likelihood-based Pseudo-R2 in Logistic Regression:Deriving Sample-sensitive Benchmarks.* Sociological Methods & Research, 2018. **47**(3): p. 507-31.

71. Charrad, M., N. Ghazzali, V. Boiteau, and A. Niknafs, *NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.* Journal of Statistical Software, 2014. **61**(6): p. 1 - 36.

72. Hammond, R.K., M.C. Pahl, C. Su, D.L. Cousminer, M.E. Leonard, S. Lu, et al., *Biological constraints on GWAS SNPs at suggestive significance thresholds reveal additional BMI loci.* Elife, 2021. **10**.

73. Lander, E. and L. Kruglyak, *Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results.* Nat Genet, 1995. **11**(3): p. 241-7.

74. Chen, Z., M. Boehnke, X. Wen, and B. Mukherjee, *Revisiting the genome-wide significance threshold for common variant GWAS.* G3 Genes|Genomes|Genetics, 2021. **11**(2).

75. Yang, S., J. Wen, S.T. Eckert, Y. Wang, D.J. Liu, R. Wu, et al., *Prioritizing genetic variants in GWAS with lasso using permutation-assisted tuning.* Bioinformatics, 2020. **36**(12): p. 3811-7.

76. Feig, D.G., *RIDGE REGRESSION: WHEN BIASED ESTIMATION IS BETTER.* Social Science Quarterly, 1978. **58**(4): p. 708-16.

77. Le Cessie, S. and J.C. Van Houwelingen, *Ridge Estimators in Logistic Regression.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1992. **41**(1): p. 191-201.