DL-PRS: a novel deep learning approach to polygenic risk scores

Sijia Huang
University of Pennsylvania
Xiao Ji
GlaxoSmithKline (United States)
Michael Cho
Brigham and Women's Hospital
Jaehyun Joo
University of Pennsylvania
Jason Moore (jhmoore@upenn.edu)
University of Pennsylvania

Research Article

Keywords: COPD, Heterogenous Disease Machine Learning Methods, Gene-gene Interaction Modeling, Predictive Performance

Posted Date: April 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-423764/v1

Abstract

Background COPD is a complex heterogeneous disease influenced by both environmental and genetic risk factors. Traditional genome wide association studies (GWAS) have been successful in identifying many reproducible risk variants of moderate to small effect. Polygenic risk scores (PRS) were developed as way to aggregate risk alleles weighted by their effect size to produce a score which could be used in clinical practice to identify individuals at high risk of disease. A limitation of both GWAS and PRS is that they make the important assumption that the effect of each allele is independent and not modified by other genetic or environmental factors. Machine learning methods such as deep learning (DL) neural networks complement the GWAS and PRS paradigm by making fewer assumptions about the nature of the genetic effects being modeled. For example, the hidden layers of a DL model have the potential to model gene-gene interactions with non-additive effects on disease risk. The goal of the present study was to develop a DL neural network approach to GWAS and PRS and to compare it to the prevailing paradigm based on modeling independent effects. We applied our DL-PRS method to genetic association data from several GWAS studies of chronic obstructive pulmonary disease (COPD).

Results We developed a DL learning algorithm for modeling the relationship between genetic variation from GWAS and risk of COPD in several population-based studies. We then developed a DL-PRS based on nodes and associated weights from the first and second layer of the DL neural network. Our DL-PRS framework has overall satisfactory performance in the prediction of COPD and provides significant contribution to prediction in addition to the current PRS methods. Moreover, regarding the clinical relevance of COPD, our DL-PRS has a consistent and closer relationship regarding individual deciles and lung functions such as FEV1/FVC and predicted FEV1%.

Conclusions Not only does DL-PRS show favorable predictive performance with current benchmark PRS methods, but it also extends the ranges of PRS deciles in predicting different stages of COPD. Moreover, our DL-PRS results were replicated in an independent cohort. This study opens the door to the use of machine learning for developing risk scores from models developed using fewer assumptions about the nature of the genetic effects.

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a common inflammatory airways disease with the fourth highest mortality rate among U.S. adults. It leads to increased acute exacerbations and substantial health care costs [1, 2]. It is well known that environmental risk factors, especially smoking status and occupational exposure to hazardous chemicals and/or pollutions, are primary causes of COPD [3]. However, not all the people who have been exposed to toxic environmental factors develop airflow obstruction, and this observation has led to evaluation of genetic factors [4, 5].

Genetic risk factors have been extensively explored and hundreds of COPD and lung function related risk variants have been identified in different cohorts [6–8]. Previous studies revealed that the individual

impact of genetic variants on complex diseases such as COPD is usually very small suggesting that their utility for assessing risk is limited. Polygenic risk scores (PRS) are appealing for assessing risk of COPD because they consider many risk variants in aggregate [9, 10]. A PRS is constructed by adding together the individual risk alleles with each weighted by their effect sizes. In practice, the risk alleles and their effect sizes are estimated in a genome-wide association study (GWAS) in one or more population-based studies, and the derivative PRS are assessed for predictive accuracy in one or more independent studies. An important assumption of GWAS and the resulting PRS is that the genetic effects are completely independent from one another. In other words, PRS does not account for the possibility that some of the genetic effects might be dependent on context defined by demographic, environmental, or other genetic factors. This assumption may not be realistic for a disease as complex and heterogeneous as COPD. This motivates the development and evaluation of machine learning methods for identifying genetic risk scores which make fewer assumptions about genotype to phenotype relationships.

The goal of the present study is to evaluate whether a deep learning neural network algorithm can improve the predictive accuracy of PRS above and beyond that provided by the simple additive approach. Our working hypothesis is that the hidden layer nodes of the deep learning (DL) neural network will capture nonlinear dependencies (i.e. non-additive genetic effects) which might improve risk prediction in independent data. We also hypothesize that considering genetic dependencies or interactions might help to address the heterogeneity of COPD [1]. We developed a DL-PRS approach and compared its performance to the state-of-art PRS methods for prediction of COPD. We also evaluated the association of DL-PRS with clinical features. All results were replicated in independent data.

Methods

Study cohorts/Data processing workflow

The workflow of our analysis workflow is summarized in Figure 1.

We selected our study population from the UK Biobank cohort under the following procedures: For spirometry data, individuals who have complete information on age (field ID: 21003), sex (field ID: 31), height (field ID: 50) and smoking status (field ID: 20160) and have passed the spirometry quality control requirements were included [8, 11] (Supplementary Figure 1). Individuals having FEV1/FVC smaller than 0.7 and predicted FEV1 smaller than 0.8 were defined as COPD cases. Likewise, individuals having FEV1/FVC larger than or equal to 0.7 and predicted FEV1 larger than or equal to 0.8 were defined as COPD controls [12].

For genotype data, we included individuals with European ancestry according to genotype grouping (field ID: 22006) and having genetic kinship of at least a third degree away from each other (field ID: 22021). We excluded individuals identified as outliers in heterozygosity and missing rates (field ID: 22027). Additionally, we performed genotype QC using plink2 [13] with the following criteria: 1) missing genotype rate (–geno) < 0.01; 2) minor allele frequency (–maf) > 0.0001; 3) Hardy–Weinberg equilibrium (–HWE) p-value < 1e-7; 4) information metric (INFO) > 0.3.

We randomly divided UK Biobank COPD cases into training (33%), validation (33%) and testing (34%) sets. Considering the extremely imbalanced distribution of cases and controls, we randomly selected controls matching the main demographic features (age, sex and smoking status) with cases. The finalized UK Biobank data consist of training (7336 cases, 7336 controls), validation (7336 cases, 7336 controls) and testing (7558 cases, 7558 controls) sets.

To validate the transferability of PRS approaches, COPDgene phase I European ancestry data (2826 cases and 3249 controls) was included as an external testing dataset [14]. We identified the joint variants across the UK Biobank and COPDgene genetic data and finalized the datasets for further analysis.

We performed a genome-wide association analysis (GWAS) of the binary COPD phenotype based on the training set. In the GWAS analysis, we fit a logistic regression of COPD status on each genetic variant and the key demographic and genetic features (age, genetic sex, smoking history, genotype array and the first 10 principal components). The p-values of the variants in the association analysis were used for further feature selection and thresholding.

Deep polygenic risk score derivation and validation

Since our primary hypothesis is that taking into account gene-gene interactions will improve the prediction of PRS, we incorporate a unique inner hidden layer structure in the DL neural network to capture nonlinear or non-additive dependencies. We propose the DL-PRS method through pytorch (version 1.4.0) and conducted the following steps: 1) A fully-connected deep learning network was built with selected variants as input and COPD-phenotype as binary output. 2) We selected binary cross-entropy loss function to adapt to binary outcome. 3) We selected two inner layers, dropout functions (dropout rate = 0.75), learning rate ($1 * 10^{-5}$), weight decay rate ($1 * 10^{-2}$) and batch size (n = 50) through parameter grid search process with validation set. 4) DL-PRS is calculated based on the linear combination of the inner layer nodes weights and nodes values (Supplementary Figure 2). Through consideration of the hidden layer nodes and coefficients to calculate PRS, DL-PRS is built with the non-linear relationship of input nodes through the non-linearity nature of deep neural network structure.

PRS benchmark methods

We evaluated the predictive performance of four benchmarking PRS methods including traditional PRS method implemented by PRSice without clumping, logistic regression, pruning and thresholding (P+T) and LDpred.

The raw PRS method is implemented by PRSice without clumping (–no-clump) [15]. This method calculates PRS through the weighted sum of alternative allele and its weight (beta coefficient) given by the association test.

Logistic regression is another popular PRS method which weights the alternative alleles through applying a multinomial logistic regression on the alleles to predict for the phenotype [16]. We employed this using

the LogisticRegression function from the Python-based sklearn machine learning library (version 0.22) [17].

Pruning and thresholding (P+T) applies two filtering steps compared to the raw PRS method. The variants are firstly clumped using linkage disequilibrium (LD) information and are removed based on a significant p-value threshold in the secondary thresholding step. Only the most representative variant with the strongest GWA signal and weak correlation with other variants are kept in clumping. We used PRSice to calculate the pruning and thresholding PRS. A key hyperparameter with clumping is the threshold for r2 (–clump-r2) that represents the square of the correlation coefficient between two variants. We ran PRS with clumping with an array of thresholds for r2 (0.1, 0.2, 0.5, 0.8, 0.85, 0.9, 0.95) and found that with the threshold of 0.1 we were able to achieve best overall predictive performance in the validation dataset (Supplementary Table 1).

LDpred is a Bayesian PRS method inferring posterior mean effect sizes by taking into account linkage disequilibrium (LD) among markers [18]. This method adjusts the effect sizes from GWA summary statistics by assuming a prior for the genetic effect size and uses LD information from a reference panel. Instead of using an external LD reference panel, we used the UK Biobank training set to estimate the LD structure and tuned the hyperparameters for LDpred by evaluating its predictive performance in the UK Biobank validation set. Since the UK Biobank training set will represent the most accurate and homogenous LD structure for the UK Biobank validation set, we believe this approach will maximize the expected performance of LDpred. We set the hyperparameter of LD radius (–ldr) to the number of variants / 3000 as recommended by the authors and the fraction of causal variants used in Gibbs sampler (–f) was set to default, where an array of values (1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001) were tested.

Lassosum sets its unique PRS structure through applying L1 penalized regression framework on GWA summary statistics and a LD reference panel [19, 20]. Similar to LDpred, we also chose the UK Biobank training set as the LD reference panel in the Lassosum setting. There are two main hyperparameters of Lassosum, the weight for the LD matrix (s) and the weight for L1 regularization (l) in the objective function of Lassosum. We applied the default Lassosum pipeline where an array of s values (0.2, 0.5, 0.9, 1) and an array of l values were tested and the best combination of s and l that yields an optimal predictive performance in the validation set was reported (s= 0.2, l = 0.001).

Statistical analysis

All the statistical tests were performed using R (version 4.0.2). We employed t.test() and performed paired two sample t-test to search for associations between PRS deciles and continuous lung function phenotypes. Specifically, linear regressions lm() were built on the PRS deciles to predict for continuous phenotypes and the coefficients of the deciles together with confidence intervals were drawn from the regression and plotted to present the association. Likelihood ratio test chisq.test() and multinomial logistic regressions multinom() were performed to test the association between categorical phenotype variables and PRS deciles. The AUC comparison test is built using pROC function roc.test() [21].

Results

Population characteristics

Table 1 presents the characteristics of UK Biobank and COPDgene population respectively. The majority of characteristics of the two populations are different because of the differences in the recruitment strategies and inclusion criteria. The population in UK Biobank had fewer females, younger age, and lower weight compared to the COPDgene dataset. Specifically, the proportion of participants in COPDgene with smoking history was 100% while the proportion in UK Biobank was 74.1%.

Characteristics of DL-PRS approach

Our working hypothesis is that the DL-based PRS will benefit from a DL neural network architecture which incorporates latent representations with the ability to model nonlinear or non-additive effects among genetic variants. Without these hidden layer structures, the DL model is expected to perform very similar to traditional logistic regression model. To test this, we constructed a neural network model with no hidden layers and compared the predictive accuracy under different thresholding conditions (100%-3690 snps, 75%-2767 snps, 50%-1844 snps and 25%-924 snps of input SNPs based on GWAS test significance). The performance of the neural network model and logistic regression were close to each other on UK Biobank testing set (Supplementary Figure 3).

To derive the best performing DL-PRS model, we performed sensitivity analysis which tests the PRS prediction accuracy with incremental SNP thresholding conditions and different inner layer nodes to construct the PRS. From the accuracy plot shown in Supplementary Figure 4, we discovered that the PRS built on the 2^{nd} inner layer nodes performs consistently better compared to the PRS built on the 1^{st} inner layer nodes and the combined nodes from both 1^{st} and 2^{nd} inner layers. The sensitivity plot also shows that the prediction accuracy is going up briefly and going downwardly after a few epochs as the number of SNPs goes up incrementally. We decided to use the PRS based on the 2^{nd} inner layer nodes with top 25% GWAS significant input SNPs underlying the DL-PRS method in the following sections.

Characteristics of the best PRSs

In order to infer the optimal PRS, we compared our DL-PRS result with various PRS benchmarking methods: raw PRS, multinomial logistic regression, pruning and thresholding and LDpred in predicting COPD phenotype in UK Biobank testing data and COPDgene testing set. The prediction results are shown in Figure 2 and table 2. Logistic regression has the best prediction performance (AUC = 0.569, 95% CI = 0.559~0.578, R^2 = 0.015, 95% CI = 0.010~0.019) in UK Biobank testing set, followed by Lassosum (AUC = 0.568, 95% CI = 0.560~0.576, R^2 = 0.0150, 95% CI = 0.011~0.018), DL-PRS (AUC = 0.564, 95% CI = 0.555~0.573, R^2 = 0.013, 95% CI = 0.009-0.016) and pruning and thresholding (AUC = 0.564, 95% CI = 0.554~0.573, R^2 = 0.013, 95% CI = 0.010-0.017). In the independent COPDgene testing set, Lassosum ranks 1^{st} in prediction performance (AUC = 0.565, 95% CI = 0.551~0.579, R^2 = 0.012, 95% CI = 0.008~0.018), followed closely by the DL-PRS method (AUC = 0.560, 95% CI = 0.546~0.575, R^2 = 0.010,

95% CI = 0.005-0.015) and pruning and thresholding (AUC = 0.554, 95% CI = 0.539~0.568, R^2 = 0.009, 95% CI = 0.005~0.015). Raw PRS and LDpred both have worse performance compared to the other methods (Figure 2, Table 2, Supplementary Table 2). To test the difference of the COPD prediction among various PRS methods, we did significance tests of AUCs using DL-PRS as reference and the other PRS methods as comparison method (Supplementary Table 2). In UK Biobank testing set, the best performing PRS methods are Lassosum, DL-PRS, P+T and logistic regression and these are not significantly different from each other. In COPDgene testing set, lassosum performs better but similar with DL-PRS and PRSice, and these three are significantly better compared with the other PRS methods.

The addition of DL-PRS significantly increased the explanatory power of the COPD risk prediction model (p = 3.79e-13, likelihood ratio test). Regarding other PRS benchmark methods, lassosum is the leading method with the most significant increased power (p = 2.87e-18), followed by pruning and thresholding (P+T) (p=4.56e-15). However, DL-PRS is still significant (p = 2.59e-3) with likelihood ratio test in COPDgene testing data) comparing the prediction model with both DL-RPS and lassosum PRS model to the model with only lassosum PRS, and achieved p=2.08e-5 in likelihood ratio test comparing prediction model with both DL-PRS and pruning thresholding based PRS to the prediction model with only pruning thresholding PRS. These results show that DL-PRS is providing extra predicting value even when comparing with the state-of-the-art methods. Other PRS methods has less significant improvement compared to lassosum, P+T and DL-PRS (Table 3). Regarding the COPD risk model, DL-PRS is the third strongest predictor (p = 3.90e-13) following pack-years of smoking (p < 2.2e-16) and age (p < 2.2e-16) (Table 4). Compared with the COPD risk model with the other clinical predictors, DL_PRS significantly improved the predictive performance of the COPD-risk (p = 3.79e-13, likelihood ratio test).

The performance of DL-PRS on predicting COPD risk and lung function is significantly better

Given the performance of DL-PRS, logistic regression, lassosum and pruning and thresholding (P+T) are very similar to each other in binary phenotypic prediction, we next wanted to know the extent to which these PRS scores are correlated with clinical phenotypes such as lung function. When we stratified the COPDgene testing population according to the PRS deciles from these three methods, we found that the DL-PRS score has the highest correlation with lung function and COPD-related risks compared to lassosum PRS, logistic PRS and P+T PRS (Figure 3 and Figure 4, table 5). The gradient of DL-PRS deciles is strongly correlated with pulmonary function such as FEV1/FVC and predicted FEV% compared to P+T PRS and logistic PRS (Figure 3 A-H). Moreover, to quantify the differences of PRS methods deciles' clinical relevance, we performed linear regression across the deciles with FEV1/FVC and predicted FEV1% and measured the main statistics of fit (Supplementary Table 3). DL-PRS deciles is most significantly related to lung functions (p = 1.02e-5, adjusted R^2 = 0.913 for FEV1/FVC and p = 4.22e-6, adjusted R^2 =0.930 for pred FEV1%) compared to lassosum PRS (p = 1.26e-5, adjusted R^2 = 0.908 for FEV1/FVC and p=1.60e-5, adjusted R^2 = 0.902 for predicted FEV1%), logistic PRS (p = 1.2e-3, adjusted R^2 = 0.718 for FEV1/FVC and p=8.22e-4, adjusted R^2 = 0.743 for predicted FEV1%) and Pruning and Thresholding PRS (p = 2.63e-4, adjusted R^2 = 0.805 for FEV1/FVC and p = 9.81e-5. Adjusted R^2 = 0.847 for predicted FEV1%). The DL-PRS deciles are more negatively related with lung functions compared to PRS scores

based on lassosum, logistic regression and pruning and thresholding method. Last the but not least, regarding the consistency of the trend of PRS deciles, DL-PRS performs better compared to lassosum and pruning and thresholding PRS, with smaller residual standard errors.

More importantly, the DL-PRS top decile represents a drastically differentiated COPD subgroup comparing to the bottom decile/other deciles. The top decile DL-PRS is also more significantly enriched in the COPD patients with severe (GOLD stage 3 coefficient = 0.48, p = 2.3e-3) and very severe (GOLD stage 4 coefficient = 0.43, p=1.65e-2) pulmonary dysfunction (Figure 4, Table 5). Regarding COPD controls, the top decile DL-PRS is less enriched but not significant (GOLD stage 0 coefficient = -0.22, p = 0.13). Compared to logistic regression, lassosum and pruning and thresholding, DL-PRS is more significantly enriched with larger coefficients (Supplementary Table 4). Another relative phenotype is the gold stage predicted by St. George's Respiratory Questionnaire score (SGRQ), which measures the overall health quality for respiratory disease patients. DL-PRS is most significantly enriched in the most severe SGRQ status (stage D), compared to lassosum, logistic PRS and P+T PRS (Supplementary Table 4, Figure 4).

Discussion

We developed a new PRS using hidden layers nodes and weights from a DL neural network and evaluated its ability to predict risk of COPD using GWAS data from the UK Biobank and COPDgene cohorts where we also evaluated clinical relevance with measures of lung function. The predictive performance of our DL-PRS model compares favorably with current benchmark PRS methods in association with lung function and other COPD-related risks. The COPD prediction results also show that DL-PRS has extended the prediction value of the other state-of-the-art methods through the comparison to the prediction model based only on lassosum (p = 2.59e-3) and Pruning and Thresholding (p=2.08e-5). These results suggest that the consideration of genetic interactions does contribute significantly to the phenotype prediction and can be combined with the traditional PRS methods for further clinical assistance. Regarding the trends across deciles, our DL-PRS based approach has a much consistent downward trend with smaller residuals compared to pruning and thresholding and Lassosum PRS, and the slope is steeper compared to the logistic regression PRS. This discovery could due to that our PRS model is more stable with consideration the interaction effects, which is more likely to be the real scenario in biology. Our results further inform that that subgroups differentiated by DL-PRS deciles are highly related with current standards of COPD grading such as GOLD stages [1, 22]. While many PRS studies have focused on identifying the highest risk group, our results have expanded this to a wider spectrum, where lower PRS predicts a less severe clinical stage and better lung function measurements [23]. Given the heterogeneous nature of COPD, detailed characterization of COPD subgroups would be beneficial for personalized treatment, intervention and consequently improvement in overall health status [24]. Moreover, the prediction and clinical relevance are both tested in a complete independent population, which enhances the replicability of our PRS risk model.

Although our results have shown the merits of DL-PRS in comparison with the current PRS protocols, our method has several limitations. Firstly, we didn’t make a significant better improvement itself in prediction

of COPD compared to Lassosum and Pruning and thresholding. This could be due to the redundancy and extra noise from the input snps, both Lassosum and Pruning and Thresholding deals elegantly selecting very significant and non-redundant snps from the input, either from a L1 penalized regression and clumping to select the most significant snps without redundancy. To manage this issue, we plan to further expand our deep learning structure with penalization method in the near future. Interpretability is also a common issue in the DL neural network field and there is a clear trade-off with prediction [25]. We made a substantial effort to simplify the necessary hyperparameters and built a fully connected feed-forward model, where others can probe the model through doing a weight-based connection calculation [26]. Another limitation lies in the computational cost of the algorithm, our approach requires the input of individual's genetic information, and the DL neural network process has a requirement on both the hardware and computational time. We tackled this issue by conducting an initial variant feature selection for the DL neural network input variants, thus maximizing the prediction efficiency while minimizing the dimension of the inputs. However, for disease like COPD which has very complex traits coming from numerous genetic variants and the interactions among them, our method is underpowered with a drop of the variants which are considered uninformative in the GWA analysis.

In conclusion, we introduced a novel PRS method which considers interactions among genetic variants through a DL neural network architecture. DL-PRS not only predicts COPD disease status as well or better than traditional methods, but it also more closely correlates with measures of lung function and risk factors. Our study has demonstrated that machine learning can complement traditional PRS methods and may have improved clinical utility for common diseases such as COPD. This is but a first step, and there are many algorithms, models, and variations to explore.

Declarations

Ethics approval, accordance and consent for publication

Not applicable

Availability of data and materials

The genome-wide association results for FEV1, FVC and FEV1/FVC are available from UK Biobank at http://www.ukbiobank.ac.uk/ ;

UK Biobank genetic data are released in http://www.ukbiobank.ac.uk/scientists-3/genetic-data/;

The sources of COPDgene data used in this study are available at www.copdgene.org.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

JHM envisioned the project, SH and designed the workflow. SH and JHM wrote the manuscript. All authors have read, revised, and approved the final manuscript.

Acknowledgements

The authors would like thank Soumitra Ghosh at Human Genetics, GSK R&D for his many contributions. They also thank Abhinav Suri for his useful discussions.

Funding

The development of this tool was supported by National Institutes of Health research grants [R01 LM010098].

References

1. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. Am J Respir Crit Care Med. 2017;195:557–82.
2. Sakornsakolpat P, Prokopenko D, Lamontagne M, Reeve NF, Guyatt AL, Jackson VE, et al. Expanded genetic landscape of chronic obstructive pulmonary disease reveals heterogeneous cell type and phenotype associations. bioRxiv. 2018;:355644.
3. Sama SR, Kriebel D, Gore RJ, DeVries R, Rosiello R. Environmental triggers of COPD symptoms: a case cross-over study. BMJ Open Respir Res. 2017;4. doi:10.1136/bmjresp-2017-000179.
4. Diemen CC van, Postma DS, Aulchenko YS, Snijders PJLM, Oostra BA, Duijn CM van, et al. Novel strategy to identify genetic risk factors for COPD severity: a genetic isolate. Eur Respir J. 2010;35:768–75.
5. Rennard SI, Vestbo J. COPD: the dangerous underestimate of 15%. Lancet Lond Engl. 2006;367:1216–9.
6. El-Zein RA, Young RP, Hopkins RJ, Etzel CJ. Genetic Predisposition to Chronic Obstructive Pulmonary Disease and/or Lung Cancer: Important Considerations When Evaluating Risk. Cancer Prev Res (Phila Pa). 2012;5:522–7.
7. Hall R, Hall IP, Sayers I. Genetic risk factors for the development of pulmonary disease identified by genome-wide association. Respirology. 2019;24:204–14.
8. Wain LV, Shrine N, Artigas MS, Erzurumluoglu AM, Noyvert B, Bossini-Castillo L, et al. Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. Nat Genet. 2017;49:416–25.
9. Moll M, Sakornsakolpat P, Shrine N, Hobbs BD, DeMeo DL, John C, et al. Chronic obstructive pulmonary disease and related phenotypes: polygenic risk scores in population-based and case-control cohorts. Lancet Respir Med. 2020;8:696–708.

10. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460:748–52.
11. Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne CA, et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat Genet. 2019;51:481–93.
12. Hobbs BD, de Jong K, Lamontagne M, Bossé Y, Shrine N, Artigas MS, et al. Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. Nat Genet. 2017;49:426–32.
13. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4. doi:10.1186/s13742-015-0047-8.
14. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, et al. Genetic Epidemiology of COPD (COPDGene) Study Design. COPD. 2010;7:32–43.
15. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. GigaScience. 2019;8. doi:10.1093/gigascience/giz082.
16. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet. 2019;104:21–34.
17. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. ArXiv13090238 Cs. 2013. http://arxiv.org/abs/1309.0238. Accessed 3 Feb 2021.
18. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015;97:576–92.
19. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. Genet Epidemiol. 2017;41:469–80.
20. Moll M, Lutz SM, Ghosh AJ, Sakornsakolpat P, Hersh CP, Beaty TH, et al. Relative contributions of family history and a polygenic risk score on COPD and related outcomes: COPDGene and ECLIPSE studies. BMJ Open Respir Res. 2020;7. doi:10.1136/bmjresp-2020-000755.
21. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
22. Donohue JF, Bollu VK, Stull DE, Nelson LM, Williams VS, Stensland MD, et al. Long-term health-related quality-of-life and symptom response profiles with arformoterol in COPD: results from a 52-week trial. Int J Chron Obstruct Pulmon Dis. 2018;13:499–508.
23. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50:1219–24.
24. Agusti A, Calverley PM, Celli B, Coxson HO, Edwards LD, Lomas DA, et al. Characterisation of COPD heterogeneity in the ECLIPSE cohort. Respir Res. 2010;11:122.

25. Freitas AA. Comprehensible classification models: a position paper. ACM SIGKDD Explor Newsl. 2014;15:1–10.
26. Olden JD, Jackson DA. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. Ecol Model. 2002;154:135–50.

Tables

Tables 1-5 are available in the Supplementary Files.

Figures

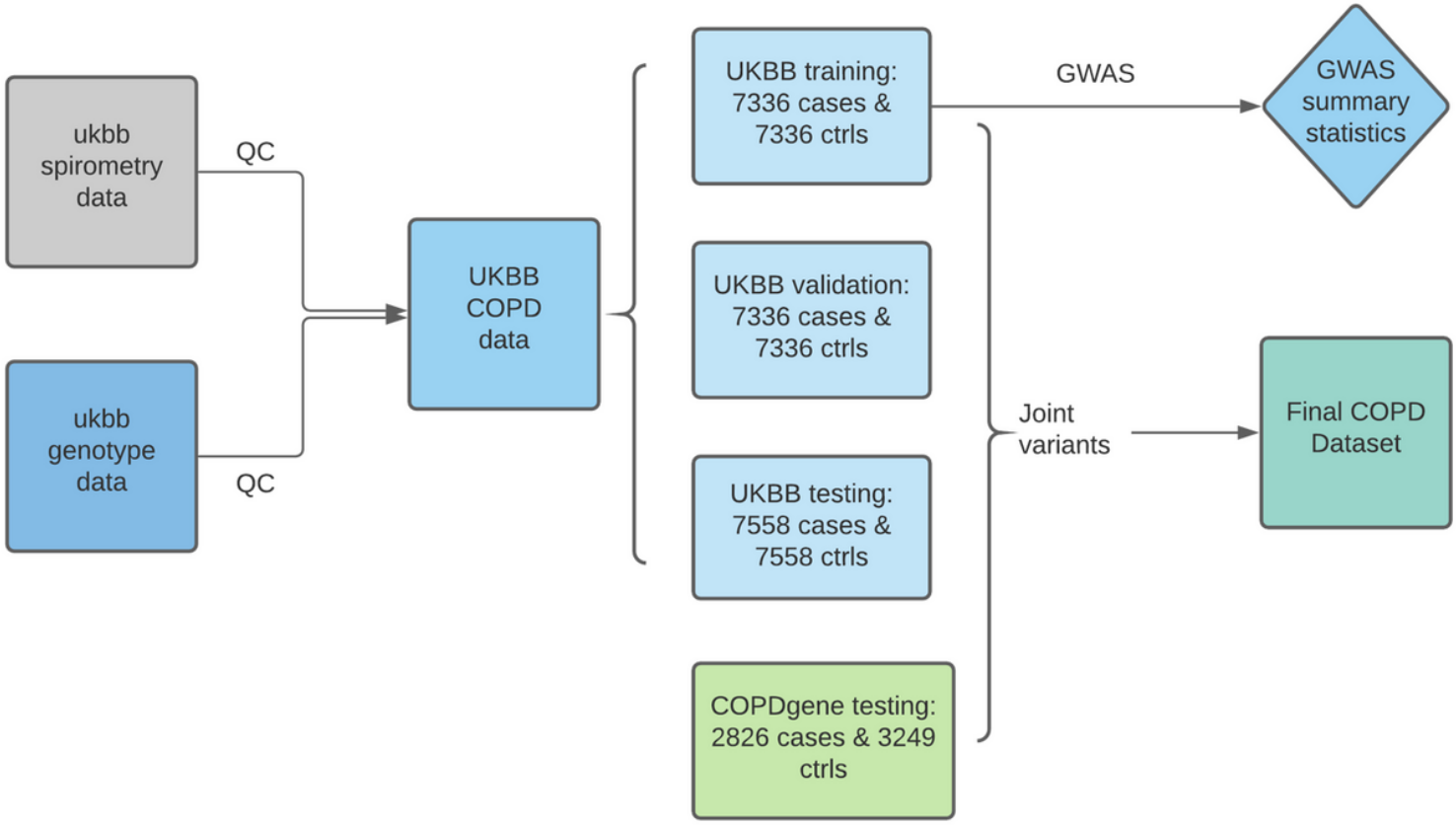

Figure 1

Data preprocessing workflow for UK biobank and COPDGene cohorts.

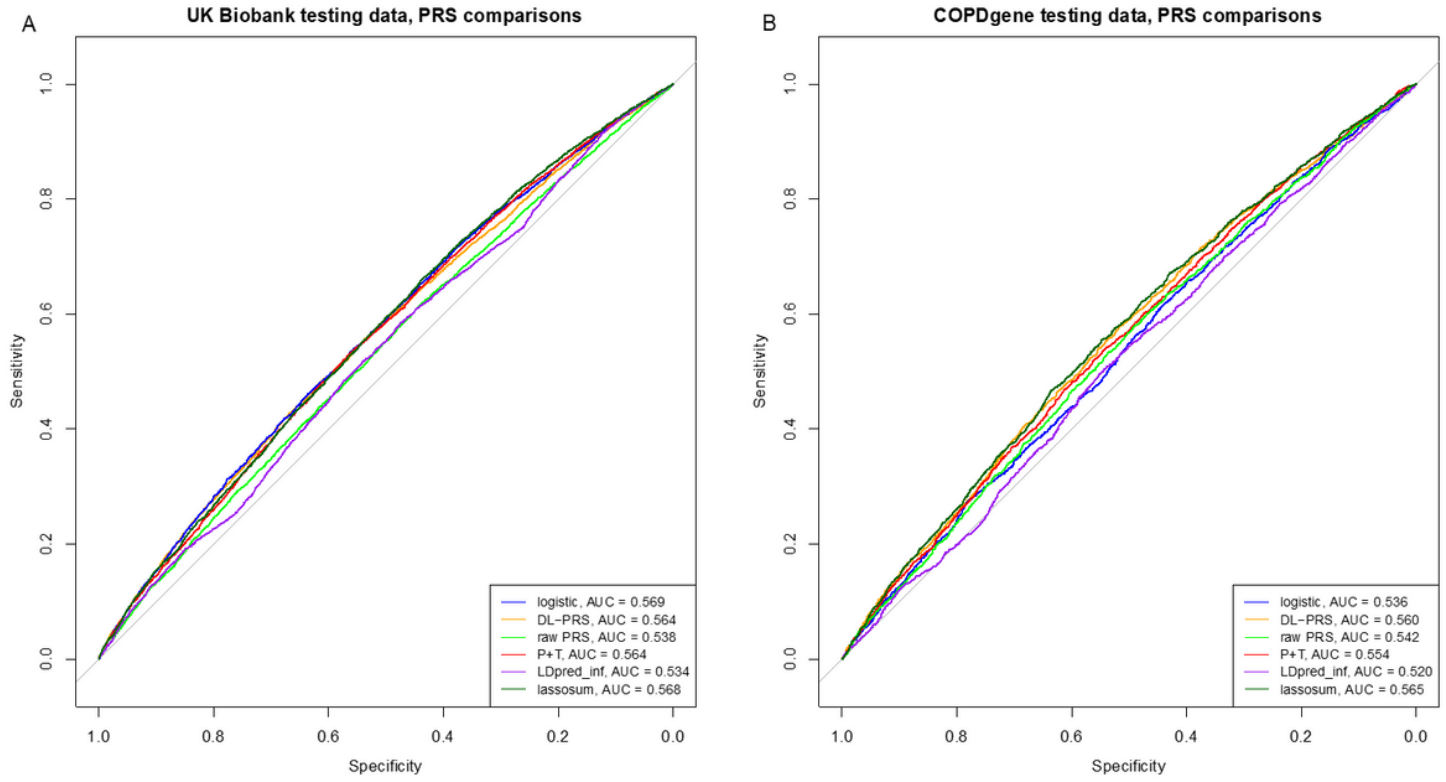

Figure 2

The predictive performances of five PRS methods including PRSice with or without clumping, logistic regression, LDpred and DL-PRS were evaluated by area under receiver operating characteristic (AUC)

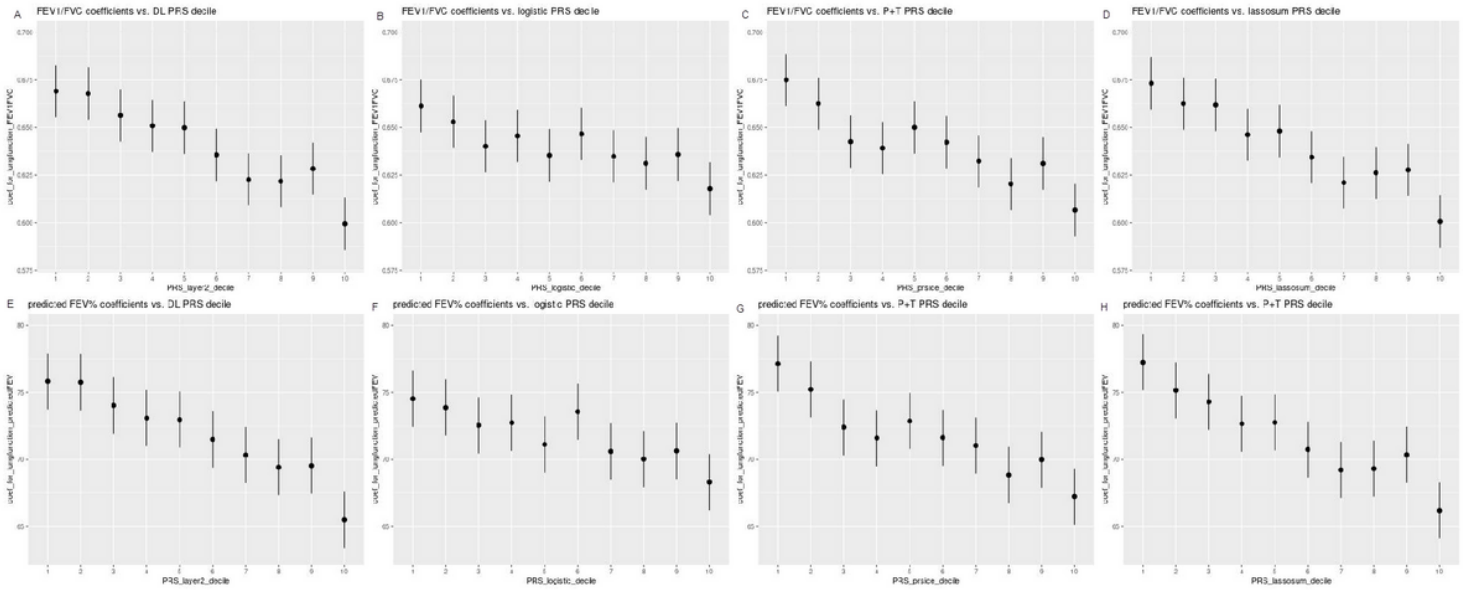

Figure 3

COPDGene subjects were stratified according to their PRS deciles (1-10 from low to high DL-PRS, logistic PRS ,Pruning and thresholding PRS and lassosum PRS). Subjects from decile 1-10 were compared for

their mean change (measured by regression coefficient) in pulmonary function measures including FEV1/FVC (A-D) and percent predicted FEV1 1(E-H).

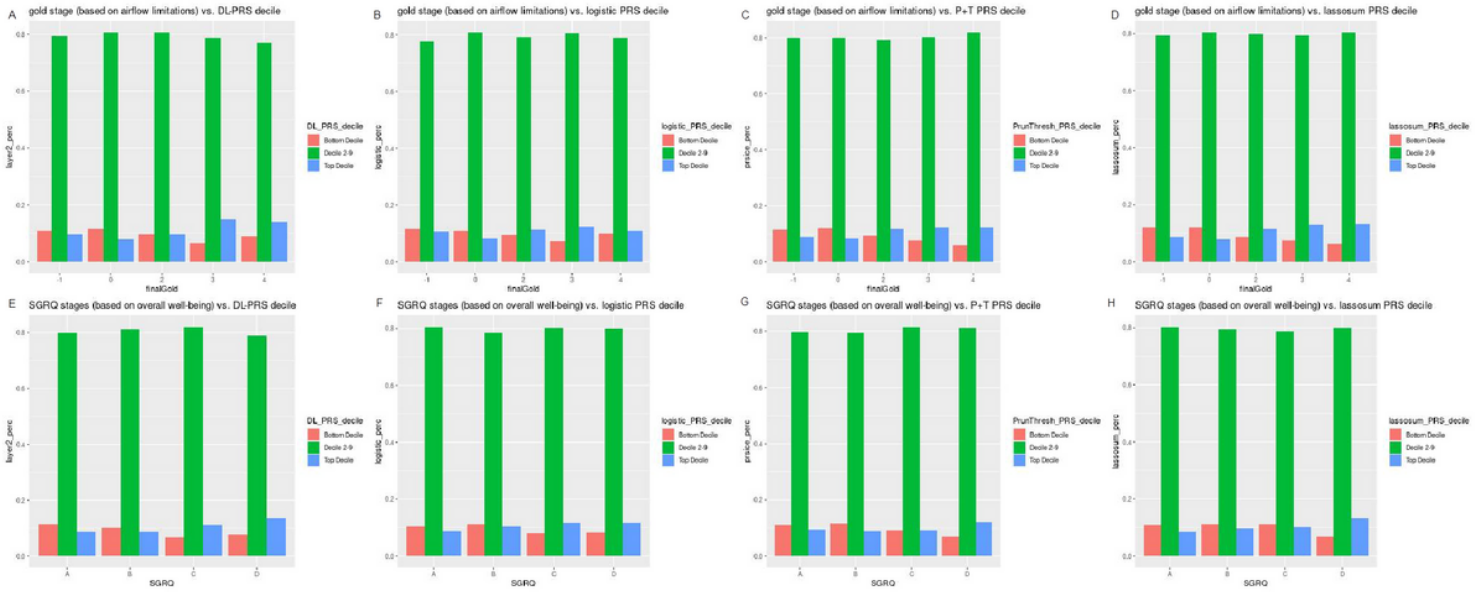

Figure 4

The enrichment of COPD patients with top PRS decile (DL, logistic, P+T and lassosum) vs. COPD patients with bottom PRS decile within each GOLD stage (A-D) and SGRQ stage(E-H) were tested.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- deepPRSTables.docx
- SupplementaryFigures.docx
- SupplementaryTables.docx