Online Supplement

Emphysema Detection in Smokers: DLNO Beats DLCO-based models — Supplementary Methods, Tables, and Figures

Gerald S. Zavorsky, PhD, RRT[1] Roberto W. Dal-Negro, MD[2] Ivo van der Lee, MD, PhD[3] Alexandra M. Preisser, MD[4]

[1]Department of Physiology and Membrane Biology, University of California, Davis, Davis, California, United States

[2]CESFAR – National Center for Respiratory Pharmacoeconomics and Pharmacoepidemiology, Verona, Italy

[3]Department of Pulmonology, Spaarne gasthuis, Haarlem, The Netherlands

[4]Institute for Occupational and Maritime Medicine, University Medical Center Hamburg–Eppendorf, Hamburg, Germany

Table of Contents

Supplementary Methods

Study Design and Population

We conducted an individual participant data (IPD) meta-analysis pooling raw, participant-level data from four European hospital cohorts[1]. IPD enabled harmonization of variables, uniform quality control (QC), and re-analysis using consistent definitions across studies[1]. The pooled dataset contained 496 participants. Body weight was missing in two cohorts[2,3] and was imputed using a regression derived in 230 adults aged 50–69 years:

weight (kg) = 0.59 × height (cm) + 7.97 × sex (1 = male; 0 = female) − 33.57; $R^2 = 0.51$; SEE = 8.62 kg.

In Moinard & Guénard (1990)[2], raw data were embedded in the article; 10 COPD cases lacked explicit emphysema status. Based on clinical presentation (mean $FEV_1/FVC = 0.56$; arterial $PO_2 = 58$ mmHg; DLCO z-score= −3.04), these were classified as emphysema for the present analyses. After QC filtering (see below), three cohorts remained[2-4], comprising 408 participants: 85 with CT-confirmed emphysema and 323 smokers without emphysema. Most were current or former smokers (86%; IQR 14–43 pack-years). The fourth cohort used a 5-s NO–CO breath-hold time (BHT)[5]. Although acceptable under the DLNO ERS technical standards[6], it was excluded to avoid protocol heterogeneity. Attempts to obtain additional NO–CO double-diffusion datasets were unsuccessful; one prominent group declined participation despite repeated outreach[7]. The final harmonized dataset is available in a cloud repository[8].

Construction of Data Quality Table

For each cohort, G.S.Z. abstracted whether the following pre-specified items were available and analysable: (1) presence of both COPD (Disease = 1) and non-COPD (Disease = 0) groups; (2) pack-years; (3) percent emphysema by CT volume; (4) smoking history; (5) mMRC dyspnoea score; (6) sex; (7) height; and (8) weight. Technical quality (Item 9) was assessed record-by-record using harmonized criteria for the simultaneous 10 ± 2 s NO–CO protocol. A record failed quality control if any of the following were present: BHT outside 8.0–12.0 s; $VA/TLC > 1.00$; $RV/TLC < 0.20$; inspired volume/FVC (IV/FVC) < 0.85; or other documented protocol deviations. Study-level quality control was the proportion of records passing all checks; studies with >95% pass rate was marked ✓, otherwise ✗ with the pass percentage. The 5-s BHT cohort[5] was not included in QC tallies to maintain protocol homogeneity (90% of its records met the other QC criteria). In Table S1, ✓ denotes availability/analytic readiness; ✗ denotes absence/unusable data. The "Total number of checkmarks" is the count of ✓ across the nine items. See **Table S1**.

Variable Standardization and Analytic Quality Control

Spirometry, lung volumes, and diffusing capacity indices were standardized to z-scores (GLI for spirometry[9], lung volumes[10], $DLCO_{10s}$[11,12], VA_{10s}[11,12], and KCO_{10s}[11,12]). Device-appropriate DLNO reference equations[13] and 10-s breath-hold DLNO/KNO equations[14] were applied to ensure cross-study comparability. When DLNO and DLCO were measured simultaneously, DLCO z-scores were derived from the same sources as the DLNO reference equations[13,14].

Analytic QC applied the following exclusions prior to modelling: cases that had breath-hold outside 8–12 s; $VA/TLC \geq 1.0$; $FEV_1/FVC \geq 1.0$; $RV/TLC < 0.20$; and inspired volume/FVC < 0.85. Analyses were restricted to complete cases after QC. Age and sex distributions of the final dataset are shown in **Figure S1**.

LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) is a powerful regularization technique used in regression analysis to simultaneously perform variable selection and shrinkage, enhancing model interpretability and prediction accuracy, particularly in datasets with many predictors or multicollinearity[15,16]. LASSO modifies the standard linear or logistic regression objective by adding an L1 penalty, the sum of the absolute values of the coefficients, to the loss function. This penalty, controlled by a tuning parameter λ, shrinks the coefficients of less predictive variables toward zero, effectively excluding them from the model, while retaining and estimating the coefficients of the most relevant predictors[15,16]. For binary outcomes like emphysema (yes emphysema or no emphysema), LASSO is applied within logistic regression, minimizing the negative log-likelihood plus the λ times the sum of absolute coefficients, encouraging sparsity by setting some coefficients to zero. This sparsity is particularly useful for fitting the z-scores — TLC, RV/TLC, FEV_1, FVC, FEV_1/FVC, $DLCO_{10s}$, $DLNO_{10s}$, VA_{10s}, KCO_{10s}, and KNO_{10s} — as predictors, allowing LASSO to identify a subset of these lung function metrics most strongly associated with emphysema.

LASSO was implemented to identify the most important predictor variables for emphysema, leveraging a dataset comprising 408 subjects with complete cases (after filtering). The process began with preparing the data, ensuring the predictors (z-scores standardized relative to reference equations like GLI or GAMLSS) were numeric, and the outcome (emphysema) was binary. A LASSO logistic regression model was fit, transforming the predictors into a matrix and the outcome into a vector, then applying cross-validation (e.g., 10-fold) to determine the optimal λ that minimizes prediction errors, such as log-loss or misclassification errors. This optimal λ was used to fit the final LASSO model, which shrank some coefficients to zero, thereby selecting the most relevant z-scores—potentially those variables most physiologically linked to airflow limitation, gas transfer, and hyperinflation in emphysema—while excluding less informative variables. With non-zero coefficients, the resulting model provides a sparse, interpretable set of predictors, improving standard logistic regression by addressing multicollinearity among lung function metrics and reducing overfitting. The selected predictors, their coefficients, and model performance metrics (e.g., area under the ROC curve, accuracy) would then be used in binary logistic regression analyses and compared to other methods like principal component analysis (PCA) or hierarchical partitioning to validate findings and highlight LASSO's role in simplifying the model for emphysema prediction.

Binary Logistic Regression

Once LASSO determined possible predictors of emphysema, binary logistic regression was used to determine the best model. Both generalized linear models (GLM) and generalized linear mixed-effects models (GLMM) with a random intercept for "Study" were employed to account for potential clustering effects. Models were evaluated based on BIC (frequentist)[11,17] method – and the LOOIC (Bayesian)[18,19], which implemented

Markov Chain Monte Carlo (MCMC) sampling to generate posterior distributions for model parameters[20,21]. Weakly informative priors were used: a normal distribution with mean zero and standard deviation one for all regression coefficients, a Student *t*-distribution with three degrees of freedom (location zero, scale 2.5) for the intercept, and an exponential distribution with rate one for the standard deviation of the study-level random effect. Models were fit with four MCMCs, each with 15,000 thousand iterations, including 5,000 warm-up iterations; the target acceptance probability was set to approximately 0.99999 and the maximum tree depth was fifteen. Predictive performance was compared using Pareto-smoothed importance-sampling leave-one-out cross-validation with moment matching, and relative support with model weights summarized based on stacking and pseudo-Bayesian model averaging.

We also fit a focused three-predictor model—forced expiratory volume in one second (standardized score), total lung capacity (standardized score), and diffusing capacity for nitric oxide from the generalized additive models for location, scale, and shape framework (standardized score)—with and without a study random intercept, and compared models by the difference in expected log predictive density, choosing the simpler model when its performance was within about one standard error of the best model. In addition, we examined projection-predictive variable selection with forward selection (falling back to a generalized linear model if the generalized linear mixed-model procedure failed), computed variance inflation factors and standard generalized-linear-model fit metrics (including the Akaike information criterion and the Bayesian information criterion). The Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are not directly applicable to the LOOIC[18]. The frequentist (BIC) and Bayesian (LOOIC) analytical methods assessed model superiority, with detailed explanations provided in **Tables S4-S5**.

Principal Component Analysis and Hierarchical Partitioning Analysis

To explore the underlying structure of lung function predictors and their association with emphysema, we performed Principal Component Analysis (PCA) on standardized z-scores of the best predictors and compared it with the additional of fitted DLCO z-scores obtained from GLI equations [11,12]. PCA was conducted to reduce the dimensionality of these correlated predictors into principal components (PCs) that explain the maximum variance, with no additional scaling required due to their z-score standardization. The resulting PCs were then used as predictors in a binary logistic regression model to assess their predictive power for emphysema (coded as 0/1 in the Disease variable), with model fit evaluated using pseudo-R² metrics (McFadden's R^2) and information criteria (AIC, BIC). To quantify the relative contribution of each predictor to the variance explained in emphysema, we applied hierarchical partitioning. This involved fitting a logistic regression model with all predictors, partitioning the marginal R² into unique and shared contributions, reporting percentages of the total explanatory power for each predictor, adjusted for their collinearity. Both analyses were performed on a dataset of 408 subjects, with missing data handled by listwise deletion to ensure complete cases. A comparison of PCA vs hierarchical partitioning is provided in **Table S6**. Results of PC analyses are provided in Tables **S11-S14 and Figure S4**, with Hierarchical partitioning results in **Tables S15-S16.**

AUROC, MCC and Kappa Statistics

Classification performance was measured using the area under the receiver operating characteristic curve (ROC) and Matthews Correlation Coefficient (MCC). The 95% CI for the ROC models was calculated using DeLong's method for imbalanced datasets [22,23]. The false discovery rate was controlled among AUROCs at 0.01 using the Benjamini-Hochberg procedure [24].

While AUROC assesses discriminatory ability, MCC provides a balanced evaluation in datasets with class imbalances [25-27], making it a preferred metric for binary classifications. The 95% Confidence Interval (CI) for the MCC were generated from 100,000 bootstrapped samples. In addition, other metrics were used to classify those that had progressed to emphysema compared to those that did not. The name, definitions, and formulas for other classification metrics used, are present in **Tables S7-S9**.

Kappa statistical analysis was conducted to evaluate the level of agreement between three prespecified logistic models:

- **Model A (3 predictors):** FEV_1 z-scores [GLI] + TLC z-scores [GLI] + $DLCO_{10s}$ z-scores [GLI]
- **Model B (4 predictors):** Model A + $DLNO_{10s}$ z-scores [GAMLSS]
- **Model C (3 predictors):** FEV_1 z-scores [GLI] + TLC z-scores [GLI] + $DLNO_{10s}$ z-scores [GAMLSS]

The discordance between equations was calculated as 1−κ where κ reflects the concordance[28]. Discordance categories were defined as follows: $(1-\kappa) < 0.1$ = Negligible discordance; $0.1 \leq (1-\kappa) \leq 0.20$ = Very low discordance; $0.21 \leq (1-\kappa) \leq 0.40$ = Low discordance; $0.41 \leq (1-\kappa) < 0.60$ = Moderate discordance; $0.61 \leq (1-\kappa) < 0.79$ = High discordance; $(1-\kappa) \geq 0.8$ = Very high discordance. The Kappa analysis quantified differences in LLN classifications derived from Zavorsky & Cao (2022)[13] compared to van der Lee *et al.* (2007) [14] and GLI equations. These comparisons provided insights into the variability of LLN thresholds and potential impacts on clinical classifications.

Net Reclassification Index (NRI) & Integrated Discrimination Index (IDI)

We compared three prespecified logistic models:

- **Model A (3 predictors):** FEV_1 z-scores [GLI] + TLC z-scores [GLI] + $DLCO_{10s}$ z-scores [GLI]
- **Model B (4 predictors):** Model A + $DLNO_{10s}$ z-scores [GAMLSS]
- **Model C (3 predictors):** FEV_1 z-scores [GLI] + TLC z-scores [GLI] + $DLNO_{10s}$ z-scores [GAMLSS]

Because patients derived from multiple studies, we evaluated a study-level random intercept (1|Study) via a prespecified decision rule. For each model we fit (i) a simple logistic regression and (ii) a Generalized Linear Mixed-Model (GLMM) with a Study random intercept. We retained the GLMM only if (a) the random-effect variance was >0 (non-singular fit) and (b) AIC improved by >2 versus the Generalized Linear Model (GLM). Otherwise, we used the GLM. (In our data, all three final models were GLMs).

For each fitted model we converted predicted probabilities to yes/no decisions using the Youden's J–optimal threshold (threshold that maximizes sensitivity + specificity − 1). The threshold was re-optimized within every bootstrap resample so that uncertainty in the operating point was propagated into all interval estimates.

At the Youden-optimized operating point we computed accuracy, balanced accuracy, sensitivity, specificity, PPV, NPV, FPR, FNR, FDR, FOR, F1, Cohen's κ, Matthews correlation coefficient (MCC), likelihood ratios (+LR/−LR), diagnostic odds ratio (DOR), and discordance. Model-to-model differences (B−A and C−A) were obtained with a paired bootstrap (see below); we labelled a difference "statistically different" only when the 95% bootstrap CI excluded 0**. Tables S17-S18** presents the comparison results for Models A, B, and C**.**

We quantified reclassification between models two ways:

1. Threshold-based reclassification at the Youden-optimized cut-points (**Table S19**):
   - NRI = NRI+ + NRI−**,** where NRI+ = Pr(Up | Case) − Pr(Down | Case) and NRI− = Pr(Down | Control) − Pr(Up | Control).
   - We also report the four component proportions: Pr(Up | Case), Pr(Down | Case), Pr(Down | Control), Pr(Up | Control).
   - IDI (Integrated Discrimination Index) is the difference in mean predicted risk between cases and controls for the two models.
2. Category-free reclassification: the same NRI components and IDI computed without any fixed risk categories (**Table S20**).

All reclassification results are shown for B−A (adding DLNO to Model A) and C−A (replacing DLCO with DLNO) (**Tables S19-S20**).

For each model we assessed calibration by logistic recalibration:

$$\operatorname{logit}\{P(Y=1)\} = \alpha + \beta \operatorname{logit}(\hat{p}),$$

reporting the calibration intercept α and calibration slope β. Intercept ≈ 0 and slope ≈ 1 indicate good calibration. CIs were obtained by bootstrap. Uncertainty was quantified with 50,000 bootstrap resamples**.** For GLMs we used subject-level resampling; had a GLMM been retained we would have used a **cluster** (Study) bootstrap. For model comparisons (B−A, C−A) we used a paired bootstrap that resampled the same subjects (or the same clusters) for both models within each draw. For each quantity we report the bootstrap standard error and the percentile 95% CI (2.5th–97.5th percentiles across resamples).

Random-Intercept Screening and Model Ranking Methods

Bayesian fits with weakly informative priors was used—Normal(0, 1) for fixed effects; Student-t(3, 0, 2.5) for the intercept; Exponential(1) for the standard deviation of random intercepts—run with four Markov chain Monte Carlo (MCMC) chains, up to 20,000 iterations, and Pareto-smoothed importance-sampling leave-one-out cross-validation (PSIS-LOO) with moment matching. If any Pareto k diagnostic exceeded 0.7, we re-fit those cases using exact leave-one-out refits (reloo). We compared random-intercept and fixed-effects versions using the difference in expected log predictive density (ΔELPD) and its standard error (SE); the random-intercept variant was chosen when ΔELPD > SE. If Bayesian fitting failed, we fit generalized linear models / generalized linear mixed models (GLM/GLMM) and selected the variant with the lower Bayesian Information Criterion (BIC).

For each chosen variant we computed the area under the receiver operating characteristic curve, the Matthews correlation coefficient (MCC), and BIC; MCC was optimized over the classification threshold. For out-of-sample robustness, MCC was estimated using stratified 5-fold cross-validation (CV), selecting the threshold on training folds and evaluating on held-out folds.

All 34 models were ranked using **PSIS-LOOIC** and **BIC** (lower values indicate better fit/parsimony) and **AUROC** and **MCC** (higher values indicate better discrimination). “Equal-weight” rankings averaged ranks across the available metrics (weights renormalized if any metric was missing; PSIS-LOOIC was excluded if unavailable). "Weighted" rankings followed **six prespecified schemes**:

- **Weighted Average** *(LOOIC 0.40, AUROC 0.30, MCC 0.20, BIC 0.10).* Prioritizes generalization (most significant weight on LOOIC) while still valuing discrimination (AUROC, MCC) and parsimony (BIC). BIC is modest to avoid double-penalizing complexity, given that LOOIC already favours simpler, better-generalizing models.
- **Generalization-Emphasis** *(LOOIC 0.55, AUROC 0.20, MCC 0.15, BIC 0.10).* For use when out-of-sample performance is paramount (e.g., multi-site deployment, transportability).
- **Discrimination-Emphasis** *(AUROC 0.45, MCC 0.35, LOOIC 0.10, BIC 0.10).* For clinical contexts where case–control separation matters most (screening/triage). AUROC captures threshold-free separation; MCC reflects performance at an operating point. LOOIC/BIC remain as safeguards.
- **BIC-Omitted** *(BIC = 0; remaining weights renormalized).* Sensitivity analysis to confirm parsimony penalties are not driving results, recognizing that LOOIC already disfavours gratuitous complexity.
- **AUROC-Emphasis** *(AUROC 0.35, LOOIC 0.30, MCC 0.25, BIC 0.10.* Mirrors the common AUC-first evaluation to show that conclusions are robust even when discrimination is given extra prominence.
- **Average Rank (all schemes).** For presentation, we also report the **unweighted mean** of the six scheme-specific ranks for each model. This derived "Average Rank" was **not** resampled in inferential procedures.

Ties were broken by favoring **lower PSIS-LOOIC/BIC**, then **higher AUROC/MCC**, then **fewer predictors**. An **optional** collinearity penalty was prespecified with a hard flag for VIF > 10); when applied, penalized ranks were obtained by adding this penalty to the equal/weighted rank. Ties favoured lower LOOIC/BIC, then higher AUROC/MCC, and then fewer predictors. An optional collinearity penalty was prespecified as max (0, max VIF − 5) × 0.25 (with an optional hard flag for VIF > 10); when applied, penalized ranks were obtained by adding this penalty to the equal/weighted Rank.

Scheme Robustness of comparison of model ranks

To assess whether model comparisons were robust to the weighting of evaluation metrics**,** we prespecified the six ranking schemes above and used the best-ranked z-score model [MODEL C: TLC & FEV$_1$ (GLI), DLNO (GAMLSS)] as the baseline. For each comparator and for each scheme we computed the **rank difference**

$$\Delta\text{rank} = \text{rank}_{\text{comparator, scheme}} - \text{rank}_{\text{MODEL C, scheme}},$$

So that positive values indicate the comparator ranks worse than MODEL C under that scheme. We then calculated, for each comparator, the mean Δ rank across the six schemes**.**

To reflect sensitivity to weighting choice (not patient-level sampling variability)**,** we obtained 95% bootstrap intervals **by** resampling the six schemes with replacement **(**n = 6**,** B = 10,000**)** and recomputing the across-scheme mean Δ rank for each bootstrap replicate. We did not resample or use the derived "Average Rank" column in this procedure; intervals were based on the six scheme-specific differences. Intervals entirely > 0 indicate the comparator is consistently ranked worse than MODEL C across the prespecified schemes. Results for the top 10 of 34 models are displayed in **Figure 3** of the main article.

Decision Curve Analysis

To evaluate the clinical utility of the emphysema classification models, we conducted decision curve analysis (DCA) using out-of-fold predictions from repeated cross-validation. Net benefit was plotted against threshold probability values ranging from 0 to 0.25 (**Figure S7).** Each model's net benefit was compared against two reference strategies: Treat All (light grey line): assumes all patients are treated; Treat None (dashed black line): assumes no patient is treated. The net benefit was computed using the standard DCA formula:

$$\text{NB}(p_t) = \frac{\text{TP}}{n} - \frac{\text{FP}}{n} \cdot \frac{p_t}{1 - p_t},$$

Where TP = true positives, FP = false positives, *pt* = threshold probability, and *n* = total number of patients.

Laptop Specifications Used for Code Execution

The central processing unit (CPU) used is the Intel Core i9-12950HX, featuring 16 cores (8 performance and eight efficiency) and 24 threads, with a base clock of 2.3 GHz that boosts up to 5 GHz. The CPU is paired with 128 GB of RAM (running at 1795.6 MHz DRAM frequency). FEV_1

Package function mapping

Part 1: Workflow Summary (General Readership)

All analyses were conducted in R (version 4.4.2), using a combination of base functions and specialized packages. The process followed a structured pipeline:

1. Data Handling & Preprocessing

SPSS files were imported and converted into tidy R data frames. Variables were cleaned, reshaped, recoded, and standardized across studies. Factor levels were harmonized; labelled vectors were coerced to base types. Model outputs were tidied into consistent formats for analysis and reporting. Visualization tools were used to generate calibration plots, biplots, and publication-quality figures. Tables and figures were exported to Word and Excel formats. Parallelization was used to speed up bootstraps and cross-validations, with runtime and reproducibility tightly controlled.

2. Modeling, Evaluation & Inference

- LASSO regression was used to select lung function predictors most associated with emphysema.
- Logistic regression models—both frequentist and Bayesian—were fit with and without study-level random effects.
- Model selection was guided by AIC, BIC, and LOOIC, depending on the framework.
- Classification performance was evaluated using AUROC, MCC, and metrics like sensitivity, specificity, and F1-score.
- Bootstrapping and stratified cross-validation were applied to assess stability and uncertainty.
- PCA and hierarchical partitioning were used to reduce dimensionality and understand variable importance.

- Collinearity was checked using variance inflation factors (VIFs), and calibration diagnostics were used to assess model fit.

All R packages used are presented in **Tables S2-S3**.

Table S1. Data quality summary by study

Data Quality Item	Dal Negro *et al.* (2024)[5]	Diener *et al.* (2021)[4]	Moinard & Guénard (1990)[2]	van der Lee *et al.* (2009)[3]
1. Both COPD (Disease = 1) and non-COPD subjects (Disease = 0) were provided in the study	✓	✓	✓	✓
2. Pack Years Included	✓	✓	✓	✓
3. Percent emphysema by CT Volume	✓	✗	✗	✓
4. Smoking history included	✗	✓	✗	✗
5. mMRC dyspnoea scores included	✓	✗	✗	✗
6. Sex of subject included	✓	✓	✓	✓
7. Height of subjects included	✓	✓	✓	✓
8. Weight of subjects included	✓	✓	✗	✗
9. Met technical quality standards (i.e. > 95% of cases met quality control requirements)	✗ (2%) *	✗ (77.6%)	✓ (100%)	✓ (99.2%)
Total number of checkmarks (out of 9 possible)	7	6	5	6

Failing quality control means that one of the following was found in a case: Breath-hold time was not between 8.0-12.0 s; VA/TLC ratio ≥ 1.0; FEV_1/FVC ratio ≥ 1.0; RV/TLC ratio < 0.20; or inspired volume to FVC ratio < 0.85. Each case that failed quality control was eventually removed before statistical analyses.

*The low percentage reflects the study's 5-s breath-hold. We excluded that study to harmonize breath-hold time across studies; without that criterion, 90% of cases in that study met quality control requirements.

Dal Negro *et al.* (2024)[5] used the Hyp'Air Compact device (Medisoft®, Belgium) to measure DLNO and DLCO via electrochemical NO and CO sensors. In contrast, Moinard & Guénard (1990)[2] and van der Lee *et al.* (2009)[3] employed chemiluminescence-based analyzers (Thermo Electron Corporation, MA, USA; and CLD 77AM, Eco Physics, Zurich, Switzerland, respectively) for DLNO assessment. Diener *et al.* (2021)[4] utilized the Jaeger MasterScreen PFT Pro (CareFusion, Hochberg, Germany), which also used an electrochemical NO and CO sensors for DLNO / DLCO measurements.

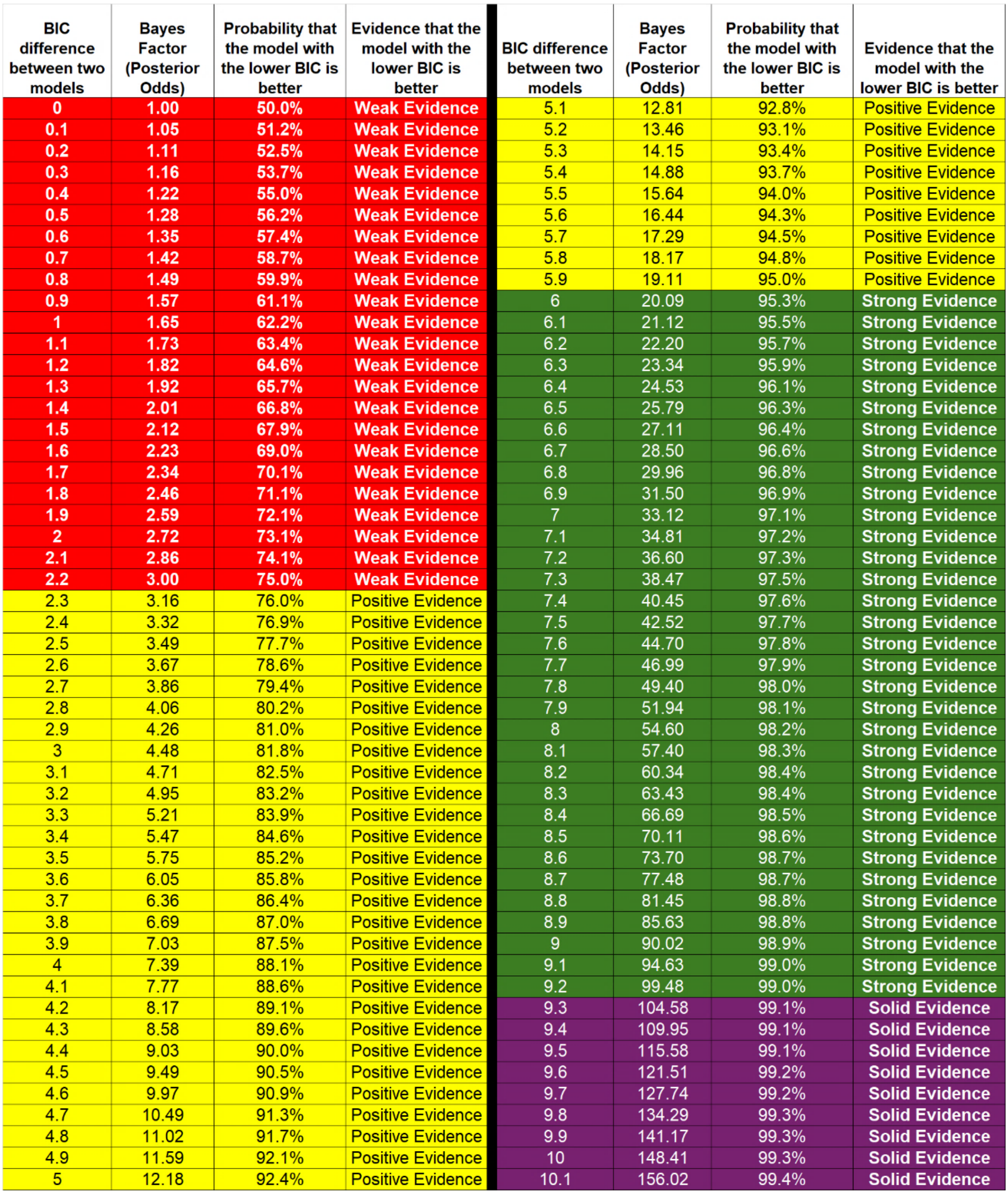
Table S2. Package–Function Mapping – Data Handling and preprocessing

Functionality	**Package(s)**	**Notes**
Data import (Excel)	readxl	Import .xlsx; used for all figures.
Data import (SPSS)	haven	Import .sav files.
Data wrangling	dplyr, tidyr, vctrs, forcats, stringr	Standardize, reshape, encode, and clean variables.
Visualization (labels)	ggrepel	Overlap-avoiding text/markers (e.g., asterisks near endpoints).
Visualization (rich text)	ggtext	HTML/Markdown text in plots (legend/annotations).
Visualization (core)	ggplot2	Publication-ready plots.
Figure assembly/layout	cowplot, grid	Compose multi-panel figures; layout control; grid::unit() for sizing.
Graphics device / export	ragg	High-resolution TIFF/PNG via agg_tiff(); improved anti-aliasing.
Model tidying	broom	Convert model objects into tidy data frames.
Table export	officer, flextable, openxlsx	Generate Word and Excel tables.
Reproducibility & tooling	tictoc, conflicted, parallelly	Runtime tracking, namespace resolution, parallel tools.
Collinearity diagnostics	car	Variance inflation factor (VIF) calculations.

Table S3. Package–Function Mapping – Modelling, Evaluation, Inference

Functionality	Package(s)	Notes
LASSO regression	glmnet	Penalized variable selection.
Bayesian modelling	brms, rstan, cmdstanr	Logistic models; Stan backend; robust PSIS-LOO.
Frequentist modelling	stats::glm, lme4::glmer	GLMs and mixed-effects (study random intercept when used).
Model tidying (GLMs/GLMMs)	broom, broom.mixed	Tidy summaries for reporting.
Information criteria	loo	PSIS-LOOIC (−2×elpd), Pareto-k diagnostics, model comparison.
Discrimination & ROC	pROC	AUROC, DeLong CIs, Youden-J threshold.
Cross-validation	custom K-fold; future, future.apply	Stratified K-fold CV; thresholds learned in training; parallel execution.
Bootstrapping & resampling	base R (sample, quantile); optional: future	Paired/bootstrap resampling incl. mean Δrank across schemes (B=10,000).
PCA	stats::prcomp	Dimensionality reduction; variance explained.
Hierarchical partitioning	MuMIn, glmm.hp	Partition McFadden's R^2 into unique/joint contributions.
Mixed-effects & GLMs	lme4, stats	Fit fixed/random-effects logistic models.
Model comparison & pseudo-R^2	pscl	McFadden's R^2 and related indices.
Rank tests & post-hoc	stats::friedman.test, PMCMRplus	Friedman test; Conover post-hoc comparisons.
Multiple testing	stats::p.adjust(method = 'BH')	Benjamini–Hochberg FDR control.
Decision-curve analysis	ggplot2, dplyr (or mda / dcurves)	Net benefit vs threshold using OOF predictions; 95% bootstrap CIs.

Table S4. Estimating model superiority based on BIC differences: interpreting the probability that the lower-BIC model is better



BIC difference between two models	Bayes Factor (Posterior Odds)	Probability that the model with the lower BIC is better	Evidence that the model with the lower BIC is better	BIC difference between two models	Bayes Factor (Posterior Odds)	Probability that the model with the lower BIC is better	Evidence that the model with the lower BIC is better
0	1.00	50.0%	Weak Evidence	5.1	12.81	92.8%	Positive Evidence
0.1	1.05	51.2%	Weak Evidence	5.2	13.46	93.1%	Positive Evidence
0.2	1.11	52.5%	Weak Evidence	5.3	14.15	93.4%	Positive Evidence
0.3	1.16	53.7%	Weak Evidence	5.4	14.88	93.7%	Positive Evidence
0.4	1.22	55.0%	Weak Evidence	5.5	15.64	94.0%	Positive Evidence
0.5	1.28	56.2%	Weak Evidence	5.6	16.44	94.3%	Positive Evidence
0.6	1.35	57.4%	Weak Evidence	5.7	17.29	94.5%	Positive Evidence
0.7	1.42	58.7%	Weak Evidence	5.8	18.17	94.8%	Positive Evidence
0.8	1.49	59.9%	Weak Evidence	5.9	19.11	95.0%	Positive Evidence
0.9	1.57	61.1%	Weak Evidence	6	20.09	95.3%	Strong Evidence
1	1.65	62.2%	Weak Evidence	6.1	21.12	95.5%	Strong Evidence
1.1	1.73	63.4%	Weak Evidence	6.2	22.20	95.7%	Strong Evidence
1.2	1.82	64.6%	Weak Evidence	6.3	23.34	95.9%	Strong Evidence
1.3	1.92	65.7%	Weak Evidence	6.4	24.53	96.1%	Strong Evidence
1.4	2.01	66.8%	Weak Evidence	6.5	25.79	96.3%	Strong Evidence
1.5	2.12	67.9%	Weak Evidence	6.6	27.11	96.4%	Strong Evidence
1.6	2.23	69.0%	Weak Evidence	6.7	28.50	96.6%	Strong Evidence
1.7	2.34	70.1%	Weak Evidence	6.8	29.96	96.8%	Strong Evidence
1.8	2.46	71.1%	Weak Evidence	6.9	31.50	96.9%	Strong Evidence
1.9	2.59	72.1%	Weak Evidence	7	33.12	97.1%	Strong Evidence
2	2.72	73.1%	Weak Evidence	7.1	34.81	97.2%	Strong Evidence
2.1	2.86	74.1%	Weak Evidence	7.2	36.60	97.3%	Strong Evidence
2.2	3.00	75.0%	Weak Evidence	7.3	38.47	97.5%	Strong Evidence
2.3	3.16	76.0%	Positive Evidence	7.4	40.45	97.6%	Strong Evidence
2.4	3.32	76.9%	Positive Evidence	7.5	42.52	97.7%	Strong Evidence
2.5	3.49	77.7%	Positive Evidence	7.6	44.70	97.8%	Strong Evidence
2.6	3.67	78.6%	Positive Evidence	7.7	46.99	97.9%	Strong Evidence
2.7	3.86	79.4%	Positive Evidence	7.8	49.40	98.0%	Strong Evidence
2.8	4.06	80.2%	Positive Evidence	7.9	51.94	98.1%	Strong Evidence
2.9	4.26	81.0%	Positive Evidence	8	54.60	98.2%	Strong Evidence
3	4.48	81.8%	Positive Evidence	8.1	57.40	98.3%	Strong Evidence
3.1	4.71	82.5%	Positive Evidence	8.2	60.34	98.4%	Strong Evidence
3.2	4.95	83.2%	Positive Evidence	8.3	63.43	98.4%	Strong Evidence
3.3	5.21	83.9%	Positive Evidence	8.4	66.69	98.5%	Strong Evidence
3.4	5.47	84.6%	Positive Evidence	8.5	70.11	98.6%	Strong Evidence
3.5	5.75	85.2%	Positive Evidence	8.6	73.70	98.7%	Strong Evidence
3.6	6.05	85.8%	Positive Evidence	8.7	77.48	98.7%	Strong Evidence
3.7	6.36	86.4%	Positive Evidence	8.8	81.45	98.8%	Strong Evidence
3.8	6.69	87.0%	Positive Evidence	8.9	85.63	98.8%	Strong Evidence
3.9	7.03	87.5%	Positive Evidence	9	90.02	98.9%	Strong Evidence
4	7.39	88.1%	Positive Evidence	9.1	94.63	99.0%	Strong Evidence
4.1	7.77	88.6%	Positive Evidence	9.2	99.48	99.0%	Strong Evidence
4.2	8.17	89.1%	Positive Evidence	9.3	104.58	99.1%	Solid Evidence
4.3	8.58	89.6%	Positive Evidence	9.4	109.95	99.1%	Solid Evidence
4.4	9.03	90.0%	Positive Evidence	9.5	115.58	99.1%	Solid Evidence
4.5	9.49	90.5%	Positive Evidence	9.6	121.51	99.2%	Solid Evidence
4.6	9.97	90.9%	Positive Evidence	9.7	127.74	99.2%	Solid Evidence
4.7	10.49	91.3%	Positive Evidence	9.8	134.29	99.3%	Solid Evidence
4.8	11.02	91.7%	Positive Evidence	9.9	141.17	99.3%	Solid Evidence
4.9	11.59	92.1%	Positive Evidence	10	148.41	99.3%	Solid Evidence
5	12.18	92.4%	Positive Evidence	10.1	156.02	99.4%	Solid Evidence

Bayes Factor (Posterior Odds): The posterior odds are calculated using the BIC differences. It quantifies how much more likely one model is better compared to another. The Bayes factor can be estimated from the BIC difference by the formula: **Bayes Factor = $e^{\Delta BIC \div 2}$**. Here, the ΔBIC is the difference in BIC scores between the two models. This exponentiation reflects how changes in BIC scores can exponentially affect the likelihood ratio between the two models. The probability that the model with lower BIC is better can be derived from the Bayes Factor. If the Bayes factor is *B*, the probability *P* that the model with the lower BIC is better – after considering the observed data – can be estimated as: $\boldsymbol{P = \beta \div (1+\beta)}$**.** This formula assumes equal prior probabilities for the two models. Evidence Strength (e.g., weak, positive, strong, solid) is based on thresholds of the Bayes factor or the BIC differences. Commonly, larger Bayes factors indicate stronger evidence for one model over another. Specific thresholds for these categories can vary, but, this table is based on the suggestions by Raftery (1995)[29] but without rounding. The table stops at a ΔBIC of 10.1 as the evidence remains "Solid" at any point larger than 9.2. The differences between models using the Leave-One-Out Information Criterion (LOOIC) are interpreted the same way.

Table S5. Comparison of Bayesian Information Criterion (BIC) and Leave-One-Out Information Criterion (LOOIC)

Criterion	BIC	LOOIC
Definition	The BIC examines model performance. The BIC is a criterion for model selection that balances goodness-of-fit and model complexity.	The LOOIC is used to evaluate predictive accuracy with a focus on model generalizability. The LOOIC is a cross-validation-based metric that assesses model predictive accuracy.
Focus	Balancing model fit and complexity (parsimony).	Model generalizability and predictive accuracy.
Interpretation	Lower BIC values indicate a better trade-off between model fit and complexity.	Lower LOOIC values suggest better predictive accuracy on unseen data.
Complexity Penalty	Penalizes models more heavily for added parameters to avoid overfitting.	Penalizes overfitting implicitly by estimating prediction errors via cross-validation.
Underlying Assumption	Assumes the data are from a parametric model (often normal distribution).	Relies on fewer distributional assumptions and uses resampling.
Use Case	Suitable for comparing nested models or simpler parametric models.	Ideal for complex models, including hierarchical or non-nested models.
Data Requirements	Requires a well-defined likelihood function.	Can work with a broader range of model types, including Bayesian models.
Computational Cost	Relatively low, as it requires only the model likelihood and parameter count.	Computationally intensive due to resampling or approximation techniques.
Strengths	Simple, fast, and effective for straightforward parametric models.	Robust and adaptable, particularly for evaluating predictive performance.
Limitations	May perform poorly with complex or hierarchical models.	Computationally expensive and sensitive to the choice of approximation methods.

Table created from the works of Vehtari *et al.* (2017) [18], Schwarz *et al.* (1978)[30] & Burnham *et al.* (2004) [31]

Table S6. Comparison principal component analysis and hierarchical partitioning

Aspect	Principal Component Analysis	Hierarchical Partitioning
Definition	A dimensionality-reduction technique that transforms correlated variables into uncorrelated principal components.	A method to quantify each predictor's relative importance by partitioning model R^2 into unique and shared contributions (via R^2 partitioning).
Focus	Variance structure within predictors (e.g., z-scores for TLC, FEV$_1$, DLNO$_{10s;}$ GAMLSS), reducing dimensionality for modelling.	Variance in the outcome (emphysema = disease) explained by each predictor, accounting for collinearity and shared effects.
Interpretation	PC1, PC2, PC3, and PC4 are orthogonal linear combinations of the z-score predictors. In our data, PC1 loaded strongly on DLNO$_{10s}$ and explained ~52% of predictor variance; PC2 represents hyperinflation (~17%), and PC3 represents airway obstruction (~12%), and PC4 was non-significant).	McFadden's R^2 = 0.663. Individual, standalone contributions (% of total) was ~78% for FEV$_1$ z-scores, ~48% for DLNO$_{10s}$ (GAMLSS) z-scores, and ~13% for TLC z-scores.
Complexity Penalty	None directly; adding PCs to regression increases model complexity (e.g., BIC/AIC trade-offs).	Adjusts via R^2 partitioning. High collinearity can complicate interpretation; in our data VIFs were modest (~1.09 to 1.62) for the three-predictor model of z-scores for TLC, FEV$_1$, DLNO$_{10s}$). However, VIFs ranged from 1.12 to 4.44 in the four-predictor model of z-scores for TLC, FEV$_1$, DLNO$_{10s}$, and DLCO$_{10s.}$
Underlying Assumption	Linear relationships among predictors; variance is informative. Outcome is not used to derive PCs.	Predictors contribute to the outcome in a linear fashion; a fitted model (e.g., logistic regression) is valid for R^2 partitioning.
Use Case	Simplify multicollinear predictors (z-scores; typical VIFs ~1.4) for regression; identify key variance patterns.	Assess predictor importance for disease, especially with collinear predictors; prioritize z-scores of TLC, FEV$_1$, and DLNO for contributions.
Data Requirements	Numeric continuous predictors (z-scores); no missing values after listwise deletion (n = 408).	Numeric predictors and a binary outcome (disease: yes/no); complete cases; fitted logistic regression.
Computational Cost	Low to moderate; matrix decomposition scales with number of predictors.	Moderate; depends on model fitting and number of R^2 partitions; scales with predictors and sample size (n = 408).
Strengths	Reduces dimensionality; handles collinearity; identifies major variance patterns.	Quantifies unique and shared variance; adjusts for collinearity; provides clear predictor importance (e.g., strong effect for FEV$_1$).
Limitations	Less interpretable than original predictors; assumes linear relations; does not model the outcome directly.	Sensitive to model specification; requires a valid outcome model; interpretations can be affected by shared variance.

Table S7. Definitions and alternatives names for classification metrics

1. Kappa Statistic (κ): The Kappa statistic measures the agreement between two raters or classification methods, accounting for the agreement that could occur by chance. It ranges from -1 (no agreement) to 1 (perfect agreement), with 0 indicating chance-level agreement.

2. F_1 Score: The F1 Score is the harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives. It is commonly used for imbalanced datasets.

3. Accuracy: Accuracy is the proportion of correctly classified observations (both true positives and true negatives) out of all observations. It is sensitive to class imbalance.

4. Balanced Accuracy: Balanced accuracy is the average of sensitivity and specificity, providing a performance measure that accounts for class imbalance. It is particularly useful when the dataset is skewed.

5. Sensitivity (Recall, True Positive Rate, TPR): Sensitivity measures the proportion of actual positives correctly identified as positive. Also called recall or the true positive rate (TPR).

6. Specificity (True Negative Rate, TNR): Specificity is the proportion of actual negatives correctly identified as negative. It is also known as the true negative rate (TNR).

7. **Positive Predictive Value (PPV, Precision):** PPV indicates the proportion of positive test results that are true positives (TP). Also called precision, it assesses how reliable positive classifications are.

8. Negative Predictive Value (NPV): NPV measures the proportion of negative test results that are true negatives (TN), indicating the reliability of negative classifications.

9. False Omission Rate (FOR): FOR is the proportion of false negatives (FN) among all negative predictions, indicating how often a negative prediction is incorrect.

10. False Positive Rate (FPR, Fall-out): FPR is the proportion of false positives (FP) among all actual negatives. It is also called the fall-out rate and represents the chance of a false alarm.

11. False Negative Rate (FNR, Miss Rate): FNR measures the proportion of actual positives that are incorrectly classified as negatives. It is also known as the miss rate.

12. False Discovery Rate (FDR): FDR is the proportion of false positives (FP) among all positive predictions, indicating how often a positive prediction is incorrect.

13. Positive Likelihood Ratio (**+LR**): +LR quantifies how much more likely a positive test result is for someone with the condition compared to someone without the condition.

14. Negative Likelihood Ratio (−LR): −LR quantifies how much less likely a negative test result is for someone with the condition compared to someone without the condition.

15. Matthews Correlation Coefficient (MCC): MCC evaluates the correlation between observed and predicted classifications, considering all confusion matrix elements. It ranges from -1 (inverse prediction) to 1 (perfect prediction).

16. Diagnostic Odds Ratio (DOR): DOR combines sensitivity and specificity to describe the odds of a positive test result in those with the condition versus those without it. A higher value indicates better test performance.

Table S8. Formulas for key classification metrics

1. **Kappa Statistic** (κ): $\kappa = \frac{p_o - p_e}{1 - p_e}$ Where: • $p_o = \frac{TP+TN}{TP+TN+FP+FN}$ (Observed agreement) • $p_e = \frac{(TP+FP)(TP+FN)+(TN+FN)(TN+FP)}{(TP+TN+FP+FN)^2}$ (Expected agreement)	9. **False Omission Rate (FOR):** $\text{FOR} = \frac{FN}{FN + TN}$
2. **F1 Score:** $F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ Where: • $\text{Precision} = \frac{TP}{TP+FP}$ • $\text{Recall} = \frac{TP}{TP+FN}$ (Sensitivity)	10. **False Positive Rate (FPR):** $\text{FPR} = \frac{FP}{FP + TN}$
3. Accuracy and Discordance: $\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$ Discordance: = 1 – Accuracy	11. **False Negative Rate (FNR):** $\text{FNR} = \frac{FN}{FN + TP}$
4. **Balanced Accuracy:** $\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$	12. **False Discovery Rate (FDR):** $\text{FDR} = \frac{FP}{FP + TP}$
5. **Sensitivity (Recall):** $\text{Sensitivity} = \frac{TP}{TP + FN}$	13. **Positive Likelihood Ratio (+LR):** $+\text{LR} = \frac{\text{Sensitivity}}{1 - \text{Specificity}}$
6. **Specificity:** $\text{Specificity} = \frac{TN}{TN + FP}$	14. **Negative Likelihood Ratio (-LR):** $\text{-LR} = \frac{1 - \text{Sensitivity}}{\text{Specificity}}$
7. **Positive Predictive Value (PPV):** $\text{PPV} = \frac{TP}{TP + FP}$	15. **Matthews Correlation Coefficient (MCC):** $\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
8. **Negative Predictive Value (NPV):** $\text{NPV} = \frac{TN}{TN + FN}$	16. **Diagnostic Odds Ratio (DOR):** $\text{DOR} = \frac{\text{Sensitivity} \cdot \text{Specificity}}{(1 - \text{Sensitivity}) \cdot (1 - \text{Specificity})}$ Alternatively: $\text{DOR} = \frac{TP \cdot TN}{FP \cdot FN}$

Table S9. Additional definitions and alternatives names for classification metrics

- **Unique (Independent) contribution:** The Unique contribution is the share of model fit (McFadden's R^2) attributable solely to a predictor after apportioning overlap with the others. It is computed as the average marginal increase in McFadden's R^2 when adding the predictor to every possible subset of the remaining predictors (Shapley value). By construction it is non-negative**.** Larger values indicate a stronger independent effect.

- **Average share:** The part of fit associated with a predictor that is shared with other predictors (overlap/redundancy) or that arises from mild suppression/synergy. Defined operationally as **Average share = Standalone − Unique.** It can be positive (overlap/redundancy) or negative (suppression).

- **Individual (Standalone) contribution:** The McFadden's R^2 from a single-predictor model that includes only that predictor. It reflects explanatory power in isolation**,** without adjusting for shared effects with other predictors.

- **Unique (% of total):** The predictor's Unique share as a percentage of the **full model's** McFadden's R^2: Unique (%) = 100 × (Unique / Total R^2). These percentages across predictors should sum to ~**100%** (up to rounding) and provide a decomposition-consistent ranking.

- **Average share (% of total):** The Average share as a percentage of the full model's McFadden's R^2. **Average share (%) = 100 × ((Standalone − Unique) / Total R^2).** Values may be positive (overlap) or negative (suppression).

- **Individual (Standalone) (% of total):** The Standalone contribution as a percentage of the full model's McFadden's R^2: Standalone (%) = 100 × (Standalone / Total R^2). These do not sum to 100% across predictors (each single-predictor model captures overlapping signal).

- **Joint (shared) component:** The portion of fit not uniquely assignable to any single predictor: Joint = Total R^2 − Σ Unique. A negative Joint indicates mild redundancy/overlap among predictors (the sum of Unique slightly exceeds the total), which is common and typically small.

- **Tjur's R^2 (Coefficient of Discrimination, D):** Mean predicted probability among cases minus mean predicted probability among controls: $\mathbf{D = E[p \mid y=1] - E[p \mid y=0]}$. Tjur's R^2 measures how well the model separates cases from controls in absolute risk space. Threshold-free and easy to interpret: **0** means no separation; **1** means perfect separation. Sensitive to calibration (shifts that change mean predicted risks will change D).

- **Efron's R^2:** Fraction of variance in the binary outcome explained by predicted probabilities: $\mathbf{R^2_E = 1 - \Sigma(y - p)^2 / \Sigma(y - \bar{y})^2}$. Efron's R^2 is a mean-squared-error–based pseudo-R^2 comparing the model to a null model that always predicts the prevalence $\bar{y}$. Can be negative if the model is worse than the null; approaches **1** with perfect predictions. Reflects both discrimination and calibration.

- **Brier Score:** Mean squared error between predicted probabilities and the actual outcomes (0 or 1). The **reference** is usually a naive model that always predicts the **outcome prevalence**

Table S10. BIC and LOOIC values for the 29 models examined.

Model	Is a "Study Level Random Intercept Needed?	BIC	BIC Rank	LOOIC	LOOIC Rank
Model C [TLC & FEV$_1$ (GLI), DLNO (GAMLSS)]	No	164.6	1	149.7	2
Four predictor z-scores [TLC & FEV$_1$ & KCO (GLI), DLNO (GAMLSS)]	No	168.5	2	149.8	3
Model B [TLC & FEV$_1$ & DLCO (GLI), DLNO (GAMLSS)]	No	169.6	3	150.9	4
Four predictor z-scores [TLC & FEV$_1$(GLI), DLNO & DLCO (GAMLSS)]	No	170.4	4	151.7	6
Three predictor z-scores [TLC & FEV$_1$ & DLCO (GLI)]	No	170.6	5	155.4	7
Five predictor z-scores [(TLC, FEV$_1$ & FEV$_1$/FVC & KCO (GLI), DLNO (GAMLSS)]	No	172.1	6	149.1	1
Model A [TLC & FEV$_1$ (GLI), DLCO (GAMLSS)]	No	175	7	159.9	8
Three predictor z-scores [TLC & FEV$_1$ (GLI), DLCO (van der Lee)]	No	177.4	8	162.4	9
Six predictor z-scores [TLC & FEV$_1$ & FEV$_1$/FVC & KCO (GLI); KNO (van der Lee), DLNO (GAMLSS)]	No	178.1	9	151.6	5
FEV$_1$/FVC z-scores (GLI)	No	209.6	10	201.7	10
FEV$_1$ z-scores (GLI)	No	214.5	11	206.7	11
RV/TLC z-scores (GLI)	Yes	285.9	12	266.1	12
Summed z-scores (DLNO + DLCO, GAMLSS)	No	288.8	13	281	13
KNO z-scores (van der Lee)	No	292	14	284.1	17
Two predictor z-scores [DLNO (GAMLSS), DLCO (GLI)]	No	294.1	15	282.5	14
Two predictor z-scores (DLNO & DLCO, GAMLSS)	No	294.7	16	282.9	15
DLNO z-scores (GAMLSS)	No	297.7	17	289.9	18
DLCO z-scores (GLI)	No	300.2	18	292.3	19
Summed z-scores (DLNO+DLCO, SLR)	No	301.1	19	293.2	20
DLCO z-scores (GAMLSS)	No	301.9	20	294	22
DLCO z-scores (SLR)	No	304.1	21	296.1	24
Summed z-scores (DLNO+DLCO, van der Lee)	No	304.3	22	296.5	25
KCO z-scores (GLI)	Yes	305	23	283.8	16
Two predictor z-scores (DLNO, DLCO, SLR)	No	305.6	24	293.6	21
Two predictor z-scores (DLNO, DLCO, van der Lee)	No	308.3	25	296.5	25
DLCO z-scores (van der Lee)	No	309.2	26	301.4	27
KCO z-scores (van der Lee)	Yes	315.8	27	295.4	23
DLNO z-scores (SLR)	Yes	320.1	28	312.1	28
DLNO z-scores (van der Lee)	Yes	322.2	29	314.4	30
FVC z-scores (GLI)	Yes	330.4	30	313.8	29
TLC z-scores (GLI)	Yes	376	31	356.9	31
VA z-scores (GLI)	Yes	413.9	32	395.4	32
VA z-scores (GAMLSS)	Yes	415.3	33	395.4	32
VA z-scores (SLR)	No	428.1	34	420.2	34

SLR = Segmented Linear Regression (Zavorsky & Cao 2022) ; GAMLSS = Generalized Additive Models of Location Scale & Shape (Zavorsky & Cao 2022); GLI = Global Lung Function Initiative Equations; van der Lee = Equations of van der Lee et al., 2007.

Table S11. Principal Component (PC) Analyses Loadings (PCs used in model)

Variable	PC1	PC2	PC3	PC4
FEV$_1$ z-scores, GLI equations, [Quanjer et al. 2012]	0.24	0.04	**0.43**	0.18
FVC z-scores, GLI equations, [Quanjer et al. 2012]	0.18	–0.02	**0.42**	0.36
FEV$_1$/FVC z-scores, GLI equations, [Quanjer et al. 2012]	0.22	0.06	0.27	–0.06
TLC z-scores, GLI equations, [Hall et al. 2021]	0.02	**–0.49**	–0.16	0.26
RV/TLC z-scores, GLI equations, [Hall et al. 2021]	–0.12	–0.30	**–0.48**	–0.17
DLCO z-scores, SLR, [Zavorsky & Cao 2022]	**0.32**	0.04	–0.14	0.01
DLCO z-scores, GAMLSS, [Zavorsky & Cao 2022]	**0.32**	0.05	–0.12	0.00
DLCO z-scores, GLI equations, [Stanojevic et al. 2017]	**0.32**	0.04	–0.15	–0.03
DLCO z-scores, [van der Lee et al. 2007]	**0.32**	0.02	–0.13	–0.02
DLNO z-scores, SLR, [Zavorsky & Cao 2022]	**0.31**	–0.13	–0.10	0.03
DLNO z-scores, GAMLSS, [Zavorsky & Cao 2022]	**0.32**	–0.10	–0.11	0.01
DLNO z-scores, [van der Lee et al. 2007]	**0.31**	–0.14	–0.11	0.01
VA z-scores, SLR, [Zavorsky & Cao 2022]	0.06	**–0.35**	0.29	**–0.52**
VA z-scores, GAMLSS, [Zavorsky & Cao 2022]	0.06	**–0.34**	0.30	**–0.52**
VA z-scores, GLI equations, [Stanojevic et al. 2017]	0.17	**–0.44**	–0.02	0.29
KCO z-scores, GLI equations, [Stanojevic et al. 2017]	0.21	0.35	–0.14	–0.24
KNO z-scores, [van der Lee et al. 2007]	0.26	0.23	–0.11	–0.24

Principal Component 1 (PC1) represents alveolar-capillary gas transfer (DLNO, DLCO). Principal Component 2 (PC2) represents hyperinflation (TLC, VA). Principal Component 3 (PC3) represents airway obstruction (FEV$_1$, FVC) and air trapping (RV/TLC). Principal Component 4 (PC4)

Table S12. Logistic Regression Coefficients (Principal Component Analyses Model with PC1-PC4)

Term	Estimate	Standard Error	Statistic	Odds Ratio	*p*-value
Intercept	–2.47	0.26	–9.46	0.084	<0.00001
PC1	–0.89	0.10	–8.68	0.411	<0.00001
PC2	–0.63	0.13	–4.86	0.532	<0.00001
PC3	–0.85	0.15	– 5.58	0.427	<0.00001
PC4	0.09	0.17	0.54	1.094	0.587

The baseline odds of COPD when all PCs are zero (mean values) are 0.084, or about 8% of the odds of being a control, reflecting a strong baseline tendency toward the control group. A one-standard-deviation increase in PC1 (better gas alveolar-capillary transfer reduces the odds of COPD by approximately 59% (1 – 0.411). A one-standard-deviation reduction in PC2 (reduced hyperinflation and air trapping) reduces the odds of COPD by about 47% (1 – 0.532). A one-standard-deviation increase in PC3 (increases in FEV_1, FVC spirometry and decreases in RV/TLC) reduces the odds of COPD by about 57% (1 – 0.427). A one-unit increase in PC4 (higher VA variability, FVC) increases the odds of COPD by about 9% (1.094 – 1), but this is not significant, indicating no reliable effect.

Table S13. Logistic Regression Coefficients (Reduced PCA Model with PC1-PC3)

Term	Estimate	Standard Error	Statistic	Odds ratio	*p*-value
Intercept	–2.45	0.26	-9.50	0.086	<0.00001
PC1	–0.89	0.10	-8.70	0.411	<0.00001
PC2	–0.63	0.13	-4.86	0.532	<0.00001
PC3	–0.84	0.15	-5.59	0.432	<0.00001

The baseline odds of COPD when all PCs are zero (mean values) are 0.064, or about 8% of the odds of being a control, reflecting a strong baseline tendency toward the control group. A one-standard-deviation increase in PC1 (better alveolar-capillary gas transfer) reduces the odds of COPD by approximately 59% (1 – 0.411). A one-standard-deviation reduction in PC2 (reduced hyperinflation and air trapping) reduces the odds of COPD by about 47% (1 – 0.532). A one-standard-deviation increase in PC3 (increases in FEV_1, FVC spirometry and decreases in RV/TLC) reduces the odds of COPD by about 57% (1 – 0.432).

Table S14. Comparison of Model fit

Model	AIC	BIC
Full Model (PC1-PC4)	170.3	190.4
Reduced Model (PC1-PC3)	168.6	184.6

The odds ratios for PC1, PC2, and PC3 are nearly identical to those of the full model, confirming that the exclusion of PC4 does not affect the significant predictors. The slight difference in PC3 (0.427 vs. 0.432) is negligible. The reduced model maintains the same protective effects: better gas transfer (PC1), less hyperinflation (PC2), and improved spirometry (PC3) significantly reduce COPD odds

Table S15. Hierarchical partitioning results for the three main predictors of emphysema: (FEV_1 z-scores, TLC z-scores and $DLNO_{10s}$ z-scores)

Variable	Unique contribution (R^2) [95%CI]	Average Share (R^2) [95%CI]	Individual (Standalone) Contribution (R^2) [95% CI]	Variance Inflation Factor
FEV_1 Z-scores, GLI equations, [Quanjer et al. 2012]	0.354 [0.272, 0.451]	0.161 [0.120, 0.204]	0.515 [0.409, 0.637]	1.58
$DLNO_{10s}$ Z-scores, GAMLSS, [Zavorsky & Cao 2022]	0.213 [0.154, 0.287]	0.103 [0.059, 0.148]	0.316 [0.228, 0.417]	1.62
TLC Z-scores, GLI equations, [Hall et al. 2021]	0.111 [0.061, 0.179]	–0.026 [–0.063, 0.015]	0.084 [0.028, 0.166]	1.09

Variable	Unique Contribution (% of total) [95% CI]	Average Share (% of total) [95%CI]	Individual (Standalone) Contribution (% of total) [95% CI]
FEV_1 Z-scores, GLI equations, [Quanjer et al. 2012]	53.4% [43.2, 62.5]	24.2% [19.0, 28.5]	77.6% [66.0, 86.8]
$DLNO_{10s}$ Z-scores, GAMLSS, [Zavorsky & Cao 2022]	32.1%[23.7, 41.2]	15.5% [9.2, 21.0]	47.6% [35.5, 59.2]
TLC Z-scores, GLI equations, [Hall et al. 2021]	16.7% [9.0, 26.6]	–4.0% [–9.4, 2.2]	12.7% [4.2, 24.6]

Component	Joint R^2 [95%CI]	% of Total [95%CI]
Joint (shared) contribution	–0.015 [–0.025,–0.005]	–2.2% [–3.6, –0.7]
Variance-like Summaries		
Total McFadden R^2[95% CI]	0.663 [0.577, 0.774]	100.0
Tjur (R^2) [95% CI] (coefficient of discrimination)	0.691 [0.618, 0.760]	N/A
Efron R^2 [95% CI]	0.679 [0.574, 0.773]	N/A
Brier Score [95% CI]	0.051 [0.035, 0.069]	N/A

95% CIs are from 20,000 bootstrap resamples. The Independent contribution (Shapley) is the average marginal gain in McFadden's R^2 when a predictor is added across all possible subsets; it is the portion of fit uniquely attributable to that predictor after apportioning overlap. The Joint (shared) component is the residual fit not uniquely assignable to any single predictor; a negative value indicates mild redundancy/suppression among predictors (the sum of unique parts slightly exceeds the total, so the joint term is negative). In this table, the Unique Contributions sum to 0.354 + 0.213 + 0.111= 0.678, while the total McFadden's R^2 is 0.663, yielding Joint = −0.015 (≈ −2.3% of total). Redundancy is therefore small. The ranking by unique importance is **FEV_1 > $DLNO_{10s}$ > TLC**, and all three also show meaningful standalone contributions (0.515, 0.316, 0.084), supporting inclusion of all three predictors. (Minor differences reflect rounding.) Specifically, McFadden's R^2=0.663 indicates that the 3-predictor DLNO logistic model reduces deviance (or equivalently improves log-likelihood) by ~66% relative to an intercept-only (null) model. Note that McFadden's R^2 is a relative fit measure, not "variance explained" (variance isn't defined the same way for a binary outcome).

Table S16. Hierarchical partitioning results when adding $DLCO_{10s}$ z-scores (GLI) to the 3-predictor $DLNO_{10s}$ (GAMLSS) model of emphysema

Variable	Unique contribution (R^2) [95%CI]	Average Share (R^2) [95%CI]	Individual (Standalone) Contribution (R^2) [95% CI]	Variance Inflation Factor
FEV_1 Z-scores, GLI equations, [Quanjer et al. 2012]	0.286 [0.211, 0.377]	0.229 [0.174, 0.288]	0.515 [0.409, 0.637]	1.58
$DLNO_{10s}$ Z-scores, GAMLSS, [Zavorsky & Cao 2022]	0.122 [0.087, 0.170]	0.194 [0.133, 0.260]	0.316 [0.228, 0.417]	4.44
TLC Z-scores, GLI equations, [Hall et al. 2021]	0.113 [0.062, 0.183]	–0.029 [–0.071, 0.018]	0.084 [0.028, 0.166]	1.12
$DLCO_{10s}$ Z-scores, GLI, [Stanojevic et al. 2017]	0.103 [0.071, 0.147]	0.207 [0.148, 0.271]	0.310 [0.222, 0.411]	3.93

Variable	Unique Contribution (% of total) [95% CI]	Average Share (% of total) [95%CI]	Individual (Standalone) Contribution (% of total) [95% CI]
FEV_1 Z-scores, GLI equations, [Quanjer et al. 2012]	42.9% [32.9, 52.4]	34.4% [27.6, 39.8]	77.4% [66.0, 86.8]
$DLNO_{10s}$ Z-scores, GAMLSS, [Zavorsky & Cao 2022]	18.3%[13.3, 24.1]	29.2% [20.6, 36.5]	47.5% [35.5, 59.2]
TLC Z-scores, GLI equations, [Hall et al. 2021]	17.0% [9.1, 26.9]	–4.3% [–10.5, 2.6]	12.7% [4.2, 24.6]
$DLCO_{10s}$ Z-scores, GLI, [Stanojevic et al. 2017]	15.5% [10.7, 21.3]	31.1% [22.9, 38.2]	46.5% [34.1, 58.4]

Component	Joint R^2 [95%CI]	% of Total [95%CI]
Joint (shared) contribution	0.042 [0.022, 0.064]	6.3% [3.3, 9.0]
Variance-like Summaries		
Total McFadden R^2[95% CI]	0.666 [0.582, 0.778]	100.0
Tjur (R^2) [95% CI] (coefficient of discrimination)	0.704 [0.632, 0.772]	N/A
Efron R^2 [95% CI]	0.687 [0.581, 0.780]	N/A
Brier Score [95% CI]	0.052 [0.035, 0.070]	N/A

95% CIs are from 20,000 bootstrap resamples. The Independent contribution (Shapley) is the average marginal gain in McFadden's R^2 when a predictor is added across all possible subsets; it represents the portion of fit uniquely attributable to that predictor after apportioning overlap. The **Joint (shared)** component is the residual fit not uniquely assignable to any single predictor; a **positive** value indicates shared information/synergy among predictors (part of the fit is common across predictors). In this table, the Unique Contributions sum to 0.286 + 0.122 + 0.113 + 0.103 = 0.624 while the total McFadden's R^2 0.666, yielding **Joint = 0.042** (≈ **6.3%** of total). This implies modest shared structure and limited redundancy. The ranking by unique importance is

FEV_1 > $DLNO_{10s}$ > TLC > $DLCO_{10s}$, and all four also show meaningful standalone contributions (0.515, 0.316, 0.084, 0.310), supporting inclusion of all predictors. (Minor differences reflect rounding.) Specifically, McFadden's R^2=0.666 indicates the 4-predictor logistic model reduces deviance (i.e., improves log-likelihood) by ~67% relative to a null model; note McFadden's $R^2$is a **relative fit** index—not variance explained for a binary outcome. VIFs ≤ 4.44 suggest no severe multicollinearity.

Comparison with Table S14 (3-predictor model): In **Table S15**, total McFadden's R^2 changes only slightly **(**0.666) compared to **Table S14** (vs 0.663); trivial difference within CIs). Discrimination metrics improve modestly: Tjur's R^2 0.704 vs 0.691 and Efron's R^2 0.687 vs 0.679**;** the Brier score is essentially unchanged (0.052 vs 0.051). Adding $DLCO_{10s}$ contributes small, but unique information (0.103≈15.5% of total) and increases the **s**hared (Joint) fit from −0.015 (≈ −2.2%) to + 0.042 (≈ 6.3%), indicating overlap/synergy with the other predictors rather than harmful redundancy. VIFs up to 4.44 suggest moderate, but not severe, collinearity (notably between $DLNO_{10s}$ and $DLCO_{10s}$).

Although z-scores for $DLNO_{10s}$ (GAMLSS) and $DLCO_{10s}$ (GLI) were strongly correlated (pairwise R^2 ≈0.73), VIFs were <5 ($DLNO_{10s}$ = 4.44; $DLCO_{10s}$ = 3.93), indicating moderate but not severe multicollinearity. Coefficient SEs are inflated (~2×), but prediction remained stable; hierarchical partitioning showed $DLCO_{10s}$ adds both unique ($\Delta R^2 \approx 0.103$) and shared information.

Table S17. Model Comparison (B – A)

Metric	Model B (4-predictors) (FEV$_1$ z-scores (GLI) + TLC z-scores (GLI) + DLNO$_{10s}$ z-scores [GAMLSS] + DLCO$_{10s}$ z-scores [GLI]) [95% CI]	Model A (3-predictors) (FEV$_1$ z-scores (GLI) + TLC z-scores (GLI) + DLCO$_{10s}$ z-scores [GLI]) [95% CI]	Δ (B − A) [95% CI]	Different?
MCC	0.770 [0.717, 0.893]	0.800 [0.698, 0.892]	-0.031 [-0.073, 0.086]	No
Kappa	0.762 [0.699, 0.893]	0.799 [0.677, 0.892]	-0.037 [-0.088, 0.099]	No
Discordance	0.086 [0.034, 0.110]	0.069 [0.034, 0.123]	0.017 [-0.042, 0.039]	No
F1 score	0.817 [0.768, 0.915]	0.843 [0.752, 0.915]	–0.026 [–0.062, 0.072]	No
FOR	0.023 [0.003, 0.043]	0.032 [0.007, 0.049]	–0.009 [–0.026, 0.006]	No
Accuracy	0.914 [0.890, 0.966]	0.931 [0.877, 0.966]	–0.017 [–0.039, 0.042]	No
Balanced accuracy	0.915 [0.892, 0.957]	0.913 [0.883, 0.951]	0.002 [–0.006, 0.032]	No
Sensitivity	0.918 [0.841, 0.987]	0.882 [0.819, 0.973]	0.035 [–0.027, 0.102]	No
Specificity	0.913 [0.875, 0.981]	0.944 [0.861, 0.984]	–0.031 [–0.073, 0.057]	No
PPV	0.736 [0.651, 0.924]	0.806 [0.631, 0.932]	–0.071 [–0.168, 0.121]	No
NPV	0.977 [0.957, 0.997]	0.968 [0.951, 0.993]	0.009 [–0.006, 0.026]	No
FPR	0.087 [0.019, 0.125]	0.056 [0.016, 0.139]	0.031 [–0.057, 0.073]	No
FNR	0.082 [0.013, 0.159]	0.118 [0.027, 0.181]	–0.035 [–0.102, 0.027]	No
FDR	0.264 [0.076, 0.349]	0.194 [0.068, 0.369]	0.071 [–0.121, 0.168]	No
+LR	10.6 [7.6, 47.0]	15.8 [6.8, 52.7]	–5.3 [–25.3, 15.9]	No
-LR	0.1 [0.0, 0.2]	0.1 [0.0, 0.2]	–0.0 [–0.1, 0.0]	No
DOR	117.4 [77.1, 767.8]	127.1 [67.8, 571.3]	–9.7 [–142.2, 404.4]	No

Both models were fit as simple logistic regressions because the "Study" random intercept was not retained (Model A: AICGLM = 148.6 vs AIC$_{GLMM}$=150.6; Model B: 149.5 vs 151.54; in both cases $\tau^2 \approx 0$ and adding the random intercept worsened AIC by ~2). Operating thresholds were chosen by optimal Youden's J on model-specific predictions (Model A: J-optimal p = 0.276; Model B: p = 0.123) and re-optimized within each bootstrap draw. CIs come from 10,000 subject-level bootstrap resamples; a difference is called "Different?" only when the 95% CI for Δ(B−A) excludes 0. No metric showed a statistically significant difference between models (all Δ(B−A) 95% CIs overlapped 0). Point estimates suggest that adding DLNO$_{10s}$ (Model B) trades a small increase in sensitivity (Δ ≈ +0.047) for a decrease in specificity (Δ ≈ −0.043) and PPV (Δ ≈ −0.105), leaving accuracy (~0.94 vs ~0.914) and balanced accuracy (~0.915 vs ~0.914) essentially unchanged. MCC (Δ ≈ −0.047), κ (Δ ≈ −0.055), likelihood ratios (Δ+LR ≈−9.5; Δ−LR ≈− 0.05), and DOR (Δ ≈−31.1) were numerically lower with the 4-predictor model, but with wide CIs that include no effect. Overall, the 4-predictor model does not provide a statistically demonstrable improvement over the 3-predictor model at the Youden-optimized thresholds.

Table S18. Model comparison (C – A)

Metric	Model C (3-predictors) (FEV$_1$ z-scores (GLI) + TLC z-scores (GLI) + DLNO$_{10s}$ z-scores [GAMLSS]) [95% CI]	Model A (3-predictors) (FEV$_1$ z-scores (GLI) + TLC z-scores (GLI) + DLCO$_{10s}$ z-scores [GLI]) [95% CI]	Δ (C – A) [95% CI]	Different?
MCC	0.817 [0.707, 0.892]	0.800 [0.698, 0.892]	0.017 [–0.080, 0.079]	No
Kappa	0.817 [0.688, 0.891]	0.799 [0.677, 0.892]	0.018 [–0.095, 0.092]	No
Discordance	0.061 [0.034, 0.118]	0.069 [0.034, 0.123]	–0.007 [–0.039, 0.042]	No
F1 score	0.855 [0.759, 0.915]	0.843 [0.752, 0.915]	0.013 [–0.068, 0.066]	No
FOR	0.034 [0.007, 0.046]	0.032 [0.007, 0.049]	0.003 [–0.024, 0.014]	No
Accuracy	0.939 [0.882, 0.966]	0.931 [0.877, 0.966]	0.007 [–0.042, 0.039]	No
Balanced accuracy	0.914 [0.887, 0.953]	0.913 [0.883, 0.951]	0.000 [–0.016, 0.027]	No
Sensitivity	0.871 [0.831, 0.976]	0.882 [0.819, 0.973]	-0.012 [–0.055, 0.095]	No
Specificity	0.957 [0.868, 0.982]	0.944 [0.861, 0.984]	0.012 [–0.074, 0.059]	No
PPV	0.841 [0.642, 0.930]	0.806 [0.631, 0.932]	0.034 [–0.168, 0.130]	No
NPV	0.966 [0.954, 0.993]	0.968 [0.951, 0.993]	–0.003 [–0.014, 0.024]	No
FPR	0.043 [0.018, 0.132]	0.056 [0.016, 0.139]	–0.012 [–0.059, 0.074]	No
FNR	0.129 [0.024, 0.169]	0.118 [0.027, 0.181]	0.012 [–0.095, 0.055]	No
FDR	0.159 [0.070, 0.358]	0.194 [0.068, 0.369]	–0.034 [–0.130, 0.168]	No
+LR	20.1 [7.2, 49.6]	15.8 [6.8, 52.7]	4.3 [–25.2, 21.5]	No
-LR	0.14 [0.0, 0.2]	0.13 [0.0, 0.2]	0.01 [–0.1, 0.1]	No
DOR	148.5 [71.2, 612.2]	127.1 [67.8, 571.3]	21.4 [–184.1, 265.9]	No

Both models were analysed with ordinary logistic regression because adding a "Study" random intercept did not improve fit (Model C: AIC_{GLM} =148.6 vs AIC_{GLMM} =150.6, $\tau^2 \approx 0$; Model A: AIC_{GLM}=154.6 vs AIC_{GLMM} =154.9, $\tau^2 = 0.213$). Operating thresholds were chosen by maximizing Youden's J on each model's predicted probabilities (Model C: $p = 0.276$; Model A: $p = 0.258$) and re-optimized within every bootstrap replicate. Confidence intervals are percentile 95% CIs from 50,000 paired subject-level bootstrap resamples; a difference is flagged "Different?" only when the 95% CI for Δ(C−A) excludes 0. No metric met this criterion. Point estimates suggest Model C yields slightly higher specificity (Δ ≈ +0.012) and PPV (Δ ≈ +0.034) with a small reduction in sensitivity (Δ ≈ −0.012); overall accuracy (Δ≈+0.007) and balanced accuracy (Δ ≈ 0.000) are essentially unchanged. Differences in MCC (Δ ≈ +0.017), κ (Δ ≈ +0.018), likelihood ratios (Δ +LR ≈ +4.25; Δ −LR ≈ +0.011), and DOR (Δ ≈ +21.4) are modest with wide CIs that include no effect. Thus, at Youden-optimized thresholds, Model C does not provide a statistically significant improvement over Model A**.**

Table S19. Reclassification summary at Youden-optimized thresholds (B – A; C – A)

	Metric	B – A Estimate	B – A SE	C – A Estimate	C – A SE
1	Net Reclassification Improvement (NRI) (overall net improvement in reclassification)	0.004 [–0.013, 0.064]	0.020	0.001 [–0.032, 0.054]	0.021
2	Net proportion of true positives (NRI+) reclassified to higher risk	0.035 [–0.027, 0.101]	0.035	–0.012 [–0.054, 0.096]	0.038
3	Net proportion of true-negatives (NRI–) reclassified to lower risk	–0.031 [–0.072, 0.058]	0.032	0.012 [–0.075, 0.059]	0.033
4	The fraction of true positives that gets “upped” in risk [Pr (Up | Case)]	0.035 [0.000, 0.102]	0.032	0.000 [0.000, 0.098]	0.030
5	The undesirable fraction for cases (lowered risk when they have COPD) [Pr(Down | Case)	0.000 [0.000, 0.030]	0.009	0.012 [0.000, 0.056]	0.017
6	The fraction of true-negatives (controls) that gets “down-rated” [Pr (Down | Control)]	0.003 [0.000, 0.065]	0.018	0.019 [0.000, 0.069]	0.020
7	The undesirable fraction of true-negatives (controls) that gets “up-rated” [Pr (Up | Control)]	0.034 [0.000, 0.072]	0.020	0.006 [0.000, 0.077]	0.022
8	Integrated Discrimination Index (IDI). This is the average predicted-risk gap between cases and controls	0.014* [0.001, 0.047]	0.012	0.013 [–0.010, 0.045]	0.014

SE = standard error; CI = confidence interval.

Model A (3-predictors): FEV_1 z-scores (GLI)+ TLC z-scores (GLI) + $DLCO_{10s}$ z-scores [GLI]

Model B (4-predictors): FEV_1 z-scores (GLI) + TLC z-scores (GLI)+ $DLNO_{10s}$ z-scores [GAMLSS] + $DLCO_{10s}$ z-scores [GLI]

Model C (3-predictors): FEV_1 z-scores (GLI) + TLC z-scores (GLI) + $DLNO_{10s}$ z-scores [GAMLSS]

Operating decision thresholds were the Youden-J optima from each model’s ROC curve (Model A: 0.230; Model B: 0.254; Model C: 0.320), reflecting a single, pre-specified decision rule per model. Calibration intercept and slope are computed on model-predicted probabilities and do not depend on the decision threshold

Reclassification metrics use the “up”/“down” convention: “Up” = higher predicted risk under the right-hand model (B or C) vs Model A; “Down” = lower predicted risk. NRI+ = Pr(Up | Case) − Pr(Down | Case); NRI− = Pr(Down | Control) − Pr(Up | Control); NRI = NRI+ + NRI−. IDI is the change in the average predicted-risk gap between cases and controls. Estimates and 95% CIs come from 50,000 paired bootstrap resamples. Statistical significance is inferred when the 95% CI excludes 0 (asterisk)*. In these data, only the IDI for B − A showed a small but statistically significant improvement (0.014 [0.001, 0.047]); all other reclassification components for B − A and C − A had CIs spanning 0.

Table S20. Reclassification Summary at category-free reclassification (B – A; C – A)

Metric	B – A	B – A	C – A	C – A
	Estimate	SE	Estimate	SE
1 Net Reclassification Improvement (NRI) (overall net improvement in reclassification)	0.341 [–0.123, 0.962]	0.275	0.290 [–0.467, 0.882]	0.337
2 Net proportion of true positives (NRI+) reclassified to higher risk	0.176 [–0.100, 0.562]	0.169	0.200 [–0.265, 0.525]	0.200
3 Net proportion of true-negatives (NRI–) reclassified to lower risk	0.164 [–0.087, 0.457]	0.141	0.090 [–0.258, 0.413]	0.167
4 The fraction of true positives that gets “upped” in risk [Pr (Up | Case)]	0.588* [0.450, 0.781]	0.084	0.600* [0.368, 0.763]	0.100
5 The undesirable fraction for cases (lowered risk when they have COPD) [Pr(Down | Case)	0.412* [0.219, 0.550]	0.084	0.400* [0.237, 0.632]	0.100
6 The fraction of true-negatives (controls) that gets “down-rated” [Pr (Down | Control)]	0.582* [0.456, 0.729]	0.070	0.545* [0.371, 0.706]	0.083
7 The undesirable fraction of true-negatives (controls) that gets “up-rated” [Pr (Up | Control)]	0.418* [0.271, 0.544]	0.070	0.455* [0.294, 0.629]	0.083
8 Integrated Discrimination Index (IDI). This is the average predicted-risk gap between cases and controls	0.014* [0.001, 0.046]	0.012	0.013 [–0.010, 0.046]	0.014

SE = standard error; CI = confidence interval.

Model A (3-predictors): FEV_1 z-scores (GLI)+ TLC z-scores (GLI) + $DLCO_{10s}$ z-scores [GLI]. The calibration intercept was 0.00[–0.44, 0.44], and calibration slope was 1.00 [0.80, 1.21].

Model B (4-predictors): FEV_1 z-scores (GLI) + TLC z-scores (GLI)+ $DLNO_{10s}$ z-scores [GAMLSS] + $DLCO_{10s}$ z-scores [GLI]. The calibration intercept was 0.00[–0.45, 0.45], and calibration slope was 1.00 [0.79, 1.21].

Model C (3-predictors): FEV_1 z-scores (GLI) + TLC z-scores (GLI) + $DLNO_{10s}$ z-scores [GAMLSS]. The calibration intercept was 0.00[–0.45, 0.45], and calibration slope was 1.00 [0.79, 1.21].

SE = standard error. $DLCO_{10s}$ z-scores (GLI) = fitted DLCO z-scores using GLI references equations for Whites; FEV_1 z-scores (GLI) = fitted FEV_1 z-scores using GLI references equations for Whites; TLC z-scores (GLI) = fitted TLC z-scores using GLI references equations for Whites; $DLNO_{10s}$ z-scores (GAMLSS) = fitted DLNO z-scores using the GAMLSS reference equations for Whites (Zavorsky & Cao, 2022). Model A is the 3-predictor model of $DLCO_{10s}$ z-scores (GLI), FEV_1 z-scores (GLI) and TLC z-scores (GLI). Youden’s J–optimal threshold was derived from the ROC curve. This simulates a real-world scenario in which one must commit to a single decision rule across all models. Youden’s J threshold for the Model A was 0.230, for the 4-predictor model was 0.254, and for the 3-predictor model with $DLNO_{10s}$ z-scores [$DLNO_{10s}$ z-scores (GAMLSS), FEV_1 z-scores (GLI), TLC z-scores (GLI), Youden’s J Threshold was 0.32. NRI+ = Pr(Up | Case) - Pr(Down | Case). NRI– = Pr(Down | Ctrl) - Pr(Up | Ctrl) for numbers 1 to 11 above, Statistical significance is inferred when the 95% confidence interval **does not** cross zero*. 50,000 bootstrap samples were used to generate the 95% CI.

Figure S1. The age and sex breakdown between those with emphysema and those without emphysema in the pooled dataset, after filtering (n=323 controls; n = 85 with emphysema)

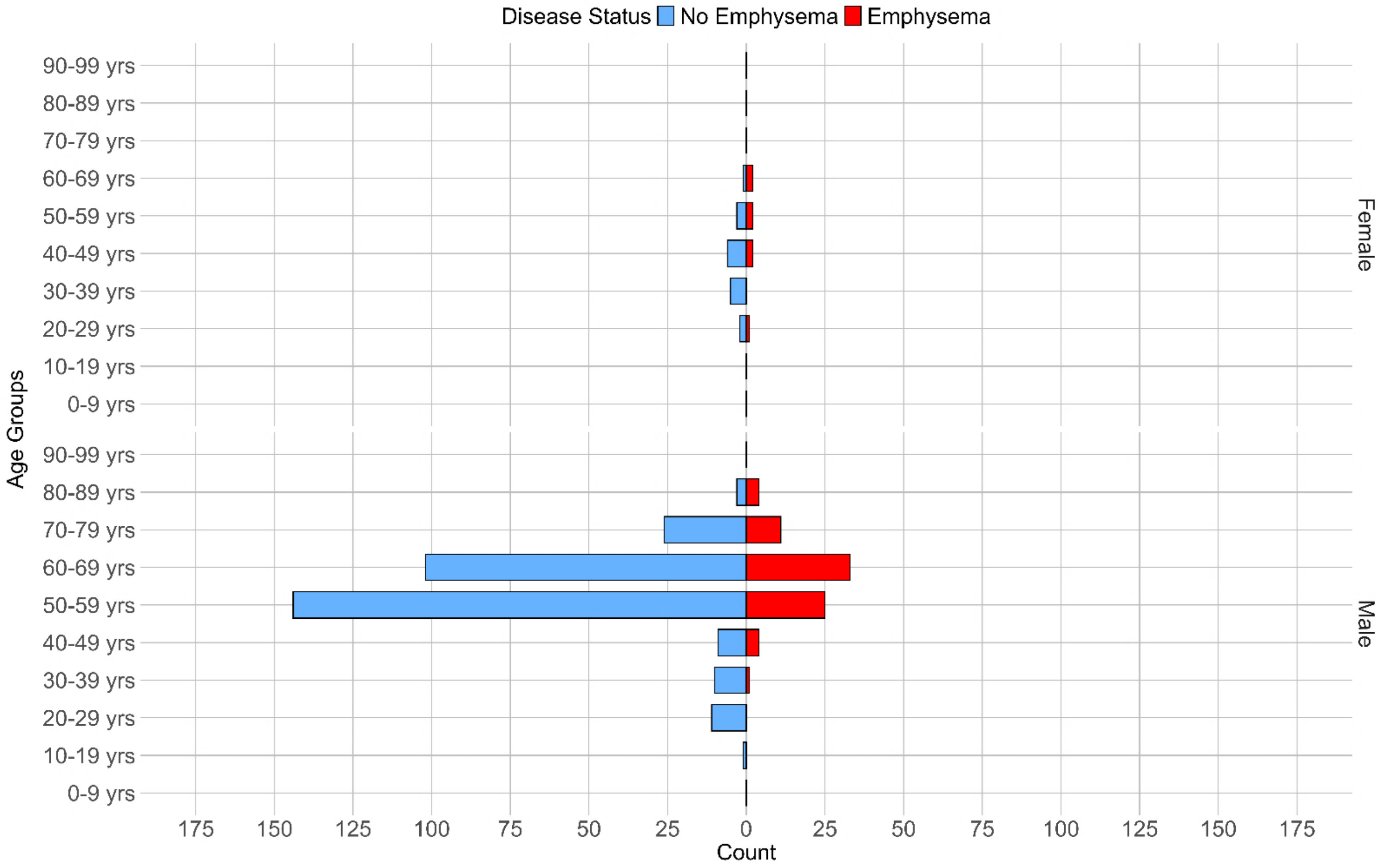


Figure S2. The association between $DLNO_{10s}$ z-scores and $DLCO_{10s}$ z-scores in the filtered pooled dataset

Reference equations of Zavorsky & Cao (2022) [13] were used to generate z-scores. The reference equations of Zavorsky & Cao (2022)[13] were used as they account for the pulmonary function device used to measure $DLNO_{10s}$ and $DLCO_{10s}$. Breath-hold time was 9.6 ± 0.6 s. solid black line is the best fit line. The dashed black lines are the 95% prediction CI. The purple dashed lines are the location of the lower limit of normal (i.e. z-score = –1.645). red circles = smokers with emphysema. Green circles represent smokers without emphysema.

$DLNO_{10s}$ z-scores = 0.871·($DLCO_{10}$ z-scores) – 0.126, $R^2 = 0.71$, standard error of the estimate = 0.692, $p < 0.0001$. The 95% CI for the slope = 0.816 to 0.925, n = 422 (323 smokers without emphysema and 85 smokers with emphysema). Segmented, "Piecewise" equations were used.

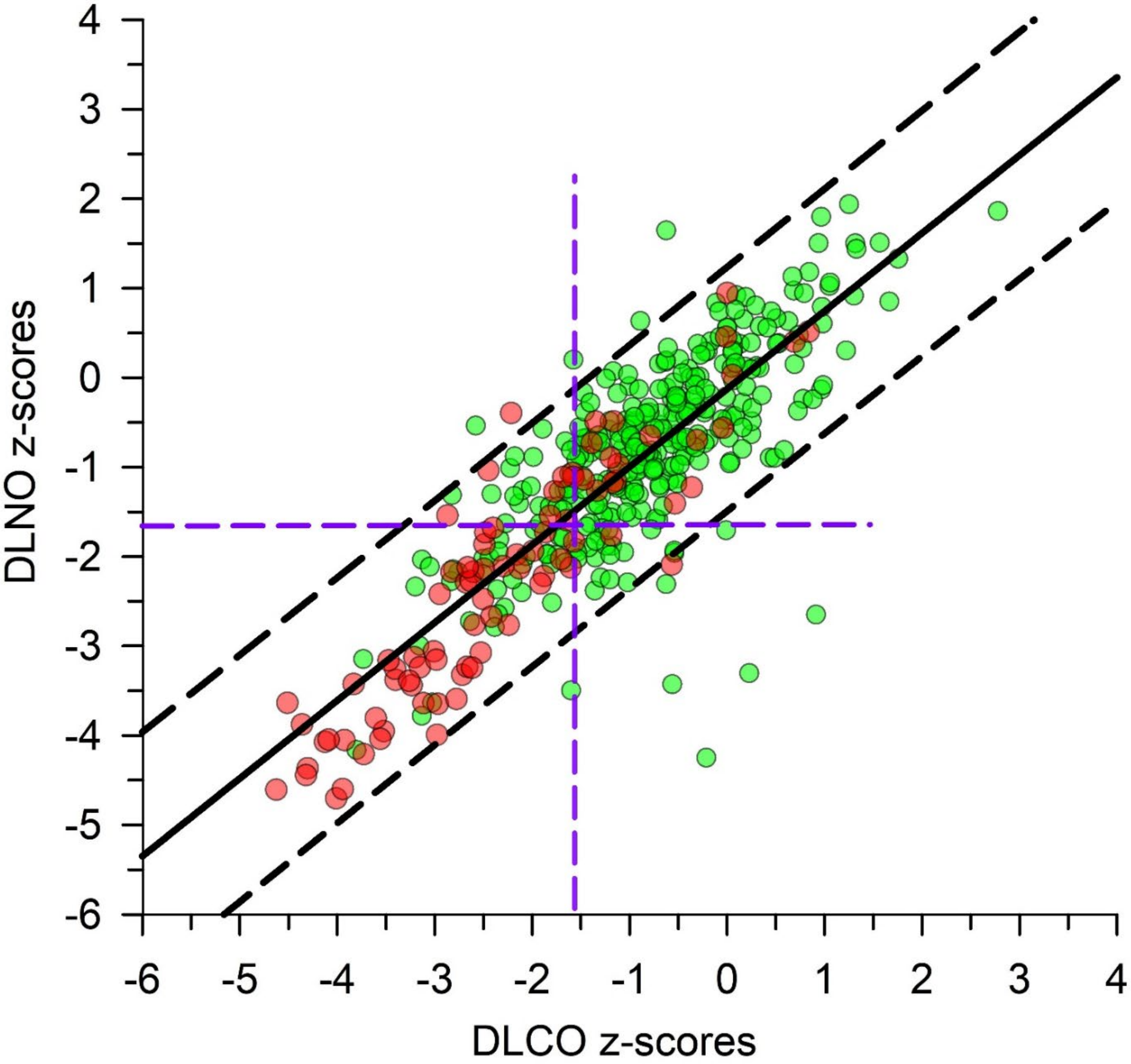


Figure S3. Bayesian Information Criterion (BIC) and Leave-One-Out Information Criterion (LOOIC) of models for emphysema prediction and generalizability.

Models are ranked by their difference in BIC (black circles) and LOOIC (white circles) relative to the top-performing model (top = best). **This figure is a continuation of Figure 1** in the manuscript in which the top 17 models are presented. Here, the first ranked model (Model C, the top ranked model) is there for perspective, and then 18th through 34th ranked models are presented below. BIC penalizes model complexity; LOOIC evaluates predictive performance via cross-validation.

- **Red zone (BIC or LOOIC difference ≤ 2.2):** Models nearly as good as the best model.
- **Yellow zone (BIC or LOOIC difference 2.3–5.9):** Models with substantial but acceptable performance differences compared to the best model.
- **Green zone (BIC or LOOIC difference 6.0–9.2):** Models with considerably weaker performance compared to the best model.
- **Purple zone (BIC or LOOIC difference ≥ 9.3):** Models with significantly poorer fit compared to the best model.

The x-axis shows the difference from the best model—smaller is better. The best-performing model is the three-predictor z-score model of TLC + FEV_1 + $DLNO_{10s}$ (GAMLSS) derived from the GLI equations [9,10] and DLNO z-scores from the GAMLSS equations [13] (n=323 smokers with without emphysema; n= 85 smokers with emphysema).

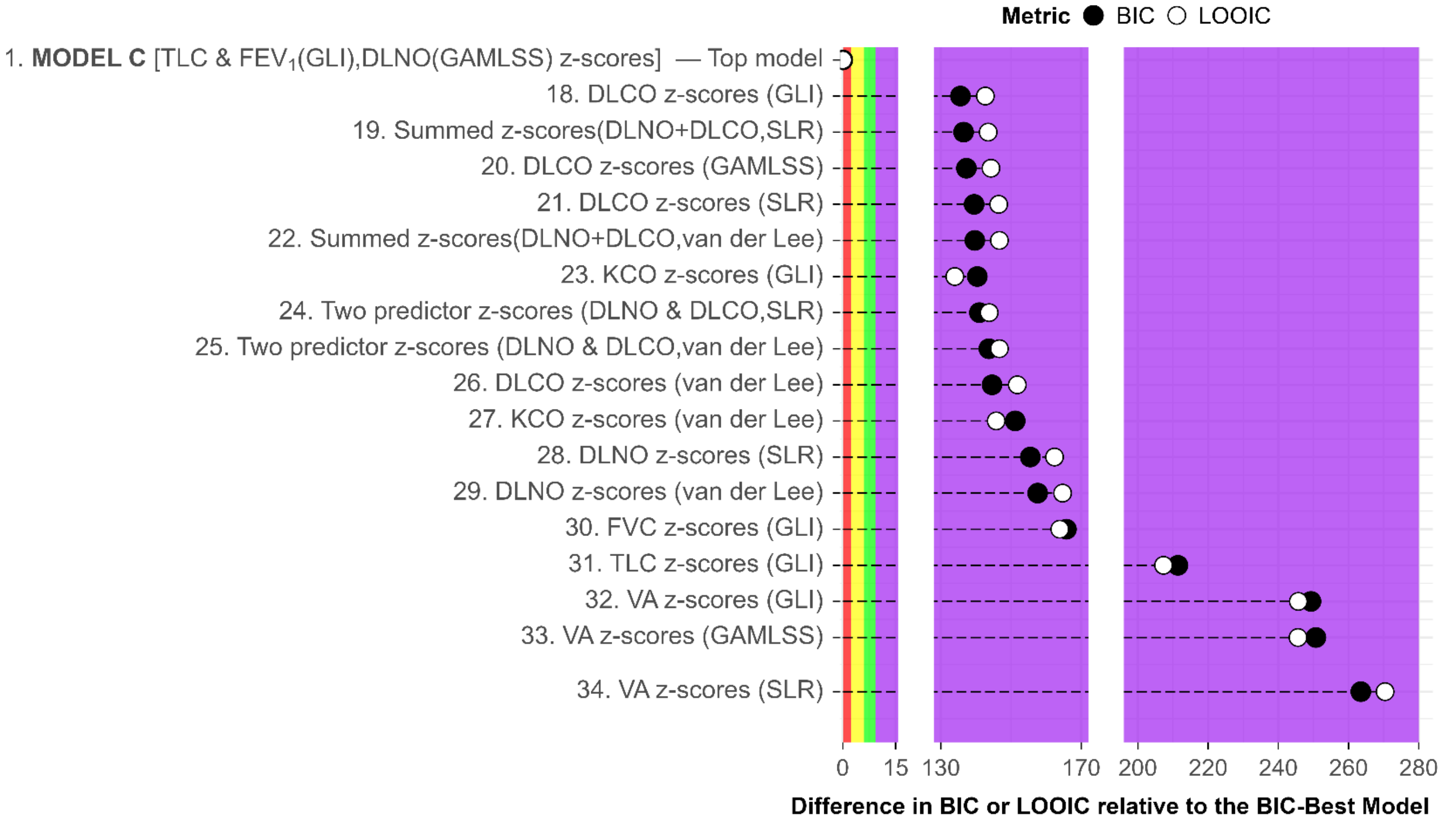


Figure S4. ROC and PR Curves with cross-validation performance for the three-predictor z-score model of TLC z-scores + FEV_1 z-scores + $DLNO_{10s}$ (GAMLSS)

(A) Receiver Operating Characteristic (ROC) curve showing the trade-off between sensitivity (true positive rate) and 1-specificity (false positive rate) across 5×5 repeated cross-validation. (B) Precision-Recall (PR) curve illustrating the relationship between precision (positive predictive value) and sensitivity. Shaded areas reflect variability across folds based on 100,000 bootstrapped samples; the red dot marks the optimal operating point on each curve.

The ROC curve shows strong performance with high sensitivity and low false positive rates (AUROC = 0.96; 95% CI: 0.95–0.97). The PR curve starts with precision near 1.0 at low recall but declines to ~0.4 at full recall, with wider variability at higher recall values. Youden's J = 0.82; threshold = 0.30. Other metrics include PR = 0.91 (95% CI: 0.88–0.93), sensitivity = 0.87 (95% CI: 0.83–0.90), specificity = 0.95 (95% CI: 0.95–0.96), and MCC = 0.81 (95% CI: 0.78–0.84). Results are based on 323 smokers without and 85 with emphysema. Logistic regression outputs for this model are reported in Table 2.

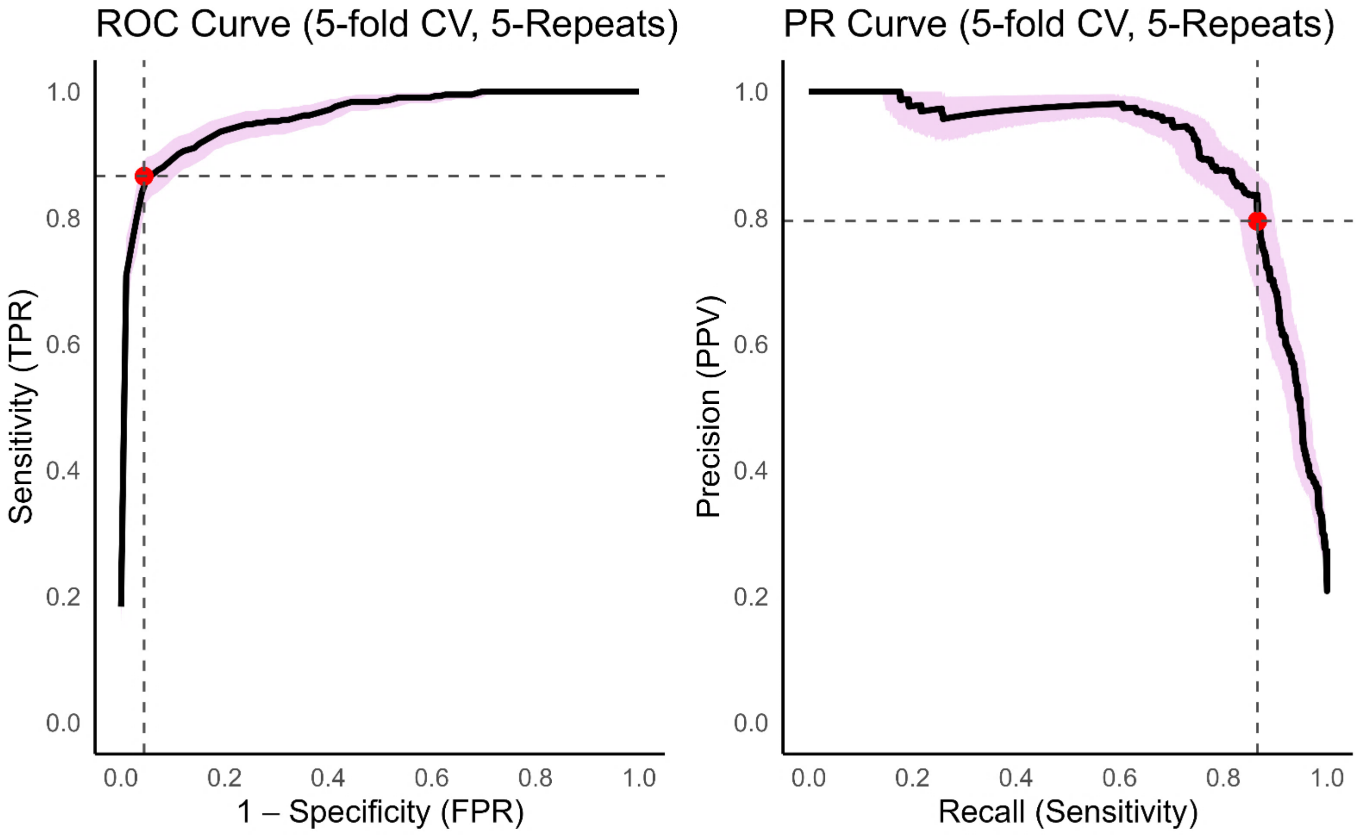
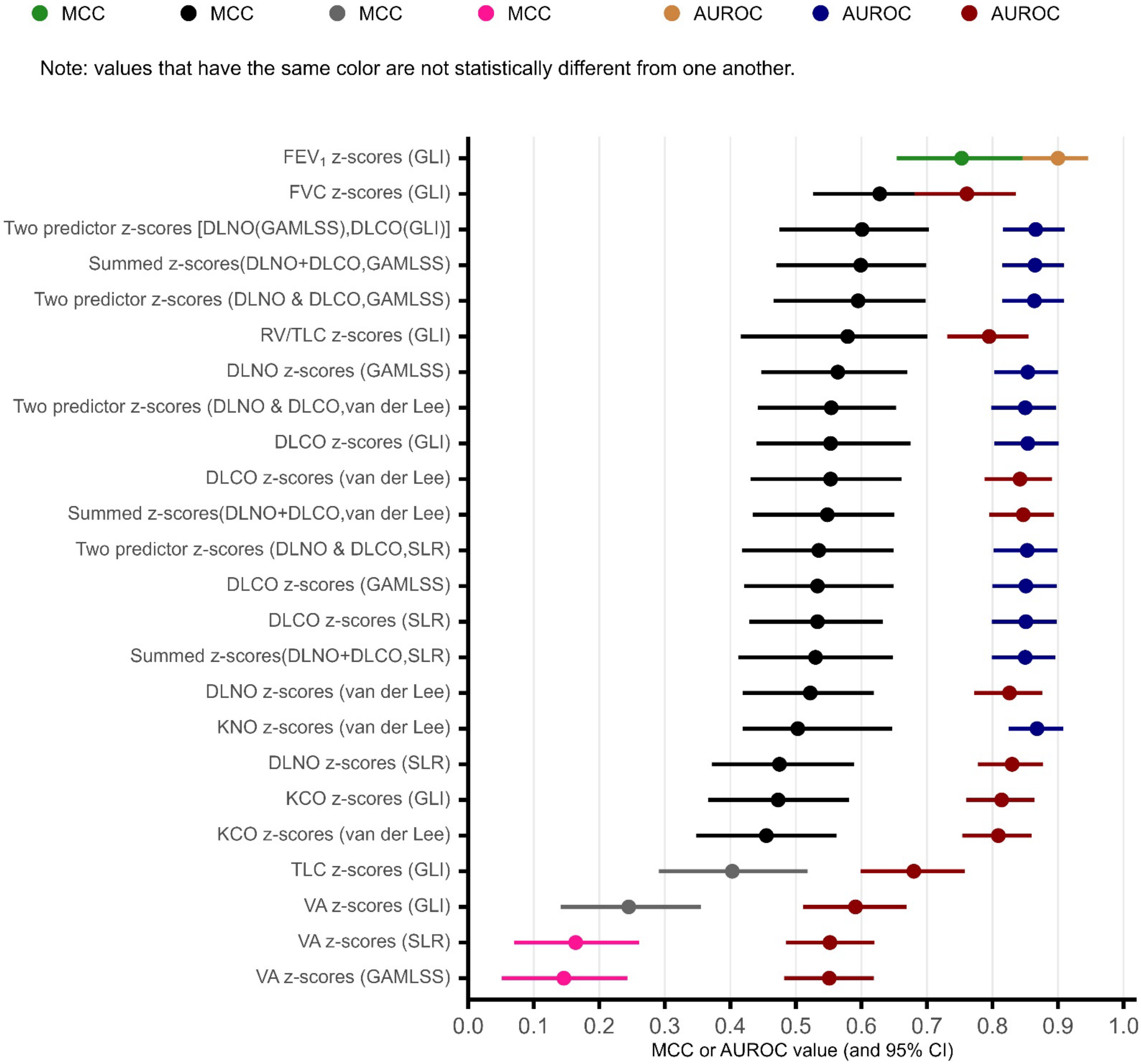


Figure S5. Discriminatory classification performance for the bottom 24 predictive models.

The Matthews correlation coefficient (MCC) and the area under the receiver operating characteristic curve (AUROC) are shown for models 11 through 34 (the bottom 24 models). Models sharing the same color for their point estimates are not statistically different from one another, based on 10 000 bootstrapped samples (2-sided) and after correction for multiple comparisons at a false-discovery-rate of 5% using the Benjamini-Hochberg procedure.



Note: values that have the same color are not statistically different from one another.

Figure S6. Scree plot of variance explained by principal components in COPD analysis

This scree plot displays variance explained per principal component (orange line/points) and cumulative variance (blue line). PC1–PC4 explain 51.5%, 16.6%, 11.6%, and 9.5% of the variance, respectively, for a cumulative total of 89.2%. The sharp drop after PC1 suggests an "elbow" point, with most variance captured in the first three components.

Although including PC4 would raise total explained variance above the common 85% threshold, logistic regression showed no meaningful association between PC4 and COPD status. The coefficient (β = 0.09) represents the change in the log odds of having COPD for each one-unit increase in PC4, controlling for other components. This effect was small and non-significant (SE = 0.17; z = 0.54; p = 0.587), with an odds ratio of 1.09 (95% CI ≈ 0.78–1.53), indicating no clear relationship. For parsimony and interpretability, only PC1–PC3 were retained. See Tables S11–S14 for full results.

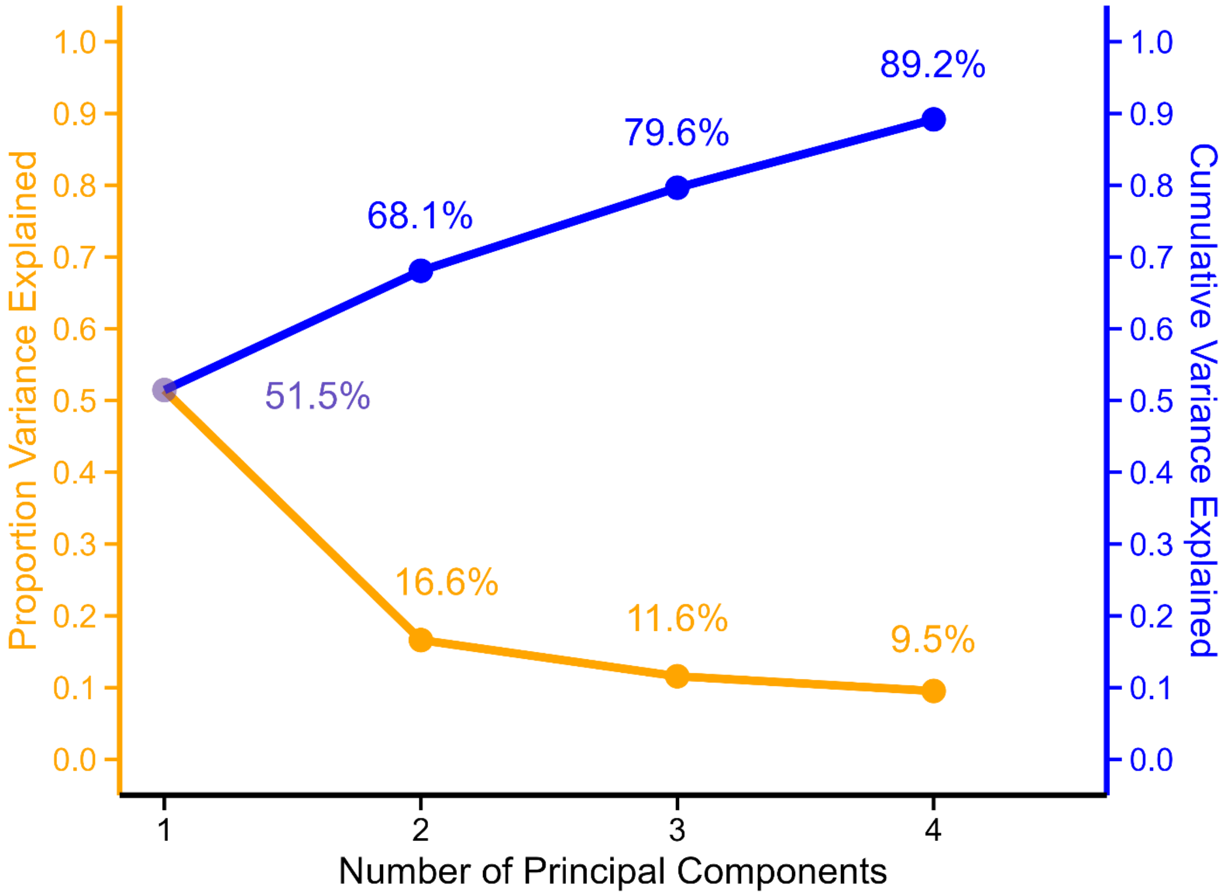
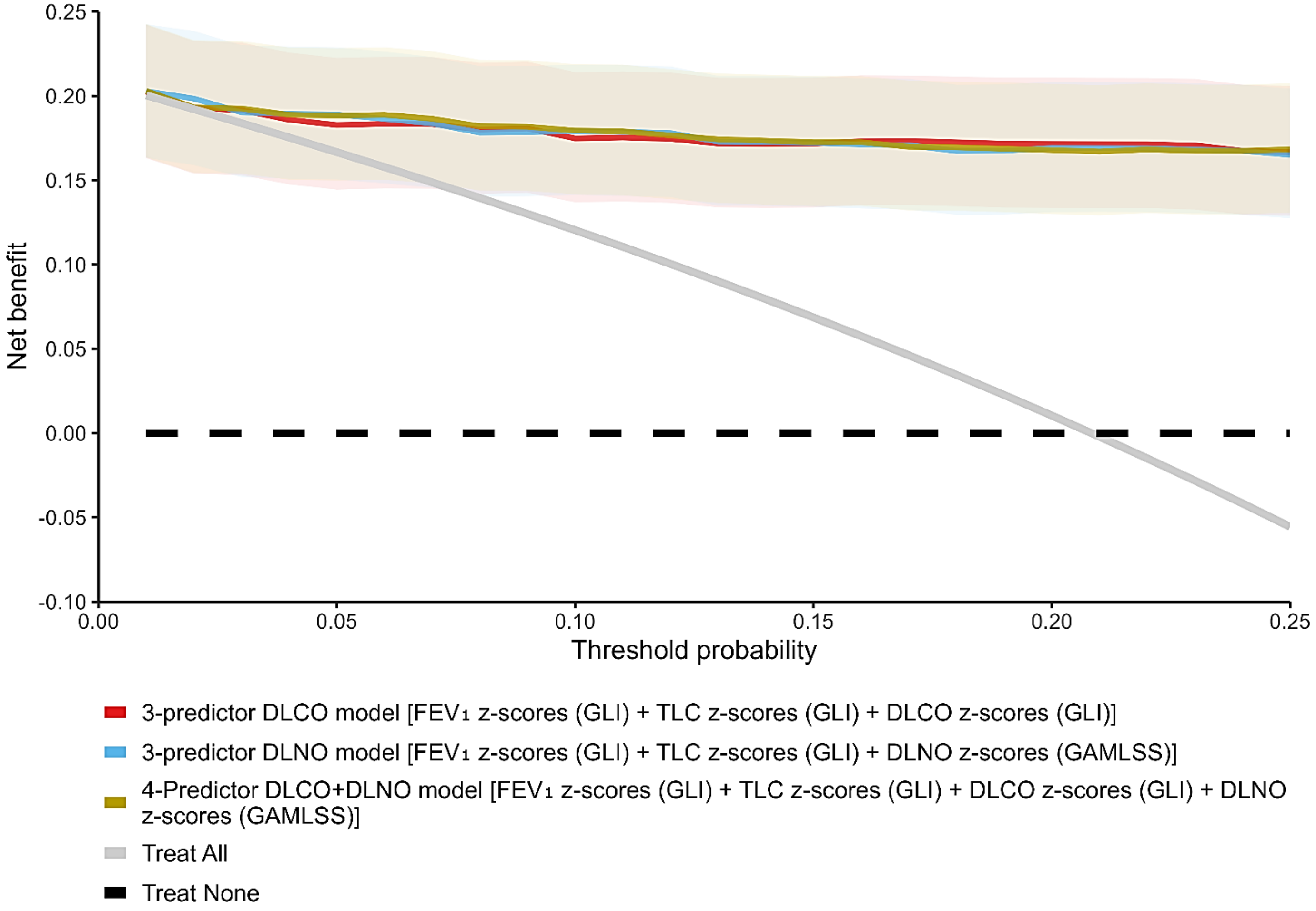


Figure S7. Decision curve analysis for COPD classification models

Net benefit is plotted against the threshold probability (0–0.25) using out-of-fold predictions from repeated cross-validation. Curves compare three models with two references: *Treat All* (light grey) and *Treat None* (dashed black). Shaded bands represent 95% bootstrap CIs. Across the clinically relevant range (probability threshold = 0–0.25), all three models outperform *Treat None*, and—except at very low thresholds—also outperform *Treat All*. Once the probability threshold exceeds the disease prevalence (~0.21), the net benefit of *Treat All* drops to zero, while the models remain positive. The 4-predictor model performs similarly to both 3-predictor models; their curves overlap with minimal differences across thresholds. Net benefit remains around 0.16–0.20—translating to 16–20 more true positives per 100 patients compared to treating none. In practice, either 3-predictor model (DLCO_{10s} or DLNO_{10s}) offers nearly the same clinical utility as using both, with little added value from including both measures.



References

1. Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. Cochrane Working Group. *Stat Med*. Oct 15 1995;14(19):2057–79. doi:10.1002/sim.4780141902
2. Moinard J, Guénard H. Determination of lung capillary blood volume and membrane diffusing capacity in patients with COLD using the NO-CO method. *Eur Respir J*. Mar 1990;3(3):318–22.
3. van der Lee I, Gietema HA, Zanen P, et al. Nitric oxide diffusing capacity versus spirometry in the early diagnosis of emphysema in smokers. *Respir Med*. Dec 2009;103(12):1892–7. doi:10.1016/j.rmed.2009.06.005
4. Diener L, Herold R, Harth V, Preisser A. DLNO/DLCO in different patient collectives; influencing factors have different effects on DLNO and DLCO. From the 61st Congress of the German Society for Pneumology and Respiratory Medicine. [Abstract]. *Pneumologie*. Apr 2021;75:S3–S3. doi:10.1055/s-0041-1723272
5. Dal Negro RW, Turco P, Povero M. Single-Breath Simultaneous Measurement of DL(NO) and DL(CO) as Predictor of the Emphysema Component in COPD - A Retrospective Observational Study. *Int J Chron Obstruct Pulmon Dis*. 2024;19:2123–2133. doi:10.2147/COPD.S467138
6. Zavorsky GS, Hsia CC, Hughes JM, et al. Standardisation and application of the single-breath determination of nitric oxide uptake in the lung. *Eur Respir J*. Feb 2017;49(2):1600962. doi:10.1183/13993003.00962-2016
7. Schulz U, Langwieler S, Riedel S, Schreiber J. [Pulmonary capillary blood volume and membrane components of pulmonary diffusion capacity in patients with chronic obstructive bronchitis (COPD)]. *Pneumologie*. Apr 2014;68(4):266–9. Pulmonales kapillares Blutvolumen und Membrankomponente der pulmonalen Diffusionskapazitat bei Patienten mit chronisch obstruktiver Bronchitis (COPD). doi:10.1055/s-0034-1365056
8. Zavorsky GS, Dal-Negro R, van der Lee I, Preisser A. Data from: "Nitric oxide uptake in the lungs in smokers with emphysema ", Mendeley Data, V2, doi: 10.17632/g4jp8wd6f6.2. 2025. Deposited Febrary 26, 2025.
9. Quanjer PH, Stanojevic S, Cole TJ, et al. Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *Eur Respir J*. Dec 2012;40(6):1324–43. doi:10.1183/09031936.00080312

10. Hall GL, Filipow N, Ruppel G, et al. Official ERS technical standard: Global Lung Function Initiative reference values for static lung volumes in individuals of European ancestry. *Eur Respir J*. Mar 2021;57(3)doi:10.1183/13993003.00289-2020

11. Stanojevic S, Graham BL, Cooper BG, et al. Official ERS technical standards: Global Lung Function Initiative reference values for the carbon monoxide transfer factor for Caucasians. *Eur Respir J*. Sep 2017;50(3)doi:10.1183/13993003.00010-2017

12. Stanojevic S, Graham BL, Cooper BG, et al. [Corrigendum] Official ERS technical standards: Global Lung Function Initiative reference values for the carbon monoxide transfer factor for Caucasians. *Eur Respir J*. 2020;56:1700010.

13. Zavorsky GS, Cao J. Reference equations for pulmonary diffusing capacity using segmented regression show similar predictive accuracy as GAMLSS models. *BMJ Open Respir Res*. Feb 2022;9(1)doi:10.1136/bmjresp-2021-001087

14. van der Lee I, Zanen P, Stigter N, van den Bosch JM, Lammers JW. Diffusing capacity for nitric oxide: reference values and dependence on alveolar volume. *Respir Med*. Jul 2007;101(7):1579–84. doi:10.1016/j.rmed.2006.12.001

15. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. Feb 28 1997;16(4):385–95. doi:10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

16. Tibshirani R. Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met*. 1996;58(1):267–288. doi:DOI 10.1111/j.2517-6161.1996.tb02080.x

17. Bowerman C, Bhakta NR, Brazzale D, et al. A Race-neutral Approach to the Interpretation of Lung Function Measurements. *Am J Respir Crit Care Med*. Mar 15 2023;207(6):768–774. doi:10.1164/rccm.202205-0963OC

18. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput*. Sep 2017;27(5):1413–1432. doi:10.1007/s11222-016-9696-4

19. Bürkner P-C. brms: An R package for Bayesian multilevel models using Stan. *J Stat Softw*. 08/29 2017;80(1):1 – 28. doi:10.18637/jss.v080.i01

20. Hoffman MD, Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J Mach Learn Res*. 2014;47(47):1593–1623.

21. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner P-C. Rank-Normalization, folding, and localization: An improved for R ̂ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*. 2021;16(2):667–718.

22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. Sep 1988;44(3):837–45.

23. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. Mar 17 2011;12:77. doi:10.1186/1471-2105-12-77

24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc B*. 1995;57(1):289–300.

25. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. Jan 2 2020;21(1):6. doi:10.1186/s12864-019-6413-7

26. Chicco D, Totsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min*. Feb 4 2021;14(1):13. doi:10.1186/s13040-021-00244-z

27. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One*. 2017;12(6):e0177678. doi:10.1371/journal.pone.0177678

28. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276–82.

29. Raftery AE. Bayesian model selection in social research. *Sociological Methodology*. 1995;25:111–163. doi:Doi 10.2307/271063

30. Schwarz G. Estimating the Dimension of a Model. *The Annals of Statistics*. 1978;6(2):461–464.

31. Burnham KP, Anderson DR. Multimodel inference - understanding AIC and BIC in model selection. *Sociol Method Res*. Nov 2004;33(2):261–304. doi:10.1177/0049124104268644