Journal of Statistics on Health Decision

EXTENDED ABSTRACT

Modelling the impact of the disease on people with COPD – a comparison of feature selection methods

Jorge Cabral[1], Pedro Macedo[1], Alda Marques[2,3], Vera Afreixo[1]

[1]Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Aveiro, Portugal
[2]Respiratory Research and Rehabilitation Laboratory (Lab3R), School of Health Sciences (ESSUA), University of Aveiro, Aveiro, Portugal
[3]Institute of Biomedicine (iBIMED), School of Health Sciences, University of Aveiro, Aveiro, Portugal

Introduction:

Linear models (LMs) aim to predict outcomes given p features [1]. The following measure can be used to quantify the fit: (i) mean squared error (MSE); (ii) coefficient of determination (R^2); (iii) adjusted R^2; (iv) Akaike's information criterion (AIC) [2]; (v) Bayesian information criterion (BIC) [3].

Criteria to choose the most appropriate methods to select features in datasets are unclear [4–6]. One approach is the automatic stepwise selection which removes one feature at a time. Another is the

Least Absolute Shrinkage and Selection Operator (LASSO) which adds a penalty term λ that reduces the magnitude of coefficients [7].

Information theory provides criteria for setting up probability distributions on the basis of partial knowledge [8]. Normalized entropy [9] measures the information content of a particular model or feature. It was defined based on the consistent and asymptotically normal generalized maximum entropy estimator [10]. Features with normalized entropy approximately equal to one should be excluded from the model.

Chronic obstructive pulmonary disease (COPD) is a progressive, treatable and preventable respiratory disease [11]. The 2020 imposed lockdown due to the Coronavirus Disease 2019 (COVID-19) pandemic is likely to have influenced the daily life of people with COPD.

We aimed to compare feature selection (FS) methods and describe the effect of the COVID-19 lockdown, sociodemographic and clinical features on the impact of the disease on people with COPD.

Methods:

Sociodemographic, anthropometric and clinical data from stable people with COPD recruited in GENIAL (PTDC/DTP-PIC/2284/2014) and PRIME (PTDC/SAU-SER/28806/2017) projects were used.

The COPD assessment test (CAT) was performed at baseline and 5 month after (post) and evaluated the disease impact [12], [13]. The minimal clinically important difference (MCID) is 2 points [14].

Change of CAT score (dCAT) was considered the outcome.

FS was performed in numeric data standardized by subtracting the mean and dividing by the standard deviation: (i) the λ used in LASSO was the one that produced the lowest 5-fold cross-validation MSE from a grid of 15000 values; (ii) the AIC/BIC based stepwise automatic selection consisted of a backward elimination of terms from a LM with all features in order to obtain the lowest AIC/BIC [15]; (iii) normalized entropy procedure with optimization of the support [10].

Ordinary least squares (OLS) LMs and fit measures were applied with the features selected. A LM with the features selected by the entropy algorithm that returned the highest leave-one-out crossvalidation R^2 (LOOCV R^2) was computed with non-standardized data. An α=0.05 was considered.

Results:

A total of 42 participants with mean age 66.3 years (sd 7.8), 3 to 4 comorbidities (64.3%) and a median CAT score of 9.0 ([Q1,Q3]=[5.3,11.0]) were included, 24 (57.1%) of whom in the prelockdown group. No significant differences were found between groups (Table S1) nor median CAT scores at different assessments (Figure 1).

The MSE was minimized with λ=1.26 and selected CCI and respiratory emergencies (Figure S1, Table 1).

The AIC algorithm removed 18 features. With decreasing order of importance CCI, AECOPD and SGRQ were kept. Using BIC, CCI and respiratory emergencies remained. (Table 1).

CCI had the lowest normalized entropy (0.901) followed by the SGRQ (0.929). Respiratory emergencies, pack-years and BMI registered a value under 0.95. (Table 1, Figure S2).

Keywords: COPD, COVID-19, Feature Selection, LASSO, Normalized Entropy, Stepwise Selection

Corresponding author: Jorge Cabral jorgecabral@ua.pt

Supplementary material: CabralJ_EA35_SuppMat.pdf

Conflict of interest: The authors declare no conflict of interests.

Clinical study registration number: NCT03701945

First published: 20JUL2022

Table 1 - Feature's importance according to LASSO, AIC based stepwise automatic selection (StepAIC), BIC based stepwise automatic selection (StepBIC) and entropy estimation algorithms.

Features	LASSO	StepAIC	StepBIC	Entropy (NE)	Mean Importance
CCI	1	1	1	1 (0.901)	1
SGRQ	4	3	4	2 (0.929)	3.25
Respiratory emergencies	2	11	2	3 (0.941)	4.5
AECOPD	3	2	3	13 (0.990)	5.25
FEV1 % predicted	5	4	5	9 (0.974)	5.75
Group	7	6	7	16 (0.996)	9
mMRC	11	9	10	8 (0.972)	9.5
Respiratory hospitalizations	6	13	13	6 (0.963)	9.5
FEV1/FVC	12	5	6	18 (0.997)	10.25
BORG Fatigue	9	7	8	19 (0.997)	10.75
LTOT	8	8	9	20 (0.999)	11.25
Sex	10	10	11	15 (0.995)	11.5
Age	14	12	12	12 (0.987)	12.5
NIV	13	15	15	7 (0.965)	12.5
Body mass index	19	16	16	5 (0.944)	14
Pack years	17	20	20	4 (0.943)	15.25
Smoking no. of years	18	14	14	17 (0.997)	15.75
BPAAT Moderate	21	17	17	10 (0.981)	16.25
Smoking status	15	21	21	11 (0.987)	17
BORG Dyspnoea	16	19	19	14 (0.993)	17
BPAAT Vigorous	20	18	18	21 (1.000)	19.25

Abbreviations: AECOPD, acute exacerbation of COPD; BPAAT, brief physical activity assessment tool; BMI, body mass index; CCI, Charlson comorbidity index; COPD, chronic obstructive pulmonary disease; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; LTOT, long-term oxygen therapy; mMRC, modified medical council dyspnoea scale; NIV, non-invasive ventilation; SGRQ, St. George's respiratory questionnaire; NE, normalized entropy. Green indicates that features were selected.

The LM using the features selected by LASSO and the BIC method was the same and had the lowest AIC and highest and LOOCV R^2 (0.12). The LM generated by the AIC method and the entropy algorithm with 3 features achieved the highest R^2 (0.27). No significant differences between models were found (Table 2).

The LM with 3 features from the entropy algorithm shows that participants with severe CCI are expected to have a decreased dCAT by 6.47 point when compared with participants with mild CCI (CI95=[2.49,10.45]). Participants with respiratory emergencies tend to have an increased dCAT by 3.22 points. If at the same time, they have a mild or moderate CCI score they tended to recover above the MCID. Those without emergencies but with a severe CCI are expected to worsen above the MCID (Figure 2).

Table 2 - Linear model's coefficients and p values for the COPD assessment test score difference using 135 as predictors the features selected by LASSO, AIC based stepwise automatic selection (StepAIC), BIC 136 based stepwise automatic selection (StepBIC) and entropy estimation algorithms (n=42).

	LASSO		StepAIC		StepBIC		Entropy 1 feature		Entropy 2 features		Entropy 3 features		Entropy 4 features		Entropy 5 features	
	Beta	*p*	*Beta*	*p*	*Beta*	*p*	*Beta*	*p*	*Beta*	*p*	*Beta*	*p*	*Beta*	*p*	*Beta*	*p*
AECOPD	-	-	0.32	**0.035**	-	-	-	-	-	-	-	-	-	-	-	-
CCI	-0.42	**0.003**	-0.45	**0.002**	-0.42	**0.003**	-0.41	**0.006**	-0.41	**0.007**	-0.43	**0.003**	-0.43	**0.004**	-0.43	**0.007**
BMI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.01	0.941
Pack Years													0.00	0.978	0.00	0.987
Emergencies	0.29	**0.038**			0.29	0.038	-	-			0.30	**0.035**	0.30	**0.038**	0.30	**0.041**
SGQR			-0.23	0.120	-	-	-	-	-0.12	0.398	-0.13	0.339	-0.13	0.376	-0.13	0.381
R^2	0.254		0.271		0.254		0.167		0.182		0.271		0.271		0.271	
adjusted R^2	0.216		0.215		0.216		0.147		0.141		0.215		0.194		0.173	
LOOCV R^2	0.124		0.082		0.124		0.073		0.040		0.091		0.063		0.033	
AIC	111.892		112.889		111.892		114.490		115.730		112.894		114.893		116.887	
AICc	112.524		113.970		112.524		114.798		116.362		113.975		116.560		119.287	
BIC	117.105		119.839		117.105		117.966		120.943		119.845		123.581		127.313	
log-L	-52.946		-52.444		-52.946		-55.245		-54.865		-52.447		-52.447		-52.443	

Abbreviations: AECOPD, acute exacerbation of COPD; BMI, body mass index; CCI, Charlson comorbidity index; COPD, chronic obstructive pulmonary disease; SGRQ, St. George's respiratory questionnaire; *p*, *p* value; log-L, log-likelihood; LOOCV, leave-one-out cross-validation.

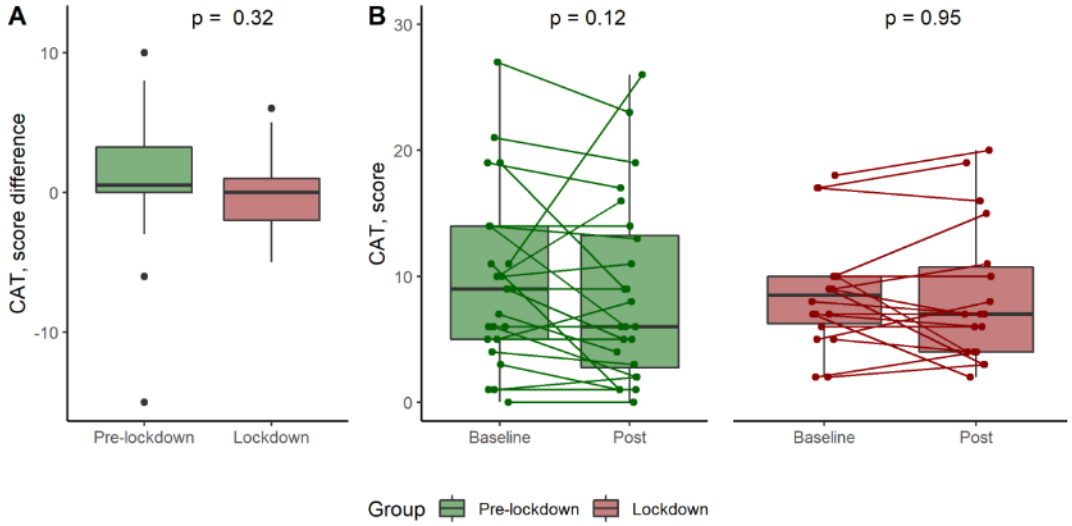

Figure 1 - Distribution of the participants' COPD assessment test (CAT) score: (A) score difference according to the group; (B) scores in the different assessments according to the group. p, p value for the Wilcoxon rank sum test (A) or Wilcoxon signed rank test (B).

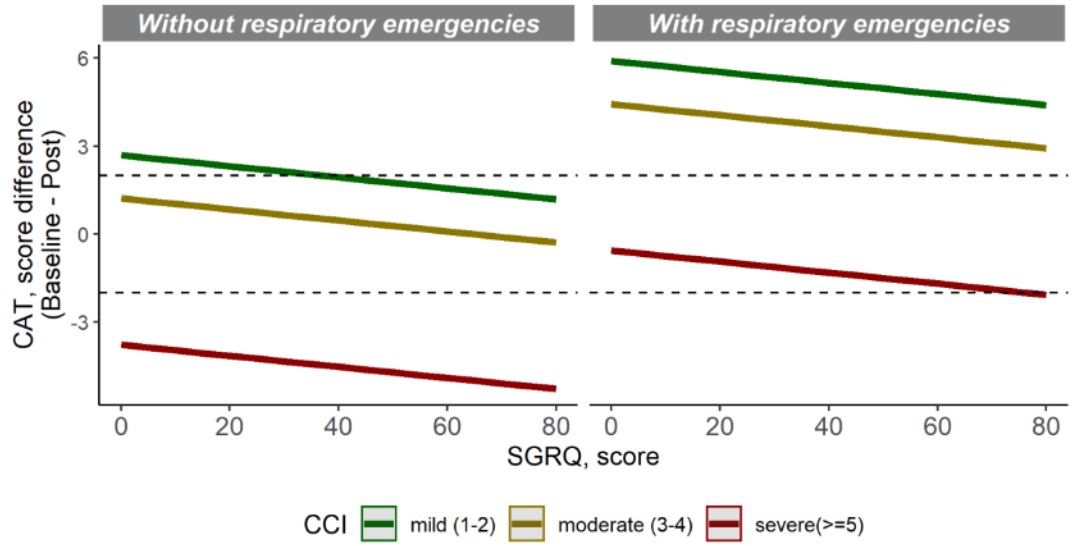

Figure 2 - Predicted difference between baseline and post COPD assessment test score according to the Charlson comorbidity index (CCI), the existence of respiratory emergencies in the previous year and the St. George respiratory questionnaire' score. Dashed line represents the minimal clinically important difference.

Discussion:

In regression models with many features where does not exist relationships between features and dependent variable, some features can be considered relevant by significance tests [16]. Elimination algorithms can overestimate the effect size of features and should not be used if $p > n$ [17,18]. OLS may be biased [19] and lead to unstable solutions because they cannot deal with limited information, small samples and collinearity. Normalized entropy estimation is appealing because it imposes no structure on data [10]. Nevertheless, the LM obtained with 3 features selected by the entropy approach was at least not worse than the remaining.

Our model suggests that lockdown had no influence in COPD impact but those with comorbidities but no emergencies tended to recover poorly from the pandemic.

Ethics committee and informed consent:

Five independent Ethics Committees (Centro Hospitalar do Médio Ave ref. 09/2016 and 10/2018; Unidade Local de Saúde de Matosinhos ref. 10/CES/JAS 17/02/2017 and 73/CE/JAS 12/10/2018; Centro Hospitalar Baixo Vouga ref. 777638 and 086892; Hospital Distrital da Figueira da Foz ref. 1807/2017 and 27/05/2019; Administração Regional de Saúde do Centro ref. 64/2016 and 85/2018) approved the study. Written informed consent was obtained from all participants before data collection. Data protection was ensured by the National Committee for Data Protection (no. 7295/2016) and followed the General Data Protection Regulation.

Acknowledgements:

This work was supported by FCT through CIDMA and projects UIDB/04106/2020 and UIDP/04106/2020. This research was also supported by Programa Operacional de Competitividade e Internacionalização—POCI, through Fundo Europeu de Desenvolvimento Regional—FEDER (POCI01-0145- FEDER-028806), by Fundação para a Ciência e Tecnologia (PTDC/SAU-SER/28806/2017) and under the project UIDB/04501/2020.

References:

1. Bishop CM. Pattern Recognition and Machine Learning. Springer; 2006.
2. Akaike H. Maximum likelihood identification of gaussian autoregressive moving average models. Biometrika. 1973;60(2):255-265. https://doi.org/10.1093/biomet/60.2.255
3. Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics. 1978;6(2):461-464. https://doi.org/10.1214/aos/1176344136
4. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media; 2009.
5. Abu-Mostafa YS, Magdon-Ismail M, Lin HT. Learning from Data. Vol 4. AMLBook New York, NY, USA; 2012.
6. George EI. The Variable Selection Problem. J Am Stat Assoc. 2000;95(452):1304-1308. https://doi.org/10.1080/01621459.2000.10474336
7. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;58(1):267-288. http://www.jstor.org/stable/2346178 https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
8. Jaynes ET. Information Theory and Statistical Mechanics. Physical Review. 1957;106(4):620630. https://doi.org/10.1103/PhysRev.106.620
9. Golan A, Judge GG, Miller D. Maximum Entropy Econometrics: Robust Estimation with Limited Data. Wiley; 1996.
10. Mittelhammer R, Cardell N, Marsh T. The Data-Constrained Generalized Maximum Entropy Estimator of the GLM: Asymptotic Theory and Inference. Entropy. 2013;15(12):1756-1775. https://doi.org/10.3390/e15051756
11. Global Initiative for Chronic Obstructive Lung Disease. GOLD Report 2022. Global Initiative for Chronic Obstructive Lung Disease. Published online 2022.
12. George F. Diagnóstico e Tratamento Da Doença Pulmonar Obstrutiva Crónica. 028/2011. Direção Geral da Saúde; 2013.
13. Jones PW, Harding G, Berry P, Wiklund I, Chen WH, Kline Leidy N. Development and first validation of the COPD Assessment Test. European Respiratory Journal. 2009;34(3):648. https://doi.org/10.1183/09031936.00102509
14. Kon SSC, Canavan JL, Jones SE, et al. Minimum clinically important difference for the COPD Assessment Test: a prospective analysis. The Lancet Respiratory Medicine. 2014;2(3):195-203. https://doi.org/10.1016/S2213-2600(14)70001-3
15. Zuur A, Ieno E, Walker N, Saveliev A, Smith G. Mixed Effects Models and Extensions in Ecology With R.; 2009. https://doi.org/10.1007/978-0-387-87458-6_1
16. Freedman DA. A Note on Screening Regression Equations. The American Statistician. 1983;37(2):152-155. https://doi.org/10.1080/00031305.1983.10482729
17. Gareth J, Hastie T, Tibshirani R, Witten D. An Introduction to Statistical Learning: With Applications in R. Springer Science + Business Media, LLC; 2013.
18. Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP. Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology. 2006;75(5):1182-1189. https://doi.org/10.1111/j.1365-2656.2006.01141.x
19. Ryan TP. Modern Regression Methods. Wiley; 2009. https://doi.org/10.1002/9780470382806