**Title**: A simple algorithm for the identification of clinical COPD phenotypes

**Authors**: Pierre-Régis Burgel[1,2;*], M.D., Jean-Louis Paillasseur[3;*], Wim Janssens[4], M.D., Jacques Piquet[5;#], M.D., Gerben ter Riet[6], Judith Garcia-Aymerich[7Take] , M.D., Borja Cosio[8], M.D., Per Bakke[9], M.D., Milo A Puhan[10], M.D., Arnulf Langhammer[11], M.D., Inmaculada Alfageme[12], M.D., Pere Almagro[13], M.D., Julio Ancochea[14], M.D., Bartolome R Celli[15], M.D., Ciro Casanova[16], M.D., Juan P de-Torres[17], M.D., Marc Decramer[4], M.D., Andrés Echazarreta[18] , M.D., Cristobal Esteban[19], M.D., Rosa Mar Gomez Punter[20], M.D., MeiLan K Han[21], M.D., Ane Johannessen[22], M.D., Bernhard Kaiser[23], Ph.D., Bernd Lamprecht[24], M.D., Peter Lange[25], M.D., Linda Leivseth[26], PhD., Jose M Marin[27], M.D., Francis Martin[28;#], M.D., Pablo Martinez-Camblor[29], Ph.D., Marc Miravitlles[30], M.D., Toru Oga[31], M.D., Ana Sofia Ramírez[32], M.D., Don D Sin[33], M.D., Patricia Sobradillo[34], M.D., Juan J Soler-Cataluña[35], M.D., Alice M Turner[36], M.D., Francisco Javier Verdu Rivera[37], M.D., Joan B Soriano[38],M.D., and Nicolas Roche[1,2,*], M.D. on behalf of Initiatives BPCO, EABPCO, Leuven and 3CIA study groups

[1] University Paris Descartes (EA2511), Sorbonne Paris Cité, Paris, France; pierre-regis.burgel@aphp.fr

[2] Dpt of Respiratory Medicine, Cochin Hospital, AP-HP, Paris, France

[3] Effi-Stat, Paris France; jean-louis.paillasseur@effi-stat.com

[4] Respiratory Division, University Hospital Gasthuisberg, K.U. Leuven, Leuven, Belgium; wim.janssens@uzleuven.be

[5] Dept of Respiratory Medicine, Le Raincy-Montfermeil Hospital, Montfermeil, France; jpiquet@ch-montfermeil.fr

[6] Dept. General Practice - Academic Medical Center, Amsterdam, The Netherlands; g.terriet@amc.uva.nl

[7] ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Universitat Pompeu Fabra (UPF), CIBER Epidemiología y Salud Pública (CIBERESP) Barcelona, Spain; jgarcia@creal.cat

[8] Unidad de Investigación, Servicio de Neumología, Hospital Universitario Son Espases, Palma de Mallorca, Spain; borja.cosio@ssib.es

[9] Dept. of Clinical Science, Faculty of Medicine and Dentistry, University of Bergen, Norway; Per.Bakke@uib.no

[10] Epidemiology, Biostatistics und Prevention Institute (EBPI), University of Zurich, Switzerland; miloalan.puhan@uzh.ch

[11] Department of Public Health and General Practice, HUNT Research Centre, Norwegian University of Science and Technology, Levanger, Norway; arnulf.langhammer@ntnu.no

[12] Universidad de Sevilla, Spain; ialfageme@separ.es

[13] Internal Medicine, Hospital Universitari Mutua de Terrassa, Universitat de Barcelona, Barcelona, Spain; 19908pam@comb.cat

[14] Pneumology Service, La Princesa Institute for Health Research (IP), Hospital Universitario de la Princesa, Madrid, Spain; juli119@gmail.com)

[15] Brigham and Women's Hospital, Boston, Massachusetts, United States; BCelli@copdnet.org

[16] Hospital Nuestra Señora de la Candelaria, Tenerife, Spain; casanovaciro@gmail.com

[17] Clınica Universidad de Navarra, Pamplona, Spain; jupa65@hotmail.com

[18] Servicio de Neumonología Hospital San Juan de Dios de La Plata, Buenos Aires. Argentina; aechaza@ciudad.com.ar

[19] Hospital Galdakao-Usansolo, Galdakao, Bizkaia, Spain; cristobal_esteban@yahoo.es

[20] Servicio de Neumología, Hospital Universitario La Princesa, Madrid, Spain; rosamar_gp@hotmail.com

[21] University of Michigan, Ann Arbor, MI, USA; mrking@med.umich.edu

[22] Centre for Clinical Research, Haukeland University Hospital, Bergen, Norway; ane.johannessen@helse-bergen.no

[23] Department of Pulmonary Medicine, Paracelsus Medical University Hospital, Salzburg, Austria; bernhardkaiser@gmx.at

[24] Department of Pulmonary Medicine, General Hospital Linz (AKH), Linz, Austria; bernd.lamprecht@akh.linz.at

[25] Section of Social Medicine, Department of Public Health, Copenhagen University, Copenhagen Denmark; peter.lange@sund.ku.dk

[26] Centre for Clinical Documentation and Evaluation, Northern Norway Regional Health Authority, Tromso, Norway; linda.leivseth@helse-nord.no

[27] Hospital Universitario Miguel Servet, Zaragoza, Spain; jmmarint@unizar.es

[28] Pneumologie, Centre Hospitalier de Compiègne, Compiègne, France, F.MARTIN@ch-compiegnenoyon.fr

[29] Hospital Universitario Central de Asturias (HUCA), Oviedo, Spain; and Universidad Autónoma de Chile, Chile; pmcamblor@hotmail.com

[30] Pneumology Department. Hospital Universitary Vall d'Hebron. CIBER de Enfermedades Respiratorias (CIBERES), Barcelona, Spain; mmiravitlles@vhebron.net

[31] Department of Respiratory Care and Sleep Control Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan; ogato@kuhp.kyoto-u.ac.jp

[32] Facultad de Medicina UASLP, San Luis Potosí, México; asofiarmz@gmail.com and asr.3cia@gmail.com

[33] James Hogg Research Centre, University of British Columbia; Division of Respiratory Medicine, Department of Medicine, St Paul's Hospital, Vancouver, Canada; don.sin@hli.ubc.ca

[34] Hospital Universitario Araba, Sede Txagorritxu, Vitoria, Spain; psobradillo@separ.es

[35] Servicio de Neumología, Hospital Arnau de Vilanova, Valencia, Spain; jjsoler@telefonica.net

[36] Queen Elizabeth Hospital Research Laboratories, Birmingham, UK; a.m.wood@bham.ac.uk
[37] H.U. Son Espases, Palma de Mallorca, Spain; franciscoj.verdu@ssib.es

[38] Instituto de Investigación Hospital Universitario de la Princesa (IISP) , Universidad Autónoma de Madrid , Madrid , Spain; jbsoriano2@gmail.com

**Corresponding author:**

Pierre-Régis Burgel, MD PhD
Service de Pneumologie, Hôpital Cochin
27 Rue du Faubourg St Jacques
75014 Paris, France
E-mail: pierre-regis.burgel@cch.aphp.fr
Tel: 33 1 58 41 23 67
Fax: 33 1 46 33 82 53

**Word count:** 2902

**Figures and Tables:** 3 tables, 5 figures

**References:** ~~28~~ 29

## ABSTRACT (193 words)

This study aimed at identifying simple rules for allocating COPD patients to clinical phenotypes identified by cluster analyses.

Data from 2409 COPD patients from French/Belgian COPD cohorts were analysed using cluster analysis, resulting in the identification of subgroups for which clinical relevance was determined by comparing 3-year all-cause mortality. Classification and regression trees (CARTs) were used to develop an algorithm for allocating patients to these subgroups. This algorithm was tested in 3651 patients from the COPD Cohorts Collaborative International Assessment (3CIA) initiative.

Cluster analysis identified five subgroups of COPD patients with different clinical characteristics (especially regarding severity of respiratory disease and presence of cardiovascular comorbidities and diabetes). CART-based algorithm indicated that the variables relevant for patient grouping differed markedly between patients with isolated respiratory disease ($FEV_1$, dyspnoea grade) and those with multimorbidity (dyspnoea grade, age, $FEV_1$ and body mass index). Application of this algorithm to the 3CIA cohorts confirmed that it identified subgroups of patients with different clinical characteristics, mortality rates (median, from 4% to 27%) and age at death (median, from 68 to 76 years).

A simple algorithm, integrating respiratory characteristics and comorbidities, allowed the identification of clinically-relevant COPD phenotypes.

**Take home message**: An algorithm integrating respiratory characteristics and comorbidities identifies clinical COPD phenotypes.

## INTRODUCTION

Airflow limitation is the hallmark of chronic obstructive pulmonary disease (COPD) and forced expiratory volume in one second ($FEV_1$) has long been used as the main criteria for characterization of disease severity [1, 2]. Analyses of observational cohorts (e.g., the ECLIPSE cohort) have revealed that COPD patients with similar levels of $FEV_1$ experienced different degrees of disease burden reflected by dyspnoea levels, exacerbations rates, health-related quality of life (HRQoL) impairment and exercise limitation [3]. Accordingly, the current classification of COPD proposed by the Global initiative for chronic Obstructive Lung Disease (GOLD) incorporates not only $FEV_1$ but also dyspnoea or HRQoL, and previous occurrence of COPD exacerbations and/or hospitalization [1]. Although this classification is not fully evidence-based, it has the advantage of taking into account some of the clinical heterogeneity of COPD with the aim of predicting future risk and proposing corresponding treatment choices. A limitation of this classification is that it does not account for age, an important determinant of prognosis in patients with COPD [4]. Further, the GOLD classification does not account for comorbidities, which are frequent and contribute to prognosis [5-7].

Several groups have used clusters analyses to explore clinical heterogeneity in cohorts of patient in COPD [8-10]. These studies have identified consistent clinical COPD phenotypes at high risk of mortality, including (i) younger patients with severe respiratory disease, few cardiovascular co-morbidities, and poor nutritional status, and (ii) older patients with moderate respiratory disease, metabolic and cardiovascular co-morbidities, and obesity [11]. They have also identified patients with mild disease and good prognosis [12, 13]. However, all published studies had limitations related to relatively small sample size and lack of further validation in independent samples [11, 13]. Further, the results of cluster analyses are difficult to translate for use in daily practice, since they provide no tool for individual patient allocation in the identified phenotypes.

In the present study, our aim was to develop and validate an algorithm, based on easily available clinical data, for assigning patients with COPD to clinically-relevant phenotypes.

## METHODS

### Overall design

Data from three French/Belgian COPD cohorts were used to identify clinical COPD phenotypes using cluster analysis. Classification and Regression Trees (CARTs) [14] analysis was then used to develop an algorithm for allocating individual COPD patients recruited in these French/Belgian cohorts to specific subgroups. This algorithm was further tested in an independent sample of patients with COPD, using data from the COPD Cohorts Collaborative International Assessment (3CIA) initiative [15].

### COPD patient cohorts

The French/Belgian COPD cohorts are composed of three cohorts: the INITIATIVES BPCO cohort [8], the French College of General Hospital Respiratory Physicians (CPHG) cohort [16], and the Leuven cohort [12]. Patients within these cohorts have a diagnosis of COPD based on post-bronchodilator $FEV_1/FVC<0.70$ and were recruited in stable state in university hospitals (INITIATIVES BPCO and Leuven cohorts) [8, 12] or at the time of hospitalization for COPD exacerbations (CPHG cohort) [16], as previously described. The COPD Cohorts Collaborative International Assessment (3CIA) initiative contains pooled individualized data from 22 cohorts of patients with COPD, who were recruited in publicly-funded hospitals or in population based-studies [15]. All cohorts were approved by local Ethics Committee and all subjects provided informed written consent.

### Statistical analysis plan

First, COPD patients recruited in the French/Belgian cohorts were classified into subgroups based on the results of a cluster analysis of data obtained at inclusion in the cohorts. The clinical relevance of the identified subgroups was established by examining their association with 3-

year all-cause mortality. Next, CARTs were used for the development of an algorithm, assigning COPD patients to the subgroups identified by cluster analysis. The clinical value of this algorithm was examined using 3-year all-cause mortality in the French/Belgian cohorts. Finally, the algorithm was tested for external validation using data from the 3CIA initiative database [15]. Mortality risks among subgroups were analysed using Kaplan-Meier curves and Cox models. Concordance Probability Estimate (CPE) was used to evaluate the discriminatory power of classifications for mortality prediction. Data are presented as median (interquartile range, IQR) or n (%). Analyses were performed using SAS 9.2 (SAS Institute Inc., Cary, NC, USA) and Tanagra 1.4 (Lyon, France) softwares. Additional information on the methods can be found in the online supplement to this manuscript.

### Cluster analysis of the French/Belgian COPD cohorts

Variables were selected for inclusion in the cluster analysis based on their previous association with future risk and prognosis in COPD patients [1, 6] and included age, body mass index (BMI), $FEV_1$ (% predicted), modified Medical Research Council (mMRC) dyspnoea scale, number of exacerbations in the previous 12 months, and presence/absence of cardiovascular comorbidities (hypertension, coronary artery disease and/or left heart failure) and/or diabetes. Identification of subgroups of patients with COPD associated with survival was achieved using factor analysis for mixed data (FAMD) [17, 18], followed by classification of patients using Ward’s agglomerative hierarchical cluster analysis [8, 12]. The clinical relevance of the identified subgroups was examined by comparing their all-cause mortality at three years, as described previously [8, 12]. These subgroups (phenotypes) were labelled using Roman numbers.

### Development of an algorithm for assigning COPD patients to specific subgroups in the French/Belgian cohorts

The development of an algorithm for assigning COPD patients to the subgroups identified by cluster analysis was achieved using CART analysis [14, 19], a non-parametric decision tree learning technique [19]. Variables included in this analysis were those selected for the cluster analysis (see above). Threshold values for these variables were based on those obtained by CART analysis and were slightly modified for improved practicality (see online supplement for a detailed explanation).

## External validation of the algorithm

The algorithm established in the French/Belgian cohorts was then tested in an independent group of patients with COPD from the 3CIA database. Patients in this database (n=16332) were considered eligible for the study if data necessary to apply the algorithm (age, BMI, $FEV_1$ % predicted, mMRC scale, presence/absence of cardiovascular comorbidities and diabetes) and information on vital status at three years were available. Patients with appropriate data (n=3651) were classified by the algorithm into the five classes described above (labelled using Arabic digits), and these classes were compared according to their clinical characteristics, all-cause mortality at three years and age at death.

## RESULTS

### Patients and overall study design

Study design is explained in **Figure 1** and characteristics of patients with COPD at inclusion in the French/Belgian cohorts (n=2409 patients) and in the 3CIA database (n=3651 patients) are presented in **Table S1**. Their 3-year all-cause mortality rates were 30.8% and 11.6%, respectively. Patients included in the French/Belgian cohorts were characterized by older age, more severe airflow limitation and higher rates of cardiovascular comorbidities and/or diabetes. Further, 57% of patients in the French/Belgian cohorts were recruited at the time of hospitalization for COPD exacerbations (as part of the CHPG cohort) [16].

### Cluster analysis of the French/Belgian cohorts

**Table 1** shows the five subgroups (labelled **I to V**) identified in the French/Belgian COPD cohorts using cluster analysis (*see online supplement, **Table S2 to S6 and Figure S1***). **Table 2** summarizes the main descriptors of these subgroups according to increasing rates of 3-year all-cause mortality. Subgroup V (mortality rate, 2.5%) was characterized by mild respiratory disease and low rates of comorbidities. Subgroup II (mortality rate, 21.8%) was characterized by moderate to severe respiratory disease and low rates of comorbidities. Subgroup III (mortality rate, 30.0%) was characterized by older age than subgroup II, with high prevalence of comorbidities and obesity. Subgroup IV (mortality rate, 47.0%) was characterized by very severe respiratory disease with low rates of cardiovascular comorbidities and diabetes. Subgroup I (mortality rate, 50.9%) had less severe respiratory disease than subgroup IV but had older age and very high rates of cardiovascular comorbidities and diabetes.

### Use of CART for the development of an algorithm assigning COPD patients to subgroups of patients identified by cluster analysis in the French/Belgian cohorts

CART analysis provided an algorithm that allowed assigning up to 80% of the patients to the subgroups identified by cluster analysis (*see online supplement,* ***Table S7 and S8***). This algorithm is presented in **Figure 2** and clinical characteristics of patients according to the five classes obtained by applying this algorithm are presented in **Table 3.** Kaplan-Meier survival curves by cluster analysis-defined subgroups (**Figure 3A**) and by CART-defined classes (**Figure 3B**) showed comparable results. Concordance probability estimates were 0.61 (95% CI; 0.59 – 0.63) for cluster analysis-defined subgroups and 0.60 (95% CI; 0.58 – 0.62) for CART-defined classes, confirming that both methods had comparable discriminatory power for the identification of subgroups with different prognosis.

## Evaluation of the algorithm using data from the 3CIA initiative database

The algorithm developed in the French/Belgian cohorts was then tested using data obtained in COPD patients from the 3CIA database. Characteristics of the 3651 patients distributed into classes according to this algorithm are presented in **Table 3**. Kaplan-Meier survival curves by classes are presented in **Figure 3C**. The concordance probability estimates was 0.62 (95% CI; 0.59 – 0.64).

## Comparison of mortality rates among classes in the French/Belgian COPD cohorts vs. 3CIA database

Because 3-year mortality rates varied widely between French/Belgian COPD cohorts and the 3CIA database, we used Cox analysis to examine hazard ratios of mortality among patients in the five classes defined by our algorithm in the French/Belgian and 3CIA cohorts, respectively. Forest plots corresponding to these analyses are presented in **Figure 4.** Although absolute rates of death were markedly higher in the French/Belgian cohorts, hazard ratios of mortality among the five classes were rather comparable in the French/Belgian cohorts and in the 3CIA initiative.

Distribution by GOLD grades of severity of airflow limitation [1] in patients who died during follow-up is presented in **Figure 5**. When comparing classes with high rates of all-cause mortality, patients without cardiovascular comorbidities/diabetes (class 4) who died were predominantly in GOLD 4 whereas patients with cardiovascular comorbidities/diabetes (class 1) who died had less severe airflow limitation (predominantly GOLD 2 and 3). Comparable observations were made when comparing patients in class 2 vs. class 3 (intermediate mortality rates).

## DISCUSSION

In the present study, we first performed a cluster analysis in a pool of French/Belgian COPD cohorts, which identified five subgroups (phenotypes) of patients with different rates of all-cause mortality at 3 years and different age at death. We then used CART analysis in this pool of French/Belgian cohorts to develop an algorithm that allowed allocation of patients into five classes, corresponding to the subgroups identified by cluster analysis. This simple algorithm was based on clinical variables (including cardiovascular comorbidities and/or diabetes and respiratory characteristics) routinely available in daily practice. Classification of COPD patients using this algorithm allowed the identification of subgroups of patients differing on 3-year all-cause mortality and age at death in the pool of French/Belgian cohorts, providing internal validation of the approach. It provided comparable results in patients included in the 3CIA initiative database, which contained an independent group of patients with COPD recruited in multinational cohorts, providing external validation. This algorithm identifies clinical phenotypes relevant to prognosis in patients with COPD, which could help exploring underlying pathophysiological mechanisms and developing novel strategies of care.

The algorithm described in the present study is the first to integrate comorbidities (cardiovascular diseases, diabetes, and obesity) and age to more classical respiratory variables ($FEV_1$ and dyspnoea) for improving the characterization of patients with COPD. An important yield of this algorithm is to identify patients belonging to two subgroups with poorer prognosis, i.e. class 1 and 4, and to highlight the corresponding determinants, i.e., the severity of the respiratory component (as assessed by the degrees of lung function impairment and dyspnoea) and the presence of major cardiovascular comorbidities or metabolic risk factors (diabetes). This data confirm previous studies showing that (A) cardiovascular and metabolic comorbidities contribute to worsen outcomes (e.g., mortality, hospitalization and exacerbation) in patients with COPD [6, 20] and (B) two very different phenotypes of COPD patients at poor prognosis exist (those with severe respiratory disease, often occurring at a younger age and those with multimorbidity including cardiovascular and metabolic diseases, often characterized by an older age) [9, 12]. Importantly, this study extends previous data by studying larger numbers of patients (including larger numbers of women) recruited in multiple countries and provides a simple algorithm that can be used in the clinic for classifying the patients. One yield of the algorithm is to highlight the variables on which clinicians and researchers should focus during follow-up and treatment adaptation. Whether specific strategies need to be developed for all or some of the identified phenotypes now needs to be tested prospectively. Similarly, future studies should aim at determining whether these phenotypes are associated with specific biomarkers reflecting underlying pathophysiological mechanisms.

The main strengths of the present study were the application of exploratory statistical analyses complemented by clinical knowledge in large cohorts of patients, the validation of findings in an external pool of cohorts and the use of a robust variable (mortality) for validation. We also recognize that the present study has limitations. Our assessment of comorbidities was based on physician diagnoses, not taking into account occult conditions, which are reported to occur in COPD patients [21]. To limit such underestimation of the impact of undiagnosed cardiovascular diseases, the definition of cardiovascular comorbidities was relatively loose and included hypertension (a risk factor for cardiovascular disease rather than a disease itself). This definition also corresponds to what happens in real-life daily practice, where many patients do not benefit from systematic screening for cardiovascular comorbidities. Although COPD patients are at high risk of lung cancer, which is associated with poor prognosis, patients with active lung cancer were generally excluded from the present cohorts, limiting our findings to COPD patients without active lung cancer. Specific causes of mortality were not available in the cohorts used in the present analyses and the prognostic value of the phenotypes was confirmed using all-cause mortality. Previous studies showed that causes of mortality in COPD populations differ between patients with mild vs. severe airflow obstruction, with a higher relative weight of cancer and cardiovascular causes in patients with less severe airflow impairment, and more respiratory causes in the more severe [22]. Among patients who died, differences in the GOLD grades of airflow obstruction (see Figure 4) between phenotypes with comparable survival rates (e.g., class 1 vs. class 4 and class 2 vs. class 3) suggest that patients with high rates of cardiovascular comorbidities and/or diabetes (e.g., class 1 and 3) were more likely to die from extrapulmonary causes. Importantly, even if one of its properties is to identify populations with different mortality rates, the algorithm is not intended at representing a prognostic index, since the determinants of a given prognosis might differ markedly between patients of a given group. The large difference in mortality rates between the two groups of cohorts largely relates to the fact that 57% of patients in the French/Belgian cohorts were recruited at the time of hospitalization for a COPD exacerbation (CPHG cohort) [16], reflecting the prognostic impact of hospitalizations. Although hospitalization appears an important prognostic factor, it should be considered a marker of disease severity rather than a phenotype *per se.* This was the basis for not including previous hospitalisation as a variable in the cluster analysis. However, COPD exacerbations (which are important in the characterization of patients with COPD) [23] were included in the cluster analysis and in the CART analysis. The finding that exacerbations were not retained in our final algorithm should not be misinterpreted as exacerbations remain important events in the life of patients with COPD [24]; it merely reflects that non-hospitalized exacerbations were not significantly related to prognosis. The performance of classification trees could also be improved by the integration of biomarkers reflecting inflammatory (fibrinogen, white blood cell count, CRP, eosinophils…) [25-27] and cardiovascular (BNP, copeptin, pro-adrenomedullin…) [28] biological phenomena.

The field of COPD phenotypes was once considered “the future of COPD” [29], but moving from exploratory research studies to the clinic has proven difficult. The algorithm described in the present study offers a new way of combining and hierarchizing well-known prognostic criteria (including comorbidities, age and symptoms) to identify COPD phenotypes in the clinic. This approach may serve as a basis for developing phenotype-specific therapeutic strategies by recruiting appropriate at-risk target populations in clinical trials. We speculate that our algorithm may also help in unravelling specific biological pathways that were previously missed due to mixing of various phenotypes in the current classifications of COPD.

## Funding sources

The analyses reported here were supported by an unrestricted grant from Boehringer Ingelheim France, which played no role in study design, data collection, analysis and interpretation of data, writing of the manuscript nor decision to submit it for publication.

The Initiatives BPCO study was supported by unrestricted grant from Boehringer Ingelheim France and (until 2015) Pfizer.

None of the funding sources of the individual trials were involved in any aspect of the 3CIA initiative, including the design, data collection and analysis, decision to publish, or preparation of the manuscript. PM-C was supported by research grant MTM2011-23204 from the Spanish Ministerio de Ciencia e Innovación (FEDER support included). JG-A has a researcher contract from the Instituto de Salud Carlos III (CP05/00118), Ministry of Health, Spain.

## References

1. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report: GOLD Executive Summary. Eur Respir J. 2017;49(3).
2. Celli BR, Decramer M, Wedzicha JA, Wilson KC, Agusti A, Criner GJ, et al. An official American Thoracic Society/European Respiratory Society statement: research questions in COPD. Eur Respir J. 2015;45(4):879-905.
3. Agusti A, Calverley PM, Celli B, Coxson HO, Edwards LD, Lomas DA, et al. Characterisation of COPD heterogeneity in the ECLIPSE cohort. Respir Res. 2010;11:122.
4. Puhan MA, Garcia-Aymerich J, Frey M, ter Riet G, Anto JM, Agusti AG, et al. Expansion of the prognostic assessment of patients with chronic obstructive pulmonary disease: the updated BODE index and the ADO index. Lancet. 2009;374(9691):704-11.
5. Divo M, Cote C, de Torres JP, Casanova C, Marin JM, Pinto-Plata V, et al. Comorbidities and risk of mortality in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2012;186:155-61.
6. Mannino DM, Thorn D, Swensen A, Holguin F. Prevalence and outcomes of diabetes, hypertension and cardiovascular disease in COPD. Eur Resp J. 2008;32(4):962-9.
7. Fabbri LM, Boyd C, Boschetto P, Rabe KF, Buist AS, Yawn B, et al. How to integrate multiple comorbidities in guideline development: article 10 in Integrating and coordinating efforts in COPD guideline development. An official ATS/ERS workshop report. Proc Am Thorac Soc. 2012;9(5):274-81.
8. Burgel PR, Paillasseur JL, Caillaud D, Tillie-Leblond I, Chanez P, Escamilla R, et al. Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. Eur Respir J. 2010;36(3):531-9.
9. Garcia-Aymerich J, Gomez FP, Benet M, Farrero E, Basagana X, Gayete A, et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. Thorax. 2011;66(5):430-7.
10. Rennard S, Locantore N, Delafont B, Tal-Singer R, Silverman E, Vestbo J, et al. Identification of five copd subgroups with different prognoses in the ECLIPSE cohort using cluster analysis. Annals of the American Thoracic Society. 2015.
11. Pinto LM, Alghamdi M, Benedetti A, Zaihra T, Landry T, Bourbeau J. Derivation and validation of clinical phenotypes for COPD: a systematic review. Respir Res. 2015;16:50.
12. Burgel PR, Paillasseur JL, Peene B, Dusser D, Roche N, Coolen J, et al. Two distinct chronic obstructive pulmonary disease (COPD) phenotypes are associated with high risk of mortality. Plos ONE. 2012;7:e51048.
13. Burgel PR, Paillasseur JL, Roche N. Identification of clinical phenotypes using cluster analyses in COPD patients with multiple comorbidities. Biomed Res Int. 2014;2014.
14. Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. Monterey, CA.: Wadsworth & Brooks; 1984.
15. Soriano JB, Lamprecht B, Ramirez AS, Martinez-Camblor P, Kaiser B, Alfageme I, et al. Mortality prediction in chronic obstructive pulmonary disease comparing the GOLD 2007 and 2011 staging systems: a pooled analysis of individual patient data. Lancet Respir Med. 2015;3:443-50.
16. Piquet J, Chavaillon JM, David P, Martin F, Blanchon F, Roche N. High-risk patients following hospitalisation for an acute exacerbation of COPD. Eur Respir J. 2013;42:946-55.
17. Wikipedia. Factor analysis of mixed data 2016 [cited 2016 january 25th]. Available from: https://en.wikipedia.org/wiki/Factor_analysis_of_mixed_data.
18. Pagès J. Multiple Factor Analysis for Mixed Data [article in French]. Rev Statistique Appliquée. 2004;52:93-111.
19. Wikipedia. Predictive analytics 2016 [cited 2016 January 25th]. Available from: https://en.wikipedia.org/wiki/Predictive_analytics#Classification_and_regression_trees.
20. Briggs A, Spencer M, Wang H, Mannino D, Sin DD. Development and validation of a prognostic index for health outcomes in chronic obstructive pulmonary disease. Archives of internal medicine. 2008;168(1):71-9.
21. Rutten FH, Cramer MJM, Grobbee DE, Sachs APE, Kirkels JH, Lammers JWJ, et al. Unrecognized heart failure in elderly patients with stable chronic obstructive pulmonary disease. Eur Heart J. 2005;26(18):1887-94.
22. Sin DD, Anthonisen NR, Soriano JB, Agusti AG. Mortality in COPD: role of comorbidities. Eur Resp J. 2006;28:1245-57.
23. Hurst JR, Vestbo J, Anzueto A, Locantore N, Müllerova H, Tal-Singer R, et al. Susceptibility to exacerbation in chronic obstructive pulmonary disease. N Engl J Med. 2010;363:1128-38.
24. Wedzicha JA, Seemungal TA. COPD exacerbations: defining their cause and prevention. Lancet. 2007;370(9589):786-96.

25. Duvoix A, Dickens J, Haq I, Mannino D, Miller B, Tal-Singer R, et al. Blood fibrinogen as a biomarker of chronic obstructive pulmonary disease. Thorax. 2013;68:670-6.
26. Thomsen M, Dahl M, Lange P, Vestbo J, Nordestgaard BG. Inflammatory biomarkers and comorbidities in chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2012;186:982-8.
27. George L, Brightling CE. Eosinophilic airway inflammation: role in asthma and chronic obstructive pulmonary disease. Ther Adv Chronic Dis. 2016;7:34-51.
28. Stolz D, Meyer A, Rakic J, Boeck L, Scherr A, Tamm M. Mortality risk prediction in COPD by a prognostic marker panel. Eur Resp J. 2014;44:1557-70.
29. Han MK, Agusti A, Calverley PM, Celli BR, Criner G, Curtis JL, et al. Chronic obstructive pulmonary disease phenotypes: the future of COPD. Am J Respir Crit Care Med. 2010;182:598-604.

**Table 1. Characteristics and 3-year mortality in COPD patients (n=2409) recruited in the French/Belgian COPD cohort according to the five subgroups identified using cluster analysis**

| Variable | Subgroup I<br>n= 609 | Subgroup II<br>n= 871 | Subgroup III<br>n= 327 | Subgroup IV<br>n= 287 | Subgroup V<br>n= 315 |
|---|---|---|---|---|---|
| *Male gender* | 82% (499) | 76% (662) | 80% (263) | 75% (215) | 78% (246) |
| *Age, years* | 77 [71 ; 81] | 64 [57 ; 73] | 74 [69 ; 80] | 64 [57 ; 71] | 61 [57 ; 67] |
| *BMI* | 25 [22 ; 28] | 24 [20 ; 27] | 30 [27 ; 34] | 20 [18 ; 23] | 26 [24 ; 29] |
| *Obesity >30 $kg/m^2$* | 14.8% (90) | 9.6% (84) | 49.5% (162) | 2.8% (8) | 19.4% (61) |
| *$FEV_1$ % predicted* | 36 [29 ; 46] | 45 [34 ; 56] | 59 [49 ; 68] | 27 [22 ; 33] | 78 [68 ; 92] |
| *GOLD grade* | | | | | |
| *GOLD 1* | 0 % | 3 % | 7 % | 0 % | 44 % |
| *GOLD 2* | 18 % | 35 % | 67 % | 4 % | 51 % |
| *GOLD 3* | 43 % | 43 % | 21 % | 31 % | 5 % |
| *GOLD 4* | 39 % | 19 % | 4 % | 65 % | 0 % |
| *mMRC dyspnoea scale* | 3 [3 ; 4] | 2 [1 ; 2] | 2 [1 ; 2] | 3 [3 ; 4] | 1 [0 ; 1] |
| *Exacerbations/patient/year* | 2 [1 ; 3] | 1 [0 ; 2] | 1 [0 ; 2] | 2 [2 ; 4] | 0 [0 ; 1] |
| *Hospitalizations/patient/year* | 1 [0 ; 2] | 0 [0 ; 1] | 0 [0 ; 1] | 1 [0 ; 2] | 0 [0 ; 0] |
| *Any cardiovascular disease or diabetes* | 91% (554) | 19% (165) | 96% (315) | 6% (17) | 26% (83) |
| *Specific comorbidities* | | | | | |
| *Coronary artery disease* | 39% (235) | 6% (56) | 39% (127) | 3% (10) | 12% (39) |
| *Hypertension* | 55% (334) | 11% (99) | 64% (208) | 2% (6) | 11% (36) |
| *Left heart failure* | 24% (148) | 3% (22) | 20% (66) | 1% (4) | 2% (9) |
| *Diabetes* | 7% (43) | 2% (18) | 12% (40) | 0% (1) | 8% (25) |
| *3-year mortality, % (n)* | 50.9% (310) | 21.8% (190) | 30.0% (98) | 47.0% (135) | 2.5% (8) |
| *Age at death, years* | 79 [74 ; 84] | 70 [61 ; 79] | 79 [74 ; 84] | 69 [63 ; 75] | 65 [60 ; 71] |

**Table 2. Main descriptors of the 5 COPD phenotypes identified by cluster analysis in the French/Belgian COPD cohort***

| | **GOOD prognosis** | **INTERMEDIATE prognosis** | | **POOR prognosis** | |
|---|---|---|---|---|---|
| *Phenotype number* | **V** | **II** | **III** | **IV** | **I** |
| *3-year mortality rate* | **2.5%** | **21.8%** | **30.0%** | **47.0%** | **50.9%** |
| *Phenotype name* | **Mild respiratory** | **Moderate to severe respiratory** | **Moderate to severe comorbid/obese** | **Very severe respiratory** | **Very severe comorbid** |
| *Airflow limitation* | Mild to moderate | Moderate to very severe | Mild to severe | Severe to very severe | Moderate to very severe |
| *Median body mass index* | 26 | 24 | 30 | 20 | 26 |
| *Clinical manifestations* | | | | | |
| *Dyspnea* | Mild | Moderate | Moderate | Severe | Severe |
| *Exacerbations* | 0 | Unfrequent | Unfrequent | Frequent | Frequent |
| *Hospitalizations* | 0 | Unfrequent | Unfrequent | Frequent | Frequent |
| *Rates of cardiovascular comorbidities/diabetes* | Low | Low | Very high | Very low | Very high |
| *Median age (years)* | 61 | 64 | 74 | 64 | 77 |

*the order is chosen based on 3-yr mortality rates.

**Table 3. Characteristics and 3-year mortality rates in COPD patients recruited in the French/Belgian COPD cohorts or in the 3CIA initiative database according to the five classes identified using the CART-based algorithm**

| **French/Belgian cohorts (n=2409)** | | | | | |
|---|---|---|---|---|---|
| | **Class 1** | **Class 2** | **Class 3** | **Class 4** | **Class 5** |
| | **n= 648 (27%)** | **n= 981 (41%)** | **n= 283 (12%)** | **n= 267 (11%)** | **n= 230 (10%)** |
| ***Male gender*** | 81 % | 76 % | 83 % | 78 % | 76 % |
| ***Age, years*** | 75 [70; 80] | 65 [58; 73] | 74 [66; 79] | 66 [59; 74] | 61 [57; 68] |
| ***BMI*** | 25 [22; 29] | 24 [21; 28] | 30 [25; 33] | 21 [19; 25] | 25 [23; 28] |
| ***Obesity >30 kg/$m^2$*** | 19 % | 11 % | 46 % | 9 % | 10 % |
| ***$FEV_1$ % predicted*** | 38 [29; 47] | 46 [37; 57] | 60 [54; 71] | 27 [22; 31] | 78 [70; 93] |
| ***GOLD grade*** | | | | | |
| ***GOLD 1*** | 1 % | 4 % | 11 % | 0 % | 47 % |
| ***GOLD 2*** | 18 % | 36 % | 75 % | 0 % | 53 % |
| ***GOLD 3*** | 45 % | 42 % | 12 % | 27 % | 0 % |
| ***GOLD 4*** | 36 % | 17 % | 3 % | 73 % | 0 % |
| ***mMRC dyspnoea scale*** | 3 [3 ; 3] | 2 [1 ; 2] | 2 [1; 2] | 3 [3; 4] | 1 [0; 1] |
| ***Exacerbations/patient/year*** | 2 [1; 3] | 1 [0; 2] | 1 [0; 2] | 2 [1; 3] | 0 [0; 1] |
| ***Hospitalizations/patient/year*** | 1 [0 - 2] | 0 [0 - 1] | 0 [0 - 1] | 1 [0 - 2] | 0 [0 - 0] |
| ***Any cardiovascular comorbidity or diabetes*** | 100 % | 21 % | 100 % | 0 % | 0 % |
| ***3-year mortality*** | 50 % | 23 % | 24 % | 45 % | 3 % |
| ***Age at death, years*** | 78 [73; 83] | 72 [64; 80] | 79 [71; 84] | 71 [65; 76] | 71 [61; 72] |
| | | | | | |
| **3CIA initiative database (n=3651)** | | | | | |
| | **Class 1** | **Class 2** | **Class 3** | **Class 4** | **Class 5** |
| | **n=452 (12%)** | **n= 1614 (44%)** | **n=398 (11%)** | **n=150 (4%)** | **n=1037 (29%)** |
| ***Male gender*** | 66 % | 66 % | 73 % | 71 % | 53 % |
| ***Age, years*** | 71 [63; 76] | 63 [57; 68] | 72 [64; 76] | 64 [58; 71] | 59 [51; 66] |
| ***BMI*** | 26 [23; 30] | 25 [22; 28] | 30 [25; 32] | 24 [21; 27] | 25 [22; 27] |
| ***Obesity >30 kg/$m^2$*** | 24 % | 11 % | 48 % | 7 % | 10 % |
| ***$FEV_1$ % predicted*** | 44 [34; 59] | 53 [42; 66] | 69 [59; 80] | 27 [21; 31] | 80 [70; 93] |
| ***GOLD grade*** | | | | | |
| ***GOLD 1*** | 4 % | 11 % | 26 % | 0 % | 49 % |
| ***GOLD 2*** | 33 % | 48 % | 65 % | 0 % | 51 % |
| ***GOLD 3*** | 44 % | 35 % | 8 % | 35 % | 0 % |
| ***GOLD 4*** | 19 % | 6 % | 1 % | 65 % | 0 % |
| ***mMRC dyspnoea scale*** | 4 [2 ; 4] | 1 [1 ; 2] | 1 [0 ; 2] | 4 [4 ; 4] | 0 [0 ; 1] |
| ***Any cardiovascular comorbidity or diabetes*** | 100 % | 35 % | 100 % | 0 % | 0 % |
| ***3-year mortality*** | 23 % | 11 % | 14 % | 27 % | 4 % |
| ***Age at death, years*** | 76 [70; 81] | 68 [63; 74] | 75 [71; 79] | 71 [62; 78] | 68 [61; 73] |

## LEGENDS OF FIGURES

**Figure 1. Study design.** Patients with COPD recruited in the French/Belgian cohorts were classified into subgroups (phenotypes) based on the results of a cluster analysis of clinical data obtained at inclusion in the cohorts. Next, CARTs were used on the same data to determine the best variables and thresholds necessary for the development of an algorithm for assigning COPD patients to the subgroups identified by cluster analysis in the French/Belgian cohorts. This analysis lead to the development of a simple algorithm for allocating patients with COPD into 5 classes. This algorithm was then tested for external validation using data from the 3CIA initiative database (n=16332). This latter analysis was only possible in patients with available data (n=3651), i.e. with all the variables contained in the algorithm. In each analysis, the clinical relevance of the identified subgroups/classes was established by examining their association with 3-year all-cause mortality.

**Figure 2. Algorithm developed by CART-analysis for the classification of COPD patients.** Application to the French/Belgian and 3CIA cohorts.

**Figure 3. Kaplan-Meier analyses for assessing all-cause mortality at three years.**
**3A:** French/Belgian COPD cohorts according to the five subgroups (phenotypes, Ph) identified by cluster analysis. **3B**: French/Belgian COPD cohorts according to the five classes identified by CART analysis. **3C**: 3CIA COPD cohort according to the five classes identified by the algorithm developed in the French/Belgian cohorts. All analyses, $P<0.0001$ (Log-rank test).

**Figure 4. Relative mortality risks at three years among COPD patients in the French/Belgian COPD cohorts (A) and in the 3CIA initiative (B).** COPD patients were classified into five classes according to the algorithm. Horizontal bars show hazard ratios and

95% confidence intervals of mortality risks between classes. For example, in the French/Belgian COPD cohorts, subjects in class 4 have a 23.2-fold (95% CI 10.2–52.7) increased risk of mortality when compared with subjects in class 5.

**Figure 5. Distribution of airflow limitation severity by GOLD grade at inclusion in the cohorts in patients who died during follow-up. A. French/Belgian cohorts. B. 3CIA initiative**. Data are presented as % of the total number of death in each class. Absolute numbers of deaths (n) in each class are also presented.