Large scale international validation of the ADO index in subjects with COPD: an individual subject data analysis of 10 cohorts

ONLINE SUPPLEMENT

Appendix 1. Details of cohort studies

The **Barmelweid cohort (Switzerland)** enrolled, between 2004 and 2005, patients with moderate to very severe COPD in a secondary care hospital that provides both acute and rehabilitative care.[1] All patients were recruited after they had followed a pulmonary rehabilitation program. The Barmelweid cohort served as derivation cohort for the original ADO index.

The **Basque study (Spain)** is a is a prospective, observational study that included ambulatory patients visited from March 2004 to June 2005 in an outpatient clinic of Hospital Cruces, a university-affiliated, tertiary referral hospital in Bilbao.[2] Patients were followed up for a mean period of 4.5 years. In November 2009, mortality was assessed by exhaustive revision of clinical reports and by telephone contact.

The **Cardiovascular Health Study (CHS, USA)** is a population-based, longitudinal study of coronary heart disease and stroke in adults aged 65 years and older.[3] Eligible participants were sampled from Medicare eligibility lists in four areas in the US. All participants of the CHS were eligible for this analysis.

The **Copenhagen City Heart Study (CCHS, Denmark)** involves the study of an on-going prospective, random and age-stratified cohort of adults recruited in four waves from the general population (1976-1978, 1981-1983, 1991-1994 and 2000-2003).[4] The basis for this paper was all participants included in the second and third wave.

The **Jackson Heart Study (JHS, USA)** is a 12-year single-site observational study initiated in 2000 to investigate the etiology of cardiovascular, renal and respiratory disease in African American adults.[5, 6] All JHS participants were recruited from a tri-county area in central Mississippi, including Hinds, Rankin and Madison counties; participants with standardized spirometry (collected between 2000 and 2004) were eligible for this analysis.

The **Lung Health Study (LHS, USA)** was a multicenter (10 centers) randomized clinical trial carried out from October 1986 to April 1994, aimed to determine whether a program of smoking intervention and use of an inhaled bronchodilator could slow the rate of decline in

pulmonary function over a 5-year follow-up period.[7, 8] At baseline, all patients were active smokers between the ages of 35 and 60 with mild to moderate airflow obstruction defined as an FEV_1/FVC ratio less≤0.7, and an FEV_1 between 50-90% predicted.

The **National Emphysema Treatment Trial (NETT, USA)** was a randomized trial that compared lung volume reduction surgery versus medical care in 1,218 patients with severe COPD who had followed a pulmonary rehabilitation program.[9] In this analysis, we included the entire cohort (n=2,252) of patients who underwent spirometry as part of the eligibility assessment for NETT (1998 to 2002). Thus we included patients that were randomized subsequently (n=1,218) as well as patients with COPD who were not randomized (n=1,034).

The **Phenotype and Course of COPD Study (PAC-COPD, Spain)** is a prospective longitudinal study of 342 COPD patients hospitalized for the first time because of a COPD exacerbation in nine teaching hospitals in Spain between January 2004 and March 2006.[10, 11] They were followed prospectively for cause-specific hospitalizations and all-cause mortality during a 4-years follow-up.

The **PLATINO study (Uruguay)** is a multicenter population based study on the prevalence of COPD carried out in five centers of Latin America from 2002 to 2004 among people aged 40 years and over.[12] A follow-up after five years was completed in one of the five sites of Latin America: Montevideo, Uruguay. A total of 173 Uruguayan subjects with spirometric COPD criteria in 2003 were sought in 2008, performed the same procedures as before, including spirometry pre and post bronchodilator, and are eligible for this analysis.

The **Quality of Life of Chronic Obstructive Pulmonary Disease Study Group (SEPOC, Spain)** conducted a cohort study of 321 male patients with COPD, recruited in 1993–1994 at outpatient respiratory clinics of university hospitals in Barcelona and regional hospitals in nearby cities.[13]

Appendix 2. Methods for collecting and harmonizing candidate predictors of mortality

Smoking status and respiratory symptoms (cough, sputum, wheezing) were obtained from epidemiological questionnaires. Lung function was assessed through spirometry before and after a bronchodilator (BD) administration, following standardized procedures in all cohorts. Only pre-BD spirometry was available for CHS and CCHS. The ratio of forced expiratory volume in one second (FEV_1) to forced vital capacity (FVC) ≤0.70 was used for selecting subjects and to define airways limitation.[14] The measure considered for predicting mortality was the FEV_1, expressed as a percentage of its predicted value using local prediction equations (taking into account the age, height, and sex) for each cohort.

Dyspnea was assessed in most studies by using the Medical Research Council (MRC) questionnaire or the modified MRC questionnaire, yielding a score ranging from 0 to 4. The CCHS included a set of individual questions about dyspnea in daily life that were almost identical to the MRC dyspnea questions and allowed adaptation to the MRC scale. The Barmelweid cohort used the dyspnea domain of the chronic respiratory questionnaire, which was transformed into MRC scores (MRC 0: >6.0; MRC 1: >5.0 and ≤6.0; MRC 2: >4.0 and≤5.0; MRC 3: >2.5 and ≤4.0; MRC 4: ≤2.5). The NETT cohort used the Shortness of Breath questionnaire with scores ranging from 0 to 100, which were transformed into MRC scores (MRC 0: ≤40; MRC 1: >40 and ≤60; MRC 2: >60 and ≤80; MRC 3: >80 and ≤100; MRC 4: >100).

Co-morbidities, including asthma, diabetes, ischemic heart disease, stroke, congestive heart failure, peripheral vascular disease, or hypertension, were defined either as a self-report, a self-report of a doctor-diagnosis or a doctor diagnosis after physical examination and medical charts study.

As in previous analyses[1], we explicitly excluded potential predictors of mortality which were more burdensome to measure such as exercise capacity (e.g. six-minute walk distance test), arterial blood gases or lengthy quality of life questionnaires, since these are unlikely to be available consistently in clinical practice outside academic centers. Other variables such as socioeconomic

status, working status, or drug treatments were not consistently available across cohorts, and not possible to standardize.

Appendix 3. Number and percentage of missing data for each variable and cohort

	Barmelweid cohort *Switzerland, Europe* n=231	Basque study *Spain, Europe* N=106	Cardiovascular Health Study *USA, North America* n=2,619	Copenhagen City Heart Study *Denmark, Europe* n=2,287	Jackson Heart Study *USA, North America* n=419	Lung Health Study *USA, North America* n=5,167	National Emphysema Treatment Trial *USA, North America* n=2,252	PAC-COPD Study *Spain, Europe* n=342	PLATINO study *Uruguay, South America* n=173	SEPOC study *Spain, Europe* n=318
Age	0	0	0	0	0	0	0	0	0	0
Sex	0	0	0	0	0	0	0	0	0	0
Working status	231 (100%)	106 (100%)	2619 (100%)	2287 (100%)	0	0	0	0	0	0
Smoking	43 (18.6%)	0	0	3 (0.1%)	3 (0.7%)	0	0	11 (3.2%)	0	61 (19%)
Body mass index	0	0	8 (0.3%)	12 (0.5)	0	1 (0.02%)	0	0	0	3 (0.9%)
Dyspnea	0	0	411 (15.7%)	6 (0.3%)	0	59 (1.1%)	318 (14.1%)	4 (1.2%)	0	1 (0.3%)
Cough	231 (100%)	106 (100%)	10 (0.4%)	2287 (100%)	1 (0.2%)	1908 (36.9%)	2252 (100%)	4 (1.2%)	0	318 (100%)
Sputum	231 (100%)	106 (100%)	28 (1.1%)	3 (0.1%)	1 (0.2%)	2767 (53.6%)	2252 (100%)	4 (1.2%)	0	318 (100%)
Wheeze	231 (100%)	106 (100%)	438 (16.7%)	2287 (100%)	1 (0.2%)	0	2252 (100%)	6 (1.8%)	0	318 (100%)
Pre-BD FEV_1	231 (100%)	0	302 (11.5%)	10 (0.4%)	419 (100%)	0	0	8 (2.3%)	0	0
Post-BD FEV_1	0	19 (17.9%)	2619 (100%)	2287 (100%)	0	1 (0.02%)	3 (0.1%)	0	0	0
Inhaler steroid use	231 (100%)	3 (2.8%)	3 (0.1%)	2287 (100%)	419 (100%)	5167 (100%)	0	4 (1.2%)	0	318 (100%)
6-min walking distance	0	1 (0.9%)	2619 (100%)	2287 (100%)	419 (100%)	5167 (100%)	272 (12.1%)	33 (9.7%)	173 (100%)	318 (100%)
Asthma*	0	0	19 (0.7%)	113 (4.9%)	1 (0.2%)	0	2252 (100%)	4 (1.2%)	0	0
Diabetes*	0	106 (100%)	2619 (100%)	23 (1%)	35 (8.4%)	5167 (100%)	2252 (100%)	3 (0.9%)	0	0
Cardiovascular disease*,†	0	19 (17.9%)	15 (0.6%)	0	9 (2.2%)	0	0	0	0	0
Death during 3-y follow-up	0	0	0	0	0	0	0	0	0	0

* Co-comorbidities are self-reported, self-report of a doctor diagnosis, or doctor diagnosed (according to medical chart and physical examination) depending on the cohort.

† Cardiovascular disease is defined as at least one of the following: ischemic heart disease, stroke, congestive heart failure, or peripheral vascular disease (no hypertension).

Appendix 4. Handling of missing data

For each variable and per cohort, we determined the extent and pattern(s) of missing predictor and outcome variables. We had no missing data for the outcome (death), age and sex. The extent of missing data was small for most variables (<5%) except for dyspnea (14% in NETT and 16% in CHS cohorts), and smoking status (19% in Barmelweid and 19% in SEPOC cohort). To minimize bias and loss of power, missing data were imputed using multiple imputation (10-fold, "mi impute" command, Stata 11).[15, 16] Imputation techniques are based on the correlation between each variable with missing values and all other variables as estimated from the set of complete subjects. After imputation, all datasets were merged into one individual patient data dataset. All analyses, including figures, considered the (small) variability across the ten imputed datasets, using Rubin's rule.[17, 18]

Appendix 5. Statistical analysis – complete version

The original ADO index, ranging from 0 to 10, combined age, dyspnea and airflow obstruction to predict the 3 years risk of all-cause mortality in COPD patients. It was derived in the Barmelweid cohort and validated in the PAC-COPD cohort.[1] Using the original regression coefficients of the predictors in the ADO index[1], we first assessed its discrimination (area under curve) and calibration (predicted *versus* observed risk) in all subjects except for those included in the original derivation cohort (i.e. the Barmelweid study). We then followed a systematic approach for further updating and external validating the ADO index.[19] Given the characteristics of our original derivation cohort (moderate to severe COPD patients in a specialized secondary care setting[1]) we expected that at least an update of the intercept of the (original) ADO index would be necessary because of the different underlying baseline risks across the international studies. The intercept update was not sufficient to yield better agreement between predicted and observed risks, therefore we conducted a more extensive model revision with re-estimation of regression coefficients using logistic regression analysis with death as the outcome variable and age, dyspnea and FEV_1 as predictors.

We did not use formal sample size calculations because all the cohort studies are ongoing studies. Also, there are no generally accepted approaches to estimate the sample size requirements for derivation and validation studies of risk prediction models. Some have suggested having at least 10 events per candidate variable for the derivation of a model[20,21] and at least 100 events for validation studies.[22] Since many studies to develop and validate prediction models are small a potential solution is to have large scale collaborations as ours to derive stable estimates from regression models that are likely to generalize to other populations.[23] Our sample and the number of events far exceeds all approaches for determining samples sizes and, therefore, is expected to provide estimates that are very robust.

Our aim was to obtain a risk prediction model that would be as widely applicable as possible. To develop and validate the updated ADO models, we split our ten cohorts into two

groups of five cohorts that would represent two large COPD populations that are as diverse in terms of disease severity (GOLD I to IV), settings (general population, primary care and specialized care) and geographical area as possible. A priori, i.e. without conducting any exploratory analyses for how to split up the ten cohorts, we used all subjects from CCHS, LHS, NETT, PLATINO and PAC-COPD as update (or derivation) cohort (n=10,221) and subjects from the Barmelweid study, CHS, Basque Study, JHS and SEPOC as validation cohort (n=3,693). We explicitly did not apply a random split or equivalent cross-validation procedure, as these are rather internal than external validation methods.[24,25] The current international individual patient data analysis allowed for the most optimal method of external validation, i.e. geographical validation across countries, settings and disease severity.[24-27] For obtaining the updated ADO score, we fitted a multivariable logistic regression model with all-cause death at 3-years as outcome and age, dyspnoea, and FEV_1 as predictors. We translated the final statistical model into 15 point scale[28], and obtained its associated risks of 3-year mortality with the aim of making the ADO index simple to use in practice.

To obtain additional information about accuracy of the ADO index, we performed decision curve analysis, which explores whether the overall (net) benefit is positive or negative depending on the balance between benefit and harm.[29,30] The net benefit of different treatment decisions can be obtained for a range of outcome probabilities (e.g., 3-years mortality risk), under the assumption that each threshold of outcome probability (p_t) at which a patient (or his/her physician) would opt for (or recommend) treatment is informative of how the s/he weighs the relative harms of a false-positive and a false-negative classification and its corresponding treatment decision.[29,30] Net benefit is defined as:

$$\text{Net benefit} = \frac{\text{True positive count}}{n} - \frac{\text{False positive count}}{n}\left(\frac{p_t}{(1-p_t)}\right)$$

where the true positive count is the number of subjects correctly classified by a risk tool to be at or above p_t, the false positive count is the number of subjects incorrectly classified to be at or above p_t and who would only experience the harms from treatment, and n is the total number of

subjects. The proportion of false positives is multiplied by a term that reflects the weight that is put on false positive classifications (leading to unnecessary treatment) and on false negative classifications (leading to no treatment of patients who need treatment). For example, if p_t is 50% the threshold term is =1, which means that the benefits and harms from treatment are weighted equally. If the p_t is at 10% one would weigh false negatives (= patients not treated although they would need treatment) to be nine times more important than treating a patient unnecessarily (false positive).

For the current analysis we performed the decision curve analysis with a focus on subjects with COPD at low to moderate risk for 3-year mortality (<20%) where most uncertainty about the benefit harm balance may exist. We plotted the net benefit corresponding to six hypothetical strategies for classifying patients into risk categories: (1) using only age; (2) using only dyspnea; (3) using only FEV1 measurement; (4) using only the ADO index; (5) considering all COPD patients to be above a certain risk threshold; and (6) considering all COPD patients to be below a certain risk threshold . The latter two are reference scenarios in decision analysis not necessarily reflecting clinical practice. Decision curve analysis assumes that subjects would be treated if they are at a certain risk for the outcome (e.g. ≥5% risk of 3-year mortality) and calculates how many unnecessary treatments can be avoided.

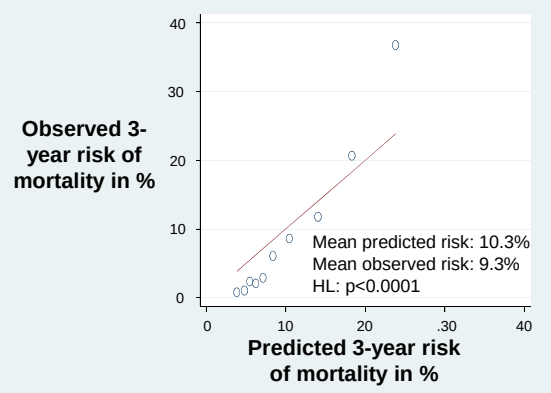
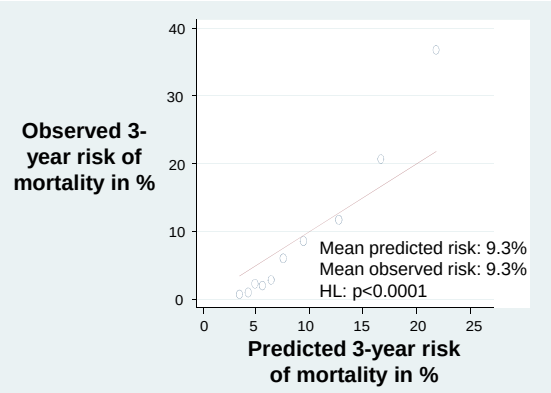
Additionally, we explored whether adding new predictors (e.g. CVD, BMI and sex) improved the updated (refitted) models' discrimination and calibration. For that purpose we repeated all the analyses above to assess if adding CVD, BMI and sex reduced risks on a continuous scale.

We conducted three sensitivity analyses to test how dependent our results were on some decisions taken. First, (sensitivity analysis 1) the analyses were repeated using multilevel (rather than conventional) logistic regression analysis. A random study-effect (related to the 3-year risk of mortality in each original study), and fixed effects for the covariates were used.[31-33] This approach takes into account that subjects within a single cohort are more likely to share some characteristics

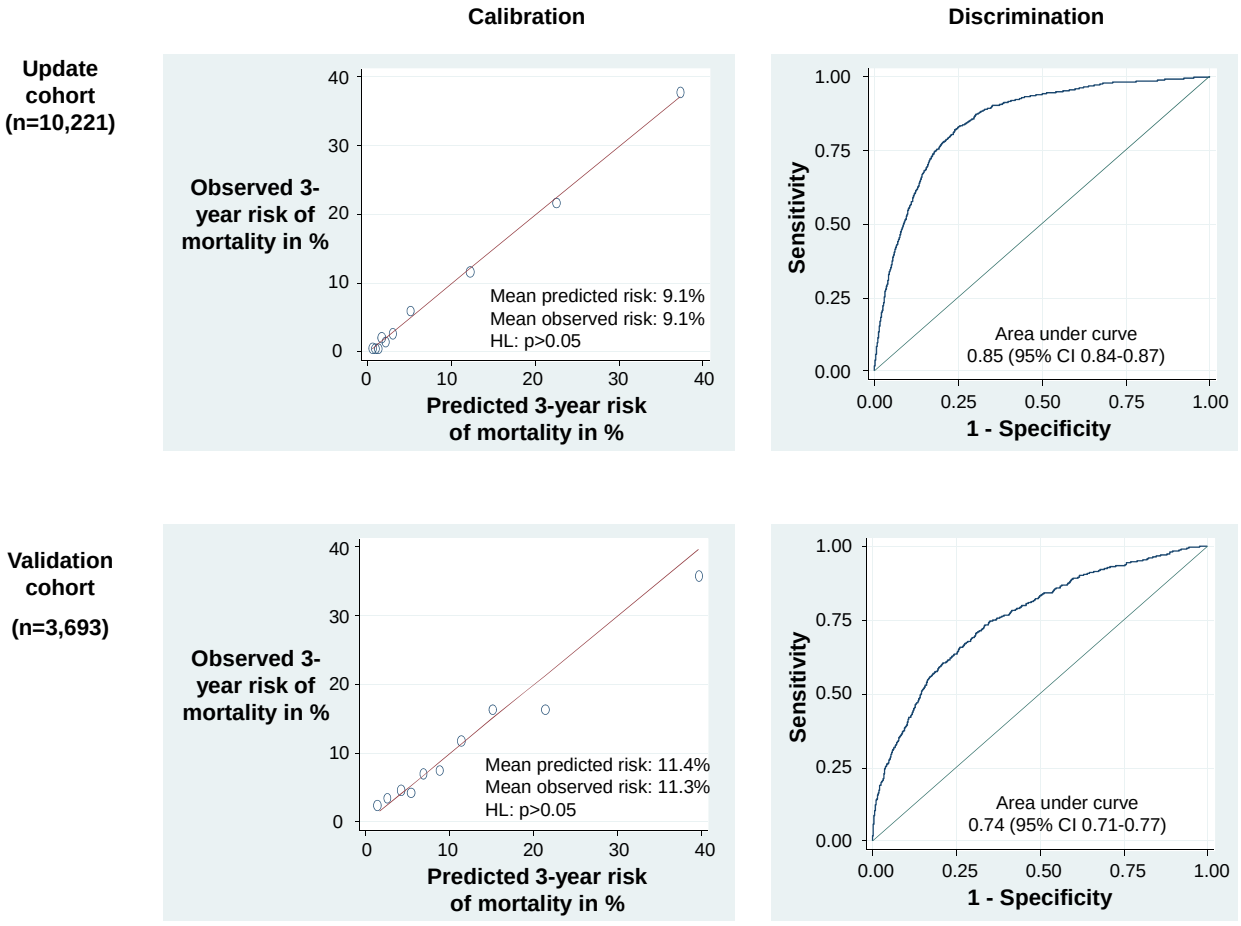
than two randomly chosen subjects from different cohorts. We compared the results from all analyses (discrimination, calibration and clinical usefulness) with those of our simpler models. Similarly, we repeated all analyses excluding subjects with GOLD stage I (sensitivity analysis 2), and excluding subjects with a physician diagnosis of asthma from cohorts where only pre-bronchodilator spirometry was available (sensitivity analysis 3). We also considered restricting analyses to subjects with a FEV_1/FVC ratio below their lower limit of normal level according to local prediction equations, thus taking into account the potential misclassification in older ages when defining COPD according to a fixed FEV_1/FVC ratio, but this analysis was not finally included because misclassification was very low (<1%). All analyses were conducted using Stata for Windows (version 11.1, College Station (TX), USA) and R 2.12 (R Foundation for Statistical Computing, Vienna, Austria, 2011).

Appendix 6. Validation of the original ADO index in 13,683 subjects with COPD

Barmelweid cohort not included because initial development cohort.

Validation cohort (n=13,683)	Calibration	Discrimination
Before re-calibration of intercept		**Area under curve (95% CI)** In all subjects (13,683): 0.82 (0.81-0.83) In derivation cohort (10,221): 0.85 (0.83-0.86) In validation cohort: (3,462) 0.70 (0.67-0.73)
After re-calibration of intercept		HL: Hosmer Lemeshow test statistic

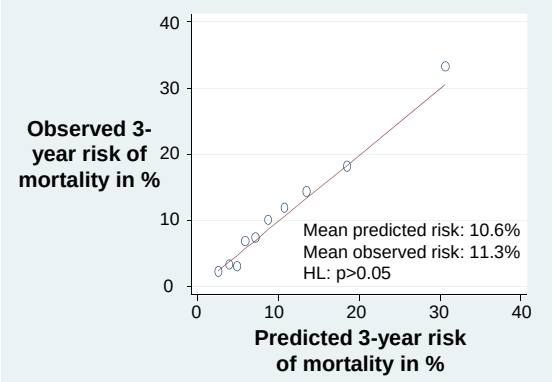
Appendix 7. Derivation and validation of the updated and extended (CVD, BMI and sex) ADO index



Appendix 8. Results of sensitivity analyses

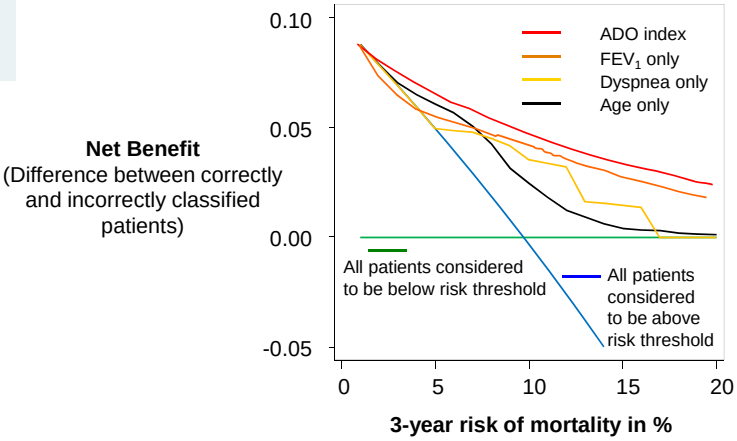
Sensitivity I: Multilevel model. Performance of updated ADO index in validation cohort (all 3,693 subjects included)

Calibration



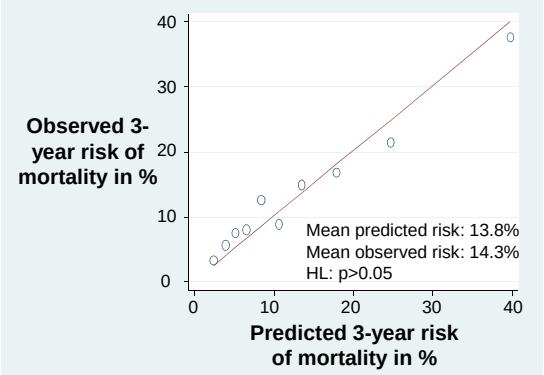
Discrimination

Area under curve
0.73 (0.70-0.75)



Sensitivity II: Performance of updated ADO index in subjects of validation cohort with GOLD stage ≥II (2,101 subjects included, 1,592 subjects with GOLD stage I excluded)

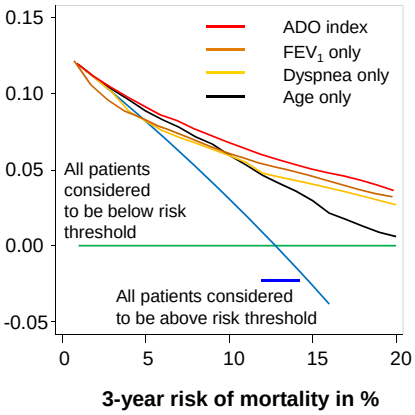
Calibration



Discrimination

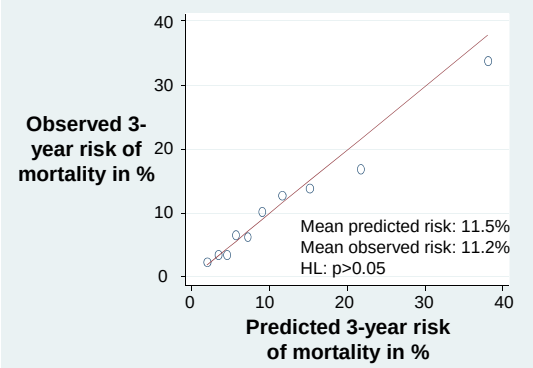
Area under curve
0.70 (0.67-0.73)

Net Benefit
(Difference between correctly and incorrectly classified patients)



Sensitivity III: Performance of updated ADO index in validation cohort after excluding subjects with a physician diagnosis of asthma from cohorts where only pre-bronchodilator spirometry was available (3,545 subjects included, 144 subjects excluded)

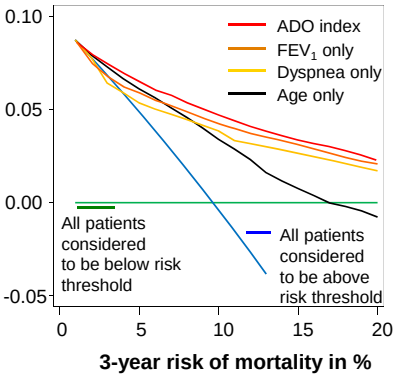
Calibration



Discrimination

Area under curve
0.73 (0.70-0.76)

Net Benefit
(Difference between correctly and incorrectly classified patients)



Appendix 9. Appendix references

1. Puhan MA, Garcia-Aymerich J, Frey M, 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**:704-11.
2. Sobradillo P, Iriberri M, Gomez B, et al. Validation of bode index as a predictor of mortality in COPD patients. In: 18th Annual Congress of the European Respiratory Society. Berlin: European Respiratory Society; 2008:P531.
3. Fried LP, Borhani NO, Enright P, et al. The Cardiovascular Health Study: design and rationale. *Ann Epidemiol* 1991;**1**:263-76.
4. Appleyard M, Hansen A, Schnohr P. The Copenhagen City Heart Study: a book of tables with data from the first examination (1976-78) and a five years follow-up (1981-1983). *Scand J Soc Med* 1989;**170**:1-160.
5. Carpenter MA, Crow R, Steffes M, et al. Laboratory, reading center, and coordinating center data management methods in the Jackson Heart Study. *Am J Med Sci* 2004;**328**:131-44.
6. Taylor HA, Jr., Wilson JG, Jones DW, et al. Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. *Ethn Dis* 2005;**15**:S6-4-17.
7. Anthonisen NR, Skeans MA, Wise RA, et al. The effects of a smoking cessation intervention on 14.5-year mortality: a randomized clinical trial. *Ann Intern Med* 2005;**142**:233-9.
8. Connett JE, Kusek JW, Bailey WC, et al. Design of the Lung Health Study: a randomized clinical trial of early intervention for chronic obstructive pulmonary disease. *Control Clin Trials* 1993;**14**:3S-19S.
9. Fishman A, Martinez F, Naunheim K, et al. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *N Engl J Med* 2003;**348**:2059-73.
10. Garcia-Aymerich J, Gomez FP, Anto JM. Phenotypic Characterization and Course of Chronic Obstructive Pulmonary Disease in the PAC-COPD Study: design and methods. *Arch Bronconeumol* 2009;**45**:4-11.
11. Garcia-Aymerich J, Gomez FP, Benet M, et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax* 2011;**66**:430-7.
12. Menezes AM, Perez-Padilla R, Jardim JR, et al. Chronic obstructive pulmonary disease in five Latin American cities (the PLATINO study): a prevalence study. *Lancet* 2005;**366**:1875-81.

13. Domingo-Salvany A, Lamarca R, Ferrer M, et al. Health-related quality of life and mortality in male patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2002;**166**:680-5.
14. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global Strategy for Diagnosis, Management, and Prevention of COPD. http://wwwgoldcopdcom/GuidelinesResourcesasp?l1=2&l2=0 2010.
15. Moons KG, Donders RA, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006;**59**:1092-101.
16. Steyerberg E. Applications of Prediction Models. In: Clinical Prediction Models - A Practical Approach to Development, Validation, and Updating. New York: Springer; 2010.
17. Donders AR, van der Heijden GJ, Stijnen T, et al. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;**59**:1087-91.
18. Schafer JL. Analysis of Incomplete Multivariate Data. Boca Raton, FL: Chapman & Hall/CRC; 1997.
19. Steyerberg EW, Borsboom GJ, van Houwelingen HC, et al. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;**23**:2567-86.
20. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;**49**:1373-1379.
21. Harrell FE Jr, Lee KL, Califf RM, et al. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984;**3**:143-52.
22. Vergouwe Y, Steyerberg EW, Eijkemans MJ, et al. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005;**58**:475-83.
23. Steyerberg EW: Study design for prediction models. In Clinical prediction models. Volume Chapter 3. New York: Springer; 2008
24. Moons KG, Royston P, Vergouwe Y, et al. Prognosis and prognostic research: what, why, and how? *BMJ* 2009;**338**:b375.
25. Royston P, Moons KG, Altman DG, et al. Prognosis and prognostic research: Developing a prognostic model. *BMJ* 2009;**338**:b604.
26. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;**19**:453-73.
27. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;**130**:515-24.

28. Sullivan LM, Massaro JM, D'Agostino RB, Sr. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med* 2004;**23**:1631-60.
29. Vickers AJ. Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. *Am Stat* 2008;**62**:314-20.
30. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;**26**:565-74.
31. Greenland S. Principles of multilevel modelling. *Int J Epidemiol* 2000;**29**:158-67.
32. Austin PC, Goel V, van Walraven C. An introduction to multilevel regression models. *Can J Public Health* 2001;**92**:150-4.
33. Urbach DR, Austin PC. Conventional models overestimate the statistical significance of volume-outcome associations, compared with multilevel models. *J Clin Epidemiol* 2005;**58**:391-400.