University of Dundee

External validation of ADO, DOSE, COTE and CODEX at predicting death in primary care patients with COPD using standard and machine learning approaches

Morales, Daniel; Flynn, Robert; Zhang, Jianguo; Trucco, Emanuele; Quint, Jennifer K.

Published in:
Respiratory Medicine

DOI:
10.1016/j.rmed.2018.04.003

Publication date:
2018

Licence:
No Licence / Unknown

Document Version
Peer reviewed version

Link to publication in Discovery Research Portal

Citation for published version (APA):
Morales, D., Flynn, R., Zhang, J., Trucco, E., & Quint, J. K. (2018). External validation of ADO, DOSE, COTE and CODEX at predicting death in primary care patients with COPD using standard and machine learning approaches. *Respiratory Medicine*, *138*, 150-155. https://doi.org/10.1016/j.rmed.2018.04.003

General rights
Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from Discovery Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 08. Feb. 2023

External validation of ADO, DOSE, COTE and CODEX at predicting death in primary care patients with COPD using standard and machine learning approaches

Daniel R. Morales, [1]Population Health Sciences Division, [2]Scottish Centre for Respiratory Research, School of Medicine, University of Dundee, UK

Rob Flynn, Medicines Monitoring Unit, School of Medicine, University of Dundee, UK

Jianguo Zhang, School of Science and Engineering (Computing), University of Dundee, UK

Emmanuel Trucco, School of Science and Engineering (Computing), University of Dundee, UK

Jennifer K. Quint, Department of Respiratory Epidemiology, Occupational Medicine and Public Health, National Heart and Lung Institute, Imperial College London, UK

Kris Zutis, Medicines Monitoring Unity, School of Medicine, University of Dundee, UK

CORRESPONDING AUTHOR

Dr Daniel R. Morales. Population Health Sciences Division, Mackenzie Building, Kirsty Semple Way, Dundee, DD2 4BF. EMAIL: d.r.z.morales@dundee.ac.uk TEL: 01382 383475

SHORT TITLE: Predicting the risk of death in people with COPD

WORD COUNT: 2988 **DESCRIPTOR**: COPD Epidemiology

© 2018. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/

ABSTRACT

Background: Several models for predicting the risk of death in people with chronic obstructive pulmonary disease (COPD) exist but have not undergone large scale validation in primary care. The objective of this study was to externally validate these models using statistical and machine learning approaches.

Methods: We used a primary care COPD cohort identified using data from the UK Clinical Practice Research Datalink. Age-standardised mortality rates were calculated for the population by gender and discrimination of ADO (age, dyspnoea, airflow obstruction), COTE (COPD-specific comorbidity test), DOSE (dyspnoea, airflow obstruction, smoking, exacerbations) and CODEX (comorbidity, dyspnoea, airflow obstruction, exacerbations) at predicting death over 1-3 years measured using logistic regression and a support vector machine learning (SVM) method of analysis.

Results: The age-standardised mortality rate was 32.8 (95%CI 32.5-33.1) and 25.2 (95%CI 25.4-25.7) per 1000 person years for men and women respectively. Complete data were available for 54879 patients to predict 1-year mortality. ADO performed the best (c-statistic of 0.730) compared with DOSE (c-statistic 0.645), COTE (c-statistic 0.655) and CODEX (c-statistic 0.649) at predicting 1-year mortality. Discrimination of ADO and DOSE improved discrimination at predicting 1-year mortality when combined with COTE comorbidities (c-statistic 0.780 ADO+COTE; c-statistic 0.727 DOSE+COTE). Discrimination did not change significantly over 1-3 years. Comparable results were observed using SVM.

Conclusion: In primary care, ADO appears superior at predicting death in COPD. Performance of ADO and DOSE improved when combined with COTE comorbidities suggesting better models may be generated with additional data facilitated using novel approaches.

KEYWORDS: COPD, mortality, epidemiology

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is a leading cause of morbidity and mortality worldwide.[1,2] Despite the large excess mortality associated with COPD, few interventions have been shown to reduce overall mortality, which include smoking cessation, long-term oxygen therapy (LTOT) and lung volume reduction surgery in small subsets of COPD patients.[3-6] Predicting death in people with COPD could help to inform patients about their disease course, and help health care professionals decide when to offer or stop therapies including when to consider palliative care. However, current guidelines on the management of COPD provide limited guidance on the prognostic assessment related to these factors, despite increasing attention to the development of risk scores for predicting future health outcomes in people with COPD.[7,8]

Risk scores are used to guide decision-making and often use a very small number of variables to predict the outcome of interest, with a trade-off between predictive performance and being user-friendly. Variables are often identified through a selection process involving modelling large amounts of data to find combinations of predictive factors with high sensitivity and specificity. Novel machine learning approaches offer several potential advantages over standard regression for risk prediction but still require validation in several applications before being accepted in mainstream clinical research.[9,10]

Several risk scores have been developed attempting to predict death in people with COPD, including ADO (age, dyspnoea and airflow obstruction), BODE (body mass index, airflow obstruction, dyspnoea and exercise capacity), COTE (COPD specific comorbidity test) and DOSE (dyspnoea, obstruction, smoking and exacerbations).[11-14] These risk scores have predominantly been developed in selected populations, often from secondary care, who may differ in terms of prognosis to those in primary care. Such characteristics include the prevalence and severity of comorbidities, which may differ between primary and secondary care COPD populations. It is therefore uncertain how generalizable the predictive performance of these risk scores are to people in primary care. The performance of ADO at predicting two year mortality has been evaluated in small population samples from primary care and have reported promising results.[15] However, larger cohorts are required to properly evaluate the predictive performance of these risk scores in primary care patients.[16]

The aim of this study was to evaluate several existing risk scores for mortality in COPD over multiple time periods with a large representative COPD population from primary care, using both traditional and a machine learning approach.

MATERIAL AND METHODS

Data source

Data from the UK Clinical Practice Research Datalink (CPRD) were used to validate several clinical prediction rules assessing mortality in people with COPD. CPRD is a large source of longitudinal UK electronic primary care medical records that has been widely used for research. It contains anonymised information on symptoms, diagnoses, tests, referrals to secondary care and death for over 13 million people, of whom 4.4 million are currently registered with a practice that is contributing data to CPRD, representing about 9% of the UK population. Data within CPRD are predominantly recorded using Read Codes, a hierarchical thesaurus of coded clinical terms used in UK primary care. General practices and patients within CPRD are required to meet defined quality standards in order to contribute data. Around 60% of the patients included in the CPRD have been linked to ONS, an administrative database containing information on all death registrations in England. Diagnoses in ONS are recorded using International Classification of Disease version 10 (ICD-10) codes.

Study approval

The study was approved by the Independent Scientific Advisory Committee for Medicines and Healthcare products Regulatory Agency database research (protocol number 15_112R).

Population

A cohort of 204473 patients with actively treated COPD present within CPRD between 01.01.2000 to 01.04.2014 was identified using a validated Read code algorithm.[17] All patients were required to have spirometry confirmed COPD with a forced expiratory volume in 1 second (FEV1)/forced vital capacity (FVC) ratio of <0.7 and be acceptable for use in research (a marker of internal quality control by CPRD). Cohort entry was defined as the latest date of the following: having been registered with a general practice providing up-to-standard data for at least 1 year prior to cohort entry; issued ≥1 COPD-related medication (defined by short or long-acting beta2-agonists, short or long-acting muscarinic antagonist, inhaled steroids, methylxanthines); and date of incident COPD Read code diagnosis. Cohort exit was defined by the earliest of: death, transfer out from the general practice, date of last data collection, or end of the study period.

Clinical prediction rules

A cross-section of the cohort was taken at 01/04/2011 for validating the clinical prediction rules. This date was selected based upon optimally recorded data in primary care whilst allowing sufficient time for follow-up. Three clinical prediction rules for predicting mortality in people with COPD were chosen for validation in primary care, namely the ADO (including the updated ADO index score which differs by allocating scores at different cutpoints of the three variables)[18], COTE and DOSE index scores. A fourth, CODEX (consisting of comorbidity, dyspnoea, airflow obstruction and the number of severe exacerbations in the previous year, which was derived to assess short- and medium-term prognosis in hospitalized COPD patients) was additionally evaluated as a post-hoc analysis.[19] BODE was not chosen because it requires data on a 6-minute walk test not routinely measured in primary care.

Percentage predicted FEV1 was either identified from the medical record or calculated based upon lung function data contained within electronic medical records.[20] COTE is a comorbidity index consisting of 13 comorbidities that potentially improve the ability to predict death in people with COPD.[13] The DOSE index score requires data on the number of acute exacerbations in the previous 12 months. Acute exacerbations were defined by 1) Read codes for acute exacerbations, and 2) Read codes for respiratory tract infections or respiratory symptoms combined with same-day oral corticosteroid prescriptions.[20] For each of these models, risk scores have been created attempting to stratify people at increased risk of mortality (supplementary table S1 and S2). For CODEX, the Charlson comorbidity score was calculated consisting of 15 comorbidities.[19]

Statistical analysis

The incidence of death in people COPD was calculated over the study period directly age-standardised to the 2013 European Standard Population. The numerator consisted of the number of deaths and the denominator the number of person years. The performance of the ADO, DOSE and COTE was validated using multivariable logistic regression, generating odds ratios and assessing each models discrimination by producing area under the receiver operator characteristic (AUC) curves and reporting the c-statistic. This was done first for each corresponding risk score and second with each of the individual variables within each clinical prediction rule. Predictive performance was assessed over a one, two and three year follow-up period, which builds on the scope of existing studies.

Sensitivity analysis

Logistic regression analysis was repeated for each of the risk scores using only percentage predicted FEV1 and MRC dyspnoea score recorded in the previous year vs. ever recorded to ensure that the timing of recording had no impact on the results of risk prediction models.

Machine learning analysis

For each risk score, results from logistic regression modelling were compared with those from a support vector machine (SVM) algorithm that provided a binary output as used in logistic regression. The LibSVM implementation of support vector machines was used.[22] All patients were grouped by their GP practice resulting in a total of 618 practices in the cohort. The data was then randomly split into two sets; a training/testing data set consisting of 80% (494) of practices, and a validation data set consisting of the remaining 20% (124) of practices. The training/testing data set was used to train the models and optimal parameter estimation. The optimal values for parameters Cost and Gamma were found by a grid search, evaluated as the optimal reported AUC curves and c-statistic, using a 10-fold cross-validation approach. The optimal kernel type was also evaluated by comparing results from Linear, Polynomial, and Radial Basis Function (RBF) kernels, with the RBF kernel performing the best. Class weighting was used to improve prediction accuracy and 95% confidence intervals were generated through bootstrapping.

RESULTS

A total of 204473 patients with COPD were identified in CPRD. The cohort contained 1.5 million person-years follow-up (mean 7.1 years per patient) during which 65878 patients died during follow-up (32.3%). The age-standardised mortality rate in people aged 40 and over was 32.8 (95%CI 32.5-33.1) per 1000 person-years for men and 25.2 (95%CI 25.4-25.7) per 1000 person years for women for the cohort. The trend in age-standardised mortality is shown in figure 1. Mortality rates in women appeared fairly level throughout the study period whilst mortality rates for older men showed a downward trend from 2006.

Validation of ADO, DOSE, COTE and CODEX in people with COPD in primary care

At 01/04/2011, 70312 eligible patients alive and registered with a GP practice were identified, of which 54879 patients (78.0%), 53707 patients (76.2%), and 52684 (75.0%) patients had complete data for analysis over a 1, 2 and 3-year period respectively. Characteristics of patients present in the cohort for the one year period as of 01/04/2011 are shown in table 1. A total of 2811 (5.1%) patients had died at 1 year, 5652 (10.5%) had died at 2 years and 8083 (15.3%) had died at 3 years. Odds ratios for each clinical prediction rule over a 1, 2 and 3-year period are shown in table 2 and supplementary tables S3 and S4 respectively. An incremental increase in the size of the odds ratio per unit increase in score was observed for ADO, DOSE and CODEX index scores but not for the COTE index score.

The discriminative performance of ADO, DOSE, CODEX and COTE for predicting the risk of death in people with COPD over a 1, 2 and 3-year period in primary care using a statistical approach are presented in table 3. The ADO index score had the best discrimination with a 1 year c-statistic of 0.730 compared with the DOSE index score (c-statistic 0.645). The updated ADO index score did not perform better than the original ADO index score. Discrimination improved slightly when actual clinical variables were modelled instead of the respective index scores (supplementary table S5). In general, discrimination did not change significantly over 1, 2 or 3-years. Discrimination of COTE modelled using only comorbidities was similar to DOSE (c-statistic 0.655 for 1 year, 0.657 for 2 year and 0.674 for 3-year models) and to CODEX (c-statistic 0.649). Discrimination of both ADO and DOSE significantly improved when modelled in combination with comorbidities used to calculate the COTE index score (c-statistic 0.730 for 1 year ADO index without COTE vs. 0.780 for 1 year ADO index with COTE, c-statistic 0.645 for 1 year DOSE index without COTE vs. 0.727 for 1 year DOSE index with COTE).

Comparative performance of support vector machine learning methodology

The discriminative performance of ADO, DOSE, CODEX and COTE for predicting the risk of death in people with COPD over a 1, 2 and 3-year period in primary care, using a SVM method of analysis is presented in table 3. Similar to the statistical approach, the ADO index score performed the best with a 1 year c-statistic of 0.723 compared with 1 year DOSE index score c-statistic of 0.654, COTE index score of 0.650 and CODEX index score of 0.651. Discrimination again improved slightly when actual features were modelled instead of index scores (supplementary table S5). Similarly, discriminative performance did not change significantly when modelled over 1, 2 or 3-year period. Predictive performance of the ADO and DOSE index scores again improved with the addition of COTE comorbidities. Overall discrimination of SVM using the limited feature set contained within ADO, DOSE, CODEX and COTE was comparable to results obtained from logistic regression.

DISCUSSION

Using electronic medical record data from primary care we were able to perform the largest external validation to date for several well-recognised clinical prediction rules derived to predict the risk of death in people with COPD. This study demonstrated that the ADO index score performed the best compared to the DOSE, COTE and CODEX index scores, and that the performance of ADO and DOSE significantly improved when COTE comorbidities were incorporated within each model. In this regard, a model with a c-statistic above 0.70 is generally considered as good and 0.80 and above is considered excellent. Although the addition of the 12 COTE comorbidities significantly improved the performance of both the ADO and DOSE index scores, model performance of COTE comorbidities alone was relatively poor and similar to that of CODEX. For all clinical prediction rules, model performance were similar whether predicting death over a 1, 2 or 3-year time period although the number of patients who died in absolute terms differed.

In a small primary care population consisting of 646 selected patients from primary care, the ADO index was shown to predict 2-year mortality with a c-statistic of 0.78 (95% CI 0.71–0.84), whilst in a larger multicentre study with many secondary care patients discrimination of ADO was substantially lower when externally validated with a c-statistic 0.70 (95%CI 0.67-0.73).[15, 18] However, ADO could be useful at discriminating survival in patients receiving

LTOT as an ADO score greater than 5 has been shown to have a 20% chance of survival at 4 years compared to 65% of people with an ADO score of 2 or 3.[23]

Although a higher DOSE index score is associated with greater risk of mortality, our study found that it consistently performed less well than ADO at predicting 1, 2 and 3-year mortality in primary care COPD patients.[24] Whilst ADO seems a better predictor of mortality than DOSE, DOSE may be a better predictor of exacerbations and hospitalisation, probably because the number of exacerbations in the previous 12 months is a component of DOSE and in many cases past events are usually highly correlated with future events.[13]

Comorbidities in the COTE index score were derived from 1664 COPD secondary care patients and had relatively poor discrimination at predicting mortality when derived with an c-statistic of only 0.66, similar to our findings.[14] Although the discriminative performance of COTE appears limited when used alone, discrimination significantly improved when combined with variables from ADO and DOSE. A similar observation occurred with CODEX which had better discrimination than the ADO and DOSE index scores for predicting 1-year mortality in COPD patients following hospital discharge, although its performance was poorer than the combination of ADO or DOSE with COTE comorbidities n our primary care population and similar to the predictive performance of 1 year mortality in the original secondary care derivation cohort (c-statistic 0.67).[19] However, the problem with many of these studies is that they contain heterogeneous, partially selected populations recruited at different stages of disease, with some cohorts at risk of bias from loss to follow-up. Heterogeneity is also seen in the time frame chosen to predict mortality, all of which make it difficult to compare models, which our study partly attempts to overcome by using a large representative real world population.

We used only patients with complete data for analysis which is a potential limitation of our study. Although methods such as multiple imputation are available they were not used because it was uncertain whether data were missing at random and the applicability of prediction models in routine care rely on such data actually being recorded. We also have no objective measure for the quality of pulmonary function testing which may vary in primary care, although evidence suggests that quality of spirometry is high with a previous validation study showing that 98.6% of spirometry traces in primary care were of adequate quality.[20] Although we use a validated approach to identifying exacerbations it is uncertain whether all were captured which could influence the performance of the DOSE index score. The accuracy of models incorporating multiple comorbidities will be influenced by the quality of diagnostic recording which may vary between data sources. However, the validity of many diagnoses within CPRD is considered high.[25-27]

The potential utility of using prediction models to drive clinical care is appealing. Prediction models such as ADO (e.g. score of 5 or more) may be useful in targeting high risk individual for interventions and trials of treatment. In the era of 'big data', machine learning approaches offer potential advantages in being able to process large amounts of clinical data such as those contained with electronic medical records. However, machine learning approaches have infrequently been applied to such data and therefore require to be validated before widespread adoption. This seems to be particularly important as our study has shown that predictive performance can be improved by incorporating more clinical information, in this example using COTE comorbidities. However, incorporating larger amounts of data may be infeasible to interpret through human factors alone. Using a limited feature set, SVM performs as well as standard logistic regression helping to validate this approach which could now be applied to a large data set that includes far more clinical data (such as blood results, prescriptions, pattern of health care access and social care data). These results suggest that other machine learning approaches such as large-scale deep learning models could also be evaluated to harness the full potential of very large data repositories. The challenge now is to create better clinical prediction models to stratify patients at greatest risk and identify ways of integrating them into routine clinical practice. One such way would be to integrate such models at annual COPD reviews that are frequently part of routine care to help target interventions and provide prognostic information to clinicians and patients including planning of palliative care services by identifying people at high risk of dying within the next year. It would also be important for helping to stratify or withdraw certain treatments helping to rationalise resources or identify vulnerable groups that should receive better access to care.

CONCLUSIONS

The ADO clinical prediction rule combined with COTE comorbidities performs relatively well at predicting the risk of death in people with COPD in primary care using data from electronic medical records. A SVM method of analysis seems a valid alternative to risk prediction modelling using such data and has potential to be used more widely with larger more complicated data sets.

ACKNOWLEDGEMENTS

We would like to acknowledge all general practices that have contributed anonymised electronic health care data to make such studies possible.

CONFLICTS OF INTEREST

Dr Quint's research group has received funding from MRC, Wellcome Trust, BLF, GSK, Insmed and AZ for other projects, none of which relate to this work. Dr Quint has received funds from Az, GSK, Chiesi and BI for Advisory board participation or travel.

CONTRIBUTORSHIP

DM conceived the work. DM, KZ, RF analysed the data. All authors contributed to: the design, analysis or interpretation of data for the work; and drafting the work or revising it critically for important intellectual content; and final approval of the version submitted for publication.

FUNDING

The study was funded by a Scottish Government Chief Scientist Office Health Informatics Challenge grant (HICG.1.30). DRM is part funded by an Academic Health Sciences Fellowship in Tayside, Scotland.

REFERENCES

1. Groenewegen KH, Schols AM, Wouters EF. Mortality and mortality related factors after hospitalization for acute exacerbation of COPD. Chest 2003;124:459–467.
2. Murray CJ, Lopez AD. Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. Lancet 1997; 349:1498–1504.
3. 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.
4. Medical Research Council Working Party. Long term domiciliary oxygen therapy in chronic hypoxic cor pulmonale complicating chronic bronchitis and emphysema. Lancet. 1981;1:681–6.
5. Nocturnal Oxygen Therapy Trial Group. Continuous or nocturnal oxygen therapy in hypoxemic chronic obstructive lung disease: a clinical trial. Ann Intern Med. 1980;93:391–8.
6. 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.
7. National Institute for Health and Clinical Excellence (2010) Chronic obstructive pulmonary disease in over 16s: diagnosis and management. NICE guideline (CG101)
8. The Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2017. Available from: http://goldcopd.org.
9. Alpaydin E. Introduction into machine learning. Second edition. 2010 The MIT Press Cambridge, Massachusetts, London, England.
10. Magoulas G.D., Prentza A. Machine Learning in Medical Applications. Lecture Notes in Computer Science 1999:2049;300-07.

11. Puhan MA, Garcia-Aymerich J, Frey M, ter Riet G, Anto JM, Agusti AG, Gomez FP, Rodriguez-Roisin R, Moons KG, Kessels AG, Held U: Expansion of the prognostic assessment of patients with chronic obstructive pulmonary disease: the updated BODE index and the ADO index. Lancet 2009; 374: 704–711.

12. Celli BR, Cote CG, Marin JM, Casanova C, Montes de Oca M, Mendez RA, Pinto Plata V, Cabral HJ: The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. N Engl J Med 2004; 350: 1005–1012.

13. Jones RC, Donaldson GC, Chavannes NH, Kida K, Dickson-Spillmann M, Harding S, Wedzicha JA, Price D, Hyland ME: Derivation and validation of a composite index of severity in chronic obstructive pulmonary disease: the DOSE Index. Am J Respir Crit Care Med 2009; 180: 1189–1195.

14. Divo M, Cote C, de Torres JP, Casanova C, Marin JM, Pinto-Plata V, Zulueta J, Cabrera C, Zagaceta J, Hunninghake G, Celli B; BODE Collaborative Group. Comorbidities and risk of mortality in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2012 Jul 15;186(2):155-61.

15. Abu Hussein N, Ter Riet G, Schoenenberger L, Bridevaux PO, Chhajed PN, Fitting JW, Geiser T, Jochmann A, Joos Zellweger L, Kohler M, Maier S, Miedinger D, Schafroth Török S, Scherr A, Siebeling L, Thurnheer R, Tamm M, Puhan MA, Leuppi JD. The ADO Index as a Predictor of Two-Year Mortality in General Practice-Based Chronic Obstructive Pulmonary Disease Cohorts. Respiration. 2014;88(3):208-14.

16. Espantoso-Romero M, Román Rodríguez M, Duarte-Pérez A, Gonzálvez-Rey J, Callejas-Cabanillas PA, Lazic DK, Anta-Agudo B, Torán Monserrat P, Magallon-Botaya R, Gerasimovska Kitanovska B, Lingner H, Assenova RS, Iftode C, Gude-Sampedro F, Clavería A; PROEPOC/COPD study group.. External validation of multidimensional prognostic indices (ADO, BODEx and DOSE) in a primary care international cohort (PROEPOC/COPD cohort). BMC Pulm Med. 2016;16:143.

17. Quint JK, Müllerova H, DiSantostefano RL, Forbes H, Eaton S, Hurst JR, Davis K, Smeeth L. Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). BMJ Open. 2014 23;4:e005540.

18. Puhan MA, Hansel NN, Sobradillo P, Enright P, Lange P, Hickson D, Menezes AM, ter Riet G, Held U, Domingo-Salvany A, Mosenifar Z, Antó JM, Moons KG, Kessels A, Garcia-Aymerich J; International COPD Cohorts Collaboration Working Group. Large-scale international validation of the ADO index in subjects with COPD: an individual subject data analysis of 10 cohorts. BMJ Open. 2012;2(6).

19. Almagro P, Soriano JB, Cabrera FJ, Boixeda R, Alonso-Ortiz MB, Barreiro B, Diez-Manglano J, Murio C, Heredia JL; Working Group on COPD, Spanish Society of Internal Medicine.. Short- and medium-term prognosis in patients hospitalized for COPD exacerbation: the CODEX index. Chest. 2014 May;145(5):972-80.

20. Rothnie KJ, Chandan JS, Goss HG, Müllerová H, Quint JK. Validity and interpretation of spirometric recordings to diagnose COPD in UK primary care. Int J Chron Obstruct Pulmon Dis. 2017;12:1663-1668.

21. Rothnie KJ, Müllerová H, Hurst JR, Smeeth L, Davis K, Thomas SL, Quint JK. Validation of the Recording of Acute Exacerbations of COPD in UK Primary Care Electronic Healthcare Records. PLoS One. 2016;11:e0151357.

22. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011;2(3):27

23. Law S, Boyd S, Macdonald J, Raeside D, Anderson D. Predictors of survival in patients with chronic obstructive pulmonary disease receiving long-term oxygen therapy. BMJ Support Palliat Care. 2014 Mar 25. doi:10.1136/bmjspcare-2012-000432.

24. Sundh J, Ställberg B, Lisspers K, Montgomery SM, Janson C. Co-morbidity, body mass index and quality of life in COPD using the Clinical COPD Questionnaire. COPD. 2011 Jun;8(3):173-81.

25. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol. 2010;69:4-14.

26. Dregan A, Moller H, Murray-Thomas T, Gulliford MC. Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study. Cancer Epidemiol. 2012;36:425-9.

27. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. Br J Gen Pract. 2010;60:e128-36.

TABLES

Table 1. Characteristics of eligible COPD patients alive and registered with a general practice during a cross-section of the cohort on 01/04/2011 for predicting 1-year mortality.

Variable	**1 year COPD cross-section**
Number of patients	54879
Age, ± SD	74.1 ± 10.3
Female, %	46.2
FEV1 % predicted, ± SD	59.5 ± 20.4
MRC dyspnoea score, %	
▪ 0	16.0
▪ 1	38.2
▪ 2	26.7
▪ 3	16.0
▪ 4	3.1
Exacerbation in the last year, %	
▪ 0-1	84.0
▪ 2	7.5
▪ 3	3.7
▪ >3	4.8
Smoking status, %	
▪ Non-smoker	68.0
▪ Smoker	32.0
COTE comorbidities, %	
▪ lung cancer	2.8
▪ oesophageal cancer	0.3
▪ pancreatic cancer	0.2
▪ breast cancer	1.0
▪ other cancers	
▪ anxiety	12.9
▪ liver cirrhosis	0.4
▪ atrial fibrillation/flutter	9.3
▪ diabetes with neuropathy	0.7
▪ pulmonary fibrosis	0.4
▪ congestive heart failure	5.7
▪ gastric/duodenal ulcers	1.4
▪ coronary artery disease	7.6

Table 2. Performance of ADO, DOSE, COTE and CODEX at predicting risk of death in people with COPD from primary care over a 1-year period using logistic regression.

CPR index score	Percent dead at 1 year (%)	Odds ratio	95% confidence interval
ADO			
0-1	0.5	Reference	-
2	1.0	1.97	1.00-3.89
3	1.7	3.37	1.78-6.38
4	3.1	6.23	3.32-11.69
5	5.7	11.61	6.20-21.72
6	7.7	16.12	8.61-30.19
7	12.6	27.93	14.91-52.34
8	18.8	44.72	23.74-84.24
9	28.0	75.13	38.85-145.3
10	39.1	124.07	52.59-292.7
DOSE			
0	2.8	Reference	-
1	3.6	1.32	1.16-1.51
2	5.4	2.00	1.76-2.28
3	6.8	2.57	2.23-2.95
4	10.8	4.26	3.69-4.93
5	12.4	4.97	4.17-5.92
6	17.3	7.36	5.83-9.30
7	20.7	9.22	6.24-13.63
8	21.7	9.80	3.62-26.52
COTE			
0	3.5	Reference	-
1	7.7	2.24	1.93-2.60
2	5.6	1.65	1.48-1.85
3	8.9	2.73	2.29-3.25
4	7.7	2.31	1.88-2.85
5	11.6	3.66	2.68-5.01
6	5.9	1.74	1.53-1.97
7	10.6	3.30	2.54-4.29
8	10.7	3.35	2.84-3.95
9	10.1	3.14	2.21-4.48
≥10	14.6	4.76	3.86-5.88
CODEX			
0	1.8	Reference	-

1	2.4	1.36	1.04-1.79
2	3.5	2.02	1.57-2.62
3	5.1	3.02	2.34-3.89
4	6.5	3.87	3.00-4.98
5	7.5	4.56	3.52-5.90
6	9.5	5.84	4.49-7.60
7	11.7	7.42	5.60-9.82
8	13.8	8.97	6.43-12.51
9	34.3	29.16	13.95-60.95
10	50.0	55.88	7.76-402.5

Index score modelled as categorical variable.

Table 3. Discriminative performance for ADO, DOSE, COTE and CODEX in people with COPD in primary care with logistic regression and support vector machine learning analysis.

Clinical Prediction Rule	1 year c-statistic (95%CI)	2 year c-statistic (95%CI)	3 year c-statistic (95%CI)
Logistic regression			
ADO Index score	0.730 (0.721-0.739)	0.734 (0.728-0.741)	0.732 (0.727-0.738)
Updated ADO Index score	0.720 (0.710-0.729)	0.725 (0.718-0.731)	0.724 (0.719-0.730)
DOSE Index score	0.645 (0.634-0.656)	0.649 (0.642-0.657)	0.645 (0.638-0.651)
COTE Index score only	0.655 (0.644-0.666)	0.657 (0.650-0.665)	0.674 (0.667-0.680)
COTE + ADO index score	0.780 (0.771-0.788)	0.789 (0.783-0.795)	0.799 (0.794-0.804)
COTE + DOSE index score	0.727 (0.717-0.737)	0.739 (0.732-0.746)	0.748 (0.742-0.754)
CODEX Index score*	0.649 (0.639-0.659)	0.656 (0.649-0.663)	0.652 (0.646-0.648)
Support Vector Machine			
ADO Index score	0.723 (0.700-0.741)	0.729 (0.714-0.743)	0.732 (0.721-0.745)
Updated ADO Index score	0.716 (0.696-0.736)	0.722 (0.706-0.736)	0.723 (0.714-0.739)
DOSE Index score	0.654 (0.630-0.678)	0.650 (0.632-0.668)	0.647 (0.631-0.660)
COTE Index score	0.650 (0.627-0.673)	0.651 (0.633-0.668)	0.670 (0.655-0.685)
COTE + ADO index scores	0.778 (0.758-0.794)	0.781 (0.768-0.795)	0.796 (0.785-0.807)
COTE + DOSE index scores	0.736 (0.711-0.757)	0.736 (0.720-0.751)	0.740 (0.726-0.754)
CODEX Index score*	0.651 (0.627-0.671)	0.657 (0.639-0.673)	0.649 (0.635-0.663)

Models with a C-statistic ≥0.7 are generally considered good and those ≥0.8 as excellent.

*Post-hoc analysis.

FIGURE LEGENDS

Figure 1. Trends in all-cause mortality in people with COPD in UK primary care by age category, A) women, B) men.