International Journal of Chronic Obstructive Pulmonary Disease Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Predicting Hospitalization Due to COPD Exacerbations in Swedish Primary Care Patients Using Machine Learning – Based on the ARCTIC Study

This article was published in the following Dove Press journal:
International Journal of Chronic Obstructive Pulmonary Disease

Björn Ställberg[1]
Karin Lisspers[1]
Kjell Larsson[2]
Christer Janson[3]
Mario Müller[4]
Mateusz Łuczko[5]
Bine Kjøller Bjerregaard[6]
Gerald Bacher[7]
Björn Holzhauer[7]
Pankaj Goyal[7]
Gunnar Johansson[1]

[1]Department of Public Health and Caring Sciences, Family Medicine and Preventive Medicine, Uppsala University, Uppsala, Sweden; [2]Integrative Toxicology, Karolinska Institutet, Stockholm, Sweden; [3]Department of Medical Sciences: Respiratory, Allergy and Sleep Research, Uppsala University, Uppsala, Sweden; [4]Department of Data Science and Advanced Analytics, IQVIA, Frankfurt Am Main, Germany; [5]Department of Data Science and Advanced Analytics, IQVIA, Warsaw, Poland; [6]Department of Real World Evidence Solutions, IQVIA, Copenhagen, Denmark; [7]Department of Clinical Development and Analytics, Novartis Pharma AG, Basel, Switzerland

Purpose: Chronic obstructive pulmonary disease (COPD) exacerbations can negatively impact disease severity, progression, mortality and lead to hospitalizations. We aimed to develop a model that predicts a patient's risk of hospitalization due to severe exacerbations (defined as COPD-related hospitalizations) of COPD, using Swedish patient level data.

Patients and Methods: Patient level data for 7823 Swedish patients with COPD was collected from electronic medical records (EMRs) and national registries covering healthcare contacts, diagnoses, prescriptions, lab tests, hospitalizations and socioeconomic factors between 2000 and 2013. Models were created using machine-learning methods to predict risk of imminent exacerbation causing patient hospitalization due to COPD within the next 10 days. Exacerbations occurring within this period were considered as one event. Model performance was assessed using the Area under the Precision-Recall Curve (AUPRC). To compare performance with previous similar studies, the Area Under Receiver Operating Curve (AUROC) was also reported. The model with the highest mean cross validation AUPRC was selected as the final model and was in a final step trained on the entire training dataset.

Results: The most important factors for predicting severe exacerbations were exacerbations in the previous six months and in whole history, number of COPD-related healthcare contacts and comorbidity burden. Validation on test data yielded an AUROC of 0.86 and AUPRC of 0.08, which was high in comparison to previously published attempts to predict COPD exacerbation.

Conclusion: Our work suggests that clinically available information on patient history collected via automated retrieval from EMRs and national registries or directly during patient consultation can form the basis for future clinical tools to predict risk of severe COPD exacerbations.

Keywords: COPD, machine learning, exacerbation, hospitalization

Introduction

Chronic obstructive pulmonary disease (COPD) affects more than 300 million people worldwide and leads to premature mortality.[1] In Sweden, the prevalence of COPD is 8% in the population aged above 50 years and approximately 3000 people die annually due to COPD.[2] Patients with severe COPD account for only 6% of the COPD population, but yet account for 30% of the economic burden in Sweden.[3,4] A large proportion of COPD patients develop exacerbations, periods of enhanced symptoms, often caused by infections which leads to increased costs of hospitalizations.[5]

Correspondence: Björn Ställberg
Department of Public Health and Caring Sciences, Family Medicine and Preventive Medicine, Uppsala University, Box 564, Uppsala, SE-75122, Sweden
Tel +46-070-3149944
Email b.stallberg@telia.com

In Sweden patients with mild to moderate COPD are mostly diagnosed, treated and followed up in primary care, while those with severe disease and severe exacerbations are mainly managed in secondary care.[6,7] The aim of treatment is to improve symptoms and quality of life, prevent exacerbations, improve physical condition and maintain lung capacity, and is tailored to disease and exacerbation severity.[8]

Risk of exacerbation has been associated with history of previous exacerbations[9–12] and comorbidities, specifically ischemic heart disease, heart failure, other respiratory diseases, gastroesophageal reflux and depression/anxiety.[11,13–17] To estimate future risk for exacerbations, prediction models that integrate multiple risk factors can be useful tools in clinical practice, to support early intervention, healthcare resource planning, reducing the burden of inpatient care and improve quality of life for patients. To date, these prediction models have been limited in value and validity. Previous models have been heterogeneous in terms of number and type of predictors, statistical methods and assessment of model performance[9,18] and only a few studies conducted an external[19,20] or internal[21] validation.[22,23] Another recent large Canadian study observed that severe COPD exacerbations can be predicted within two months, by using administrative health data.[24] Notably, the majority of previous investigations had a relatively long follow-up (most commonly one year).

To better understand imminent risk of exacerbations, the present study aimed to develop a prediction model to assess risk of hospitalization within ten days due to COPD exacerbations. Machine learning algorithms were trained and tested on the real-world data collected from Swedish primary and secondary healthcare settings.

Patients and Methods

Study Design

We performed a retrospective, observational cohort study including incident COPD patients (aged ≥40 years) in Sweden between 1st January 2000 and 31st December 2013, using data from the ARCTIC study.[25] The study was designed to develop a prediction model for identification of factors that cause hospitalization due to COPD exacerbations using machine learning algorithms. Data from electronic medical records (EMRs) were collected for patients from 52 primary care centers across Sweden (Figure 1), using an established software system (Pygargus Customized eXtraction Program [CXP] 3.0).[26] The study selection covered a representative sample of the COPD population and healthcare centers by a mix of rural and urban areas with both large and small cities. Patients with COPD are mainly treated and managed in primary care in Sweden.

EMR data were linked by the Swedish National Board of Health and Welfare (SNBHW) to national registers with mandatory reporting. Linkage of individual-level information was enabled across the national registers through unique personal identification numbers (pseudonymized by SNBHW),[26] available for each Swedish resident from birth or immigration. The national registers included:

Longitudinal Integration Database for Health Insurance and Labour Market Studies (LISA), socio-demographic data (educational level, marital status, occupational status, retirement, economic compensation social benefits).[27] The National Patient Register (NPR) including diagnoses according to International Classification of Diseases (ICD) codes, from inpatient and outpatient specialist care. The NPR started in 1964 with complete coverage across all counties in Sweden in 1987, with comprehensive outpatient specialist care information available beginning in 2001.[28]

National Prescribed Drug Register (NPDR), which, since 2005, tracks full details of all dispensed medications (ATC codes).[28] In this study, respiratory medications and medications against comorbidities are grouped (Supplemental Tables 3 and 4) and medications not in the groups are referred to as "other medications".

National Cause of Death Register (NCDR), including information related to date and cause of death (primary and underlying).[28]

Study Cohort and Prediction Period

Patients aged ≥40 years who received a first COPD diagnosis (ICD-10: J44)) regardless of an asthma diagnosis (ICD-10: J45/J46) during the study period 1st January 2000 and 16th December 2013, were included in the study (Figure 2). The index date was defined as the date of first COPD diagnosis during the study period. Patients with <365 days of information in the registers before index date were excluded as were those with incomplete socioeconomic information.

The start of the prediction period was set to June 2006 or later to ensure a minimum of 365 days from the first available prescription data from the NPDR (Figure 3). The prediction period extended until the earliest of the following occurrences: (1) death of a patient; (2) last record in the EMR or NPR; (3) end of the study (31-Dec-2013).

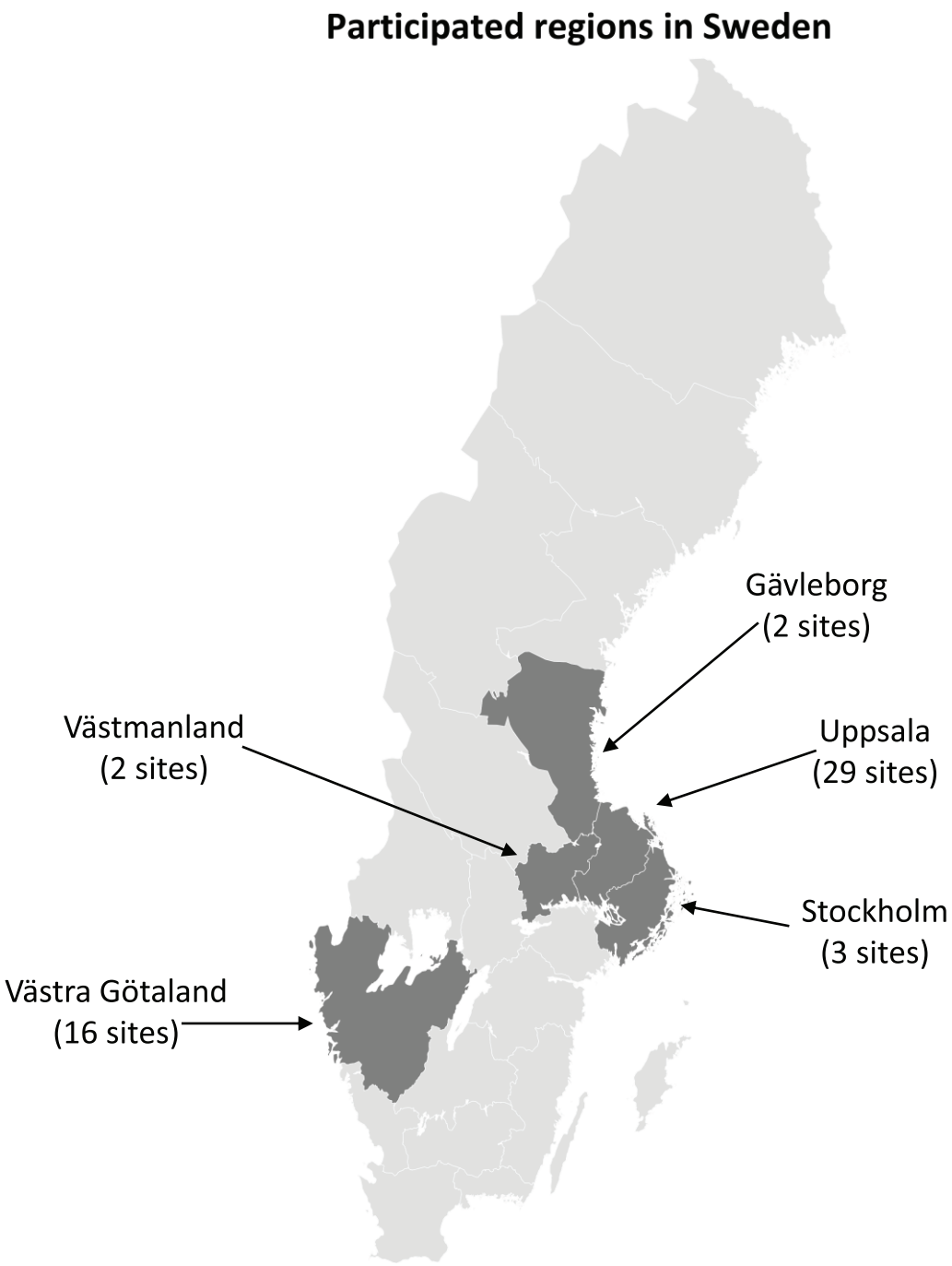

Figure 1 Overview of the included primary care centers (sites) from five regions across Sweden; these sites were from five regions in Sweden; Gävleborg (2 sites), Stockholm (3 sites), Uppsala (29 sites), Västmanland (2 sites) and Västra Götaland (16 sites). In total, 52 primary care centers were included.

Outcome: Exacerbations

The outcome of our study was severe COPD exacerbations that needed hospitalization within a prediction window. Severe exacerbations were defined as a record of a COPD-related hospitalization (ICD-10: J44 as a primary diagnosis or ICD-10: J44.0/J44.1 as a secondary diagnosis) in a secondary care setting. If another severe exacerbation occurred within 10 days of the last visit, it was handled as only one severe exacerbation.

Prediction Variables and Prediction Windows

All available variables (except those in free-text structure) from the ARCTIC dataset were tested as potential predictors

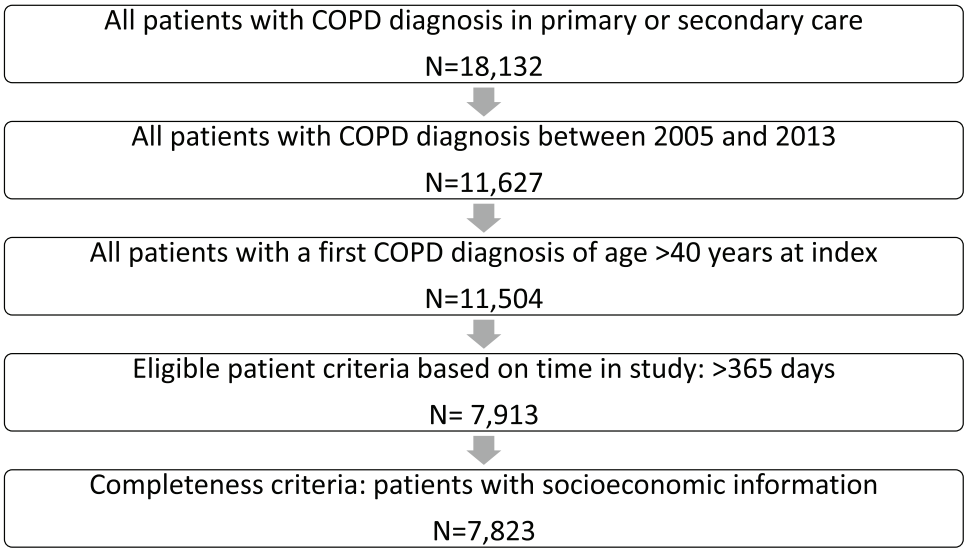

Figure 2 Flowchart represents the patient selection procedure of the study. In the ARCTIC study, there was a total of 18,132 COPD patients. Of these, 7823 patients had their first COPD diagnosis between 2005 and 2013 AND were over 40 years old at index AND were enrolled in the study for more than 365 days AND had available socioeconomic information. These patients fulfilled all inclusion criteria and were eligible for the study.

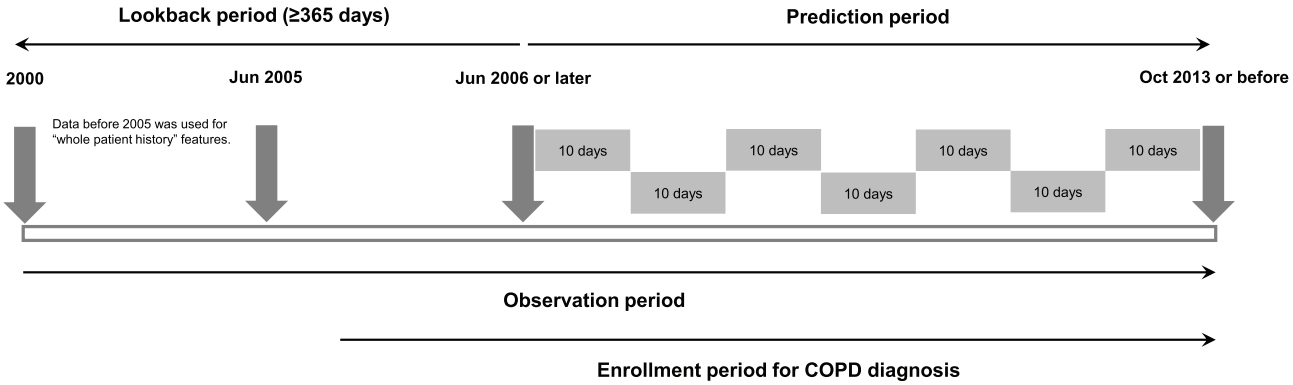

Figure 3 Overall prediction period was defined from June 2006 to October 2013, assuring the patients can be observed across data sources used for the study. Patient specific prediction periods were defined based on events captured in the data sources.

in this study, to find out which variables could be used to predict an exacerbation. Information from free-text case notes was only used to create variables for BMI and smoking status. Charlson Comorbidity Index (CCI) was used as an overall measure of comorbidities.[29] A complete list of predictor variables ie number of contacts with primary and secondary care, number of exacerbations, comorbidities, and medication is provided in Supplemental Tables 1–4 Previous exacerbations as predictor variables included both severe (defined above) and moderate exacerbations defined as a record of a dispensed prescription of systemic corticosteroids (ATC-code: H02AB) and/or respiratory antibiotics (ATC-code: J01AA, J01CA) without a record of a COPD-related hospitalization.

The prediction period was divided into non-overlapping 10-day prediction windows for each patient (Figure 3). Before the start of each 10-day prediction window, lookback periods of 10, 30, 60, 90,180, 365 days or the entire patient history were set up depending on the variable. For details refer to Supplemental Table 1.

Statistical Analysis

Several machine-learning methods were applied and assessed to predict exacerbations of adult COPD patients. These methods included logistic regression with multiple regularization methods (lasso, ridge and elastic net), random forest and gradient boosted trees models (XGBoost). The list covers both simple, classical approaches commonly found in the existing statistical literature (logistic regressions) and state-of-the-art tree-based models that automatically detect non-linear interactions amongst

patient characteristics, therefore being highly relevant for complex medical challenges like predicting short-term exacerbations. To reduce the problem of imbalanced outcome distribution, different class imbalance corrections were tested, and random under-sampling proved to show the best performance in this study. This set of models (along with support vector machines and neural networks, which were not taken into consideration being more challenging to interpret) are considered gold standard for machine learning classification studies done on tabular data. Resampling was applied during cross-validation, making sure that only training folds of each cross-validation iteration are affected, and the effect of resampling is tested on the non-resampled test fold in each cross-validation iteration. Different methods and steps for predicting factor selection were implemented to determine which factors to include, such as prefiltering zero variance and highly correlated predictors ($r > 0.8$) as well as selection based on intermediate, cross validated random forest model of the full predictor set. Model performance was assessed using the Area under the Precision-Recall Curve (AUPRC). Since this metric describes best model performances for the positive class, in this case prediction of exacerbations. Other metrics like AUROC take into account the performance on the negative class, which can be misleading, especially when the negative class represents the majority of the cases (imbalance problem). To compare performance with previous similar studies, the Area Under Receiver Operating Curve (AUROC) was also reported. The model with the highest mean cross validation AUPRC was selected as the final model and was in a final step trained on the entire training dataset.

Models were trained on 75% of the patients (the training set) and tested on the remaining 25% (the test set) as an internal validation of the prognostic model. To account for the problem of repeated measures with correlated patient observations, the data were split into training and test set using a group-based split, so all data for a single patient were either in the training or the test set. To improve robustness and avoid overfitting (lack of generalization) when making modelling choices for each algorithm, we used 4-fold cross-validation on the training data. Imputation of missing values for predictors was done using median for logistic regression models and Random Forest while XGBoost can handle missing values automatically without imputation. Analyses were performed using SAS version 9.4 (SAS Institute) and R version 3.4.4.

Results

Patient Demographics and Exacerbations

In total, 18,132 patients had a reported COPD diagnosis in the EMRs and of these, 7823 patients were eligible for this study (Figure 2). In Table 1, patient demographics are subdivided according to exacerbation history which included the observation period (look-back and prediction period). The mean age at index date was similar for patients without (66.5 [SD 10.1] years) and with a history of exacerbations (66.9 [SD 10.5] years), as were the proportion of women (56% and 58%, respectively). A higher proportion of patients with a history of exacerbations than those without exacerbations had their first COPD diagnosis in secondary outpatient care (35% vs 17%) or primary care setting (34% vs 26%).

Patients with exacerbations had more days per year receiving sick leave benefits, had a lower mean income, were less likely to be working, and had a higher CCI than patients without exacerbations. However, the educational level did not differ substantially between the patient groups (30% with high school 2 years in both groups). Patients with exacerbations seemed to have a higher prevalence of current smoking (19% vs 10%), although information on smoking status were missing for 72% of the patients with no exacerbations and for 63% of the patients with exacerbations.

Model Selection and Performance

Out of the tested models, gradient, gradient boosted trees (XGBoost,[30] (Apache License)) with undersampling was selected for the final model because it had the highest mean cross validation (CV) score for AUPRC of 0.11 (Table 2). AUPRC was 0.17 (95% CI: ± 0.001) for the whole training set and 0.08 (95% CI: ± 0.001) for the testing set and AUROC was 0.88 (95% CI: ± 0.001) on the whole training set and 0.86 (95% CI: ± 0.001) on the testing set. On the testing sets, the recall was 0.16 (95% CI: ± 0.001) and precision was 0.11 (95% CI: ± 0.001).

Prediction Factors

The 20 most important predictors are presented in Table 3. Several prediction factors were related to previous exacerbations, eg, number of severe exacerbations at the different time point (that is, 1–180 days, all time, 1–60 days, 1–180 days year before), and number of moderate exacerbations at 1–30 days. The association between exacerbations and hospitalization is described in Figure 4. A first COPD diagnosis in inpatient care was a stronger predictor of

Table 1 Sociodemographic Characteristics

	No Exacerbation[a]	Exacerbation[a]
	N=5654 (72%)	N=2169 (28%)
Baseline characteristics		
Age at Index (Mean years)	66.5	66.9
Female n (%)	3166 (56)	1,258 (58)
First COPD diagnosis: Inpatient n (%)	3,223 (57)	672 (31)
First COPD diagnosis: Outpatient n (%)	961 (17)	759 (35)
First COPD diagnosis: Primary Care n (%)	1,470 (26)	737 (34)
Socioeconomic characteristics [b]		
Work and Finance		
Number of days per year with sick leave benefits (mean)	7.6	12.1
Number of hours with sick leave benefits per year*100 (mean)	0.08	0.10
Income from social transferals, n (%)	1300 (23)	542 (25)
Income (mean)	445.1	299.8
Income from social security, n (%)	283 (5)	130 (6)
Employment – Working, n (%)	961 (17)	260 (12)
Employment – Not working, n (%)	4354 (77)	1670 (77)
Employment – No Information, n (%)	339 (6)	239 (11)
Health		
CCI (mean)	1.2	2.0
Any sick leave, n (%)	170 (3)	108 (5)
Smoking – No Information, n (%)	4071 (72)	1366 (63)
Smoking – No smoker, n (%)	396 (7)	108 (5)
Smoking – Ex Smoker, n (%)	622 (11)	282 (13)
Smoking – Current smoker, n (%)	565 (10)	412 (19)
Education		
Education – No Information	339 (6)	239 (11)
Education - Primary School < 9 years	1583 (28)	586 (27)
Education - Primary School 9 years	622 (11)	239 (11)
Education – High School 2 years	1696 (30)	651 (30)
Education – High School 12 years	509 (9)	174 (8)
Education – Post High School <3 years	396 (7)	108 (5)
Education – Post High School ≥3 years	396 (7)	108 (5)
Education – Research Education	57 (1)	22 (1)

Notes: [a]Exacerbations occurring during the observation period (look-back period and prediction period). [b]Socioeconomic information retrieved every year during the prediction period.
Abbreviation: CCI, Charlson Comorbidity Index.

Table 2 Model Performance and Best Models by Different Setting

Model Performance	Setting	AUPRC (95% CI)	AUROC (95% CI)	Recall (95% CI)	Precision (95% CI)
XGBoost, undersampling	Trainset	0.17 (0.001)	0.88 (0.001)	–	–
	Test set	0.08 (0.001)	0.86 (0.001)	0.16 (0.001)	0.11 (0.001)
	CVS	0.11	–	–	–

Abbreviations: AUPRC, area under the precision-recall curve; AUROC, area under receiver operating curve; CI, confidence interval; CVS, cross validation score; XGBoost, extreme gradient boosting.

Table 3 Top 20 Most Important Features of Prediction Hospitalization Due to COPD, by Using Machine Learning Models

Rank	Feature	Importance[a]
1	Number of severe exacerbations (last 180 days)	0.33
2	Number of severe exacerbations (whole history) – standardized by the number of days	0.11
3	Number of COPD – related contacts (whole history) – standardized by the number of days	0.066
4	Whether first COPD diagnosis was classified as "inpatient"	0.054
5	Charlson Comorbidity Index (CCI) from the year before the prediction	0.047
6	Number of medications from "other" group[b] (last 365 days)	0.019
7	Whether first COPD diagnosis was classified as "outpatient"	0.016
8	Number of moderate exacerbations (last 30 days)	0.012
9	Number of prescriptions for Antibiotics (whole history) – standardized by the number of days	0.010
10	Number of severe exacerbations (last 180 days, 1 year before prediction date)	0.009
11	Number of COPD – related contacts (last 180 days)	0.008
12	Number of visits (person not defined) in inpatient care (last 180 days)	0.007
13	Number of diagnoses of ischemic heart diseases in inpatient care (whole history) – standardized by the number of days	0.007
14	Number of severe exacerbations (last 60 days)	0.006
15	Number of prescriptions for COPD Medication (last 365 days)	0.006
16	Number of non-COPD – related contacts (whole history) – standardized by the number of days	0.006
17	Number of diagnoses of respiratory disease in inpatient care, all time (whole history) – standardized by the number of days	0.005
18	Number of diagnoses from "other" group in outpatient care (whole history) – standardized by the number of days	0.005
19	Number of diagnoses from "other" group in inpatient care (last 30 days, 1 year before prediction date)	0.005
20	Number of prescriptions for Oral steroids (last 30 days)	0.004

Notes: [a]The value implies the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model. [b]Medications against comorbidities (Supplemental Table 3) and respiratory medications (Supplemental Table 4) and were divided into main groups and sub-groups. Medications not in the groups are referred to as "other medications".

hospitalization compared to first COPD diagnosis in outpatient care. Number of COPD-related healthcare contacts over the entire follow-up time was more important compared with the number of non-COPD-related healthcare visits. Features related to COPD medications, number of other prescriptions 1–365 days and number of prescriptions for antibiotics during patient's whole observation period (look-back and prediction period) were important predictive features. CCI from the year before the prediction was also important.

Discussion

In this observational cohort study, we developed a machine-learning algorithm to understand predictors for severe COPD exacerbations. We discuss model performance and the top-20 factors used by the model, which were the most predictive for hospitalizations due to COPD exacerbation. These factors can be divided into five main groups: history of exacerbations, comorbidities, medications, setting of first COPD diagnosis and healthcare contacts.

Model Performance

The test set AUROC of 0.86 was relatively high,[18–24] but our models may be of limited value for prediction of severe exacerbations due to high false positive rate. Instead of a focus on accuracy and AUROC, the sensitivity, positive predictive value and AUPRC should also be considered. For instance, if our predictive model would be used to assess risk of hospitalization in the next 10 days, for 16% of patients who are hospitalized (sensitivity), 89%

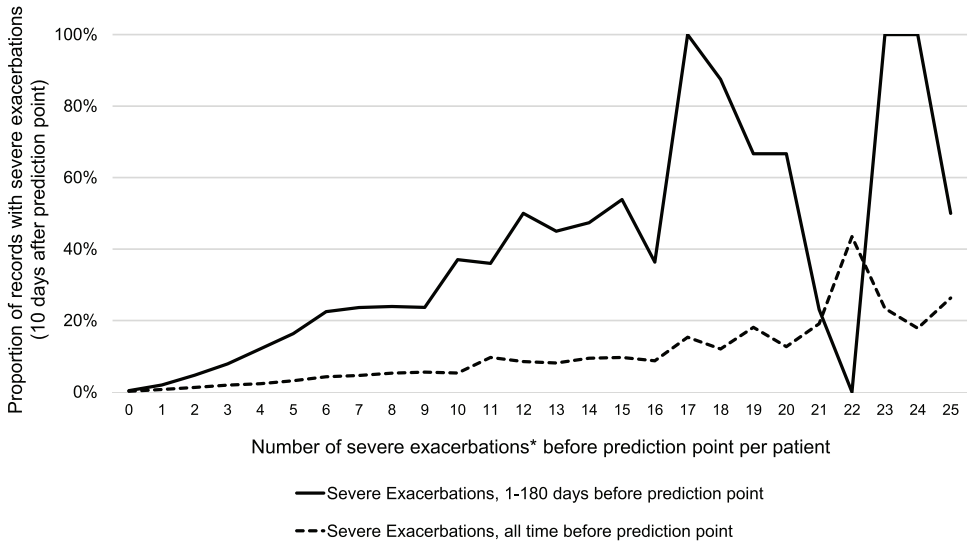

Figure 4 Relationship between the history of severe exacerbations and probability of hospitalization for severe exacerbations, within 1–10 days. The number of previous severe exacerbations, especially within 180 days before prediction point, drastically increases the probability of having a severe exacerbation within the next 10 days. *Severe exacerbations were defined as exacerbation where a hospital stay was required.

of patients who the model assigned a risk of hospitalization would have no hospitalization event in the next 10 days (11%-positive predictive value).

Furthermore, the heterogeneity of previously published models in terms of predictors, statistical methods and assessment of model performance, hampers comparisons with our study.[9,18–24] Regarding time horizons, we could only identify one small study in outpatient care[31] with such a short follow-up as used in our models.

Finally, the large number of predictors might constitute a barrier to using the present models in clinical practice unless the predictors are recorded in EMRs or are substantially simplified.

History of Exacerbations

A history of exacerbations was the most important predictor of future COPD exacerbations. The risk of exacerbation increased if severe exacerbations occurred within the last 180 days and increased with the number of severe exacerbations in the patient's entire health history. Previous moderate exacerbations were also strong predictors of a future hospitalization if they occurred within 1–30 days before the prediction point. Previous studies have also identified prior exacerbations as highly associated with the risk of future exacerbations.[9–12,24,32]

Comorbidities

The CCI (summary measure of number/severity of comorbidities) was a good predictor of hospitalization due to COPD exacerbation. This is consistent with a previous study showing that the more comorbidities patients have, the stronger a predictor it is for future exacerbations and possible hospitalizations.[14] Another observational study of 213 COPD patients showed that 54% of the patients suffered from at least four comorbidities.[13]

The most common comorbidities of future exacerbations that were observed among COPD patients in previous research were ischemic heart disease, heart failure, other respiratory diseases, gastroesophageal reflux disease, CVD and depression/anxiety.[11,14–17,24] This is in line with our findings where ischemic heart disease and respiratory disease other than COPD were strong predictors for COPD hospitalization due to acute exacerbations.

Medications

The number of prescriptions for antibiotics (mostly used for respiratory infections) and other prescriptions, were among the most important predictors, which has also been observed in other studies.[24] Medications (except antibiotics) might indicate indirectly number of comorbidities.

Setting of First COPD Diagnosis

Having a first diagnosis of COPD within secondary care (inpatient or outpatient) was a strong predictor for hospitalization due to COPD exacerbation. This suggests that patients diagnosed with COPD in secondary care may have a more severe or advanced disease at the time of diagnosis and therefore have more frequent and severe

exacerbations. An early diagnosis of COPD could have been overlooked in primary care or patient may not have been to primary care at all.[33] It is reasonable to believe that patients, with a first diagnosis in secondary care, included in our study represent two groups: (1) Patients who visited the hospital with a severe exacerbation and are diagnosed with COPD during this hospitalization (2) Patients who visited the hospital with a referral from primary care and see a pulmonary specialist who then diagnosed the COPD. An assumption can be made that COPD first discovered during an inpatient hospitalization might be more severe than COPD discovered during an outpatient visit – unless the outpatient visit is an emergency where no referral is needed.

Contacts to the Healthcare System

The third most important predictor for hospitalization due to COPD was the number of COPD-related contacts to the healthcare system in the context of the patient's entire medical history. It can be assumed that patients with more severe disease and frequent exacerbations would have more contact with the healthcare system. The contacts included severe exacerbation hospitalizations, but these were few compared to the total number of contacts. Studies have shown that medical comorbidities are common among patients with COPD[7,11,13,17] and our results show that several non-COPD-related contacts was also an important predictor in the present model.

Furthermore, when it becomes possible to remote high-risk monitor ambulatory COPD patients, such input features in the model may further improve performance of the model and may also prevent mild/moderate exacerbations from proceeding to become severe exacerbations requiring hospitalizations.

Strengths

The strengths of our study include the large sample size and long longitudinal follow-up which ensured that key predictors of severe COPD exacerbations were likely to be identified.[24] We had complete and comprehensive longitudinal data on patients, extracted from the EMRs from 52 primary care centers and linked with Swedish national health registers with a mandatory reporting for all healthcare providers. The study used all COPD-related variables such as patient information, comorbidities, medication, laboratory tests and measurements, contact to the healthcare system and seasonal variables. More than 4000 variables were created to build a model to predict which factors in a COPD patient journey in the healthcare system, could be used to predict severe exacerbations requiring hospitalization. The registers used for identification of clinical/epidemiological data are national with a high coverage of most of the conditions included in our study.[34] In addition, the completeness of the variables was high except for smoking and BMI.

Furthermore, the extraction program used for retrieval of information from EMRs has been validated in a specific study, which concluded that it is highly reliable and that appropriate and accurate information is extracted from the EMRs.[26] Finally, our dataset was large, which enabled us to perform an internal validation using a large test set not used for model training. Similarly to a Canadian study[24] we found gradient boosting to be the best performing prediction model. We picked this final model using a robust cross-validation approach, which allowed us to explore the additional value of more recently developed ML compared with more traditional logistic regression approaches. These ML models can capture complex, non-linear relationships, and interactions between predictors.

Limitations

Limitations of the study include potential misreporting of information in the data sources used (EMRs and national health registries). However, for the variables that were collected for the present study, experience from previous research shows that compliance of reporting to EMRs in primary and specialist health care is good and reporting compliance into the national registries used in this study is very high.[35] In addition, most variables are coded according to international classification systems (ICD-10 and ATC codes) limiting the risk of bias due to reporting ambiguities. However, as this study is based on patients with physician diagnosed COPD, it means we miss lung function results from substantial number of patients. Moreover, in order for a model to be functional in a real setting we have included patients with asthma to reflect the reality of patients with COPD as it is common for COPD patients to be diagnosed with asthma as well. Finally, it could be interesting for future studies to investigate a clean COPD population as a sensitivity analysis.

Conclusion

Our work suggests that clinically available information on patient history collected via automated retrieval from EMRs and national registries or directly during patient consultation can form the basis for future clinical tools to predict risk of severe COPD exacerbations.

Abbreviations

ATC, The Anatomical Therapeutic Chemical; AUPRC, Area under the Precision-Recall Curve; AUROC, Area Under Receiver Operating Curve; BMI, Body Mass Index; CCI, Charlson Comorbidity Index; CI, Confidence Interval; COPD, Chronic obstructive pulmonary disease; CVS, Cross Validation Score; CXP, Pygargus Customized eXtraction Program; EMR, Electronic Medical Records; GBM, Gradient Boosting Machines; ICD, International Classification of Diseases; LISA, Longitudinal Integration Database for Health Insurance and Labour Market Studies; NCDR, National Cause of Death Register; NPDR, National Prescribed Drug Register; NPR, The National Patient Register; SNBHW, the Swedish National Board of Health and Welfare; XGBoost, Extreme Gradient Boosting.

Data Sharing Statement

The datasets generated and/or analyzed during the current study are not publicly available due to data privacy rules and the risk of identification of a patient.

Ethics Approval and Informed Consent

This study was approved by the Swedish Ethical Review Authority as an amendment to the approval for the original ARCTIC study, granted by the local Ethical Review Board in Uppsala, Sweden (number: 2014/397, amendment number: 2019-02200).

Acknowledgments

The authors would like to thank Camilla Bengtsson (IQVIA, Stockholm, Sweden) and Pirre Emilia Räisänen (IQVIA, Espoo, Finland) for managing and writing this manuscript.

Author Contributions

B.K.B., M.M. and M. Ł. had full access to the data and conducted the analyses. K.Li., B.S., K.La., C.J. and G. J. have provided clinical expertise to the design and execution of the study. G.B., B. H., and P.G. were jointly involved in conceiving the research study, its design and also in interpretation of analyses for results and reporting purposes. All of the authors have reviewed the draft and the final results of the study and were able to request additional analyses and information as needed. All of the authors have reviewed/provided feedback on the manuscript and have approved the submitted version. All authors have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work. All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in drafting the article or revising it critically for important intellectual content; agreed to submit to the current journal; gave final approval of the version to be published; and agree to be accountable for all aspects of the work.

Funding

This study was funded by the Novartis Pharma AG (Basel, Switzerland).

Disclosure

B.S. has received honoraria for educational activities and lectures from AstraZeneca, Boehringer Ingelheim, Meda, Novartis and Teva, and has served on advisory boards arranged by AstraZeneca, Novartis, Meda, GlaxoSmithKline, Teva and Boehringer Ingelheim, and has participated in the steering committee by Novartis for this study. K.Li. has received payments for educational activities and lectures from AstraZeneca, Chiesi, Novartis and Boehringer Ingelheim, served on advisory boards arranged by Novartis, Boehringer Ingelheim, GlaxoSmithKline and AstraZeneca, and has participated in the steering committee by Novartis for this study. K. La. has, during the last 5 years, on one or more occasion served in an advisory board and/or served as speaker and/or participated in education arranged by AstraZeneca, Boehringer Ingelheim, Chiesi, GlaxoSmithKline, Orion, Mylan, Novartis, and TEVA, and has participated in the steering committee by Novartis for this study. C.J. has received honoraria for educational activities and lectures from Novartis, AstraZeneca, GlaxoSmithKline, TEVA and Boehringer Ingelheim outside the submitted work, and has participated in the steering committee by Novartis for this study. M.M., M. Ł. and B.K.B. are employed by IQVIA. G.B., B. H., and P.G. are employed by Novartis Pharma AG, Basel, Switzerland. G.J. has participated in the steering committee by Novartis for this study and served on advisory boards arranged by AstraZeneca, Novo Nordisk, and Takeda, and has participated in the steering committee by Novartis for this study. The authors report no other conflicts of interest in this work.

References

1. Lopez-Campos JL, Tan W, Soriano JB. Global burden of COPD. *Respirology*. 2016;21(1):14–23. doi:10.1111/resp.12660
2. Socialstyrelsen. *Dödsorsaker 2014*. Sveriges officiella statistik; 2015.
3. Jansson SA, Andersson F, Borg S, Ericsson A, Jonsson E, Lundback B. Costs of COPD in Sweden according to disease severity. *Chest*. 2002;122(6):1994–2002. doi:10.1378/chest.122.6.1994
4. Jansson SA, Backman H, Stenling A, Lindberg A, Ronmark E, Lundback B. Health economic costs of COPD in Sweden by disease severity–has it changed during a ten years period? *Respir Med*. 2013;107(12):1931–1938. doi:10.1016/j.rmed.2013.07.012
5. Pavord ID, Jones PW, Burgel PR, Rabe KF. Exacerbations of COPD. *Int J Chron Obstruct Pulmon Dis*. 2016;11 Spec Iss:21–30. doi:10.2147/COPD.S85978
6. Sundh J, Osterlund Efraimsson E, Janson C, Montgomery S, Stallberg B, Lisspers K. Management of COPD exacerbations in primary care: a clinical cohort study. *Prim Care Respir J*. 2013;22 (4):393–399. doi:10.4104/pcrj.2013.00087
7. Stallberg B, Janson C, Johansson G, et al. Management, morbidity and mortality of COPD during an 11-year period: an observational retrospective epidemiological register study in Sweden (PATHOS). *Prim Care Respir J*. 2014;23(1):38–45. doi:10.4104/pcrj.2013.00106
8. Singh D, Agusti A, Anzueto A, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease: the GOLD science committee report 2019. *Eur Respir J*. 2019;53(5):1900164. doi:10.1183/13993003.00164-2019
9. Guerra B, Gaveikaite V, Bianchi C, Puhan MA. Prediction models for exacerbations in patients with COPD. *Eur Respir Rev*. 2017;26 (143):143. doi:10.1183/16000617.0061-2016
10. Wan ES, DeMeo DL, Hersh CP, et al. Clinical predictors of frequent exacerbations in subjects with severe chronic obstructive pulmonary disease (COPD). *Respir Med*. 2011;105(4):588–594. doi:10.1016/j.rmed.2010.11.015
11. Husebo GR, Bakke PS, Aanerud M, et al. Predictors of exacerbations in chronic obstructive pulmonary disease–results from the Bergen COPD cohort study. *PLoS One*. 2014;9(10):e109721. doi:10.1371/journal.pone.0109721
12. Hurst JR, Vestbo J, Anzueto A, et al. Susceptibility to exacerbation in chronic obstructive pulmonary disease. *N Engl J Med*. 2010;363 (12):1128–1138. doi:10.1056/NEJMoa0909883
13. Negewo NA, Gibson PG, McDonald VM. COPD and its comorbidities: impact, measurement and mechanisms. *Respirology*. 2015;20 (8):1160–1171. doi:10.1111/resp.12642
14. Agusti A, Calverley PM, Decramer M, Stockley RA, Wedzicha JA. Prevention of exacerbations in chronic obstructive pulmonary disease: knowns and unknowns. *Chronic Obstr Pulm Dis*. 2014;1 (2):166–184. doi:10.15326/jcopdf.1.2.2014.0134
15. Kaszuba E, Odeberg H, Rastam L, Halling A. Impact of heart failure and other comorbidities on mortality in patients with chronic obstructive pulmonary disease: a register-based, prospective cohort study. *BMC Fam Pract*. 2018;19(1):178. doi:10.1186/s12875-018-0865-8
16. Westerik JA, Metting EI, van Boven JF, Tiersma W, Kocks JW, Schermer TR. Associations between chronic comorbidity and exacerbation risk in primary care patients with COPD. *Respir Res*. 2017;18 (1):31. doi:10.1186/s12931-017-0512-2
17. Stallberg B, Janson C, Larsson K, et al. Real-world retrospective cohort study ARCTIC shows burden of comorbidities in Swedish COPD versus non-COPD patients. *NPJ Prim Care Respir Med*. 2018;28(1):33. doi:10.1038/s41533-018-0101-y
18. Annavarapu S, Goldfarb S, Gelb M, Moretz C, Renda A, Kaila S. Development and validation of a predictive model to identify patients at risk of severe COPD exacerbations using administrative claims data. *Int J Chron Obstruct Pulmon Dis*. 2018;13:2121–2130. doi:10.2147/COPD.S155773
19. Almagro P, Soriano JB, Cabrera FJ, et al. Short- and medium-term prognosis in patients hospitalized for COPD exacerbation: the CODEX index. *Chest*. 2014;145(5):972–980. doi:10.1378/chest.13-1328
20. Bertens LC, Reitsma JB, Moons KG, et al. Development and validation of a model to predict the risk of exacerbations in chronic obstructive pulmonary disease. *Int J Chron Obstruct Pulmon Dis*. 2013;8:493–499. doi:10.2147/COPD.S49609
21. Moons KG, Kengne AP, Woodward M, et al. Risk prediction models: i. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*. 2012;98(9):683–690. doi:10.1136/heartjnl-2011-301246
22. Amalakuhan B, Kiljanek L, Parvathaneni A, Hester M, Cheriyath P, Fischman D. A prediction model for COPD readmissions: catching up, catching our breath, and improving a national problem. *J Community Hosp Intern Med Perspect*. 2012;2:1. doi:10.3402/jchimp.v2i1.9915
23. Brusse-Keizer M, van der Palen J, van der Valk P, Hendrix R, Kerstjens H. Clinical predictors of exacerbation frequency in chronic obstructive pulmonary disease. *Clin Respir J*. 2011;5(4):227–234. doi:10.1111/j.1752-699X.2010.00234.x
24. Tavakoli H, Chen W, Sin DD, FitzGerald JM, Sadatsafavi M; Canadian Respiratory Research N. Predicting Severe COPD exacerbations: developing a population surveillance approach with administrative data. *Ann Am Thorac Soc*. 2020. doi:10.1513/AnnalsATS.202001-070OC
25. Lisspers K, Larsson K, Johansson G, et al. Economic burden of COPD in a Swedish cohort: the ARCTIC study. *Int J Chron Obstruct Pulmon Dis*. 2018;13:275–285. doi:10.2147/COPD.S149633
26. Martinell M, Stalhammar J, Hallqvist J. Automated data extraction–a feasible way to construct patient registers of primary care utilization. *Ups J Med Sci*. 2012;117(1):52–56. doi:10.3109/03009734.2011.653015
27. Sweden S Longitudinal integration database for health insurance and labour market studies (LISA by Swedish acronym); 2004. Available from: https://www.scb.se/en/services/guidance-for-researchers-and-universities/vilka-mikrodata-finns/longitudinella-register/longitudinal-integrationdatabase-for-health-insurance-and-labourmarket-studies-lisa/. Accessed January 07, 2019.
28. Socialstyrelsen; 2017. Available from: https://www.socialstyrelsen.se/statistik-och-data/register/alla-register/. Accessed March 8, 2021.
29. University of Manitoba – Community of Health Sciences – Manitoba Center for Health Policy. Concept Descrip¬tion: charlson comorbidity index; 2019. Available from: http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?printer=Y&conceptID=1098. Accessed March 5, 2021.
30. Chen T, Guestrin C. XGBoost: a scalable tree boosting system; 2016. Available from: https://dl.acm.org/doi/pdf/10.1145/2939672.2939785. Accessed March 8, 2021.
31. Chen YJ, Narsavage GL. Factors related to chronic obstructive pulmonary disease readmission in Taiwan. *West J Nurs Res*. 2006;28 (1):105–124. doi:10.1177/0193945905282354
32. Hoogendoorn M, Feenstra TL, Boland M, et al. Prediction models for exacerbations in different COPD patient populations: comparing results of five large data sources. *Int J Chron Obstruct Pulmon Dis*. 2017;12:3183–3194. doi:10.2147/COPD.S142378
33. Larsson K, Janson C, Stallberg B, et al. Impact of COPD diagnosis timing on clinical and economic outcomes: the ARCTIC observational cohort study. *Int J Chron Obstruct Pulmon Dis*. 2019;14:995–1008. doi:10.2147/COPD.S195382
34. Ludvigsson JF, Andersson E, Ekbom A, et al. External review and validation of the Swedish national inpatient register. *BMC Public Health*. 2011;11(1):450. doi:10.1186/1471-2458-11-450
35. Kjeldsen SE, Stalhammar J, Hasvold P, Bodegard J, Olsson U, Russell D. Effects of losartan vs candesartan in reducing cardiovascular events in the primary treatment of hypertension. *J Hum Hypertens*. 2010;24(4):263–273. doi:10.1038/jhh.2009.77