**PROBAST**

Study:

PulmoListener

Step 2: Type of prediction study

**Is the study a diagnostic or a prognostic study?**

Prognostic

**Is the study a development only, development and validation or validation only study?**

Development only

**What is the model of interest?**

Deep neural network

**What is the outcome of interest?**

COPD symptom severity

Step 3: Assess risk of bias

**Domain 1: Participants**

**Describe the sources of data and criteria for participant selection**

We recruited COPD patients from three hospitals in Toronto, Canada to participate in our study for an average of 164 ±92 days.

**1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?**

Y

**1.2 Were all inclusions and exclusions of participants appropriate?**

Y

**Risk of bias introduced by selection of participants:**

Low

**Rationale of bias rating**

No further eligibility criteria

**Domain 2: Predictors**

**List and describe predictors included in the final model, e.g. definition and timing of assessment**

Participants were given a Samsung Galaxy Watch to wear during the day and charge at nighttime. The smartwatch recorded audio data at a sampling rate of 44.1 kHz.

PulmoListener constitutes an end-to-end speech processing pipeline that operates on passively collected audio. The overall pipeline, which is illustrated in Fig. 1, isolates useful speech segments from real-world audio data, converts them into meaningful feature representations, and feeds them into a machine learning model that classifies COPD symptom severity.

Within a given hour, PulmoListener discards the audio windows that do not satisfy the thresholds 𝑝speech and𝑝patient. The remaining windows are sorted according to the product of their voice activity and speaker similarity scores. The audio windows with the highest overall score for each hour are used for feature extraction and inference. PulmoListener summarizes the patient’s speech characteristics within a given day by stacking together the MFCCs of the audio windows selected to represent each of its constituent hours. Since many individuals charge their devices at night, we represent each day using 12 hours’ worth of data (9 AM to 9 PM). Whenever fewer than 12 windows are available for generating this sequence, PulmoListener imputes missing windows by selecting unused speech windows from nearby hours.

**2.1 Were predictors defined and assessed in a similar way for all participants?**

Y

**2.2 Were predictor assessments made without knowledge of outcome data?**

Y

**2.3 Are all predictors available at the time the model intended to be used?**

Y

**Risk of bias introduced by predictors or their assessment**

Low

**Rationale of bias rating**

Feature extraction is done the same way for each patient. Predictors are independent of outcome.

**Domain 3: Outcome**

**Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination:**

Participants reported the severity of their COPD symptoms every morning using the London COPD Cohort Daily Symptom Questionnaire [1, 5] — a clinically validated instrument that has been used in both clinical [51, 52]and technical [29, 50] studies to screen for COPD exacerbations. The questionnaire required patients to reflecton the severity of the eight symptoms listed in Table 1, selecting the ones that were “worse than usual” on the previous day. The questionnaire considers major symptoms as 5 points and minor symptoms as 1 point when tabulating a patient’s final symptom score.

The questionnaire scores required curation so that they could be converted to binary symptom severity labels. The London COPD Cohort Daily Symptom Questionnaire is designed to identify COPD exacerbations, which are serious but rare occurrences in most COPD patients. The questionnaire suggests that scores above 6 for two consecutive days warrant concern; however, applying such a threshold to our dataset led to significant class imbalance during model training. To binarize the scores in our dataset, we adjusted the threshold to 3 and only required patients to exceed this threshold for one day in order to assign a positive label. Although these adjustments deviate from the initial construction of our chosen instrument, their conservative nature also means that PulmoListener would need to be able to identify more subtle manifestations of COPD symptoms.

**3.1 Was the outcome determined appropriately?**

N

**3.2 Was a pre-specified or standard outcome definition used?**

PY

**3.3 Were predictors excluded from the outcome definition?**

Y

**3.4 Was the outcome defined and determined in a similar way for all participants?**

Y

**3.5 Was the outcome determined without knowledge of predictor information?**

Y

**3.6 Was the time interval between predictor assessment and outcome determination appropriate?**

Y

**Risk of bias introduced by the outcome or its determination**

High

**Rationale of bias rating**

Gold standard questionnaire was used. Predictors excluded from outcome. Appropriate timing. However, outcome was binarized and threshold for binarization was determined after recruitment and during analysis. Therefore high risk of bias.

**Domain 4: Analysis**

**Describe number of participants, number of candidate predictors, outcome events and events per candidate predictor**

Patients reported their symptoms over a combined total of 1,310 days, with an average score of 2.7 ±4.2 across all reports.

**Describe how the model was developed, predictor selection and risk group definition**

To assess whether continuous speech holds sufficient information to detect elevated symptom severity, weexamined a variety of feature extractors and model architectures for symptom severity inference:

• Machine learning: As a baseline approach, we featurised the dataset using the COMPARE feature extractor from the openSMILE toolkit [17]. This feature extractor calculates 6.3k acoustic features, including MFCCs, using diverse functionals over low-level descriptor contours [49]. For each feature, we computed seven aggregates over the day: mean, median, maximum, minimum, skewness, kurtosis, and standard deviation. Since this resulted in a large number of features, we selected the top 200 features according to their Giniimportance.(1) SVM: 𝐶 = 10, 𝛾 = 0.1, kernel = radial basis function(2) Random forest: Number of trees = 100

• Deep learning: We featurised the dataset using PulmoListener’s pipeline for MFCC generation. These sequences were then rearranged according to the input size of the network.(3) LSTM: Number of layers = 2, number of units per layer = [64, 128](4) 3D CNN (PulmoListener): Number of convolutional layers = 3, number of units per convolutional layer = [64, 128, 128], number of dense layers = 2, number of units per dense layer = [256, 2]All of our models were implemented in Python using scikit-learn2 and Keras3 with their default values unless otherwise specified.

**Describe whether and how the model was validated, either internally (cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants)**

To maximize the likelihood of success, we built personalised models for each patient by training on all other patients’ data and two weeks of the patient’s own data. We ensured that all sequences from the same day were kept in the same split to prevent information leakage.

**Describe the performance measures of the model, e.g. calibration, discrimination, classification, net benefit, and whether they were adjusted for optimism**

SEN, SPE, F1

**Describe any participants who were excluded from the analysis**

Although we recruited 28 patients, we restricted our analyses in this paper to those who had more than 4 weeks of data and exhibited severe symptoms at some point during their enrolment. Applying these inclusion criteria left us with 8 patients (3 females, 5 males), ranging in age from55–93 years (average = 66.4 ±11.7 years).

**Describe missing data on predictors and outcomes as well as methods used for missing data**

Whenever fewer than 12 windows are available for generating this sequence, PulmoListener imputes missing windows by selecting unused speech windows from nearby hours.

**4.1 Were there a reasonable number of participants with the outcome?**

Y

**4.2 Were continuous and categorical predictors handled appropriately?**

Y

**4.3 Were all enrolled participants included in the analysis?**

N

**4.4 Were participants with missing data handled appropriately?**

Y

**4.5 Was selection of predictors based on univariable analysis avoided?**

Y

**4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls)**

**accounted for appropriately?**

Y

**4.7 Were relevant model performance measures evaluated appropriately?**

Y

**4.8 Were model overfitting and optimism in model performance accounted for?**

Y

**4.9 Do predictors and their assigned weights in the final model correspond to the results**

**from multivariable analysis?**

U

**Risk of bias introduced by the analysis**

High

**Rationale of bias rating**

Considerable amount of outcomes. However, most patients were excluded from analysis and only patients with severe symptoms were included.

**Overall Risk of bias**

High