Peer Review File
Article information: http://dx.doi.org/10.21037/jtd-20-3539

Reviewer A

The authors described a prediction model based on clinical symptoms and spirometric indices. I appreciate the innovative approach. It was a two step analysis. First, a binary logistic regression was done, followed by more complex, Bayesian probabilistic networks.

I have a few concerns:
Comment 1: The study needs to be adjusted for the use of medication like inhaled steroid, LABA, Montelukast, antimuscarinic bronchodilator etc.

Reply 1: Thank you very much for your thoughtful comment. The aim of the study was to provide simple diagnostic algorithms to support the diagnoses of obstructive airway diseases irrespective of the intake of medication (as already discussed in the limitations section). Accordingly, the predictors from Figures 3 and 4 comprised simple measures such as sex, age, smoking status and history of allergic rhinitis; these factors do not depend on medication. Regarding asthma, it is also very unlikely that wheezing in the last 12 months has been completely abolished by medication and thus become worthless as indicator of the disease. Regarding COPD, it is also not to be expected that the production of phlegm is completely abolished by medication. Regarding lung function (MEF25, MEF75, MMEF), medication might have had a stronger effect in asthma compared to COPD, thereby reducing the likelihood for the diagnosis of asthma. On the other hand, it is nearly impossible to find a patient with obstructive airway disease without some sort of prior respiratory medication despite the lack of a proper diagnosis. In the revised version we more emphasized this limitation (page 15, lines 338-343) which, as mentioned, probably affects the differential diagnosis of asthma more than that of COPD.

Comment 2: In my opinion, the results section is too long and needs to be more summarized.

Reply 2: We shortened the results section by shifting two paragraphs into the supplement. The other parts were not shortened since we believe that this would affect understanding and measures of validity of our results.

Comment 3: It seems that the raw value of FEV1 was taken instead of the percent predicted, and that also could be influenced by age, gender which led to over adjustment bias.

Reply 3: Thank you for your comment. We apologize for the laziness in the respective descriptions in the figures. In all analyses, we used the z-scores of lung function parameters in order to adjust them for anthropometric characteristics as far as possible. We clarified this aspect in the manuscript (page 6, lines 122-124) and amended the figure legends.

Comment 4: Figure 4: the authors divided the subjects above and below 57 years. How did they find the cut off of 57 years?

Reply 4: This was done in the tree search procedure that identified the optimal tree. Specifically, for continuous variables, the CHAID algorithm that we used identifies the cut-off value that yields the best separation of individuals and thus the best cut-off value.

Comment 5: I understand considering allergic rhinitis for asthma, not sure why should it predict COPD?

Reply 5: Thank you for this comment. In our decision tree (figure 3) allergic rhinitis was only relevant to increase or decrease the likelihood of asthma (right part). If the z score of MMEF was high and allergic rhinitis was present, the likelihood of asthma raised (middle bars). If under this condition no allergic rhinitis was reported asthma was less likely and the likelihood of belonging to the control group raised. In contrast, in both of these conditions the likelihood of COPD was low and did not change, indicating that allergic rhinitis was not relevant for the establishment of a COPD diagnosis.

Comment 6: So FEV1 is more useful to predict COPD compared to the classic FEV1/FVC ratio. How could you explain that? Why should the GP believe in it, when FEV1/FVC ratio has been gold-standard for decades?

Reply 6: Thank you for this thoughtful comment. In the procedures used for the identification of the best lung function measures we compared FEV_1, FVC, their ratio and different flow rates. It turned out that, given the anamnestic information, FEV_1 was the superior parameter, and not FEV_1/FVC. We think that the additional presence of anamnestic information played a role for this finding. We also included a number of patients (n= 33) with the clinical diagnosis of COPD who did not have $FEV_1/FVC < 0.7$, i.e. patients of the former category GOLD 0; this fact is now mentioned in the manuscript (page 8, line 183). We believe that their inclusion was justified, as there are many data indicating that these patients have a peculiar form of COPD, mostly of restrictive type. Moreover, they are part of the patients typically found in clinical practice. A recent analysis of COSYCONET data (Mayerhofer et al., Deterioration and Mortality Risk of COPD Patients Not Fitting into Standard GOLD Categories: Results of the COSYCONET Cohort, Respiration, 2021 Jan 22:1-10) underlined that this type of patients shows all signs of COPD from a clinical perspective including COPD treatment.

Comment 7: Does it mean that the prediction model did not work for asthma?

Reply 7: The prediction models using the simple set of data evaluated in the present study do work for asthma but not as well as for COPD. This might be explained by the fact that asthma patients in our study had a low degree of airway obstruction and were similar to controls regarding lung function parameters (see table 1). Thus, the models were less useful for asthma, especially regarding the lung function parameters which can be normal in asthma outside of episodes of symptoms. This limitation refers to a wide variety of lung function parameters, and only bronchial provocation tests or the measurement of exhaled nitric oxide (FENO) seem capable of improving the diagnosis of asthma. However, even these measures are affected by previous treatment.

Comment 8: Since you are using advanced techniques, machine learning methods like XGBoost or Neural network would be much easier to follow. I doubt if a common reader will understand the key message from these decision trees. It might be easier for the readers to see each predictors' individual contribution and perhaps their rank list.

Reply 8: Thank you for your suggestion. As mentioned in the discussion section (page 14, lines 315-322), machine learning methods like XGBoost and Random Forest achieve statistical robust and reliable results. These methods, however, have the disadvantage that the final algorithm cannot be easily depicted and that only a rank order of parameters can be

given, without showing a way to implement this knowledge. We have used random forest methods previously (Kellerer, C., Klütsch, K., Husemann, K., Sorichter, S., Jörres, R. A., & Schneider, A. (2020). Capnovolumetry in combination with clinical history for the diagnosis of asthma and COPD. NPJ primary care respiratory medicine, 30(1), 1-9.) but only as an addition to and check for single decision trees that were used in order to visualize the results in a comprehensible manner. The results were fully consistent with each other. We would not like to blow up the present manuscript by these additional procedures and therefore prefer the single decision trees that can be understood without invoking algorithms that can only be realized by software objects.

Comment 9: The results of the cross-validation should be clearly mentioned in the text.

Reply 9: We now emphasize that cross-validation has been used within the procedures leading to the decision trees, i.e. only branches have been chosen which withstood the test of tenfold cross-validation (page 8, line 160).

Reviewer B

I would recommend major revision primarily for inadequate/insufficient analyses. Here are three major issues I had with the analyses:
Comment 1: The analyses only employed half of the collected questions (8/15) due to the other variables missing data. What was the full (15) list of questions? The authors do not say. It is entirely possible/probable that some of the dropped variables (e.g., smoking pack-years?) were more predictive of COPD/asthma than the eight variables that were used. I would recommend the authors go back and either use a statistical tree classification algorithm (e.g., XGBoost) that can impute missing data directly within the algorithm or employ some other missing data technique in order to include the questions/variables that were dropped from these analyses due to missing data. With all the statistical tools available today for handling missing data, there is no reason to throw away half of the available predictors due to missing data particularly when the authors claim, in the background, to have "systematically assessed the relative importance of such questions...".

Reply 1: Thank you for the comment. The 7 omitted questions comprised 3 sub-items to the questions of wheezing and cough which were dependent on a positive answer and thus available only in part of the patients. The other 4 questions comprised the presence of nocturnal symptoms typical for asthma and the presence of asthma attacks, as well as repeated respiratory infections. The questions regarding asthma were omitted as we wanted to study the situation in the absence of a previous diagnosis; please note that patients were recruited in a pneumological practice and nearly all of them had a previous diagnosis. We included these questions to check for the presence of symptoms but naturally these were not suited if simulating the situation of establishing a first diagnosis. If including the 3 sub-items on wheezing and cough, these were not among the most important 8 questions identified in univariate comparisons between groups. We therefore do not have hints that the questions that could have been reasonably included but were omitted were more informative than those included. We limited the maximum number of questions to 8 as a complex set of anamnestic items would not be realistic for clinical practice. Imputation can be very helpful in association analyses but we consider it dangerous in establishing a diagnostic algorithm and therefore rejected it. An additional argument was that missing items in our study would have some likelihood to be missing also in clinical practice and thus useless. Regarding packyears, we found in previous studies from the COSYCONET cohort that the information contained in them was surprisingly small and often inferior to the establishment of the smoking status.

Comment 2: The authors acknowledged the lack of independence between the disease groups, "There were 34 subjects with COPD and the comorbidity asthma, who were assigned to the COPD group, as this disease dominated the functional alterations." First, I'm not sure what is meant by "dominated the functional alterations." More importantly, though, I don't believe that bucketing the asthma-COPD overlap (ACOS) persons into the COPD group alone – and dropping them from the asthma group – was the proper way of handling this group. I would think these ACOS persons should either be included in both the COPD AND asthma groups (as they have both diseases) or they should be their own distinct (ACOS) group. As the authors were modeling the predictive effectiveness of certain criteria on selected respiratory diseases, removing nearly 10% of the known (i.e., diagnosed) asthma cases may have substantially impacted the subsequent asthma models.

Reply 2: Thank you for raising this point. The ACOS group (34 patients) was certainly too small to be incorporated as a separate group. Moreover, the binary decision algorithms which we implemented in the trees did not allow a simultaneous decision for asthma and COPD. When comparing asthma only with either COPD only or with asthma/COPD regarding the distribution of symptoms as well as lung function measures, the patients with asthma and COPD were more similar to COPD patients than to asthma patients. We used this decision already in previous papers regarding capnovolumetry (Kellerer, C., Klütsch, K., Husemann, K., Sorichter, S., Jörres, R. A., & Schneider, A. (2020). Capnovolumetry in combination with clinical history for the diagnosis of asthma and COPD. NPJ primary care respiratory medicine, 30(1), 1-9; Kellerer, C., Schneider, A., Klütsch, K., Husemann, K., Sorichter, S., Jörres, R. A. (2020). Correspondence between Capnovolumetric and Conventional Lung Function Parameters in the Diagnosis of Obstructive Airway Diseases. Respiration, 99(5), 389-397) and we believe that it was justified, especially as the patients with asthma were more difficult to diagnose than those with COPD. It might have been possible that a history of allergic rhinitis in patients with asthma/COPD reduced diagnostic accuracy, as allergic rhinitis was linked to asthma and not COPD (see figure 3). It also should be kept in mind that the diagnosis of concomitant asthma in COPD may be difficult even if advanced diagnostic tools are available. Again, thank you for the comment, and we included this point briefly in the discussion (page 15, lines 343-347). We also included 2 additional decision trees, namely for asthma vs control and for COPD vs control; these are shown in the supplement. When using these two trees together and requiring a positive diagnosis in both, 32 of 34 patients with asthma and COPD were recognized as such (94%). On the other hand, only 238 out of 659 patients with either asthma or COPD but not both were correctly recognized (36%). This shows a high sensitivity but insufficient specificity of the tree algorithms to identify the presence of both conditions. We also used other trees and combinations but with no useful success. Therefore, with the limited tools that we use the issue of concomitant asthma and COPD cannot be resolved, especially as the number of patients with this diagnosis (n=34) was too low to derive a specific decision algorithm.

> *Following your question, we think that this information could be of interest for the reader and therefore included in the discussion (page 15, lines 343-349) and the two corresponding trees in the supplement. Thank you very much for this hint on improving the manuscript.*

Comment 3: It was stated that "All decision trees showed the phenomenon that specific combinations of values were informative for diagnosis, while others did not carry relevant information. The informative and non-informative combinations could always be easily recognized. This variation in performance renders it questionable, whether overall measures such as specificity and sensitivity are adequate, and we therefore do not present these values."

It's not clear to me how variation in performance renders the computation of sensitivity and specificity questionable. Using binary comparisons of disease to no-disease separately for COPD, asthma, both (ACOS), and other respiratory disease, sensitivity and specificity for each category/disease could have easily been computed and presented. Given that the goal of the study was to inform disease diagnostics, it would be extremely helpful for clinicians to know how well the algorithm can identify the absence of disease in addition to detecting true disease cases.

Reply 3: In the text the overall accuracy of each of the trees is given (page 10, line 218 and page 10, line 228). If there are certain combinations of parameters that are associated with large changes in the Odds ratios of certain diseases compared to the a priori distribution, this gain in information is weakened when the results are pooled with those conditions in which there is no marked change in the Odds ratios. This does mean that overall sensitivity and specificity do not properly describe the power of a tree algorithm that allows to identify specific conditions which are only present in certain groups of patients. To illustrate this point we now compute Odds ratios of asthma vs COPD, which are given in the decision tree of asthma vs COPD in the respective nodes (Figure 3). Thank you for the comment which helped to clarify the issue and to improve the description of the results.

In no particular order, here are some other items that may need to be addressed before publication:
Comment 4: "To account for the dependence on anthropometric characteristics and eliminate these as explicit items, all variables were adjusted for age, sex and BMI." Why did these need to be removed as explicit items?

Reply 4: We removed these variables in order to keep the networks comprehensible. As functional parameters often depend on age and sex, and possibly BMI, we otherwise would have many trivial arrows from these anthropometric measures to the functional measures which would make the network very complex and cost power of the analysis.

Comment 5: Line 123 - Global Lung Function Initiative should be spelled out before using the acronym (GLI).

Reply 5: We followed this hint.

Comment 6: Lines 197/198, shouldn't it read "increased to 73.5 from 34.4" as opposed to "increased from 73.5 to 34.4"?

Reply 6: We corrected. Thank you for spotting this error (page 10, line 213).

Comment 7: In the decision trees, it was not described how the cutpoints for the continuous variables (MMEF, MEF50, MEF25, AGE) were derived.

Reply 7: This was implemented in the algorithm which we used and which consecutively tried a whole range of cut-points to identify the best one.

Comment 8: In Figure 2, the subjects with neither the diagnosis of asthma nor COPD nor control group (n=230) were excluded to give a total sample size of 1057 but, in the subsequent boxes, the 1057 participants is broken into 4 groups including this same group that was just excluded and the subsamples total to 1287, not 1057. This flowchart needs to be

consistent such that the 230 were either excluded (total n=1057) or they were included and treated as the fourth subgroup (total n=1287).

Reply 8: We corrected this error. Thank you for the careful review.

Comment 9: It is unclear whether there was overlap between the COPD/asthma groups and the "other respiratory disease" group. You may want to explicitly note the overlap (or lack thereof) for clarity.

Reply 9: Thank you for raising this issue. There was no overlap, as these patients had other diseases than asthma and COPD. If they had COPD they were attributed to the COPD group, since it would not fit the conditions of clinical practice to define "pure COPD". This was consistent with the decision to attribute patients with asthma and COPD to the COPD group. If patients had asthma without COPD, they were attributed to the asthma group.

Reviewer C

Major comments
Comment 1: This study used data from another study conducted in a pulmonary outpatient clinic and therefore authors emphasize the diagnostic reliability. The main purpose of this study is to create transferable algorithms to the general practice settings, however, there is no validation using another cohort. This validation is very important, but it is no even mentioned as a study limitation. Moreover, ease of use of the algorithms in real-world general practice settings has not been considered either.

Reply 1: Thank you for this comment. We now emphasize this point more in the revised version of the manuscript (page 15, lines 333-338 and line 351-352). It is important to note that pulmonary outpatient clinics in Germany are organized as private practices of specialists in primary care that can be visited by patients without referral. Therefore, we believe that patients recruited in general practice settings would not differ much from our study population which has relatively good lung function values (see table 1). Nevertheless, the results should be validated in a general practice setting.

> *Regarding the ease of use of the algorithms, we focused on algorithms that could be understood and followed from common clinical insight, instead of using black-box algorithms, such as neural networks or random forests, that yield a result without delineating in detail how it came about. Only a rank order of variables can be given, without specific information which variables were important in a particular patient, which can be easily seen in a decision tree. We believe that such insight is an important prerequisite for acceptance in clinical practice.*

Comment 2: Control subjects are patients without respiratory diseases who visited a pulmonary outpatient clinic, but there is no specific description of what kind of patients they are. Patients who visiting pulmonary specialists are likely to have pulmonary symptoms and may influence the study of the symptom questionnaire.

Reply 2: Control subjects were those in whom no respiratory disease could be diagnosed. The symptoms might have been due to other conditions such as cardiac disease. Moreover, this group comprised patients who visited the practice for a second time after having had complaints from a previous respiratory infection, without having chronic respiratory disease. We believe that it is unrealistic to use a group of anamnestically completely healthy subjects for comparison, when aiming at a potential application in clinical practice, and that it is

more adequate to use a group defined solely by the absence of chronic respiratory disease.

Comment 3: If MMEF was inferior to FEV1 in the path analysis, why was MMEF used for decision tree analysis?

Reply 3: In the decision tree analysis, we offered all lung function measures and MMEF was selected as the best one. This is not in contradiction to the result of the path analysis, as this involved the relationship between functional parameters and responses to questions in the different diagnoses of asthma and COPD, without making a differential diagnosis. In the path analysis, the diagnosis appeared as two different layers of the model, one of them showing additional relationships as indicated by the dashed arrows.

Minor comments
Comment 4: Line 198. Likelihood of being without respiratory disease was increased from 73.5% to 34.4%? The number seems to be the opposite.

Reply 4: Thank you for spotting this error. We corrected (page 10, line 213).

Comment 5: Reference styles are not unified. The titles of journals should be abbreviated according to the style used in Index Medicus. For reports with up to three authors, all the author names should be listed. However, if a report has more than three authors, the first three authors should be listed followed by "et al."

Reply 5: Thank you for this hint. It seems that the export from Endnote introduced an error. We corrected.