A novel statistical model for analyzing data of a systematic review generates optimal cutoff values for fractional exhaled nitric oxide for asthma diagnosis

Antonius Schneider*a, Klaus Lindea, Johannes B. Reitsma*b, Susanne Steinhauserc, Gerta Rückerd

aInstitute of General Practice, University Hospital Klinikum rechts der Isar, Technische Universität München, Orleansstraße 47, 81667 Munich, Germany
bJulius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA, Utrecht, The Netherlands
cInstitute of Medical Statistics, Informatics and Epidemiology, University of Cologne, Bachemer Str. 86, 50931 Köln, Germany
dInstitute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Stefan-Meier-Strasse 26, 79104 Freiburg, Germany

Accepted 1 September 2017; Published online 12 September 2017

Abstract

Objectives: Measurement of fractional exhaled nitric oxide (FENO) might substitute bronchial provocation for diagnosing asthma. However, optimal FENO thresholds for diagnosing asthma remain unclear. We reanalyzed data collected for a systematic review investigating the diagnostic accuracy of FENO measurement to exploit all available thresholds under consideration of pretest probabilities using a newly developed statistical model.

Study Design and Setting: One hundred and fifty data sets for a total of 53 different cutoffs extracted from 26 studies with 4,518 participants were analyzed with the multiple thresholds model. This model allows identifying thresholds at which the test is likely to perform best.

Results: Diagnosing asthma might only be possible in a meaningful manner when the pretest probability of asthma is at least 30%. In that case, FENO > 50 ppb leads to a positive predictive value of 0.76 [95% confidence interval (CI): 0.29–0.96]. Excluding asthma might only be possible, when the pretest probability of asthma is 30% at maximum. Then, FENO < 20 ppb leads to a negative predictive value of 0.86 (95% CI 0.66–0.95).

Conclusion: The multiple thresholds model generates a more comprehensive and more clinically useful picture of the effects of different thresholds, which facilitates the determination of optimal thresholds for diagnosing or excluding asthma with FENO measurement. © 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Asthma; Fractional exhaled nitric oxide; Diagnostic accuracy; Sensitivity; Specificity; Receiver operating characteristic analysis

1. Introduction

Asthma is a complex chronic inflammatory disorder of the bronchial tree, characterized by bronchial hyperresponsiveness and usually reversible airway obstruction. Many efforts continue to be undertaken to improve the diagnostic process to allow an early diagnosis, as early treatment is important for the management of the disease [1]. However, ruling-in and ruling-out the diagnosis of asthma in patients suffering from chronic symptoms like dyspnea, wheezing, and coughing is still challenging. Spirometry is considered a reference standard for diagnosing airway obstruction, but it is not possible to rule out milder forms of asthma, as obstruction is not present in these cases [2]. For these cases, bronchial provocation (BP) is a reference standard, particularly in cases with...
What is new?

Key findings
- The multiple thresholds model allows using all available information on thresholds of studies investigating biomarkers, as demonstrated with a systematic review of studies investigating the diagnostic accuracy of FE\textsubscript{NO} measurement.

What this study adds to what was known?
- The multiple thresholds model enables the calculation of summarized sensitivities and specificities of different cutoff points, and applying Bayes' theorem allows calculating the predictive values, given the prevalence of the disease.
- This allows deriving optimal thresholds related to the clinical setting where the index test is used, as defined by the pretest probability of a disease.

What is the implication and what should change now?
- A prerequisite for summarizing diagnostic test accuracy studies with the multiple thresholds model is that authors of primary studies report the pairs of sensitivities and specificities of all thresholds they examined of the index test or, even more preferably, the biomarker values of all individual patients along with their true disease status according to the reference standard.
- If reviewers encounter primary studies that report findings for multiple thresholds in sufficient detail, they should consider the multiple thresholds model or related methods.

### Results of Recent Studies

Inconclusive spirometric results, but BP is time-consuming, costly, and carries a small risk of inducing severe bronchospasm [3]. New biomarkers like fractional exhaled nitric oxide (FE\textsubscript{NO}) might help to close this diagnostic gap. Nitric oxide (NO) in orally exhaled air mainly originates from the respiratory epithelium, where it is produced by the inducible NO synthase [4]. FE\textsubscript{NO} was initially estimated by the guideline of the American Thoracic Society (ATS) as a specific marker for eosinophilic airway inflammation [5]. Results of recent studies suggested a more comprehensive concept which reclassified FE\textsubscript{NO} more accurately as a broader marker of T-helper cell type 2–mediated allergic inflammation [6,7]. Fitting to the concept of increased FE\textsubscript{NO} values correlating with allergic inflammation pattern, several studies have demonstrated a diagnostic value of FE\textsubscript{NO} measurement [8–11]. We recently performed a systematic review of all available studies investigating the diagnostic accuracy of FE\textsubscript{NO} in patients suspected to suffer from asthma. Using standard meta-analytic methods, we found a summarized area under the curve of 0.80 (95% confidence interval: 0.77–0.85) [12].

Systematic reviews of DTA studies have become a valuable tool to summarize the available evidence and generate recommendations about the use of diagnostic tests in practice [13]. If a test produces a binary result, the methods for meta-analyzing such data are now well established. In these situations, there is a single two-by-two table that can be analyzed using the bivariate or hierarchical summary approach [13]. The analysis becomes more complex if a test produces a continuous result as in the case of FE\textsubscript{NO} measurement. The actual choice of threshold now becomes an additional factor to evaluate as changing the threshold will alter the corresponding sensitivity and specificity. Clinical decision-making about how to use a test is driven by the absolute number of correct and incorrect decisions based on the results of the test and the relative weights attached to consequences due to a false-positive or false-negative test result. Of course, these weights will depend on the clinical scenario at hand, but it is clear that clinical decision-making of tests producing a continuous result has to take into account the prevalence, the choice of threshold, and the accuracy corresponding to each threshold. Therefore, primary studies often present pairs of sensitivity and specificity for multiple thresholds or a receiver operating characteristic (ROC) curve. Unfortunately, current diagnostic reviews typically select one pair of sensitivity and specificity to be analyzed using the standard bivariate meta-analysis model [14,15]. For this purpose, often the most commonly used threshold is chosen, thus ignoring all information coming from other thresholds and wasting relevant information about test performance. Therefore, it was not possible to identify optimal thresholds for medical decision-making with the traditional way of a systematic review [12].

In a recent paper [16], two authors of the present article (G.R. and S.S.) proposed a statistical approach for meta-analysis of diagnostic test accuracy (DTA) studies, henceforth called the multiple thresholds model, that allows using all available information on thresholds and the corresponding two-by-two tables from all studies. The aim was to apply the new model to reanalyze the diagnostic accuracy of FE\textsubscript{NO} measurement and to derive optimal decision rules under consideration of pretest probabilities for diagnosing asthma.

### 2. Methods

The overall methods used for our systematic review have been described previously [12]. In brief, we searched MEDLINE, Embase, and Scopus up to November 29, 2015, for potentially relevant studies. To be eligible, studies had to allow the generation of two-by-two tables for asthma diagnosis by FE\textsubscript{NO} as compared to a reference standard. The reference standard could be BP, measurement of FEV\textsubscript{1} with
brachodilation, peak flow variability, expert’s opinion, or a combination of these. Study participants were patients with suspected asthma and at least 75% had to be steroid naïve. Measurement of FENO had to be done using a mean exhalation flow rate of 50 mL/s and instantaneous flows within the range of 45–55 mL/s according to the international guidelines. Study quality was assessed with the QUADAS-2 tool. For meta-analysis, we used data for the cutoff point with the highest sum of sensitivity and specificity (i.e., the Youden Index) whenever possible.

For the analyses presented in this paper, we extracted the numbers of true test positives, true negatives, false positives, and false negatives based on each cutoff point for which these figures were available or could be calculated. The data were collected in an Excel sheet (data file in the online Appendix at www.jclinepi.com).

The multiple thresholds model creates a link between the range of thresholds and the respective pairs of sensitivity and specificity and thus allows identifying thresholds at which the test is likely to perform best [16]. Different studies may contribute a varying number of thresholds as well as different sets of thresholds. The multiple thresholds model is a multilevel random effects model. At the study level, for the group of patients without the target condition (in short disease free), the specificities at all available thresholds together provide an estimate of the cumulative distribution function (cdf) of the test results within the disease-free individuals. Likewise, for patients with the target condition (in short diseased), via the observed sensitivities at all observed thresholds, we obtain an estimate of the cdf of the test results within the diseased patients. At the meta-analytic level, the model fits the data for both groups and all available thresholds over all studies. Based on a log-logistic model, it provides estimates of the two cdfs for the two groups across all studies, accounting for the between-study heterogeneity and correlation between groups.

To be more precise, in order to fit a linear model to the data, we first applied the natural logarithm (ln) to the FENO values. We then transformed all specificities (for disease-free patients) and all one minus sensitivities (for diseased patients) using the logit transformation, logit(x) = ln(x) − ln(1 − x) and fitted a linear mixed-effects model to all data points. Each data point was weighted with the inverse variance of the respective logit-transformed proportion. The model (called DIDS* in [16], which stands for different variance of the respective logit-transformed proportion. The data points. Each data point was weighted with the inverse random intercepts and different random slopes) is given by

\[
\logit(S_{ki}) = \alpha_0 + a_{ki} + (\beta_0 + b_{ki})x_{ki} + \epsilon_{ki}
\]

\[
\logit(1 - S_{ki}) = \alpha_1 + a_{1ki} + (\beta_1 + b_{1ki})x_{ki} + \delta_{ki}
\]

where \(S_{ki}\) and \(S_{k1i}\) are the specificity and the sensitivity of the \(i\)'th threshold \(x_{ki}\) (on the log scale) in study \(k\), \(\alpha_0\) and \(\alpha_1\) are the fixed intercepts, \(\beta_0\) and \(\beta_1\) the fixed slopes for the disease-free and the diseased individuals, respectively. The terms \(a_{ki}, a_{1ki}, b_{ki},\) and \(b_{1ki}\) denote the random intercepts and slopes which are assumed to follow a common four-dimensional normal distribution, that is, to correlate across studies. With \(\epsilon_{ki}\) and \(\delta_{ki}\), we denote the within-study random errors. After using the inverse of the logit function for back-transforming the fixed effects part of the model equations, we obtain model-based distribution functions for disease-free and diseased individuals from which a model-based summary ROC curve is obtained in the usual way. The model was implemented in the free software environment R [17]. For more details of the modeling, we refer to Steinhauser et al. [16] and provide R code in the supplemental material.

We then can read off the “pooled” sensitivity and specificity values at every threshold; and a multiple thresholds summary ROC (mtsROC) curve naturally follows while preserving threshold information. Additionally, we obtain positive and negative predictive values, depending on the threshold and different possibilities with respect to the prevalence of the disease observed in the included studies. The 95% CIs of sensitivities, specificities, and predictive values are estimated using the delta method [16].

For sensitivity analysis, we investigated the influence of age by comparison of children and adults as subgroup analysis. A further subgroup analysis was performed regarding the subgroup of patients where solely BP was used as a reference standard.

### 3. Results

We included a total of 26 studies [11, 18–42] with 29 patient groups (two studies presented results for separate subgroups) and 4,518 participants (see Table 1 for characteristics).

In our original meta-analysis using the bivariate model, we found a fair diagnostic accuracy of FENO for diagnosing asthma [12]. The overall sensitivity was 0.65 (95% CI: 0.58–0.72), and the overall specificity was 0.82 (0.76–0.86). The standard sROC curve according to Rutter and Gatsonis [43] is displayed in Fig. 1A. However, valuable information on thresholds was lost because the standard bivariate approach focuses on a single pair of sensitivity and specificity from each study, thereby ignoring the impact and value of the different thresholds applied.

For the present analysis, we extracted true test positives, true negatives, false positives, and false negatives for further 121 data points in addition to the 29 used in the standard analysis. Data were now available for a total of 150 data points, the median number of cutoffs per study was 4 with interquartile range 1–7, and a maximum of 19 cutoffs. There were 53 different cutoffs used over all studies (Table 2).

The application of the multiple thresholds model [16] leads to the mtsROC curve which allows estimating the sensitivities and specificities (Fig. 1B). For example, at the threshold where the Youden index was maximized, that was \(\text{FENO} = 26 \text{ parts per billion (ppb)}\), we found a sensitivity
Table 1. Characteristics of included studies

<table>
<thead>
<tr>
<th>First author</th>
<th>Year</th>
<th>Country</th>
<th>Age group (range years or mean and SD)</th>
<th>n</th>
<th>Prevalence</th>
<th>FE_{NO} measurement device</th>
<th>Number cutoffs available</th>
<th>Reference standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arora et al. [18]</td>
<td>2006</td>
<td>USA</td>
<td>Adults (17–38)</td>
<td>172</td>
<td>0.80</td>
<td>NIOX—no further specification</td>
<td>19</td>
<td>Bronchial provocation</td>
</tr>
<tr>
<td>Cordeiro et al. [19]</td>
<td>2011</td>
<td>The Netherlands</td>
<td>Mainly adults (7–87)</td>
<td>114</td>
<td>0.37</td>
<td>NIOX Flex (c)</td>
<td>1</td>
<td>FEV1/VC + reversibility and bronchial provocation</td>
</tr>
<tr>
<td>El Halawani et al. [20]</td>
<td>2003</td>
<td>USA</td>
<td>Adults (18–40)</td>
<td>49</td>
<td>0.14</td>
<td>Sievers 280 (c)</td>
<td>4</td>
<td>Exercise challenge</td>
</tr>
<tr>
<td>Florentin et al. [21]</td>
<td>2014</td>
<td>France</td>
<td>Young adults (mean 25 SD 3)</td>
<td>178</td>
<td>0.11</td>
<td>NIOX Mino (e)</td>
<td>1</td>
<td>FEV1/VC + reversibility and PEF variability</td>
</tr>
<tr>
<td>Fortuna et al. [22]</td>
<td>2007</td>
<td>Spain</td>
<td>Adults (18–64)</td>
<td>50</td>
<td>0.44</td>
<td>SIR N-6008 (c)</td>
<td>1</td>
<td>Bronchial provocation</td>
</tr>
<tr>
<td>Fukuhara et al. [23]</td>
<td>2011</td>
<td>Japan</td>
<td>Adults (48–66)</td>
<td>61</td>
<td>0.69</td>
<td>NA623N (c)</td>
<td>1</td>
<td>FEV1/VC (+ reversibility), bronchial provocation, sputum eosinophilia</td>
</tr>
<tr>
<td>Giovannini et al. [24]</td>
<td>2014</td>
<td>Italy</td>
<td>Mainly adults (mean 38 SD 15)</td>
<td>42</td>
<td>0.50</td>
<td>HypAir FE NO (e)</td>
<td>1</td>
<td>Bronchial provocation</td>
</tr>
<tr>
<td>Heffler et al. [25]</td>
<td>2006</td>
<td>Italy</td>
<td>Mainly adults (11–75)</td>
<td>48</td>
<td>0.38</td>
<td>NIOX Flex (c)</td>
<td>17</td>
<td>FEV1/VC + reversibility and bronchial provocation</td>
</tr>
<tr>
<td>Katsoulis et al. [26]</td>
<td>2013</td>
<td>Greece</td>
<td>Adults (22–37)</td>
<td>112</td>
<td>0.43</td>
<td>NIOX Mino (e)</td>
<td>6</td>
<td>Bronchial provocation</td>
</tr>
<tr>
<td>Kostikas et al. [27]</td>
<td>2008</td>
<td>Greece</td>
<td>Young adults (mean 21 SD 2)</td>
<td>149</td>
<td>0.42</td>
<td>NIOX Mino (e)</td>
<td>6</td>
<td>FEV1/VC + reversibility, bronchial provocation</td>
</tr>
<tr>
<td>Kowal et al. [28]</td>
<td>2009</td>
<td>Poland</td>
<td>Adults (18–45)</td>
<td>540</td>
<td>0.33</td>
<td>Sievers 280i (c)</td>
<td>10</td>
<td>FEV1/VC + reversibility, bronchial provocation, PEF variability, expert’s opinion</td>
</tr>
<tr>
<td>Linkosalo et al. [29]</td>
<td>2012</td>
<td>Finland</td>
<td>Children (6–19)</td>
<td>30</td>
<td>0.60</td>
<td>Sievers 280 (c)</td>
<td>5</td>
<td>Free running test</td>
</tr>
<tr>
<td>Malinovschi et al. [30]</td>
<td>2012</td>
<td>Denmark</td>
<td>Mainly adults (14–44)</td>
<td>108</td>
<td>0.42</td>
<td>NIOX Mino (e)</td>
<td>7</td>
<td>Expert’s assessment based on FEV1/VC + reversibility, bronchial provocation, PEF variability, expert’s opinion</td>
</tr>
<tr>
<td>Never smokers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ex-smokers</td>
<td></td>
<td></td>
<td></td>
<td>62</td>
<td>0.31</td>
<td>NIOX Mino (e)</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Current smokers</td>
<td></td>
<td></td>
<td></td>
<td>112</td>
<td>0.29</td>
<td>NIOX Mino (e)</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Pedrosa et al. [31]</td>
<td>2010</td>
<td>Spain</td>
<td>Adults (mean 34 SD 13)</td>
<td>114</td>
<td>0.31</td>
<td>NIOX Mino (e)</td>
<td>1</td>
<td>Bronchial provocation</td>
</tr>
<tr>
<td>Pizzimenti et al. [32]</td>
<td>2009</td>
<td>Italy</td>
<td>Not stated (probably adults)</td>
<td>156</td>
<td>0.09</td>
<td>NIOX Mino (e)</td>
<td>1</td>
<td>Bronchial challenge, spirometry</td>
</tr>
<tr>
<td>Sato et al. [33]</td>
<td>2008</td>
<td>Japan</td>
<td>Adults (20–78)</td>
<td>71</td>
<td>0.68</td>
<td>Kimoto, Osaka, Japan (c)</td>
<td>5</td>
<td>FEV1/VC (+ reversibility) and/or bronchial provocation, symptoms</td>
</tr>
<tr>
<td>Schleich et al. [34]</td>
<td>2012</td>
<td>Belgium</td>
<td>Adults (mean 41 SD 16)</td>
<td>174</td>
<td>0.47</td>
<td>NIOX Flex (c)</td>
<td>1</td>
<td>Bronchial provocation</td>
</tr>
<tr>
<td>Schneider et al. [11]</td>
<td>2013</td>
<td>Germany</td>
<td>Mainly adults (mean 43)</td>
<td>393</td>
<td>0.39</td>
<td>NIOX Mino (e)</td>
<td>13</td>
<td>FEV1/VC + reversibility and bronchial provocation</td>
</tr>
</tbody>
</table>

(Continued)
of 0.66 (95% CI: 0.56–0.74) and a specificity of 0.79 (95% CI: 0.73–0.84). For a clinician, presenting these test indices might be still not satisfying, as he or she wants to know the probability that the patient has or has not asthma when a resulting FENO value exceeds a certain threshold. Applying Bayes’ theorem allows calculating the predictive values, given the prevalence of the disease and the sensitivities and specificities for a threshold [44]. The multiple thresholds model now includes all this information, thus allowing to calculate the positive and negative predictive values related to the full range of threshold values and for different prevalences. Fig. 2A illustrates that making the diagnosis of asthma might only be possible in a meaningful manner when the prevalence of asthma (pretest probability) is at least 30%. In that case, a FENO > 50 ppb leads to a positive predictive value (PPV) of 0.76 (95% CI: 0.29–0.96).

Fig. 2B is probably even more useful for clinical practice. Here, the posttest probability is not estimated for whether the observed value is below or above a given threshold but simply based on the specific values actually measured. Again, Bayes’ formula is used for calculating the conditional probabilities of asthma, given a patient’s individual FENO values and based on the results of the meta-analysis incorporating all relevant and reported data of the primary studies. For example, the probability of having asthma is 0.62 (95% CI: 0.20–0.91), when FENO = 50 ppb in case for a pretest probability of 30%; 0.69 (95% CI: 0.24–0.94) when FENO = 60 ppb; and 0.73 (95% CI: 0.26–0.96) when FENO = 70 ppb. Though there is some uncertainty in these numbers, as illustrated by the 95% CI, this has an interesting interpretation: instead of only using the (dichotomous) information that the

<table>
<thead>
<tr>
<th>First author</th>
<th>Year</th>
<th>Country</th>
<th>Age group (range years or mean and SD)</th>
<th>n</th>
<th>Prevalence</th>
<th>FENO measurement device</th>
<th>Number cutoffs available</th>
<th>Reference standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sivan et al. [35]</td>
<td>2009</td>
<td>Israel</td>
<td>Children (5–18)</td>
<td>113</td>
<td>0.61</td>
<td>Eco Physics CLD88 (c)</td>
<td>3</td>
<td>Expert assessment based on FEV1/VC (+/− reversibility) and/or bronchial provocation, medication</td>
</tr>
<tr>
<td>Smith et al. [36]</td>
<td>2004</td>
<td>New Zealand</td>
<td>Mainly adults (9–72)</td>
<td>47</td>
<td>0.36</td>
<td>Logan LR2000 (c)</td>
<td>4</td>
<td>FEV1/VC + reversibility, bronchial provocation</td>
</tr>
<tr>
<td>Smith et al. [37]</td>
<td>2005</td>
<td>New Zealand</td>
<td>Mainly adults (14–71)</td>
<td>52</td>
<td>0.52</td>
<td>NIOX Flex (c)</td>
<td>2</td>
<td>FEV1/FVC reversibility, meta-choline challenge, ICS response</td>
</tr>
<tr>
<td>Tilemann et al. [38]</td>
<td>2011</td>
<td>Germany</td>
<td>Adults (38–58)</td>
<td>210</td>
<td>0.41</td>
<td>NIOX Mino (e)</td>
<td>6</td>
<td>Bronchial provocation or reversibility</td>
</tr>
<tr>
<td>Voutilainen et al. [39]</td>
<td>2013</td>
<td>Finland</td>
<td>Adolescents and adults (14–31)</td>
<td>87</td>
<td>0.34</td>
<td>NIOX Flex (c)</td>
<td>1</td>
<td>Expert’s assessment based on FEV1/VC + reversibility, bronchial provocation, PEF variability</td>
</tr>
<tr>
<td>Wang et al. [40]</td>
<td>2015</td>
<td>China</td>
<td>Mainly adults (13–89)</td>
<td>515</td>
<td>0.35</td>
<td>Nano Coulomb NO analyzer, Shangwo (u)</td>
<td>1</td>
<td>Depending on FEV1 predicted positive bronchodilatation or bronchial provocation</td>
</tr>
<tr>
<td>Bronchodilatation</td>
<td></td>
<td></td>
<td></td>
<td>408</td>
<td>0.31</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Bronchial provocat.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Woo et al. [41]</td>
<td>2012</td>
<td>Korea</td>
<td>Children (8–16)</td>
<td>245</td>
<td>0.68</td>
<td>NIOX Mino (e)</td>
<td>15</td>
<td>FEV1/VC + reversibility and bronchial provocation</td>
</tr>
<tr>
<td>Zhang et al. [42]</td>
<td>2011</td>
<td>China</td>
<td>Adults (15–70)</td>
<td>106</td>
<td>0.37</td>
<td>NIOX Mino (e)</td>
<td>1</td>
<td>FEV1 with bronchodilator, bronchial challenge test</td>
</tr>
</tbody>
</table>

**Abbreviations:** SD, standard deviation; ppb, parts per billion; FENO, fractional exhaled nitric oxide; FEV1, forced expiratory volume in 1 second; VC, vital capacity; PEF, peak expiratory flow rate; reversibility, positive reaction to a bronchodilator; ICS, inhaled corticosteroids; e, electrochemical; c, chemiluminescence; u, unclear.
patient’s value exceeds a certain threshold, we use the patient’s actual value, based on the consideration that the individual’s probability of having asthma continuously increases with the actual value. The potential for excluding asthma is illustrated in Figs 2C and D. Fig. 2C shows that excluding asthma might only be possible in a useful manner, when the pretest probability of asthma is 30% at maximum. In that case, a FENO $< 20$ ppb leads to a negative predictive value of 0.86 (95% CI: 0.66–0.95). Fig. 2D illustrates that in case of a pretest probability of 30%, the probability of not having the disease is 0.79 (95% CI: 0.61–0.90) when FENO $= 20$ ppb.

Moreover, we asked whether the diagnostic performance of FENO depended on covariates such as the diagnostic reference standard (diagnosis only with provocation) and the age group included in the study (adults vs. children). We found no significant impact for the diagnostic reference standard, which is in line with our earlier analysis using the bivariate model [12]. The specificities of FENO values were significantly higher in the children group when FENO $< 25$ ppb.

4. Discussion

The multiple thresholds model generates a more complete and more clinically useful picture of the effects of different thresholds in relation to their sensitivities, specificities, and predictive values. This enhances decision making about whether and how to use diagnostic tests devices or biomarkers in clinical practice. The new model shows that an optimal threshold for making the diagnosis of asthma might be around 60 ppb, and exclusion of asthma might be possible with 20 ppb, when the pretest probability of asthma is 30%.

For calculating posttest probabilities, which are crucial for thinking through downstream consequence, we need pretest probabilities. These can be derived from the included diagnostic studies or tailored toward the specific setting in which the test will be used. For example, previous studies have shown that the pretest probability of asthma is at least 30% in primary care setting, when patients with complaints suggestive of asthma are coming for the first diagnostic workup [45]. Related to the scenario of our clinical example based on the multiple thresholds model, a primary care physician could start the treatment with inhaled corticosteroids without the necessity to confirm the diagnosis with BP when FENO $> 50$ ppb is given, corresponding to a PPV of 0.76 (95% CI: 0.29–0.96). Such a PPV is similar to established BP testing procedures, which was shown to be around 70% when the pretest probability of asthma is around 30% [3,46]. Therefore, FENO measurement might be equivalent to BP. However, the probability of asthma increases remarkably with increasing FENO values as illustrated by the calculation of the exact conditional probabilities. Further studies have to evaluate if FENO measurement might be even superior to BP in some circumstances. For example, a recent study found that patients with negative BP response might benefit from inhaled corticosteroid steroids (ICSs) when FENO $> 33$ ppb because FENO measurement turned out to be more useful for predicting ICS responsiveness than diagnosing asthma [47]. However, the authors state that the chosen criteria were likely to be “oversensitive” in detecting a response.

The ATS guideline recommends to diagnose eosinophilic inflammation when FENO $> 50$ ppb [5]. However, the multiple thresholds model under consideration of pretest probabilities clearly illustrates the impact of the clinical setting where diagnoses are made. Making the diagnosis of
asthma with FENO values lower than 60 ppb might only be possible when enhanced pretest probabilities are given, like in selected patient in the hospital setting, and even FENO = 60 ppb still gives a potential 31% chance of the diagnosis being incorrect, when the pretest probability is 30%. A valuable strategy in primary care to enhance the diagnostic added value of point-of-care-testing instruments is to combine testing results with clinical signs and symptoms. It was shown recently that the diagnostic information of C-reactive protein measurement increases remarkably when interpreted in the clinical context of medical history and clinical examination [48,49]. This might also be true for FEnO measurement, when information about wheezing and allergic rhinitis is used for clinical interpretation [45]. In this prediction model, the probability of asthma was 78% when FEnO was at least 30 ppb and both clinical signs were present. However, this prediction model needs to be validated in a confirmatory study.

Excluding asthma in primary care might be possible with FEnO < 20 ppb. It must be considered that the pretest probability of being healthy might be around 70% and is only enhanced up to 79% (95% CI: 0.61–0.90) when FEnO = 20 ppb. On the other hand, the NPV of FEnO < 20 ppb would be equal to the NPV of a 20% fall of FEV1 during BP, which can be detected with spirometry maneuvers [50]. However, there is no consensus about strategies how to deal with uncertainty in terms of accepting probabilities for clinical decision-making. Beyond this, it needs being kept in mind that FEnO had a diagnostic blind spot in noneosinophilic asthma [11,51]. This is of importance when the presence of asthma is suggested by symptoms and clinical history, but FEnO is low.

While our findings represent substantial increases to the PPV and NPV at the thresholds, respectively, they illustrate also that the chance of the diagnosis being incorrect remains sufficiently large, so that asthma would have been neither “ruled in” nor “ruled out” as a possibility. This underlines that there is still some inherent uncertainty in making the diagnosis of asthma, also depending from the course of the disease [52,53]. Consequently, from a clinical point of view, patients should be reevaluated regularly to confirm, or eventually reject, the diagnosis of asthma.

### 4.1. Limitations

Some limitations of the new model have yet to be discussed. As well as the bivariate model, the multiple thresholds model is a two-stage approach, based on the estimated study-specific sensitivities, but ignoring the uncertainty of their variances at study level. In case of zeros in the two-by-two tables, continuity correction is needed; a practice that has been criticized [54]. The estimated CIs around the posttests probabilities are relatively wide, reflecting the uncertainty that was still present within this review about the exact shape of the summary ROC curve. A limitation of our model is that it sometimes fails to estimate a valid mtsROC curve if there is large between-study heterogeneity or if too many studies report only one threshold [16]. There are other approaches that have been suggested, such as the nonparametric gamma-frailty model presented by Putter et al. [55] and the multivariante model by Hamza.
et al. [56]. Both models assume a fixed set of thresholds for all studies, which for most biomarkers is not realistic in practice. Martínez-Camblor [57] suggested a nonparametric averaging of ROC curves, not using any threshold information. A more refined approach, a multivariate regression model, has been proposed by Riley et al. [58,59]. Our model makes a parametric assumption about the distribution of the biomarker. We used a log-logistic assumption, but other parametric assumptions are possible as well.

Regarding the sensitivity analysis, we found no impact of the reference standard on the results of the multiple thresholds model, which is in line with our earlier results using the bivariate model [12]. However, there were significant differences when FE\textsubscript{NO} < 25 ppb between the groups of children and adults. It must be kept in mind that the children’s group was dominated by a single study that contributed a large number of cutoff values [41].

5. Conclusion

The multiple thresholds model allowed to generate a more complete and more clinical useful picture of the
effects of different thresholds, which alleviates the derivation of decision rules for diagnosing or excluding asthma. However, a confirmatory study using the identified FENO values as predefined thresholds would be necessary to validate these findings. A prerequisite for summarizing DTA studies with the multiple thresholds model is that authors of primary studies report the pairs of sensitivities and specificities of all thresholds they examined of the index test or, even more preferably, the biomarker values of all individual patients along with their true disease status according to the reference standard. If reviewers encounter primary studies that report findings for multiple thresholds in sufficient detail, they should consider the multiple thresholds model or related methods. This would allow a more comprehensive interpretation of the clinical value of diagnostic tests producing continuous results.

Acknowledgments

Authors’ contributions: A.S., K.L., and G.R. had the idea to apply the multiple thresholds model of S.S. and G.R. to the data of the systematic review regarding the diagnostic accuracy of FENO. G.R. calculated and generated the figures. S.S. helped with calculating, interpreting, and writing. J.B.R. contributed to the concept and writing. A.S. and K.L. wrote the first draft. A.S. and G.R. revised the manuscript. All authors revised it critically for important intellectual content and approved the final version.

Consent for publication: Not applicable.

Ethics approval and consent to participate: No ethics committee approval was needed for this systematic review.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2017.09.001.

References


[44] Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding how physicians reason. Science 1959;130:9–21.


