Comparison of Centor and McIsaac scores in primary care:

a meta-analysis over multiple thresholds

INTRODUCTION

Pharyngitis is one of the most common reasons for consulting a GP. Over the winter period, around 6% of GP consultations in the UK tend to be for patients presenting with a sore throat, which represents more than 3.5 million consultations. Although, in many cases, pharyngitis has a viral aetiology, 20–35% of cases may be caused by bacteria — specifically, group A beta-haemolytic streptococcus (GABHS). Worldwide, infection with group A streptococci (GAS) places a significant burden on global health, and around 500 million people will die from GAS-related diseases each year.

In order to stratify patients most at risk of GABHS, the Centor score was developed. Each of four clinical features — absence of cough, purulent pharyngeal exudate, anterior cervical lymphadenopathy, and temperature of >38°C — is scored with 1 or 0, depending on whether it is present.

scores range from 0 (when none of the features are present) to 4 (when all are present). In the original study, conducted in an emergency department in the US, a score of 3 was associated with a 30.1–34.1% probability of GABHS. McIsaac independently derived a prediction system based on a cohort of patients from primary care. In essence, it modifies the Centor system to include an extra variable — age. For those aged between 3 years and 14 years, 1 is added to the score, whereas, for those aged ≥15 years, 1 is subtracted from the score; hence, a patient presenting with a sore throat may have a McIsaac score of anything between −1 and 5.

Many health systems have recommended the use of Centor or McIsaac scores in their guidelines to help manage patients with acute pharyngitis.

In the UK, the Centor score is one of two prediction rules recommended by the National Institute for Health and Care Excellence (NICE). Although the extent to which these rules are used in UK general practice is unclear, a recent survey of 266 GPs in Denmark reported that approximately half used the Centor score and 15% used the McIsaac score — this was in spite of the fact that the McIsaac score is the recommended rule in Denmark for diagnosing GABHS.

The question of which rule is likely to yield the most accurate diagnosis of GABHS for patients presenting to general practice is difficult to answer based on existing research. Only one primary study to date — reported in two articles by Fine et al. — provides the data to allow a direct comparison. Furthermore, comparisons at individual thresholds are meaningless, as those thresholds are not equivalent to each other — for example, a Centor score of 3 is not equivalent to a McIsaac score of 3, as the latter is calculated with an extra variable (that of age). To compare the tests, an overall assessment across all thresholds is required, such as may be provided by a receiver operating characteristic (ROC) curve.

Although meta-analysis allows the aggregation of multiple studies, either to
produce a summary (sensitivity, false positive rate) point or a summary ROC curve, both of these methods are constrained by the inclusion of only one (sensitivity, false positive rate) data point per study, where the false positive rate = 1 – specificity. When a study reports data at multiple thresholds, an arbitrary choice has to be made on which threshold to use when extracting the data for meta-analysis. Recent developments in meta-analysis methods allow this constraint to be relaxed so, if individual studies provide data at multiple thresholds, all of the data may be included for analysis; as such, the unit of interest for each study becomes its ROC curve and not just an individual (sensitivity, false positive rate) pair. This provides the basis for generating a summary ROC (SROC) curve for the Centor and McIsaac scores based on all the data reported in the primary studies.

Two researchers independently screened the title and abstracts of all citations identified. Full texts were obtained for those articles not excluded at the screening stage, and the same two investigators independently assessed the studies for eligibility based on the above criteria. Disagreements were resolved through discussion and achieving consensus.

Data collection and quality assessment
Data were extracted on the following study characteristics:

- aim;
- test evaluated;
- start and end date;
- method of subject recruitment;
- study location;
- description of study population;
- sample size;
- reference standard;
- conclusion of study authors;
- 2 × 2 contingency table data (true positives, false positives, true negatives, and false negatives) for each reported threshold on a per-patient basis; and
- any conflicts of interest.

How this fits in
In many healthcare systems, the Centor score and McIsaac score are used by GPs and primary care professionals to diagnose group A beta-haemolytic streptococcus (GABHS); however, there is no previous meta-analysis that has compared their performances in primary care. This comparative meta-analysis demonstrates that the Centor score and McIsaac score have broadly similar performance characteristics in diagnosing GABHS infection in primary care. A score of ≤0 when using either system may have a role in ruling out GABHS infection in primary care; however, neither score is sufficiently accurate to rule in GABHS infection and, if applied as recommended, could lead to more than one in two patients being prescribed antibiotics inappropriately. Other point-of-care diagnostics that augment these scores are needed if rates of inappropriate antibiotic prescribing are to be reduced.

Because of a lack of evidence supporting its use in test accuracy reviews, however, for completeness, a Google Scholar search was also performed using the terms ‘McIsaac score’ and ‘Centor score’. The searches were limited to studies published between January 1980 and February 2019. No restrictions were placed on the language of publication. Duplicate references were discarded to get a cohesive set of studies ready to be reviewed for inclusion.

Study eligibility criteria
Studies were included if:

- the study was a cross-sectional primary study;
- the study population consisted of unselected patients presenting with a sore throat to primary care;
- the study evaluated at least one of Centor or McIsaac scores;
- the target diagnosis was GABHS;
- the reference standard was culture from a throat swab; and
- sufficient data were reported to complete the 2 × 2 table for as many thresholds as possible.

Two researchers independently screened the title and abstracts of all citations identified. Full texts were obtained for those articles not excluded at the screening stage, and the same two investigators independently assessed the studies for eligibility based on the above criteria. Disagreements were resolved through discussion and achieving consensus.

Data collection and quality assessment
Data were extracted on the following study characteristics:

- aim;
- test evaluated;
- start and end date;
- method of subject recruitment;
- study location;
- description of study population;
- sample size;
- reference standard;
- conclusion of study authors;
- 2 × 2 contingency table data (true positives, false positives, true negatives, and false negatives) for each reported threshold on a per-patient basis; and
- any conflicts of interest.

METHOD
Data sources and searches
MEDLINE, EMBASE, and PsycINFO were searched for relevant studies; the search terms used are given in Supplementary Box S1. The data were supplemented by a manual review of the references from two published meta-analyses — one by Aalbers et al, the other by Willis and Hyde. The grey literature was not specifically searched.
The quality of each included study was assessed using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool,\(^1\) which assesses the risk of bias across a number of domains. The category ‘unclear’ was used when there was insufficient information reported in the study to come to a clear decision even after discussion.

The same two researchers who screened the initial abstracts independently extracted data, and performed the appraisal and quality assessment of each study. Disagreements were resolved through discussion and achieving consensus.

**Synthesis and meta-analysis methods**

The Different random Intercept Different random Slope (DIDS) model from the R package diagmeta (https://CRAN.R-project.org/package=diagmeta) was used to fit the data from the primary studies. This fits two linear mixed models — one for the false negative rate and one for the specificity — using the study as the grouping factor and allowing data from multiple thresholds for each study. Each linear mixed model has a random intercept and random gradient term, and the four random effects are assumed to have a four-dimensional multivariate normal distribution;\(^1\) these are used to generate an SROC curve.

An SROC curve and C-statistic (area under the curve [AUC]) was generated for Centor and McIsaac scores. Positive and negative likelihood ratios were derived for each of the thresholds with bootstrap confidence intervals (CIs). Assuming a null hypothesis that there is no difference between the C-statistics, the null distribution was derived empirically using a bootstrap sample of 1000. The level of significance was set to 0.05. For each test, the summary (sensitivity, false positive rate) pair corresponding to each threshold was also derived. Calibration plots of expected probabilities versus observed probabilities were derived for positive and negative test results after fitting an additive model to the logits of these probabilities using cubic splines.\(^1\) Each plot was corrected for optimism using a bootstrap sample of 1000 as recommended by Harrell.\(^1\)

**RESULTS**

### Study selection

The searches identified 80 citations. The full selection process (outlined in Figure 1) resulted in 18 studies\(^2\) being included in the review; 10 of these used the Centor score\(^2\) and eight used the McIsaac score.\(^6\) Only one study — reported by Fine et al\(^1\) — provided sufficient data to allow a direct comparison between the two tests.

A flowchart of the primary studies’ selection decisions is given in Figure 1.

**Study characteristics**

Full study characteristics are detailed in Table 1. Of those studies using the Centor score, eight were conducted in Europe\(^2\) and two in the US.\(^1\) Of those studies using the McIsaac score, four were conducted in Europe,\(^6\) and one in Australia.\(^6\) Three studies were translated from Spanish.\(^6\) The only study to provide data on both the Centor and McIsaac scores (Fine et al\(^1\)) had a sample size that was more than 100 times larger than the next-largest study.

The median prevalence of GABHS for the studies using the Centor score was 26.4% (range: 4.7%–42.0%); for studies using the McIsaac score, it was 23.0% (range: 12.7%–44.8%). Exact half of the studies using the Centor score provided data on all thresholds, and all studies provided data for two or more thresholds. A quarter of studies using the McIsaac score provided data on all thresholds, and all studies provided data for 24 thresholds. The ROC curves for each of the studies using the Centor score are shown in Figure 2; those for each of the studies using the McIsaac score are given in Figure 3.

For two of the included studies using the McIsaac score, McIsaac was listed as the lead author.\(^6\)

### Risk of bias and applicability

There is no validated statistic for measuring between-study heterogeneity across ROC curves; however, Figure 2 and Figure 3 show that, for both tests, the ROC curves are widely distributed; this suggests there is heterogeneity between studies for both tests.

For many of the studies,\(^2\) the reporting was inadequate, which introduced uncertainty when assessing the risk of bias — for example, the method of patient selection was not always described and it was not always clear whether any subjects had been excluded. Often, it was not reported whether the reference standard was carried out blind to the test results, although it is unclear whether knowledge of the test results would have greatly influenced the results of a cultured throat swab. In general, the study populations were considered representative of those seen in the different forms of primary care.
In two studies31,37, there were discrepancies between the number of subjects recruited and the number used in analyses, thereby increasing risk of biased estimates for the statistics of interest. In addition, although in one study that used the Centor score the target condition was largely GABHS, it also included group C and group G streptococcal infection as the target condition.25 This could potentially affect the applicability of the findings of this study. QUADAS-2 results are given in Supplementary Figure S1 (McIsaac score) and Supplementary Figure S2 (Centor score).

Synthesis of results
The sensitivities, specificities, and positive and negative likelihood ratios for each threshold are given for both scores in Table 2. Figure 4 shows the SROC curves for the Centor and McIsaac scores, with points on each curve corresponding to particular thresholds; it is clear that the curves are very close to each other and this is confirmed by the C-statistic. For the Centor score, the C-statistic was 0.6888 (95% CI = 0.653 to 0.724) and for McIsaac’s score it was 0.7052 (95% CI = 0.624 to 0.778); the 95% CIs are for the sensitivity given the specificity. From the empirical distribution of the difference between C-statistics, a difference of 0.0164 has a corresponding P-value of 0.419; this suggests there is no statistically significant difference between the C-statistics for the two curves.

Two post-hoc sensitivity analyses were carried out. The first investigated the effect of excluding the largest study (that by Fine et al11,12) and resulted in the C-statistics for the Centor and McIsaac scores being 0.6724 (95% CI = 0.610 to 0.731) and 0.7167 (95% CI = 0.632 to 0.788), respectively. As such, the effect is to decrease the C-statistic for the Centor score and to increase it for the McIsaac score. Again, the difference (0.0443) was not statistically significant (P = 0.188). In the second analysis, it was noted that two of the eight included studies that used the McIsaac score were led by the researcher who proposed it (namely, McIsaac);6,37 as such, only six studies evaluated the score independently. A sensitivity analysis was conducted in which the two studies led by McIsaac were excluded to evaluate the overall effects on the C-statistic. The C-statistic for the six independent studies was 0.6700 — lower than that when all studies were included in the analysis (0.7052) and that for the Centor score (0.6888).

The calibration plot for the post-test probabilities after a positive test result (positive predictive value [PPV]) for both scores, after correcting for optimism, is shown in Figure 4. The curves broadly coincide, with overfitting being particularly evident for expected PPVs above 0.5. Supplementary Figure S3 shows the calibration plot for the post-test probabilities for a negative test result after correcting for optimism. Here, the Centor score demonstrates better calibration than the McIsaac score. For the derivation of both calibration plots, the prevalence of GABHS is assumed to be known.

Whether either test could be used to rule in, or rule out, infection is not fully addressed by the AUC. For a GABHS infection prevalence of 25%, using Bayes’ theorem the expected PPV for a McIsaac score of 5 is 59%; however, from the calibration curve this expected PPV is likely

Figure 1. Flow chart of studies selected.
<table>
<thead>
<tr>
<th>Study</th>
<th>Location</th>
<th>Score</th>
<th>Year</th>
<th>Setting</th>
<th>Age, years</th>
<th>Sample</th>
<th>Prevalence, %</th>
<th>Thresholds</th>
<th>Reference standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alper24</td>
<td>Bursa, Turkey</td>
<td>Centor</td>
<td>May 2007 to Apr 2008</td>
<td>Emek Family Practice Centre</td>
<td>7–86</td>
<td>282</td>
<td>11.4</td>
<td>0,1,2,3,4</td>
<td>Throat culture</td>
</tr>
<tr>
<td>Fine11,12</td>
<td>26 states in US</td>
<td>Centor/McIsaac</td>
<td>1 Sep 2006 to 1 Dec 2008</td>
<td>Around SBI minute clinics in CVS chain across 26 states</td>
<td>3–≥55</td>
<td>206 870</td>
<td>27.1</td>
<td>0,1,2,3,4,5</td>
<td>DNA probe and throat culture for RADT negatives, RADT for test positives</td>
</tr>
<tr>
<td>Little25</td>
<td>UK</td>
<td>Centor</td>
<td>Jan 2007 to Oct 2008</td>
<td>General practices</td>
<td>≥5</td>
<td>1086</td>
<td>33.7</td>
<td>0,1,2,3,4</td>
<td>Throat culture</td>
</tr>
<tr>
<td>Marin Cañada26</td>
<td>Madrid, Spain (Spanish)</td>
<td>Centor</td>
<td>14 Feb 2005 to 12 May 2005</td>
<td>San Fernando 2 Health Centre</td>
<td>14–81</td>
<td>140</td>
<td>24.2</td>
<td>0,1,2,3,4</td>
<td>Throat culture</td>
</tr>
<tr>
<td>Lindbæk27</td>
<td>Stokke &amp; Kongsberg, Norway</td>
<td>Centor</td>
<td>Apr 2000 to Jun 2002</td>
<td>1 general practice in Stokke and 1 in Kongsberg</td>
<td>Children and adults</td>
<td>300</td>
<td>42.0</td>
<td>0,1,2,3,4</td>
<td>Throat culture</td>
</tr>
<tr>
<td>Atlas28</td>
<td>Massachusetts, US</td>
<td>Centor</td>
<td>1 Jul 2002 to 30 Jun 2003</td>
<td>2 primary care centres in Massachusetts General Hospital</td>
<td>Adults</td>
<td>148</td>
<td>25.7</td>
<td>0,2,3,4</td>
<td>Throat culture</td>
</tr>
<tr>
<td>Chazan29</td>
<td>Nazareth, Israel</td>
<td>Centor</td>
<td>Dec 1999 to Mar 2000</td>
<td>Primary care clinics of the Clalit Health Services</td>
<td>16–80</td>
<td>204</td>
<td>24.5</td>
<td>0,2,4</td>
<td>Throat culture</td>
</tr>
<tr>
<td>Seppälä30</td>
<td>Turku, Finland</td>
<td>Centor</td>
<td>Jan 1986 to Mar 1986</td>
<td>Private health centre Pulssi</td>
<td>15–62</td>
<td>106</td>
<td>4.7</td>
<td>0,3,4</td>
<td>Throat culture</td>
</tr>
<tr>
<td>Reguerras De Lorenzo31</td>
<td>Asturias, Spain (Spanish)</td>
<td>Centor</td>
<td>Jan 2008 to May 2010</td>
<td>5 primary care centres</td>
<td>2–14</td>
<td>192</td>
<td>38.5</td>
<td>0,3</td>
<td>Throat culture</td>
</tr>
<tr>
<td>Dagnelie32</td>
<td>Utrecht, Netherlands</td>
<td>Centor</td>
<td>1990 to 1992</td>
<td>53 GPs in general practice</td>
<td>4–60</td>
<td>558</td>
<td>32.8</td>
<td>0,3</td>
<td>Throat culture</td>
</tr>
<tr>
<td>Stefaniuk33</td>
<td>Warsaw, Poland</td>
<td>McIsaac</td>
<td>Mar 2014 to May 2014</td>
<td>Ortik General Practice</td>
<td>1–≥40</td>
<td>96</td>
<td>44.8</td>
<td>1,2,3,4,5</td>
<td>Throat culture</td>
</tr>
<tr>
<td>Mistik34</td>
<td>Kayseri, Turkey</td>
<td>McIsaac</td>
<td>Jun 2013 to Jun 2014</td>
<td>Burunamin Somyurek Family Medicine Centre</td>
<td>3–85</td>
<td>624</td>
<td>12.7</td>
<td>−1,0,1,2,3,4,5</td>
<td>Throat culture</td>
</tr>
<tr>
<td>Dunne35</td>
<td>Melbourne, Australia</td>
<td>McIsaac</td>
<td>Winter/spring of 2011 and 2012</td>
<td>3 general practices &amp; ED in tertiary hospital</td>
<td>3–72</td>
<td>127</td>
<td>18.9</td>
<td>−1,2,3,4</td>
<td>Throat culture and PCR</td>
</tr>
<tr>
<td>Tanz36</td>
<td>Chicago and Cincinnati, US</td>
<td>McIsaac</td>
<td>15 Nov 2004 to 15 May 2005</td>
<td>6 community-based paediatric offices</td>
<td>3–18</td>
<td>1848</td>
<td>29.9</td>
<td>0,1,2,3,4,5</td>
<td>Throat swab culture, 6 had RADT</td>
</tr>
<tr>
<td>Flores Mateo37</td>
<td>Barcelona, Spain (Spanish)</td>
<td>McIsaac</td>
<td>Mar 2008 to May 2009</td>
<td>2 primary care centres in Castelldefels</td>
<td>1–14</td>
<td>211</td>
<td>34.1</td>
<td>0,3,4,5</td>
<td>Throat culture</td>
</tr>
<tr>
<td>McIsaac38</td>
<td>Ontario, Canada</td>
<td>McIsaac</td>
<td>Oct 1998 and Mar 1999</td>
<td>97 family physicians from 49 communities</td>
<td>Children and adults</td>
<td>580</td>
<td>17.2</td>
<td>−1,1,2,3,4</td>
<td>Throat culture</td>
</tr>
<tr>
<td>McIsaac4</td>
<td>Toronto, Canada</td>
<td>McIsaac</td>
<td>Dec 1995 to Feb 1997</td>
<td>University-affiliated family medicine centre</td>
<td>3–76</td>
<td>503</td>
<td>12.9</td>
<td>−1,1,2,3,4</td>
<td>Throat culture</td>
</tr>
</tbody>
</table>

ED = emergency department. PCR = polymerase chain reaction. RADT = rapid antigen detection test.
to translate into an observed PPV of around 49% (Figure 5). Thus, if a score of 5 is used as the threshold for prescribing antibiotics, a PPV of 49% translates into more than one in two patients receiving antibiotics unnecessarily. Although the expected PPV would increase with GABHS prevalence, the calibration curves show this would not substantially affect the observed PPV; as such, neither test is effective at ruling in GABHS.

These results lead to the question of whether these criteria can be used to rule out infection. For a McIsaac score threshold of 1, a negative test corresponds to a score of −1, or 0. Similarly, at a threshold of 0, a negative test is a score of −1. From Table 2, the negative likelihood ratios (LR−) for the McIsaac score at thresholds 0 and 1 are 0.15 and 0.23, respectively. Thus, at a prevalence of GABHS infection of 25%, from Bayes theorem, low McIsaac scores such as −1 and 0 give expected probabilities of infection of 4.8% and 7.1%. Equally, for the same prevalence, a Centor score of 0 gives an expected probability of infection of 8.1%.

For low probabilities, the corresponding calibration curves are more erratic for the McIsaac score than the Centor score (Supplementary Figure S3), so it is not
clear how well these expected probabilities translate into practice. Nonetheless, given a shared decision between GP and patient on what constitutes an acceptable risk of GABHS, a low score on either criteria may be considered as sufficient evidence not to pursue treatment or further investigation.

Table 2. Sensitivity and specificity of the Centor and McIsaac scores at different thresholds

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Sensitivity, % (95% CI)</th>
<th>Specificity, % (95% CI)</th>
<th>LR+ (95% CI)</th>
<th>LR– (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Centor score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>97.2 (96.4 to 97.8)</td>
<td>10.1 (6.3 to 15.2)</td>
<td>1.08 (1.05 to 1.14)</td>
<td>0.28 (0.23 to 0.45)</td>
</tr>
<tr>
<td>2</td>
<td>84.4 (81.4 to 87.0)</td>
<td>36.7 (28.8 to 45.1)</td>
<td>1.33 (1.19 to 1.50)</td>
<td>0.43 (0.39 to 0.55)</td>
</tr>
<tr>
<td>3</td>
<td>54.4 (48.7 to 60.0)</td>
<td>72.4 (64.4 to 79.4)</td>
<td>1.97 (1.46 to 2.40)</td>
<td>0.63 (0.58 to 0.74)</td>
</tr>
<tr>
<td>4</td>
<td>21.5 (16.6 to 27.2)</td>
<td>93.7 (89.6 to 96.4)</td>
<td>3.41 (1.83 to 4.97)</td>
<td>0.84 (0.78 to 0.90)</td>
</tr>
<tr>
<td><strong>McIsaac score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>99.7 (99.0 to 99.9)</td>
<td>2.3 (0.3 to 10.7)</td>
<td>1.02 (1.00 to 1.10)</td>
<td>0.15 (0.09 to 0.35)</td>
</tr>
<tr>
<td>1</td>
<td>97.5 (94.7 to 99.0)</td>
<td>10.8 (2.8 to 28.4)</td>
<td>1.09 (1.02 to 1.30)</td>
<td>0.23 (0.17 to 0.39)</td>
</tr>
<tr>
<td>2</td>
<td>88.0 (82.0 to 93.8)</td>
<td>31.5 (14.2 to 54.3)</td>
<td>1.30 (1.09 to 1.76)</td>
<td>0.35 (0.28 to 0.47)</td>
</tr>
<tr>
<td>3</td>
<td>68.7 (57.4 to 78.5)</td>
<td>60.8 (39.9 to 78.8)</td>
<td>1.75 (1.28 to 2.79)</td>
<td>0.51 (0.44 to 0.58)</td>
</tr>
<tr>
<td>4</td>
<td>40.0 (28.5 to 52.5)</td>
<td>84.8 (70.8 to 93.4)</td>
<td>2.64 (1.68 to 4.95)</td>
<td>0.71 (0.62 to 0.78)</td>
</tr>
<tr>
<td>5</td>
<td>16.1 (8.9 to 26.2)</td>
<td>96.3 (90.0 to 98.1)</td>
<td>4.32 (2.38 to 10.0)</td>
<td>0.87 (0.79 to 0.93)</td>
</tr>
</tbody>
</table>

CI = confidence interval. LR+ = positive likelihood ratio. LR– = negative likelihood ratio.

DISCUSSION

Summary

This is the first meta-analysis to compare the performances of the Centor and McIsaac scores in a primary care setting over multiple cut points. Although there were 10 studies that evaluated the Centor score and eight that evaluated the McIsaac score, only one primary study provided data that allowed a direct comparison of the two tests.

The meta-analysis demonstrated that the SROC curves were broadly aligned, with the curve for the McIsaac score lying slightly above that for the Centor score (see Figure 4); however, the difference was marginal and no statistically significant difference between the AUCs was found. Moreover, when those studies authored by McIsaac were excluded, a sensitivity analysis revealed that the AUC for the McIsaac score may be overstated. Nonetheless, this did not alter the conclusion that the two prediction scores have similar performance characteristics and that adding an age variable does not appear to improve the accuracy of the Centor score for diagnosing GABHS in primary care. When compared with the Centor system, the McIsaac rule changes the operating points on the SROC curve rather than improving on discrimination. In addition, with AUCs of approximately 0.7, both systems appear to be, at best, fair at differentiating those patients who have GABHS from those who do not.

The calibration of the models for both scores demonstrates over-confidence, with the expected PPVs diverging substantially from the observed PPVs for probabilities of >40%. The effect of this is that an expected PPV of 80% translates into an observed PPV of 55%. Furthermore, these plots are ‘best cases’ as they are based on the prevalence of GABHS being known for the setting. When the prevalence is unknown, the average across all studies may be used; however, in the studies that were included in this review, the prevalence of GABHS ranged between 4.7% and 44.8%, so using the average prevalence would likely lead to poorer calibration as a result.

Strengths and limitations

All of the studies provided data at ≥2 thresholds, justifying the approach of using a model that accommodates both multiple cut points and different numbers of cut points between studies. This allowed the two criteria to be compared across the whole of the ROC space. Furthermore, by using calibration plots, it was possible to
provide evidence on each criteria’s likely performance in practice and when they are most likely to be useful to clinicians.

As a reference standard, the throat swab has limitations — its performance may depend on the operator and the conditions for incubation. Alternative reference standards, such as a rising titre of the antistreptolysin O (ASO) antibody, may be used, but these also vary with age, prevalence of streptococcus, and comorbidities. ASO testing is also rarely used by investigators; none of the included studies — or those excluded due to inadequacy of a reference test — used ASO testing.

The model used in this review benefits from being able to aggregate studies that provide data at multiple thresholds; however, this needs to be weighed against the necessity for continuity corrections when there are 0 cell entries in the 2 × 2 tables. Furthermore, at present, it is not clear how the DIDS model could include study-level covariates to investigate potential sources of heterogeneity.

Some authors have recommended the use of level-specific likelihood ratios. This requires defining test positives as test results that equal the threshold score only, not the threshold score and above, as is usual practice. This alternative definition of a test positive leads to an important property of a ROC curve (monotonicity) being violated; hence, with the approach used here, it is not possible to estimate level-specific likelihood ratios.

As part of internal validation, the authors used bootstrap methods to correct the calibration plots for optimism. Other methods have been proposed that use leave-one-out cross-validation to derive a validation statistic so the internal validity of the summary estimates may be assessed; it is also possible to use other information, such as the test positive rate, to derive an estimate that is tailored to the setting of interest. However, a shortcoming with all of these methods, including the method used here, is that they are rarely subject to external validation; without this, it is difficult to make assertions on the transferability of the results.

Comparison with existing literature

A recent review of guidelines for diagnosing acute pharyngitis revealed that both the Centor and McIsaac prediction scores are incorporated into guidance for Europe and North America. The Centor score is one of two prediction rules recommended for managing patients with a sore throat in the UK, while, in Denmark and Germany, the McIsaac score is recommended. This demonstrates that these scores are considered relevant to the diagnosis of acute pharyngitis in a number of countries. Therefore, it is perhaps surprising that only two reviews have evaluated the Centor score in primary care and no systematic review has evaluated the McIsaac score in primary care. None of the reviews to date have used a model that was able to accommodate data from multiple thresholds per study in the analysis. Previous reviews have treated each threshold separately when aggregating studies, thereby ignoring potential correlations between thresholds at a primary-study level and at an SROC curve level. Furthermore, none of the reviews have sought to establish how well the prediction rules calibrate in practice.

As a comparison, the two previous reviews on the Centor score reported positive likelihood ratios for a threshold of 3 — 2.68 (95% CI = 1.92 to 3.75) and 2.35 (95% CI = 1.51 to 3.67) — and these were inflated compared with the ratios presented here. However, the negative likelihood ratios for a threshold of 1 were comparable: 0.27 (95% CI = 0.16 to 0.46) and 0.28 (95% CI = 0.23 to 0.43). NICE has recently recommended using either the Centor or the FeverPAIN score.
to assess the symptoms of patients with acute pharyngitis.\textsuperscript{10} Although the latter was derived from a UK population, to date this is the only study on FeverPAIN\textsuperscript{25} and it is yet to be replicated in other independent populations; however, it is unclear whether the FeverPAIN score would lead to a marked improvement in discrimination and calibration, particularly when it shares many of the covariates of the scores that were reviewed here.

\textbf{Implications for practice}

Although the Centor score showed better calibration than the McIsaac system for a negative result, perhaps of more relevance is that, for estimated probabilities of <20\%, the observed probabilities of GABHS in practice, given a negative test result, are consistently lower than the corresponding estimates. On this basis, a Centor score of 0 or a McIsaac score of \( \leq 0 \) is likely to correspond to an actual risk of GABHS that is lower than the expected risk of 8.5\% — as such, it is likely to be sufficient to rule out infection.

For a Centor or a McIsaac score of \( \geq 1 \), it is less clear how to proceed. In general, the probability of GABHS for these scores is likely to be too high (\( >10\% \)) to rule out infection and too low to rule in infection. NICE’s current recommendation is that a Centor score of \( \geq 3 \) is sufficient grounds to consider prescribing antibiotics either immediately or as a delayed script with advice;\textsuperscript{10} however, the evidence presented here suggests that neither score can realistically identify patients with an observed risk of GABHS of \( >50\% \), irrespective of the expected risk. There is the potential that these recommendations could lead to inappropriate prescribing of antibiotics in a large percentage of cases.

In all instances, the GP should weigh up the public-health need to reduce the number of inappropriate antibiotic prescriptions and the individual patient’s need to treat a potential infection. With this in mind, an honest discussion with the patient about the likely GABHS risk and the GP’s obligation not to prescribe antibiotics inappropriately before deciding on management seems the most reasonable way to proceed.

Any substantive improvement in the diagnosis of GABHS-related pharyngitis is likely to require either a new prediction system or the use of point-of-care technologies to augment the existing clinical prediction tools.\textsuperscript{45}

\textbf{Funding}

Brian H Willis was supported by funding from a Medical Research Council Clinician Scientist award [ref: MR/N007999/1].

\textbf{Ethical approval}

No ethical approval was required as this is a secondary analysis of data derived from published primary studies.

\textbf{Provenance}

Freely submitted; externally peer reviewed.

\textbf{Competing interests}

The authors have declared no competing interests.

\textbf{Open access}

This article is Open Access: CC BY 4.0 licence (https://creativecommons.org/licenses/by/4.0/).

\textbf{Discuss this article}

Contribute and read comments about this article: bjgp.org/letters