Appendix 1.

Number and percentage of Lancefield group A,C, or G streptococci according to each level of 3 point 'basic' score (Model 3) including the variables significant in multivariate analysis in both data sets (1 point each for short prior duration, fever in the last 24 hours, and severely inflamed tonsils). The total number at each level, and percentage of the total sample are also shown.[1]

Score	0	1	2	3	Total
First data set					
Streptococci n (%)	16 (15%)	43 (19%)	80 (43%)	37 (67%)	176 (31%)
Total n (%)	*109 (19%)*	*221(39%)*	*187(33%)*	*55 (10%)*	*572(100%)*
Second data set					
Streptococci n (%)	22 (22%)	46 (25%)	81 (52%)	18 (82%)	167 (36%)
Total n (%)	100 (22%)	183 (40%)	156 (34%)	22 (5%)	461 (100%)

[1]For example taking the first column at the top of the table: there were 109 individuals with score 0 which represents 19% of the sample, and of those 109 individuals, 16(15%) had Lancefield Group A,C or G streptococci

Appendix 2. Calibration of Centor criteria and FeverPAIN

First data set

FeverPain	N of cga events/patients	Proportion Observed (95% CI)	Proportion Predicted (95% CI)
0	7/63	0.1111 (0.0327, 0.1895)	0.0898 (0.0609, 0.1305)
1	21/155	0.1355 (0.0813, 0.1897)	0.1602 (0.1246, 0.2034)
2	45/149	0.3020 (0.2279, 0.3761)	0.2692 (0.2310, 0.3111)
3	40/103	0.3883 (0.2936, 0.4831)	0.4158 (0.3656, 0.4677)
4	43/75	0.5733 (0.4604, 0.6863)	0.5789 (0.5029, 0.6514)
5	19/25	0.7600 (0.5888, 0.9312)	0.7265 (0.6319, 0.8043)

Chi squared test statistic = 2.33, p=0.6745 – no significant difference between the observed and predicted

Centor	N with cga (C, G,or A streptococci) events/patients	Proportion Observed (95% CI)	Proportion Predicted (95% CI)
0	3/45	0.0667 (0.0001, 0.1405)	0.0605 (0.0363, 0.0992)
1	10/88	0.1136 (0.0468, 0.1805)	0.1229 (0.0887, 0.1679)
2	45/199	0.2261 (0.1677, 0.2845)	0.2337 (0.1959, 0.2763)
3	65/152	0.4276 (0.3486, 0.5067)	0.3989 (0.3518, 0.4480)
4	55/97	0.5670 (0.4677, 0.6663)	0.5909 (0.5117, 0.6656)

Chi-squared test statistic = 0.92, p=0.8218 – no` significant difference between observed and predicted

Second data set

FeverPain	N with cga events/patients	Proportion Observed (95% CI)	Proportion Predicted (95% CI)
0	9/48	0.1875 (0.0756, 0.2994)	0.1341 (0.0914, 0.1924)
1	22/121	0.1818 (0.1126, 0.4366)	0.2194 (0.1732, 0.2738)
2	46/130	0.3538 (0.3703, 0.5832)	0.3377 (0.2930, 0.3856)
3	41/86	0.4767 (0.3703, 0.5832)	0.4806 (0.4223, 0.5395)
4	38/61	0.6230 (0.5000, 0.7459)	0.6268 (0.5415, 0.7048)
5	11/14	0.7857 (0.5621, 1.000)	0.7529 (0.6494, 0.8337)

Chi squared test statistic = 2.42, p=0.6597 – no significant difference between the observed and predicted

Centor	N with cga events/patients	Proportion Observed (95% CI)	Proportion Predicted (95% CI)
0	0/15	0.00	0.1436 (0.0938, 0.2137)
1	36/114	0.3158 (0.2299,0.4017)	0.2216 (0.1709, 0.2824)
2	36/157	0.2293 (0.1632, 0.2954)	0.3260 (0.2827, 0.3724)
3	69/138	0.5000 (0.4161, 0.5839)	0.4509 (0.3993, 0.5036)

Centor	N with cga events/patients	Proportion Observed (95% CI)	Proportion Predicted (95% CI)
4	47/81	0.5802 (0.4718, 0.6887)	0.5823 (0.4991, 0.6613)

Chi-squared test statistic = 16.39, p=0.0009 – significant difference between observed and predicted

Appendix 3. Secondary analyses.

- **Sequential approach**. This approach uses the first clinical score developed in the first data set, and validates it in the second. The score for the first data set had an Area under the ROC curve (AUC) of 0.759 (95% confidence intervals 0.719 to 0.800) in the first data set and 0.651 (0.600 to 0.702) in the second data set due to the poor performance of the constituent variable in the second data set.

- **Combined datasets.** In the combined data multivariate analysis resulted in all the FeverPAIN variables being significant and also muscle aches, cervical glands and very bad sore throat (i.e. 8 variables to be used in an extended score). This score has an AUC of 0.740 (0.708, 0.773), compared with 0.713 (95% CI 0.681, 0.745) for FeverPAIN, 0.683(0.649, 0.717) for Centor, and 0.710 (0.679 to 0.741) for the first clinical score. The AUC for FeverPAIN is significantly (p<0.05) better than for Centor, and the extended score is significantly better than FeverPAIN.

Exploring the omission of rapid attendance or substitution with muscle aches. Given concerns that the variable rapid attendance might be less generaliseable to other health care contexts we explored the implications for discrimination of both excluding rapid attendance (FeverPIN), or replacing the A of attendance with A for muscle aches (i.e.FeverPaIN with 'a' for aches) since muscle aches also fulfilled the criterion for inclusion in the score: discrimination was a little lower for FeverPIN (second data set AUC 0.698 (0.649, 0.746), first data set 0.713 (0.668, 0.758)) and similar for Fever PaIN (second data set 0.703 (0.654, 0.751); first data set 0.728 (0.683, 0.773)).

The use of a simple score compared with the exact coefficients. The AUC does not alter much comparing the model with the precise logistic coefficients with the model of the simple rounded score. For FeverPAIN in the first data set the AUCs were 0.735 (0.691, 0.779) vs 0.726 (0.682, 0.770) respectively and the second data set 0.713 (0.661, 0.757) vs 0.700 (0.650, 0.748).

Reporting based on STARD initiative

1. Title, abstract and key words	Page 1-2
2. Research question or aims	Page 3
3. Describe participants, inclusion criteria	Page 4
4. Recruitment mechanisms	Page 4
5. Participant sampling (e.g. consecutive)	Page 4
6. Data collection (prospective or retrospective)	Page 4
7. Describe reference standard and its rationale	Page 4
8. Describe technical specifications	Page 4
9. Definition and rationale for cut-off points of index test and standard	Page 4
10. Describe the number training of staff performing tests and standard	Page 4
11. Were the readers of the index test blinded ?	Page 4
12. Describe the methods for calculating or comparing measures, and describing uncertainty	Pages 4-6
13. Describe methods for calculating reproducibility (if done)	N/A
14.Report when the study was done	Page 7
15. Report clinical and demographic details	Page 7
16. Report how many participants did not undergo the index test/standard	Page 7
17 Report time intervals between index test and standard	Page 4
18. Report severity of disease in those with and without target condition	Tables1-3
19. Cross tabulation of index test by results of standard	Table 1
20. Report any adverse events (test or	N/A

standard)	
21. Estimates of diagnostic accuracy and of uncertainty (CIs)	Table 1
22. Report how indeterminate results missing responses and outliers were handled	Page 5
23. report estimates of diagnostic accuracy between readers	N/A
24. report estimates of reproducibility if done	N/A
25. Discuss clinical applicability of study findings	Pages 9-10