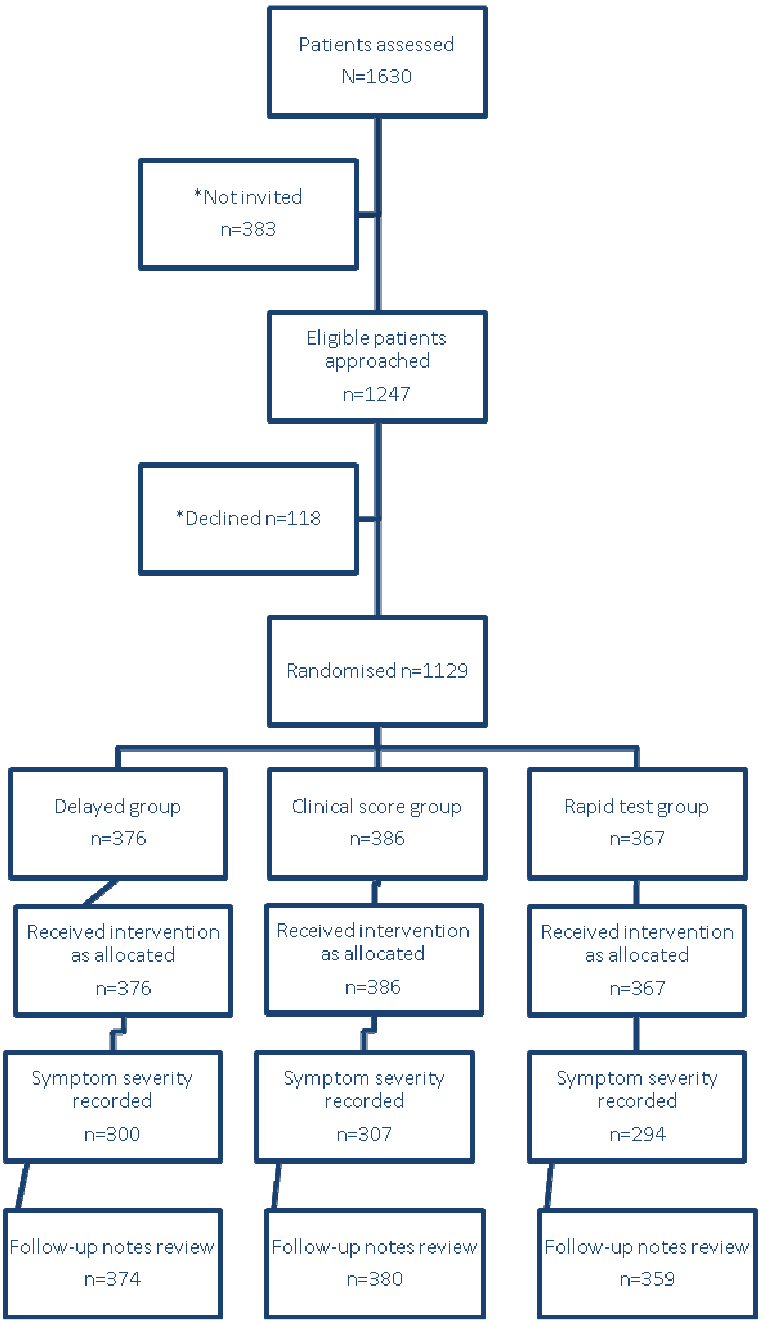
Appendix 1: [posted as supplied by author]

CONSORT trial flow diagram for first phase of the trial (Score1)



Development of the scores, evidence of differential effectiveness of the two scores and data for Score1

Patients age 5 or over presenting with acute sore throat were recruited for a diagnostic second cohort (cohort 2, n=517) consecutively after the first (cohort 1, n=606). This diagnostic part of the project involved different patients to the trial patients. All patients had a throat swab taken and sent to the microbiology laboratory and the same variables were collected in both cohorts.

An initial score (Score1) was developed from the first cohort (cohort1). The first diagnostic cohort documented that Lancefield Group C and G streptococci presented with very similar clinical features to group A streptococci. We developed a clinical score (Score1) which ranged from 0 to 6, and was based on a simple count of variables which independently predicted the presence of A,C and G streptococci (Area Under the Receiver Operator curve (AUC) 0.76): rapid attendance (short prior duration of 3 days or less), moderately bad or worse muscle aches, moderately bad or worse sore throat, the absence of a bad cough, severely inflamed tonsils and anterior cervical glands.

Score2. The 'classic' approach to the development and validation of clinical scores is a sequential approach - to develop the score in one data set, and due to the problem of over-fitting in one data set, to then validate in another data set. This was our original intention, but some variables included in Score1 did not perform well in the second data set (severity of sore throat, cervical glands) and some variables not included in the first score were significant in the second data set (fever, pus). This poor consistency resulted in poor discriminatory performance of Score1 when used in the second data set (AUC 0.65). Since one data set was clearly was insufficient to identify variables which performed consistently we used both data sets to identify variables, and used bootstrapping to overcome the problem of over-fitting. The clinical features independently predicting the presence of these streptococci in multivariate analysis in both cohorts were: rapid attendance (short prior duration of 3 days or less; multivariate adjusted odds ratio 1.92 cohort 1, 1.67 cohort 2); fever in the last 24 hours (1.69, 2.40); and doctor assessment of severity (severely inflamed pharynx/tonsils (2.28, 2.29). Absence of coryza or cough and purulent tonsils were also significant predictive variables in univariate analysis in both cohorts and in multivariate analysis in at least one cohort. Over and above the most basic model (short prior duration, severe inflammation, fever) the choice of additional variables to include (pus and 'absence of cough and coryza') was determined by consensus, including a consideration of the strength of prior evidence, but omission of key variables or substitution did not have major effects on the discrimination.

A 5 item score based on Fever, Purulence, Attend rapidly (3 days or less), severely Inflamed tonsils, and No cough or coryza (acronym FeverPAIN) had moderate discrimination (bootstrapped estimates of area under ROC curve (AUC) 0.73 cohort 1, 0.71 cohort 2) and was more consistent than the Centor criteria (AUCs cohort 1 0.65, cohort2 0.72). FeverPAIN performed well in identifying a substantial number of participants at low risk of streptococcal infection (38% in cohort 1, 36% in cohort 2 scored <=1, associated with a streptococcal percentage of 13% and 18% respectively). A Centor score of <=1 identified 23% and 26% of participants, with streptococcal percentages of 10% and 28% respectively.

The alternative approach to developing a clinical score of combining data sets to increase power provided an 8 variable score with improved discrimination, but was unwieldy for clinical purposes, and hid the considerable variability between data sets in performance of both individual variables and also the performance of the first score. Further support for the poor clinical utility of the first score also comes from the trial. Although the estimates of AUCs were bootstrapped, which provides some protection against overfitting, there is still further need to validate FeverPAIN in another large cohort.

Evidence of differential effectiveness comparing the first and second parts of the trial.
There was significantly greater improvement in symptom scores in the first 3 days following the index consultation when using the second clinical score compared to the first score (interaction term -0.38, -0.76 to -0.01, p=0.043) and lesser improvement for the rapid test group (interaction term -0.18, -0.55 to 0.19; p=0.345). There was also a significantly larger effect on symptom resolution when using the second score (interaction term hazard ratio (HR) 1.35, 1.01 to 1.79; p=0.043) but little difference when the rapid test was used (interaction term HR 1.01, 0.76 to 1.35; p=0.93). Similarly there was a significantly greater effect on antibiotic use when using the second score (interaction term odds ratio 0.43 (0.24 to 0.76);p=0.004) and a lesser effect for the rapid test group (0.74, 0.42 to 1.32;p=0.31). Thus the results of the trial when using the second score (FeverPAIN) are presented in the text as the main findings, and the results of the first score shown in appendix 1 - which show no significant differences for any outcome despite good compliance with the intended strategies in each group (see below).

Data for Score1:

Table A. Symptom severity, antibiotic use, intention to consult in the future (moderately likely or more likely), and reconsultations with sore throat for Score 1. Results are risk ratios (95% confidence intervals) or mean differences (95 % confidence intervals).

		Delayed prescribing (Control)	Clinical Score	Rapid
Mean severity of sore throat and difficulty swallowing in the 2-4 days after seeing the doctor (0 = no problem to 6 as bad as it could be)				
	Crude mean	2.95 (1.44)	3.05 (1.49)	2.83 (1.50)
	Mean difference*		0.06 (-0.15 to 0.28; p=0.560)	-0.12 (-0.34 to 0.10; ;p=0.270)
Duration of symptoms rated moderately bad or worse				
	Hazard ratio	1.00	0.95 (0.80 to 1.13; p=0.543)	1.10 (0.92 to 1.31;p=0.282)
Antibiotic use				
	Crude percentage	111/284 (39%)	137/294 (47%)	98/281 (35%)
	Risk ratio*	1.00	1.20 (0.99 to 1.42; p=0.059)	0.88 (0.69 to 1.09;p=0.265)
Belief in the need to see the doctor in future episodes				
	Crude percentage	91/278 (33%)	79/285 (28%)	76/273 (28%)
	Risk ratio*		0.85 (0.64 to 1.09;p=0.205)	0.86 (0.65 to 1.10; p=0.248)
Return to the surgery within one month with sore throat				
	Crude percentage	43/374 (11%)	34/380 (9%)	46/359 (13%)
	*Risk ratio	1.00	0.76 (0.49 to 1.16; p=0.205)	1.11 (0.74 to 1.62;p=0.618)
Return to the surgery after one month with sore throat (mean follow up 0.73 years)				
	Crude percentage	75/374 (20%)	84/380 (22%)	69/359 (19%)
	*Risk ratio	1.00	1.10 (0.83 to 1.44; p=0.488)	0.95 (0.70 to 1.27;p=0.728)

*all models control for baseline severity of sore throat and difficulty swallowing and fever during the previous 24 hours
Model for return within one month also controlled for prior antibiotic use;
Model for returns after one month additionally controlled for prior attendance with sore throat, and follow-up duration

Table B. Strategy used by clinician when using score 1.

	Control (delayed prescription)	Clinical score	Rapid test	
Strategy used by clinician				
No offer of Antibiotics	30/376 (8%)	122/385 (32%)	200/367 (55%)	
Immediate antibiotics	20/376 (5%)	74/385(19%)	53/367 (14%)	
Delayed antibiotics	326/376 (87%)	189/385 (49%)	114/367 (31%)	

Compliance with intended strategy for Score1.

As with Score2 there was good compliance with the intended strategy in each group: overall in 88% (982/119) of consultations the intended strategy was adhered to., and when delayed prescribing was advised , 468/ 612 (76%) of patients were advised to wait at least 5 days.