Supplementary Appendix

Cross-validation method

A common method of cross-validation is to split the study data into a training and validation samples. In this approach, a statistical model is developed on the training sample and its performance is then assessed against the validation sample. One limitation of this approach is that only a subset of the data is used for model building, and when sample size is limited (as is often the case) this can lead to models that are more unstable than they would have been had the whole sample been used to generate them.

Various alternative methods have been proposed to address this problem, including *K*-fold cross-validation, leave-one-out cross-validation, and bootstrapping methods that adjust for optimism.[1,2] We chose the latter approach because it is relatively simple to implement using standard statistical software and because it has been in other studies that have developed similar kinds of risk scores (see, for example, Cook at al [3]).

Cross-validating a model using bootstrapping methods that adjust for optimism involves three main steps. First, a statistical model is developed using the entire dataset (typically using logistic or linear regression analyses). A measure of fit is estimated for this model (e.g. R-squared or *c*-statistic), which may be called “the overall fit statistic”. Second, samples are drawn with replacement from the entire dataset (“bootstrap samples”) and the model developed in the previous step is re-estimated for each of these samples, with fit statistics calculated for each one (“bootstrap fit statistics”). Importantly, the coefficient values estimated from each bootstrap sample are also applied to the entire dataset and overall fit statistic is re-calculated (“population fit statistics”). The average of the difference between the population fit statistics and the bootstrap fit statistics represents the “optimism” that biases the overall fit statistic. Therefore, the final step is to subtract the value of the optimism statistic from the value of the overall fit statistic. This value is reported as the measure of model fit, “adjusted for optimism”.

Applying this methodology to our analysis, we considered three main multivariable logistic regression models as candidates for the predictive model on which to base our risk score. What follows is a specific description of the cross-validation of Model 1 (see Table 2 in the main paper); the same approach was used for the other two models.

The overall fit statistic for Model 1 was a *c*-statistic of 0.6934. After drawing 200 bootstrap samples, the average bootstrap fit statistic was a *c*-statistic of 0.6943. When the coefficients from each of the bootstrap models were applied to the entire dataset, the average population fit statistic was a *c*-statistic of 0.6935. The difference between the average population fit statistic and the average bootstrap fit statistic is 0.008 (0.6935 - 0.6943), which represents the measure of optimism in the overall *c*-statistic. Thus, after adjustment for optimism, the *c*-statistic for Model 1 is 0.6925 (0.6934 – 0.0008).

Using clinical care complaints only

In the paper we report the results of multivariate logistic regression using all complaints. Here we report the results using just those complaints that related to clinical care. The three models correspond to the three models described in the method and results section.

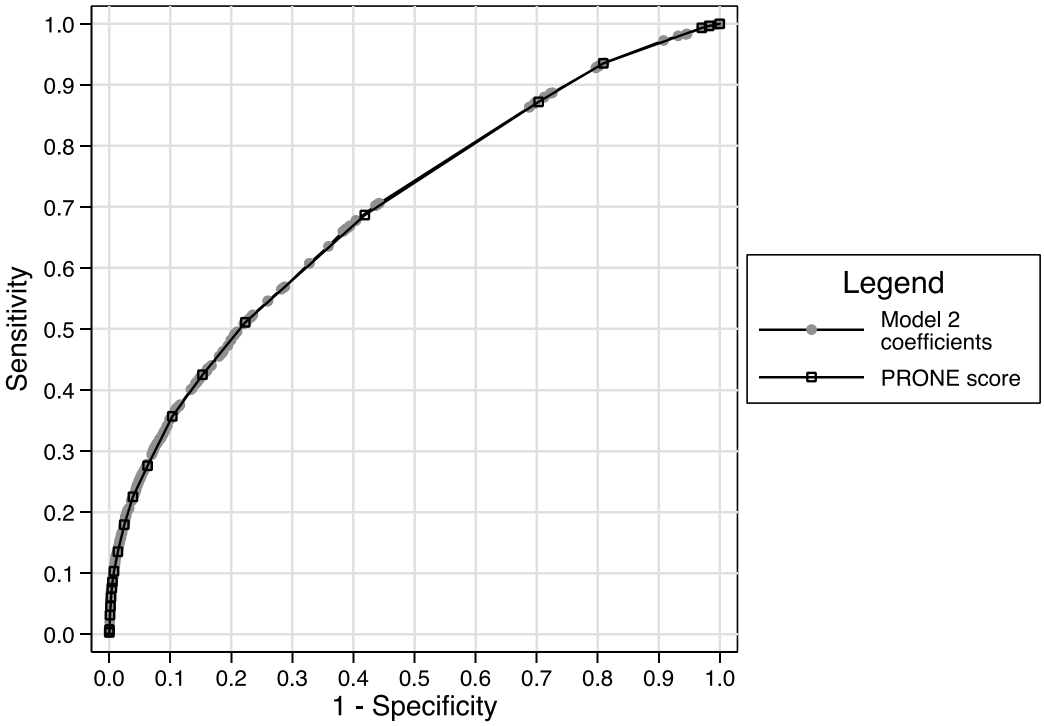
Table S1: Logistic regression models for risk of complaints within 2 years

	Model 4 **OR (95% CI)**	**Model 5** **OR (95% CI)**	**Model 6** **OR (95% CI)**
Complaint number			
1 (ref)	1.00	1.00	1.00
2	1.30 (1.09 to 1.55)	1.29 (1.09 to 1.53)	1.97 (1.74 to 2.23)
3	2.08 (1.64 to 2.63)	2.03 (1.63 t0 2.55)	3.43 (2.84 to 4.14)
4	2.73 (2.00 to 3.71)	2.83 (2.10 to 3.80)	4.95 (3.77 to 6.50)
5	5.04 (3.14 to 8.08)	4.74 (3.05 to 7.36)	8.87 (5.89 to 13.4)
6	6.46 (3.30 to 12.64)	7.01 (3.62 to 13.58)	15.68 (8.40 to 29.3)
7	5.12 (2.46 to 10.66)	4.97 (2.49 to 9.91)	10.37 (5.41 to 19.9)
8	5.34 (2.18 to 13.09)	5.87 (2.45 to 14.08)	13.22 (5.78 (30.2)
9	5.70 (1.78 to 18.26)	5.29 (1.83 to 15.31)	12.12 (4.52 to 32.5)
10+	33.82 (11.24 to 101)	28.90 (11.49 to 72)	60.6 (26.1 to 140)
Doctor’s specialty			
Anaesthesia (ref)	1.00	1.00	
Radiology	0.89 (0.37 to 2.13)	0.99 (0.43 to 2.29)	
Other specialties	1.25 (0.73 to 2.15)	1.28 (0.78 to 2.11)	
Internal medicine	1.44 (0.96 to 2.15)	1.52 (1.02 to 2.26)	
Ophthalmology	1.89 (1.17 to 3.05)	2.10 (1.32 to 3.35)	
General practice	1.84 (1.26 to 2.68)	1.98 (1.37 to 2.85)	
Psychiatry	2.28 (1.48 to 3.52)	2.30 (1.51 to 3.51)	
Orthopaedic surgery	2.48 (1.63 to 3.77)	2.71 (1.81 to 4.08)	
Other surgery	2.42 (1.60 to 3.67)	2.62 (1.75 to 3.92)	
General surgery	2.32 (1.50 to 3.59)	2.70 (1.78 to 4.09)	
Obstetrics and gynaecology	2.76 (1.81 to 4.21)	2.93 (1.95 to 4.41)	
Dermatology	3.28 (1.99 to 5.43)	3.71 (2.24 to 6.15)	
Plastic surgery	4.59 (2.93 to 7.17)	5.06 (3.30 to 7.77)	
Time since previous complaint			
1 year or more (ref)	1.00	1.00	
6 months to 1 year	1.32 (1.03 to 1.70)	1.41 (1.11 to 1.80)	
Less than 6 months	1.82 (1.46 to 2.28)	1.94 (1.57 to 2.40)	
Doctor’s sex			

Female (ref)	1.00	1.00	
Male	1.61 (1.36 to 1.90)	1.70 (1.45 to 2.00)	
Doctor’s age			
22-34 years (ref)	1.00		
35-65 years	1.66 (1.21 to 2.28)		
Location of practice			
Rural (ref)	1.00		
Urban	1.13 (0.99 to 1.29)		
C-statistic (adjusted for optimism)	0.69	0.70	0.65

Performance of PRONE score

Figure S1: Receiver-operating characteristic curves showing the performance of the logistic regression model and the 22-point Complaint PRONE score in predicting risk of complaint within 2 years



References

1. Efron B, Tibshirani R (1993) An Introduction to the Bootstrap. Chapman & Hall/CRC. 1 pp.
2. Trevor H, Robert T, Jerome F (2001) The elements of statistical learning: data mining, inference and prediction. New York: Springer-Verlag.
3. Cook NR, Buring JE, Ridker PM (2006) The effect of including C-reactive protein in cardiovascular risk prediction models for women. Ann Intern Med 145: 21–29.