Staged Identification of CAP in Fever Patients Across Epidemic Environments: Modeling &Validation

Gao Ziheng, Chen Tengfei, Ha Yanxiang, Shi Yifan, Xu Xiaolong, Li Bo, Liu Qingquan

1. Modeling Sample Size Calculation

2. R Packages Used in Work

3. External Validation Cohort's Clinical Characteristics

Supplementary Table 1 Comparison of Characteristics between Pneumonia and Non-Pneumonia in the External Validation Cohort
Figure S1 The Heatmap of IDI Analysis for the α/β Models in External Validation
Figure S2 Comparison of the ROC Curves for the Final α/β Models

4. Discussion of CAP Diagnostic Criteria

Figure S3 Series of Latent Class Analysis Charts for Automatic Subtype Classification of internal Training Cohort
Figure S4 The Correlation Heatmap among Clinical Variables in the Training Cohort

5. TRIPOD Checklist

Supplementary Table 2 TRIPOD Checklist (Prediction Model)

1. Modeling Sample Size Calculation

This study employed traditional Logistic regression and five machine learning algorithms for modeling. The sample size was calculated based on the Events per Predictor (EPP) rule. The EPP rule is commonly applied to traditional Logistic regression models to ensure model robustness and avoid overfitting. Although machine learning algorithms can learn features more deeply and require a smaller sample size, we still adopted this stricter standard to ensure the robustness of the models (Each predictor requires at least 15 events). In the modeling of the α/β model, even the traditional Logistic regression model, which performed worse than the final machine learning model, achieved an AUC value above 0.75 both in internal/external validation. This result indicates that the variables involved in this clinical problem do have objective associations. The model is constructed based on actual clinical data and the intrinsic relationships between variables, rather than being built for the sake of modeling. For the α model, which includes 7 clinical variables, there should be more than 105 = (15×7) pneumonia patients. For the β model, which introduces an additional 4 clinical variables, there should be more than 165 = (15×(7+4)) pneumonia patients. Considering all predictors and interaction terms used in the α/β models, there should be more than 285 = (15×(7+6+4+2)) pneumonia patients. Since the training cohort includes 362 pneumonia patients, the sample size requirement is met for modeling.

2. R Packages Used in Work

glmnet, ggplot2, corrplot, gridExtra, caret, pROC, xgboost, ada, ROCR, shiny, shapviz, car, DALEX, reshape2, klaR, gbm, readxl, dcurves, catboost, dplyr, compareGroups, forestplot, boot, readxl, reshape2, rms, patchwork, openxlsx, poLCA packages are used in our work with RStudio. The modeling of the five algorithms—traditional Logistic regression, Logisticnet, Randomforest, XGBoost, and AdaBoost—is based on the *caret* package and other supplementary packages. The modeling of the CatBoost algorithm is based on the *CatBoost* package.

3. External Validation Cohort's Clinical Characteristics

After completing the model construction, we further compared the clinical characteristics between pneumonia patients (N=24) and non-pneumonia patients (N=186) in the external validation cohort (**Supplementary Table 1**). It is worth noting that this external validation cohort did not screen for disease diagnosis but included all patients who visited the fever clinic. Therefore, it more closely reflects the real-world clinical environment, differentiating not only patients with pneumonia-like symptoms but also covering all patients who visited the fever clinic.

In the real-world clinical setting, medical records in emergency and fever clinics often have missing data. Although we randomly selected 300 individuals from over 2,700 visits over six months, 90 cases were excluded from the final external validation cohort due to not meeting the inclusion and exclusion criteria. This situation may be related to the purpose of the patients' visits. Some patients visited the clinic merely to obtain and purchase medications as required by regulations, rather than for actual treatment. As a result, clinicians did not document their medical records properly. This phenomenon may lead to some cases not meeting the study's inclusion and exclusion criteria, thereby affecting the composition of the final cohort.

When comparing the clinical characteristics of the 210 patients ultimately included, we found that only age, cough, and CRP showed significant differences between groups (all $P<0.05$). Compared to the 7 clinical variables that previously showed significant differences in the internal training cohort, variables that previously had high OR, such as altered mental status and dyspnea, as well as Tmax, days, and pharyngeal discomfort, which are important in the α model, did not show significant differences in this comparison. However, after further examining interaction effects, we found that the interaction terms between age and Tmax, and between age and days of illness, were significantly different between groups (all $P<0.05$).

Despite the significant differences in clinical characteristics between groups in the internal training and external validation cohorts, the α model demonstrated stable AUC values in both cohorts (both AUC=0.80), which further confirms the robustness of our model. In the modeling of both the α and β models, we utilized the CatBoost algorithm, which exhibited the smallest changes in AUC values during internal and external validation. This algorithm not only outperformed other modeling methods but also demonstrated greater stability.

Additionally, it showed significant advantages in IDI analysis (**Figure S1**). After upgrading to the β model, the AUC value in the internal validation significantly increased (DeLong Test P<0.001), while the increase in the external validation was not significant (DeLong Test P=0.748) (**Figure S2**). This discrepancy may be due to the differences in laboratory test indicators between the two cohorts. This finding suggests that in future modeling, we need more data from different epidemic environments to further optimize the model's performance.

Supplementary Table Comparison of Characteristics between Pneumonia and Non-Pneumonia in the External Validation Cohort

Characteristic		Non-pneumonia (N=186)	Pneumonia (N=24)	Odds Ratio 95%CI	P value
Age(y)		40.1 (16.4)	55.0 (19.7)	1.04 [1.02-1.07]	0.001**
Days(d)		2.23 (2.21)	4.58 (6.28)	1.17 [1.04-1.32]	0.081
Tmax(℃)		38.3 (0.69)	38.5 (0.88)	1.62 [0.90-2.91]	0.190
Pharyngeal discomfort	No=0	65 (34.9%)	9 (37.5%)	0.89 [0.37-2.25]	0.984
	Yes=1	121 (65.1%)	15 (62.5%)		
Cough	No=0	77 (41.4%)	2 (8.33%)	7.23 [2.03-49.9]	0.003**
	Yes=1	109 (58.6%)	22 (91.7%)		
Dyspnea	No=0	184 (98.9%)	22 (91.7%)	8.20 [0.82-82.1]	0.065
	Yes=1	2 (1.08%)	2 (8.33%)		
Altered mental status	No=0	183 (98.4%)	22 (91.7%)	5.59 [0.62-38.7]	0.101
	Yes=1	3 (1.61%)	2 (8.33%)		
Age*tmax		1536 (626)	2121 (779)	1.00 [1.00-1.00]	0.001**
Age*days		90.1 (106)	258 (338)	1.00 [1.00-1.01]	0.024*
Age*altered mental status		0.96 (8.11)	8.00 (27.1)	1.03 [1.00-1.05]	0.219
Age*pharyngeal discomfort		25.2 (21.9)	31.5 (28.1)	1.01 [0.99-1.03]	0.306
Age*cough		22.7 (22.5)	47.2 (21.4)	1.05 [1.02-1.07]	<0.001***
Age*dyspnea		0.71 (7.30)	7.29 (25.0)	1.03 [1.00-1.06]	0.212
NLR		5.56 (4.24)	7.76 (11.7)	1.05 [0.99-1.11]	0.368
CRP		22.2 (30.9)	45.2 (45.6)	1.01 [1.00-1.02]	0.024*
PLT		221 (64.6)	226 (70.7)	1.00 [0.99-1.01]	0.729
CRP/PLT		0.12 (0.19)	0.20 (0.21)	5.69 [1.04-31.2]	0.056

Mean(SD) ; n(%), *P<0.05, **P<0.01, ***P<0.001.

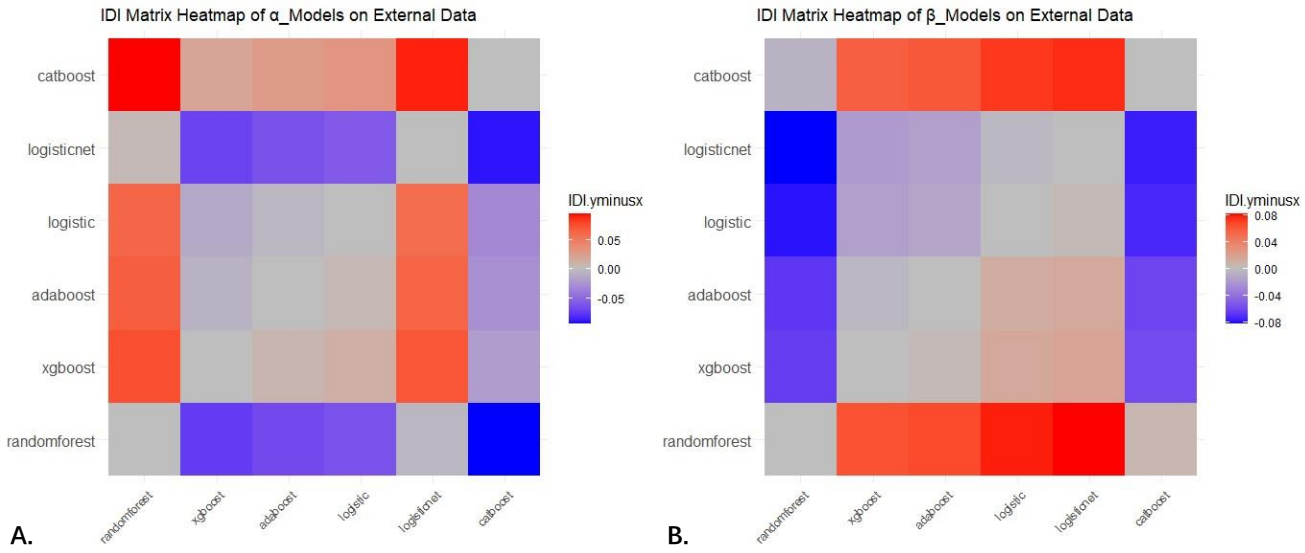

Figure S1 The Heatmap of IDI Analysis for the α/β Models in External Validation (**A.** α_model's IDI analysis; **B.** β_model's IDI analysis)

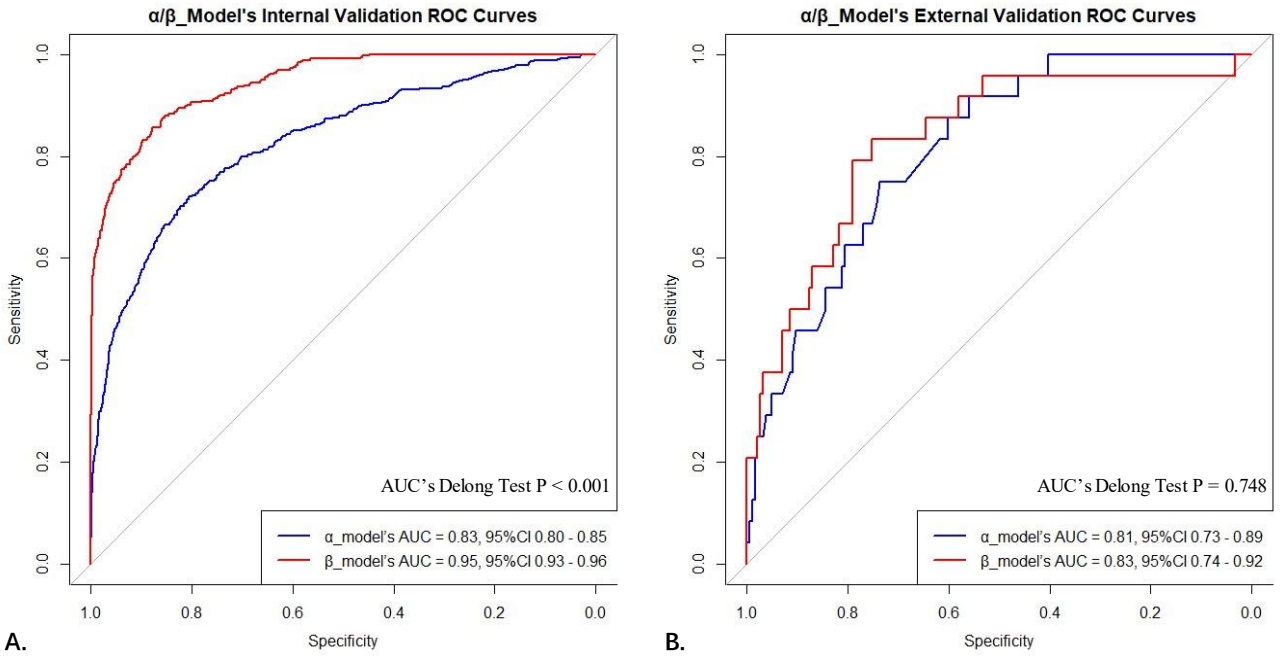

Figure S2 Comparison of the ROC Curves for the Final α/β Models (**A.** internal validation; **B.** external validation)

4. Discussion of CAP Diagnostic Criteria

The existing CAP diagnostic criteria provide vague descriptions of clinical symptoms, requiring only a few symptoms related to pulmonary infection. Unless the patients condition is severe, cough and sputum production alone are not sufficient to clinically distinguish pneumonia from patients with similar symptoms (such as upper respiratory infections). However, our model incorporates a wide range of clinical symptoms and further differentiates various clinical subtypes of CAP based on these symptoms. As a further exploration based on existing guidelines, we conducted unsupervised automatic latent class analysis in 1,781 patients from the internal training and validation cohorts (**Figure S3**). The results showed that, in addition to the obvious symptoms of altered mental status and dyspnea, pneumonia patients were distinguished from another class by the combination of clinical manifestations including fatigue, pharyngeal discomfort, head&body pain, nasal&congestion, and abdominal pain. We further constructed a heatmap of the correlations between all clinical feature variables among these patients. Many clinical symptoms, as well as laboratory tests (such as WBC), showed clear correlations with pneumonia (**Figure S4**). Thus, we suggest that in the future updates of CAP guidelines, more research on the diagnostics of clinical symptoms should be conducted, not only to identify CAP patients but also to differentiate the clinical subtypes of CAP patients.

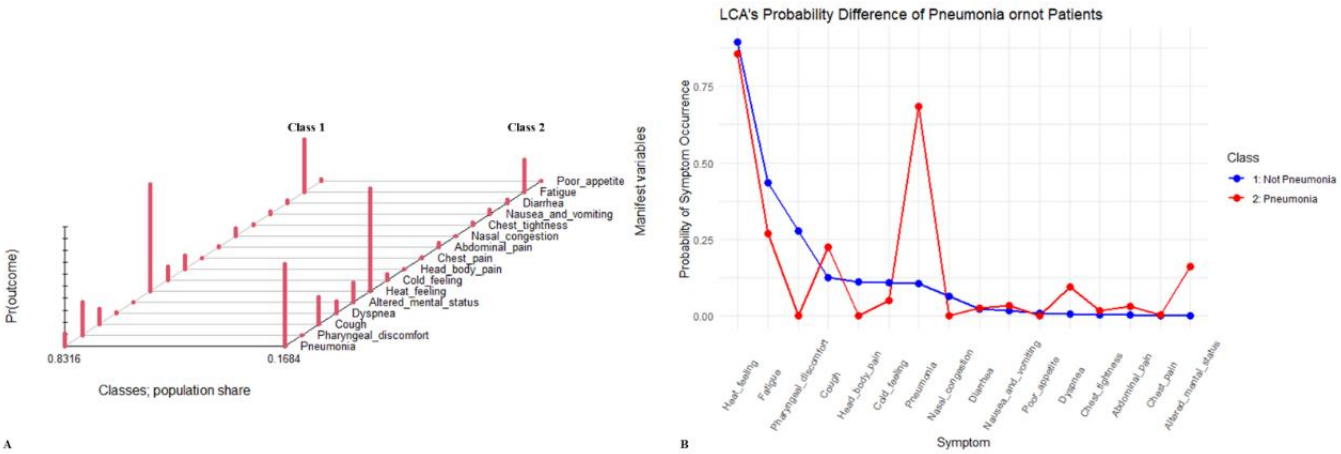

Figure S3 Series of Latent Class Analysis Charts for Automatic Subtype Classification of internal Training Cohort **(A.** The LCA plot of two latent classes demonstrating the distinction of pneumonia or not; **B.** The plot comparing LCA's probability difference of pneumonia ornot patients based on two lacent classes)

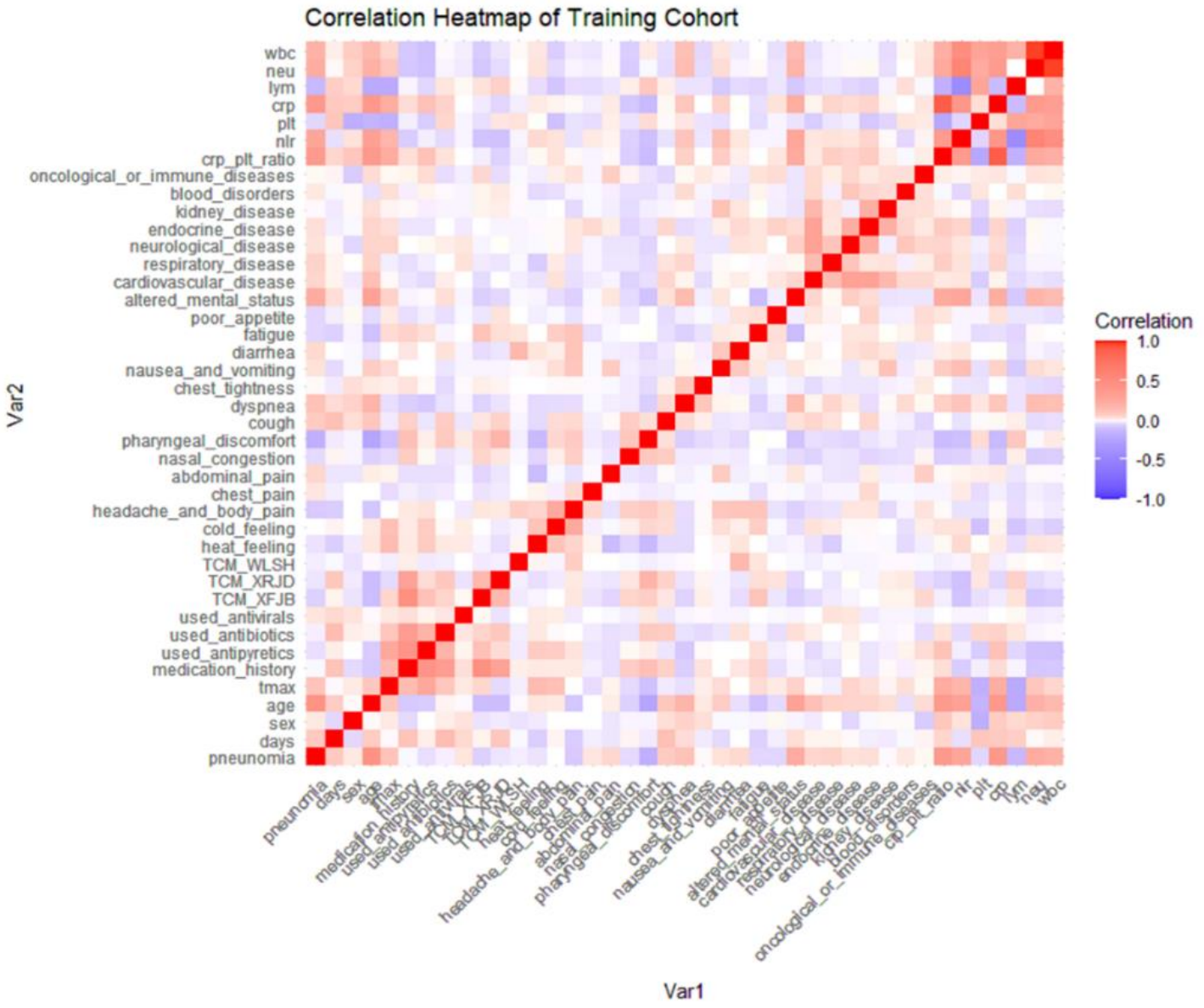

Figure S4 The Correlation Heatmap among Clinical Variables in the Training Cohort

92 **5. TRIPOD Checklist**

93 **Supplementary Table 2** TRIPOD Checklist (Prediction Model)

Section/Topic	Item	Checklist Item	Page
Title & abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	1
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	2
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	2
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	2−3
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	2−3
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	2−3
	5b	Describe eligibility criteria for participants.	2-3
	5c	Give details of treatments received, if relevant.	/
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	2,4
	6b	Report any actions to blind assessment of the outcome to be predicted.	2,4
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	2,4
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	2,4
Sample size	8	Explain how the study size was arrived at.	Supplement
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	2−4
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.	3,4
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	3,4
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	3,4
Risk groups	11	Provide details on how risk groups were created, if done.	/
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	4-6
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	4-6
Model	14a	Specify the number of participants and outcome events in each	7-8

development		analysis.	
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	7-8
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	9
	15b	Explain how to the use the prediction model.	8–9
Model performance	16	Report performance measures (with CIs) for the prediction model.	7–8, Supplement
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	9–10
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	9-10
Implications	20	Discuss the potential clinical use of the model and implications for future research.	9-10
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	Supplement
Funding	22	Give the source of funding and the role of the funders for the present study.	4

We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.