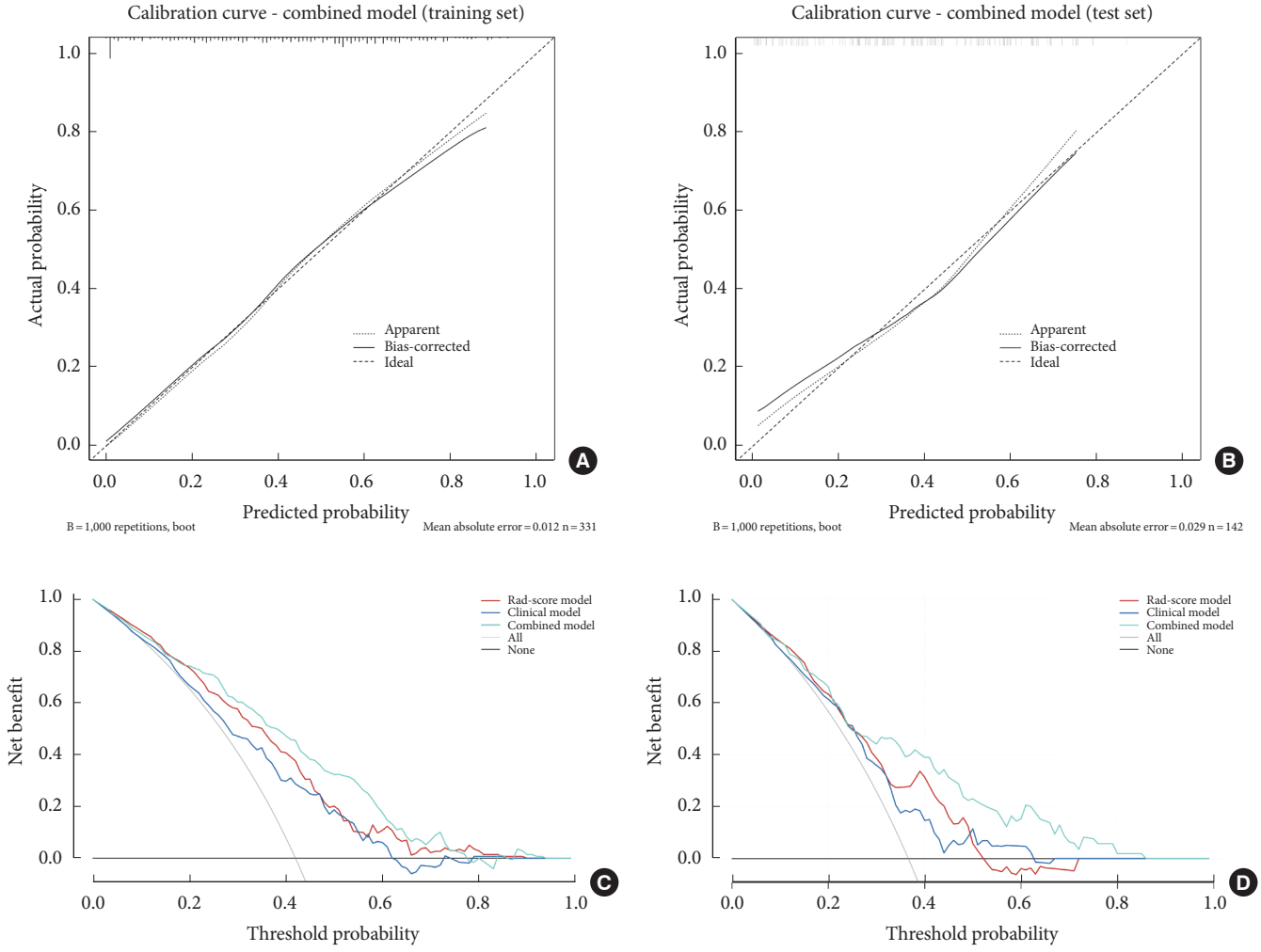

Supplementary Fig. 5. The calibration curves for the combined model on the training (A) and test sets (B). In both figures, the dashed line represents the ideal calibration curve, where predicted probabilities perfectly match the actual outcomes. The solid line shows the bias-corrected calibration, while the dotted line represents the apparent calibration. Panel A, which corresponds to the training set, demonstrates a mean absolute error (MAE) of 0.012 with 331 samples, while panel B for the test set shows a slightly higher MAE of 0.029 with 142 samples, reflecting the generalization performance of the model. Decision curve analysis for both the training (C) and test sets (D) comparing the net benefit of the Rad-score model (red), clinical model (blue), and combined model (green). The gray lines represent the net benefit of treating all patients ("all") and treating no patients ("none"). Across a wide range of threshold probabilities, the combined model (green) demonstrates superior net benefit compared to both the Rad_score and clinical models alone, particularly at lower threshold probabilities in the test set, indicating better clinical decision-making performance of the combined model in both cohorts.

https://doi.org/10.14245/ns.2448846.423