Supplementary Table 4. The results of the DeLong test for different models in the training and test sets

Comparison	Train p-value	Test p-value
Rad-score vs. clinical	0.154	0.324
Rad-score vs. combined	0.003	0.009
Clinical vs. combined	< 0.001	0.001