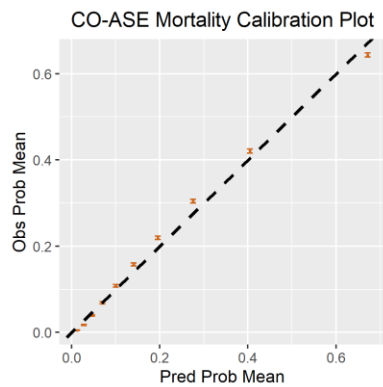


Risk Adjustment and Stratification	139	*Risk model performance	Provide empirical evidence that the risk model adequately accounts for confounding factors (e.g., assessment of model calibration and discrimination). Describe your interpretation of the results.
------------------------------------	-----	-------------------------	---

A risk-adjustment model incorporating baseline characteristics (age, sex), comorbidities, and detailed clinical data (including vital signs, laboratory values, positive blood cultures and COVID-19 tests, body mass index, and infection source per ICD-10 codes) achieved an AUROC of 0.8315 for the mortality outcome across all 3 datasets (268 hospitals). Performance was very similar when examined across each of the 3 datasets.

We assessed model calibration by O:P across deciles of risk (see figure below).



The Homer-Lemeshow test p-value was  $<0.05$ ; however, this is unsurprising given the very large sample size, and visual inspection of the calibration plots demonstrated adequate calibration.

Furthermore, the Brier Score was 0.12, indicating good calibration.