Computational Cardiology Lecture 16

Probability calibration and missing-modality evaluation

Zan Ahmad

Fall 2024

Three properties of a score

Ranking concerns ordering, classification concerns a threshold, and probability prediction concerns numerical event probabilities. Improving one need not improve all three.

Aggregate repeated observations

$$p_g = \sigma((z_{g1} + z_{g2})/2).$$

This is generally different from averaging the two probabilities.

The paired modality comparison

A-only and AB predictors are evaluated on the same unseen groups. Changing the eligible cohort would confound the effect of adding B.

Represent a missing measurement

$$B^* = \begin{cases} B, & \text{observed,} \\ m_B^{train}, & \text{missing,} \end{cases} \qquad r_B = \mathbf{1}_{\{B \text{ missing}\}}.$$

The indicator distinguishes imputation from an observed median value.

Freeze preprocessing

The training median, means and scales are reused for validation and testing. The missingness indicator records absence; it does not recover the missing information.

Temperature-scaled probabilities

$$p_T = \sigma(z/T), \qquad T > 0.$$

The supplied experiment first sharpens logits by a declared factor 2.5.

Ranking is preserved

$$z_i > z_j \quad \Longrightarrow \quad \sigma(z_i/T) > \sigma(z_j/T).$$

Positive temperature preserves ranking and ties, hence the pairwise AUC.

The one-half decision is preserved

$$\sigma(z/T) \geq \tfrac{1}{2} \quad \Longleftrightarrow \quad z \geq 0.$$

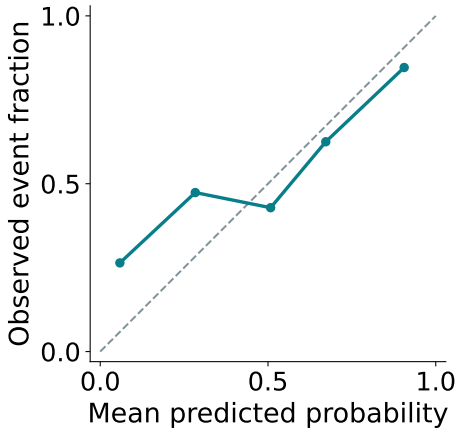
Other fixed probability thresholds need not preserve their classifications.

Fit temperature using validation likelihood

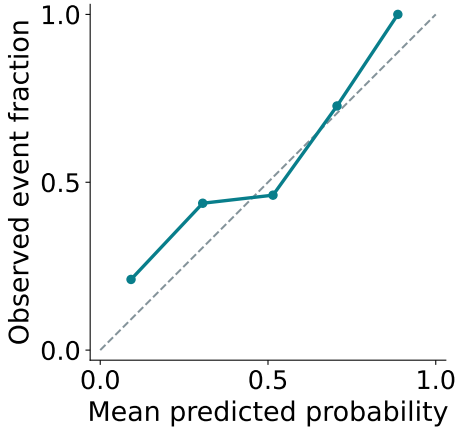
$$\widehat{T} = \arg\min_{T>0} \frac{1}{N} \sum_i [\log(1 + e^{z_i/T}) - y_i z_i / T].$$

The base model and all preprocessing remain fixed.

Sharpened AB scores



Validation-calibrated AB



The Brier score

$$\text{Brier} = \frac{1}{N}\sum_i (p_i - y_i)^2.$$

It evaluates probabilities, not only the chosen classification threshold.

Conditional probability decomposition

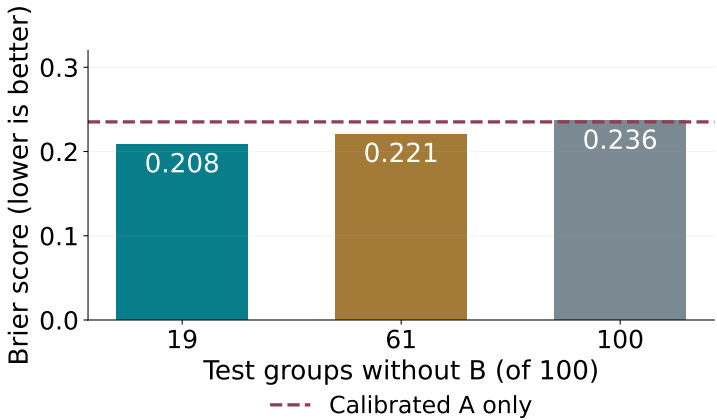
$$\mathbb{E}[(p - Y)^2 \mid X] = (p - \eta)^2 + \eta(1 - \eta).$$

The minimum occurs at the true conditional event probability η.

Reliability bins

Each bin compares its mean forecast with its observed event fraction. Counts determine how much information each point contains; an empty bin has no empirical fraction.

Brier score (lower is better)



Remove B while retaining the same groups

Replace B by the frozen missing-data rule and retain every test outcome. The change measures sensitivity to this particular information-loss intervention.

AUC and probability quality can separate

Positive temperature leaves AUC unchanged while changing Brier score and likelihood. A better probability score on one sample is an empirical result, not an algebraic guarantee.

Distribution-dependent calibration

$$P_{future}(Y = 1 \mid z) \neq P_{validation}(Y = 1 \mid z)$$

can invalidate a fitted calibration map even when the implementation is unchanged.