Probability calibration and missing-modality evaluation — review sheet

Use this after attempting the worked examples and practice. Explain each claim in words before calculating.

Key ideas

- Ranking, calibration and probability scores answer different questions.
- Calibration parameters are fitted on separate validation groups.
- A positive temperature preserves ranking and threshold-0.5 decisions.
- Evaluate modalities on the same cases and freeze the missing-data rule.
- Missingness stress tests require a stated intervention and limited claim.

Common errors and repairs

Call a lower Brier score proof of better calibration alone. Inspect reliability summaries and their counts; Brier also reflects discrimination.

Refit the imputer when test B values disappear. Use the stored training median and all fitted parameters unchanged.

Expect AB with no B to equal a separately fitted A model. The models have different coefficients and calibration parameters.

Retrieve without your notes

- What happens at temperature one?
- What does a zero-count reliability bin tell us?
- What should remain fixed during the missingness test?

Readiness check

- Distinguish ranking discrimination, probability quality and calibration.
- Fit a single temperature using validation groups while keeping test groups untouched.
- Compare one- and two-modality models on the identical held-out group population.
- Evaluate a frozen missing-data rule and state what a synthetic stress test can support.