STATISTICS PUT BIOMARKERS ON COMMON GROUND

STATISTICAL SIGNIFICANCE

A biomarker can look promising in one study and weaker in another. But the difference may not come from the biomarker - it may come from the patients. Studies often enroll populations that differ in age, disease severity, comorbidities, and other baseline characteristics. Because the area under the receiver operating characteristic curve (AUC), a common measure of diagnostic or prognostic accuracy, depends on that case mix, comparing unadjusted AUCs can create the illusion that a biomarker's performance has changed. We developed an estimand-focused statistical framework that anchors AUC evaluation to a clearly defined target population, helping researchers determine whether a biomarker will work in the people who matter and compare its performance fairly across studies.

Make Biomarker Accuracy "Run" Across Populations

WHAT DOES AUC TELL US?

The AUC measures how well a biomarker separates two groups, such as patients who do and do not experience an outcome. But the AUC is not fixed. It depends on the people being studied, so it can ***change across populations*** even when the biomarker itself has not.

DIFFERENT PATIENTS, DIFFERENT RESULTS

Biomarkers are often evaluated in carefully selected trials, while the patients who eventually use them may be older, sicker, or more diverse. This difference in patient mix, known as covariate shift, can arise from eligibility criteria, geography, or clinical practice.

As a result, a study's raw AUC may not reflect the population in which the biomarker will be used. Two studies may also report different AUCs simply because they enrolled different patients, not because the biomarker performed differently.

STATISTICAL CHALLENGE

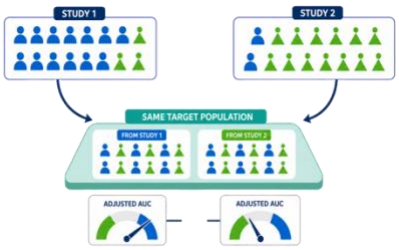
AUC is built from ***pairs of patients*** - one with the outcome and one without it. Researchers may also have only ***summary information*** about the target population, such as mean age or disease severity, rather than individual records.

SOLUTION: START WITH THE TARGET POPULATION

We start by defining the population whose AUC we want to estimate.

The question is no longer only "What is the AUC in this study?" but "How well would the biomarker work in the patients who will actually use it?"

Our ***calibration-weighting method*** reweights the study, so its patient characteristics match the target population. Underrepresented patients receive more weight; overrepresented patients receive less. The weighted comparisons then



estimate the target-population AUC.

For cross-study benchmarking, we adjust both studies ***to the same target*** before comparing them. The method can use target summaries when patient-level data are unavailable.

FROM METHOD TO MEDICINE: LUNG CANCER

We applied the framework to stair-climb power as a prognostic biomarker for 6-month survival in the POWER I and POWER II lung cancer trials. After both trials were aligned to a common population, their AUCs were more similar, suggesting that part of the original difference reflected who was enrolled rather than a true change in biomarker performance.

KEY FEATURES

- **Population-anchored**: Define AUC in a prespecified target population.
- **Fair comparison across studies**: Aligns populations first.
- **Works with summary data:** Doesn't require individual-level target data.
- **Robust and precise:** Developed doubly robust estimators.

WHY IT MATTERS

By giving every AUC a clear population anchor, the framework helps researchers **distinguish *changes in biomarker performance* from *changes in patient mix.*** That makes evidence easier to interpret and more relevant to screening, diagnosis, and trial enrollment.

ASA