Statistical Foundations of AI: From Predictive Models to Credible Evidence in Drug Development

Shrutimita Pokhariyal and Christina Geyer (presenter)

phastar

Bayes Bayesian Biostatistics Leiden, NL 2026

An AI output is a measurement, not a result

Three questions for any instrument:

1. Does it measure what we need?
2. Does the analysis account for its measurement error?
3. Does performance hold in new settings and subgroups?

Real-world example: Diabetic retinopathy grading

Development Population · Leibig 2017

0.927→ 0.956

AUC after referring most uncertain 30%

Bayesian neural network (MC dropout) estimating predictive uncertainty

Shift: US → India · Band 2021

0.94 → 0.65

AUC after referring most uncertain 70%

MC dropout on Indian images and equipment

Discrimination held; the uncertainty did not. Performance must be rechecked in each new population

phastar

Leibig et al. Sci Rep 2017; Band et al. NeurIPS Datasets & Benchmarks 2021

Proposal: model AI's error with dynamic borrowing

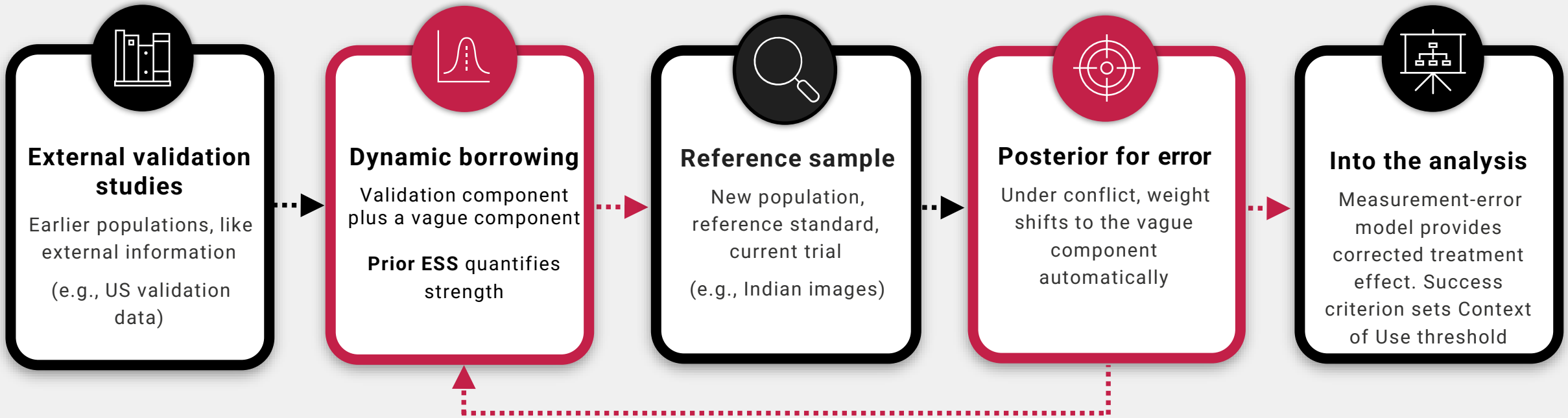
Bayesian idea	How it can strengthen AI-derived evidence
Measurement error model	AI output is data; adjust for AI error in the covariates to obtain unbiased treatment effect
Dynamic borrowing	Borrow from external validation studies; discount if reference sample conflicts
Prior effective sample size	Quantify how many new population cases the prior is worth
Prior-data conflict diagnostics	Include diagnostics on model behavior when sample conflicts
Success criterion	AI is fit for its Context of Use if Pr(error within pre-specified limits | data) > c, where c is set by the cost of a wrong decision

External validation studies are to the AI's error rates what prior information is to Bayesian models

Robust MAP prior: Schmidli et al., Biometrics 2014 · Prior ESS: Neuenschwander et al., Biometrics 2020

Bayesian dynamic borrowing for AI model error

How validation data can be borrowed, checked, and updated as an AI model moves to new populations



Retinopathy example: US studies act as „prior“. A small verification sample of Indian images would have conflicted with it, and analysis would have shifted weight away from the US data toward a robust component automatically

phastar

Prior-data conflict: Evans & Moshonov, Bayesian Anal 2006

AI-derived evidence could meet a Bayesian trial's bar



Pre-specify

- Context of Use (CoU) & Estimand
- Data-generating process and reference standard
- Prior on AI error: data sources, exchangeability, exclusion criteria
- Human-review rule
- Success criterion and prior-data conflict rule



Simulate

- Human-AI pathway
- Shifts: Case mix, device, prevalence
- Reviewers left with hardest cases
- Versus simpler model or humans
- Sensitivity analyses: robustness of assumptions, varied borrowing, imperfect reference standard



Document & Report

- Per-output metrics: model version, predictive uncertainty, review flags, reference assessments
- Version-controlled code and seeds; locked validation data
- Compliance with regulatory guidances (i.e., TRIPOD+AI, FDA draft guidance on Use of AI)

AI should not replace statistical discipline, it should be strengthened by it

By extending expectations established for external evidence in Bayesian trials to AI-derived evidence

phastar

TRIPOD+AI, BMJ 2024 · EU AI Act Arts. 10, 13, 72 · FDA draft guidance on Use of AI (Jan 2025) and Use of Bayesian Methodology (Jan 2026)