

TITLE: Causally Informed Dynamic Borrowing for Multi-Site AI Diagnostic Deployment

AUTHORS: [Bruno Boulanger](#) and Nils Boulanger

ABSTRACT:

Background. Deploying AI-based diagnostic devices across hospital sites raises a fundamental Bayesian challenge: the disease prevalence π at each site differs from the development data and varies across demographic subgroups. Since $PPV(\pi) = Se \cdot \pi / [Se \cdot \pi + (1 - Sp)(1 - \pi)]$ is an explicit function of prevalence, even a perfectly discriminating model produces clinically misleading predictions when π is misspecified. This prevalence dependence also has implications for AI fairness across demographic subgroups. A natural Bayesian response is to model the prevalence within each site $\pi_k \sim \text{Beta}(a, b)$ hierarchically and borrow across sites via robust mixture priors. However, the exchangeability assumption that justifies borrowing is rarely tested - standard approaches use scalar conflict measures that detect *that* sites differ, but not *why*.

Objective. We propose a framework that makes the exchangeability assumption formally testable by combining hierarchical Bayesian prevalence estimation with a causal directed acyclic graph (DAG) and Bayesian model comparison. The key idea is to determine *what type of distributional shift* drives the performance difference between sites - and therefore *whether* and *how* to borrow.

Method. We specify a causal DAG over the deployment structure with nodes: Site (k), demographics (G), covariates (X), true outcome (Y), and AI prediction ($\hat{Y}$). The joint distribution factorises into three candidate shift mechanisms following the Independent Causal Mechanisms principle: $P(G|k)$, $P(X|G, k)$, and $P(Y|X, G)$. We define four nested models on the DAG: M_0 (full exchangeability: all mechanisms shared across sites); M_1 (prevalence shift only: π_k varies, $P(Y|X, G)$ shared); M_2 (covariate shift: $P(G|k)$ and $P(X|G, k)$ vary, $P(Y|X, G)$ shared); M_3 (concept shift: $P(Y|X, G)$ varies across sites). Posterior model probabilities $P(M_j|\text{data})$ or Bayes factors select the dominant shift type, directly informing the borrowing strategy.

Propensity score reweighting. When covariate shift (M_2) is selected, we use propensity scores $e(x, g) = P(\text{site}=k^*|X=x, G=g)$ to compute importance weights $w(x, g) = e/(1-e)$ that reweight existing site data to match the new site's covariate distribution. The reweighted prevalence $\hat{\pi}_{rw}$ becomes the informative prior for π_{k^*} . Propensity score overlap serves as an additional exchangeability diagnostic: good overlap indicates reliable borrowing; poor overlap signals that the sites have fundamentally different patient populations, making reweighting unreliable regardless of shift type.

Borrowing decisions. The model comparison determines the strategy: **(i)** M_1 selected (prevalence shift): full borrowing via hierarchical Beta + robust mixture prior. **(ii)** M_2 selected (covariate shift): borrow with propensity score reweighting - correct for population mismatch before setting the prior. **(iii)** M_3 selected (concept shift): refuse to borrow, default to vague prior, trigger local re-estimation. The adaptive calibration then proceeds sequentially: update $\pi_k|\text{data} \sim \text{Beta}(a+y_k, b+n_k-y_k)$, evaluate $P(PPV_k > PPV_{\min}|\text{data}) > \delta$ per subgroup g , with ongoing drift monitoring triggering model re-comparison.

Novelty. Dynamic borrowing in Bayesian trials relies on conflict measures (Schmidli et al., 2014) that are “blind” to the mechanism driving the difference. Our contribution makes borrowing *causally informed*: the causal DAG decomposes site variation into interpretable mechanisms; Bayesian model comparison classifies the shift type; and propensity scores correct covariate shifts when the clinical concept is preserved. Exchangeability becomes a testable, mechanism-specific diagnostic rather than an untestable assumption. The three-phase adaptive protocol (prior specification → calibration deployment → drift monitoring) maps directly to regulatory lifecycle requirements for AI as a medical device (EU AI Act Art. 9, FDA Predetermined Change Control Plans).

PRESENTING AUTHOR BIO: Bruno Boulanger, Ph.D.

Co-founder & Senior Advisor, Trilenda

Dr Bruno Boulanger brings over 30 years of experience in pharmaceutical and medical device development, spanning discovery through clinical stages. He has held senior roles at Eli Lilly and Company and UCB across the US and Europe.

In 2003, he founded Arlenda, a pioneering firm in advanced statistical solutions. A recognized thought leader, Bruno has authored 100+ publications, edited one book in Bayesian statistics and contributed to the United States Pharmacopeia Committee on Statistics (2010–2024), and founded the BAYES conference series. Bruno also received in 2024 the ENBIS award for "Best Manager in Statistics".

At Trilenda, he brings deep expertise and strategic insight to support innovation at the intersection of statistics and healthcare.

References

- Chouldechova, A. (2017). Fair prediction with disparate impact. *Big Data*, 5(2), 153–163.
- Schmidli, H. et al. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4), 1023–1032.
- Van Calster, B. et al. (2025). There is no such thing as a validated prediction model. *BMC Medicine*, 23, 94