Conjugate and MCMC Bayesian Chain-Rule Prediction-Powered Inference for Binary Prevalence Estimation

Anonymous authors
Paper under double-blind review

Abstract

Prediction–Powered Inference (PPI) combines abundant machine predictions with scarce labels to estimate population functionals with improved efficiency. Recent Bayesian treatments of PPI have introduced general conjugate and Monte Carlo formulations. We focus on a narrower but practically important setting: binary prevalence estimation through the chain-rule functional

$$g = P(H = 1 \mid A = 1)P(A = 1) + P(H = 1 \mid A = 0)P(A = 0).$$

For the base Beta–Bernoulli model, we show that the posterior factorizes into independent Beta distributions, so uncertainty in g can be propagated by direct sampling without MCMC. We reserve NUTS only for genuinely non-conjugate extensions, including hierarchical partial pooling, logit-normal priors, K-bin score models, and joint threshold uncertainty.

We benchmark this conjugate Bayesian chain-rule estimator (CRE) against three baselines: a labeled-only Bayesian estimator, the classical difference estimator, and a prior-free analytic PPI estimator based on continuous probabilities with small-sample t critical values. Empirically, we report (i) simulation-based calibration for the conjugate engine, (ii) repeated-labeling resampling studies with $M = 500$ replications on ADNI-derived out-of-fold prediction tables, and (iii) an Alzheimer's disease MRI case study with out-of-fold threshold selection, bootstrap cut-point dispersion, and propensity-overlap diagnostics. In the full cohort, CRE achieves near-nominal coverage and narrower intervals than the labeled-only Bayesian baseline, while substantially improving stability over the difference estimator at small label budgets; gains are smaller but remain competitive in a fixed 65–70 subset. These results position conjugate Bayesian chain-rule PPI as a lightweight and auditable option for deployment-time prevalence monitoring.

1 Introduction

Motivation. High–capacity machine–learning (ML) systems now achieve strong predictive performance in high–dimensional, nonparametric regimes. Translating these predictions into *valid statistical inference*—for example, interval estimates with nominal coverage—remains challenging when models are misspecified or over–parameterized and when human labels are scarce. Likelihood–based intervals may undercover or overcover in such regimes, whereas purely imputed estimators that replace outcomes by model predictions are typically biased. Related literatures on semi–supervised and survey–calibrated estimation show that auxiliary predictions can reduce variance if their bias is corrected appropriately (Lohr, 2019).

Prediction–Powered inference (PPI). PPI leverages an accurate but imperfect predictor trained on abundant unlabeled data and then debiases it using a small labeled subset via a *rectifier* term, yielding finite–sample guarantees under weak assumptions (Angelopoulos et al., 2023; Guo & Lei, 2021). For a binary autorater $A \in \{0, 1\}$ and human label $H \in \{0, 1\}$, the quantity of interest

$$g \;=\; P(H{=}1) \;=\; P(H{=}1 \mid A{=}1)P(A{=}1) + P(H{=}1 \mid A{=}0)P(A{=}0) \;=\; \theta_A\,\theta_{H|1} + (1-\theta_A)\,\theta_{H|0}, \qquad (1.1)$$

naturally couples plentiful predictions with few labels. In practice, the benefit of this coupling is largest when unlabeled predictions are abundant and labeling is costly—precisely the setting of modern applications in medical imaging and scientific data analysis. In addition to the thresholded-A formulation, we also consider a *continuous-probability* variant in which

$$\hat{g}_{\mathrm{PPI}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} p_i}_{\bar{p}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(H_i - p_i)}_{\text{rectifier}}, \quad \mathrm{SE}(\hat{g}_{\mathrm{PPI}}) = \sqrt{\mathrm{Var}(p)/N + \mathrm{Var}(H-p)/n},$$

and we form 95% intervals with a normal critical value when $n \geq 30$ (otherwise $t_{0.975,n-1}$), yielding a *prior-free analytic* baseline complementary to our Bayesian approach (Angelopoulos et al., 2023; Efron & Tibshirani, 1993).

From prior Bayesian PPI to a narrower deployment setting. Recent Bayesian treatments of PPI have already shown that conjugate Bayesian analogues are possible in several model families and have introduced Bayesian chain-rule style estimators (Hofer et al., 2024). Our goal is narrower and more operational: we isolate the binary chain-rule prevalence setting, show that its base Beta–Bernoulli model is fully conjugate and does not require MCMC, and separate this base case from genuinely non-conjugate extensions. This distinction matters in practice because many monitoring settings require only a lightweight base estimator together with transparent diagnostics, rather than a general-purpose MCMC workflow.

This paper: conjugate CRE as the default, NUTS only when needed. We formulate a Bayesian model for equation 1.1 with independent Beta priors on $(\theta_A, \theta_{H|1}, \theta_{H|0})$. In the base Beta–Bernoulli specification, the posterior factorizes into independent Beta distributions determined by simple cell counts, so posterior uncertainty for $g = \theta_A\theta_{H|1} + (1-\theta_A)\theta_{H|0}$ follows by direct Monte Carlo. We then invoke NUTS only for extensions that truly break conjugacy—for example, hierarchical logit-normal pooling, non-Beta priors, richer score discretizations, or joint threshold uncertainty—while retaining a unified Bayesian workflow and diagnostics (Neal, 2011; Betancourt, 2017; Hoffman & Gelman, 2014).

Contributions. Relative to existing Bayesian PPI formulations, our contributions are:

- a sharp characterization of the *binary chain-rule prevalence* base case as a fully conjugate Beta–Bernoulli model with direct posterior propagation to g;
- a clean separation between this conjugate base estimator and genuinely non-conjugate extensions, for which NUTS is used only when necessary;
- a comparison against three practically relevant baselines: labeled-only Bayes, the classical difference estimator on binary A, and a prior-free analytic PPI estimator based on continuous probabilities with small-sample t critical values;
- deployment-oriented audits, including labeled–unlabeled propensity-overlap checks, out-of-fold (OOF) versus leaky threshold selection, and bootstrap dispersion of Youden's cut-point;
- an empirical evaluation built around simulation-based calibration and repeated-labeling resampling on ADNI-derived OOF prediction tables, plus an MRI case study with age-stratified operating thresholds.

Empirical overview. Our empirical findings are deliberately modest and targeted. In the full ADNI cohort, CRE achieves near-nominal coverage across label budgets and is consistently narrower than the labeled-only Bayesian baseline, while substantially improving on the instability of the classical difference estimator at very small label budgets. In the fixed 65–70 subset, the same qualitative pattern remains visible but the gains are attenuated, which is expected given the smaller cohort size. We also report OOF threshold selection, bootstrap cut-point dispersion, and exchangeability diagnostics to align the evaluation with deployment-time use rather than retrospective in-sample tuning.

Limitations. Our strongest claims are confined to the binary chain-rule prevalence setting and to repeated-labeling uncertainty under exchangeability of labeled and unlabeled samples. The current empirical study emphasizes simulation-based calibration, repeated-labeling resampling, and a single real-data imaging application; it does not claim an exhaustive synthetic stress test over all shift mechanisms or prevalence regimes. The 65–70 subset analysis is intentionally presented as a smaller-cohort sensitivity check rather than a second main victory lap. Finally, non-conjugate extensions are included for completeness, but the main empirical contribution of this paper is the conjugate base estimator and its deployment-oriented evaluation.

Broader impact. Coupling abundant machine predictions with scarce labels can reduce annotation burdens in sensitive domains (e.g., medical imaging). At the same time, decision-threshold choices and calibration must be aligned with clinical costs and equity considerations; our age-aware analysis illustrates one such alignment while emphasizing the need for domain governance and prospective monitoring.

2 Background and Related Work

2.1 Prediction–Powered inference in context: links to survey calibration, semi–supervised inference, and doubly robust estimators

Prediction–Powered Inference (PPI) exploits a high–quality but imperfect predictor trained on abundant unlabeled samples to reduce variance at fixed label budgets while preserving valid uncertainty quantification. Let $\{(X_i, Y_i)\}_{i=1}^N$ be inputs and outcomes and let f be a predictor trained chiefly on unlabeled covariates. A purely imputed estimator $\hat{\theta}_f = N^{-1}\sum_i f(X_i)$ is generally biased, whereas the classical mean $\hat{\theta}_{\text{classical}} = N^{-1}\sum_i Y_i$ is label–hungry. PPI combines the strengths by introducing a rectifier on a small labeled subset $\{(X_i, Y_i)\}_{i=1}^N$ with $n \ll N$,

$$\Delta = \frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - Y_i\big), \qquad \hat{\theta}_{\text{PPI}} = \hat{\theta}_f - \Delta,$$

and provides finite–sample guarantees under weak assumptions on the sampling of the labeled subset and mild regularity of f (Angelopoulos et al., 2023; Guo & Lei, 2021). In binary–decision settings with a machine autorater $A \in \{0, 1\}$ and a human label $H \in \{0, 1\}$, the target positivity rate $g = P(H{=}1)$ decomposes via the chain rule

$$g = P(H{=}1 \mid A{=}1)P(A{=}1) + P(H{=}1 \mid A{=}0)P(A{=}0) = \theta_A\theta_{H|1} + (1-\theta_A)\theta_{H|0},$$

which exposes how abundant predictions (informing θ_A) and scarce labels (informing $\theta_{H|a}$) jointly determine uncertainty. The same structure underlies the *difference estimator*

$$\hat{g}_{\text{diff}} = \bar{A} + \overline{(H - A)} = \frac{1}{N}\sum_{i=1}^{N} A_i + \frac{1}{n}\sum_{i=1}^{n}(H_i - A_i),$$

long studied in design–based survey sampling where auxiliary variables reduce variance without sacrificing unbiasedness (Cochran, 1977; Lohr, 2019). From this perspective, PPI can be viewed as bringing classical calibration ideas to modern ML auxiliaries. The generalized regression (GREG) and calibration estimators (Särndal et al., 1992; Deville & Särndal, 1992) adjust sampling weights (or add regression corrections) so that estimates align with known population totals; replacing "known totals" by "accurate large–N machine predictions" recovers the spirit of PPI while retaining transparent conditions for coverage.

Connections to semi–supervised inference further clarify efficiency gains. A line of work on semi–supervised means, risks, and ROC–type functionals shows that plug–in estimators based on unlabeled X can be augmented by labeled residuals to attain substantial variance reduction with valid inference (Chakrabortty & Cai, 2018; Gronsbell & Cai, 2018). In influence–function language, the rectifier acts as an augmentation term that centers the estimating equation, a device familiar from augmented inverse–probability weighting and targeted learning where *doubly robust* estimators remain consistent if either the outcome or propensity/auxiliary model is correct (Bang & Robins, 2005; van der Laan & Rose, 2011). PPI differs in emphasis: assumptions are simple and practically auditable (exchangeability of labeled/unlabeled pools and predictor

stability), and coverage statements leverage the large pool of machine predictions directly (Angelopoulos et al., 2023).

For practical deployment, we emphasize two additional ingredients often underplayed in prior narratives. First, when thresholds map scores to decisions ($A = \mathbf{1}\{S \geq t\}$), *leakage* can inflate performance if the same data inform threshold selection and evaluation; out-of-fold selection and bootstrap dispersion of Youden's cutpoint mitigate this bias (Youden, 1950; Varma & Simon, 2006; Cawley & Talbot, 2010; Efron & Tibshirani, 1993). Second, because PPI and CRE hinge on exchangeability between labeled and unlabeled pools, we advocate *propensity-overlap diagnostics* (e.g., low AUC for a labeled-versus-unlabeled classifier) as a simple precondition check (Rosenbaum & Rubin, 1983).

Practical deployment also requires calibrated probabilities when thresholding scores into $A = \mathbf{1}\{S \geq t\}$. Post–hoc calibration methods—Platt scaling (logistic/temperature scaling) and isotonic regression—improve probability reliability without changing ranking and therefore do not alter threshold–agnostic discrimination (AUC) (Platt, 1999; Zadrozny & Elkan, 2002; Niculescu-Mizil & Caruana, 2005; Guo et al., 2017). This integrates neatly with PPI: calibration sharpens decision quality, while the PPI estimator consumes only the thresholded A and labeled pairs to produce valid intervals for g. When labeled data are scarce or stratified (e.g., by age), information sharing through partial pooling or calibration across strata can stabilize estimates while respecting heterogeneity, as we adopt in the case study.

2.2 Bayesian formulations of PPI, computation, and diagnostics

A Bayesian treatment of PPI provides a coherent generative specification and direct uncertainty propagation to nonlinear functionals. In conjugate families (e.g., beta–binomial, normal–linear), recent work derives closed–form posteriors and clarifies how the PPI rectifier aligns with posterior predictive adjustments (Hofer et al., 2024). Realistic deployments often break conjugacy, especially when incorporating stratification, flexible priors, or additional structure. In these cases simulation–based inference via Hamiltonian Monte Carlo (HMC) and the No–U–Turn Sampler (NUTS) delivers efficient exploration with minimal tuning (Neal, 2011; Hoffman & Gelman, 2014; Betancourt, 2017).

In the base model with independent Beta priors and Bernoulli likelihoods for $(\theta_A, \theta_{H|1}, \theta_{H|0})$, conjugacy yields closed-form posteriors: each component is Beta with updated counts from $\mathcal{D}_A$ and the labeled 2×2 margins. Hence, one can obtain uncertainty for the functional g by drawing from these Betas and transforming. We resort to simulation-based inference (HMC/NUTS) when moving beyond conjugacy—e.g., with logit-normal hierarchies, non-Beta priors, K-bin generalizations with shared hyperparameters, or joint inference on t (Salvatier et al., 2016; Carpenter et al., 2017). Workflow discipline is crucial: rank–normalized $\hat{R} < 1.01$, adequate bulk/tail effective sample sizes (ESS), and the absence of divergences indicate reliable exploration; posterior–predictive checks (PPCs) and simulation–based calibration (SBC) ensure calibrated inferences for both primitive parameters and derived functionals like $g = \theta_A\theta_{H|1} + (1-\theta_A)\theta_{H|0}$ (Kumar et al., 2019; Vehtari et al., 2021b; Gabry et al., 2019; Talts et al., 2018; Vehtari et al., 2021a; McElreath, 2020).

Two modeling choices help in small–n or imbalanced strata. First, weakly informative priors (e.g., Jeffreys $\text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$) regularize extreme cell proportions when labeled counts for A=1 or A=0 are tiny, improving tail behavior of credible intervals with minimal bias (Gelman et al., 2013). Second, hierarchical extensions with partial pooling across subgroups g share signal while preserving between–group differences; the induced group–wise estimands $g_g = \theta_{A,g}\theta_{H|1,g} + (1 - \theta_{A,g})\theta_{H|0,g}$ inherit shrinkage–stabilized uncertainty. Model comparison within this family can be guided by PSIS–LOO expected log predictive density to avoid over–binning or over–stratification that does not improve out–of–sample fit (Vehtari et al., 2017).

Finally, distribution shift and label–missingness mechanisms warrant explicit consideration. Under MCAR (or MAR given A and coarse covariates), labeled and unlabeled pools are exchangeable for $(\theta_A, \theta_{H|a})$; sizable deviations call for stratified labeling or importance weighting under covariate–shift assumptions (Sugiyama et al., 2007). The Bayesian workflow makes such sensitivities transparent: prior–predictive checks flag implausible regions, PPCs detect lack of fit in (A, H) margins, and targeted sensitivity analyses (uniform vs. Jeffreys priors; global vs. stratified thresholds) quantify the robustness of coverage and interval width—key objectives for PPI in label–scarce regimes.

3 Methods

ADNI data statement (required). Data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). ADNI began in 2003 as a public–private partnership led by Michael W. Weiner, MD. Its goals include combining MRI, PET, biofluid biomarkers, and neuropsychological assessments to track progression of MCI and Alzheimer's disease, validating biomarkers for clinical trials, broadening cohort diversity, and providing data to the research community. See `adni.loni.usc.edu` for up-to-date details.

Data version. We downloaded ADNI data on **2025-07-14** and checked for updates prior to submission.

3.1 Generative specification, estimand, and extensions

Let $A \in \{0,1\}$ denote an autorater decision (derived from a probabilistic score $p \in [0,1]$ via a fixed operating threshold t) and let $H \in \{0,1\}$ be a human label. Abundant autorater outputs are observed as $\mathcal{D}_A = \{A_i\}_{i=1}^{N_A}$ together with a comparatively small labeled subset $\mathcal{D}_H = \{(A_i, H_i)\}_{i=1}^{N_H}$, with $N_A \gg N_H$. Define

$$\theta_A = P(A=1), \qquad \theta_{H|1} = P(H=1 \mid A=1), \qquad \theta_{H|0} = P(H=1 \mid A=0),$$

which determine the population functional

$$g(\boldsymbol{\theta}) = \theta_A \theta_{H|1} + (1-\theta_A)\theta_{H|0}, \qquad \boldsymbol{\theta} = (\theta_A, \theta_{H|1}, \theta_{H|0}) \in (0,1)^3. \tag{3.1}$$

We assume $\{A_i\}_{i=1}^{N_A} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta_A)$ and, conditionally on A_i, $H_i \mid A_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{H|A_i})$ for the N_H labeled pairs. Missingness of H is treated as MCAR outside $\{1, \ldots, N_H\}$, ensuring exchangeability of labeled and unlabeled pools for inference on $\boldsymbol{\theta}$. Autorater probabilities p_i are thresholded into $A_i = \mathbf{1}\{p_i \geq t\}$, where t is regarded as fixed during inference (deployment practice); Section 3.4 propagates uncertainty in the selection of t. For a step-by-step summary of how this generative specification is used in practice (from priors and likelihood to posterior summaries of g), see Section J.1.

Likelihood and priors.

Proposition 3.1 (Conjugate posterior for the base chain-rule model)**.** *Let* $n_A = \sum_{i=1}^{N_A} A_i$, $n_{11} = \sum_{i=1}^{N_H} \mathbf{1}\{A_i = 1, H_i = 1\}$, $n_{10} = \sum_{i=1}^{N_H} \mathbf{1}\{A_i = 1, H_i = 0\}$,
$n_{01} = \sum_{i=1}^{N_H} \mathbf{1}\{A_i = 0, H_i = 1\}$,
$n_{00} = \sum_{i=1}^{N_H} \mathbf{1}\{A_i = 0, H_i = 0\}$. *With independent* $\text{Beta}(\alpha_A, \beta_A)$, $\text{Beta}(\alpha_1, \beta_1)$, $\text{Beta}(\alpha_0, \beta_0)$ *priors, the posterior factorizes as*

$$\theta_A \mid \mathcal{D} \sim \text{Beta}(\alpha_A + n_A,\ \beta_A + N_A - n_A),$$

$$\theta_{H|1} \mid \mathcal{D} \sim \text{Beta}(\alpha_1 + n_{11},\ \beta_1 + n_{10}), \quad \theta_{H|0} \mid \mathcal{D} \sim \text{Beta}(\alpha_0 + n_{01},\ \beta_0 + n_{00}).$$

Thus $g = \theta_A \theta_{H|1} + (1-\theta_A)\theta_{H|0}$ *follows by direct Monte Carlo.*

Under these assumptions, the complete–data likelihood factorizes as

$$p(\mathcal{D}_A, \mathcal{D}_H \mid \boldsymbol{\theta}) = \prod_{i=1}^{N_A} \theta_A^{A_i}(1-\theta_A)^{1-A_i} \prod_{i=1}^{N_H} \theta_{H|1}^{H_i A_i}(1-\theta_{H|1})^{(1-H_i)A_i} \theta_{H|0}^{H_i(1-A_i)}(1-\theta_{H|0})^{(1-H_i)(1-A_i)}. \tag{3.2}$$

Unless stated otherwise, independent $\text{Beta}(1,1)$ priors are placed on $(\theta_A, \theta_{H|1}, \theta_{H|0})$; Jeffreys' $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ priors are used for sensitivity near the boundaries (Gelman et al., 2013). Weakly informative alternatives (e.g., logit–normal $\theta = \text{logit}^{-1}(\eta)$ with $\eta \sim \mathcal{N}(0, 1.5^2)$) are considered when extreme class imbalance or tiny stratum counts induce separation (Gelman et al., 2008).

Identifiability and small–cell regularization. θ_A is identified from $\mathcal{D}_A$ alone; $(\theta_{H|1}, \theta_{H|0})$ are identified from $\mathcal{D}_H$ provided each stratum $A \in \{0,1\}$ contributes at least one labeled case asymptotically. When labeled positives (or negatives) are extremely rare in a stratum, Beta priors regularize cell–probability posteriors away from 0/1; posterior predictive checks (PPCs) on the 2×2 table clarify the degree of regularization.

Continuous–score generalization. If one wishes to avoid dichotomizing p, a K–bin chain rule replaces $A \in \{0,1\}$ by $B \in \{1,\dots,K\}$ with

$$g = \sum_{k=1}^{K} P(H=1 \mid B=k)\, P(B=k).$$

and independent Beta priors on $P(H{=}1 \mid B=k)_{k=1}^{K}$ and a Dirichlet prior on $(P(B{=}1),\dots,P(B{=}K))$. This yields a strictly richer CRE at the cost of more parameters and potentially sparser labeled cells; we therefore default to $K{=}2$ and assess $K > 2$ in sensitivity analyses. Empirically, small multi-bin generalizations ($K \in \{4,5\}$) using quantile binning produced posterior means of g indistinguishable from $K{=}2$ and 95% CI widths that were effectively unchanged under both Uniform Beta$(1,1)$ and Jeffreys Beta$(\frac{1}{2},\frac{1}{2})$ priors (Table 10 in the Appendix).

Hierarchical partial pooling across strata. For age–aware deployment with strata $s \in \{1,\dots,S\}$, parameters become $\boldsymbol{\theta}_s = (\theta_{A,s}, \theta_{H|1,s}, \theta_{H|0,s})$ and we impose logit–normal hierarchies

$$\operatorname{logit}(\theta_{H|a,s}) \sim \mathcal{N}(\mu_a, \sigma_a^2), \quad \operatorname{logit}(\theta_{A,s}) \sim \mathcal{N}(\mu_A, \sigma_A^2),$$

with weakly informative hyperpriors $\mu_\cdot \sim \mathcal{N}(0, 2^2)$ and $\sigma_\cdot \sim$ Half-Normal(1), enabling partial pooling when certain strata have few labels while permitting stratum–specific deviations.

3.2 Bayesian computation, diagnostics, and uncertainty for functionals

Posterior inference targets $p(\boldsymbol{\theta} \mid \mathcal{D}_A, \mathcal{D}_H) \propto p(\mathcal{D}_A, \mathcal{D}_H \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$. For the base Beta–Bernoulli model, we do not require MCMC: we draw $(\theta_A, \theta_{H|1}, \theta_{H|0})$ directly from their Beta posteriors (Prop. 3.1) and map draws to g. For extensions that break conjugacy (hierarchical logit-normal pooling, non-Beta priors, K-bin with shared hyperparameters, joint t), we use NUTS with the same diagnostics and workflow (Section J.1).

Sampler and parameterization. We use direct Beta sampling in the base conjugate model, and Hamiltonian Monte Carlo with the No–U–Turn Sampler (NUTS) for non-conjugate extensions (Neal, 2011; Hoffman & Gelman, 2014; Betancourt, 2017). Computations are carried out in `PyMC` with automatic differentiation and mass–matrix adaptation (Salvatier et al., 2016; Carpenter et al., 2017). A typical call is

Base (conjugate): direct Beta sampling for $(\theta_A, \theta_{H|1}, \theta_{H|0})$.
Extensions (non-conjugate): `pm.sample(draws=2000, tune=1000, target_accept=0.95, chains=4)`.

with logits $\eta = \operatorname{logit}(\theta)$ as sampling parameters to improve geometry and avoid boundary pathologies. We increase `target_accept` to 0.95 or add non–centered parameterizations in the hierarchical model when divergences occur.

Diagnostics and remedial actions. Convergence is assessed via rank–normalized $\hat{R} < 1.01$, bulk/tail ESS > 400 per parameter, and absence of divergences; energy–Bayes fraction of missing information (E–BFMI) flags poor momentum resampling when < 0.3 (Vehtari et al., 2021b;a). PPCs compare replicated and observed margins of (A, H) and the induced distribution of g; simulation–based calibration (SBC) checks calibration of posteriors across synthetic draws (Gabry et al., 2019; Talts et al., 2018). If PPCs indicate misfit (e.g., skewed residual cell counts), we enlarge priors or move to the hierarchical/continuous–bin extensions above.

Posterior summaries for g and contrasts. For each retained draw $\boldsymbol{\theta}^{(s)}$, compute $g^{(s)}$ via equation 3.1 and report posterior mean, SD, and credible intervals (equal–tailed or highest–density). For strata s, we propagate to differences $g_s - g_{s'}$ by transforming paired draws; joint intervals for $(g_1,\dots,g_S)$ use empirical posterior quantiles. When t varies across strata, we re–fit per stratum so uncertainty reflects the induced $(\theta_{H|1,s}, \theta_{H|0,s})$.

Computational cost. The base (non–hierarchical) model has dimension 3 and mixes in milliseconds per iteration; hierarchical variants scale linearly in S and remain inexpensive for $S \leq 10$. All experiments run comfortably on a laptop CPU.

3.3 Baselines, frequentist connections, and coverage design

We compare against **three** baselines.

(i) Labeled–only Bayes (NB). Model $H_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta_H)$ with $\theta_H \sim \text{Beta}(1,1)$, so $g = \theta_H$ has the closed–form posterior

$$\theta_H \mid \{H_i\} \sim \text{Beta}\Big(1 + \sum_i H_i,\ 1 + N_H - \sum_i H_i\Big).$$

(ii) Difference estimator (DE). Define

$$\hat{g}_{\text{diff}} = \bar{A} + \overline{(H - A)} = \frac{1}{N_A}\sum_{i=1}^{N_A} A_i + \frac{1}{N_H}\sum_{i=1}^{N_H}(H_i - A_i).$$

We report nonparametric bootstrap confidence intervals for this estimator.

(iii) Prior-free analytic PPI (continuous p). Using probabilities p_i instead of thresholded A_i, let

$$\hat{g}_{\text{PPI}} = \bar{p} + \overline{(H - p)}, \qquad \text{SE}(\hat{g}_{\text{PPI}}) = \sqrt{\widehat{\text{Var}}(p)/N_A + \widehat{\text{Var}}(H - p)/N_H},$$

and form $100(1-\alpha)\%$ intervals with $z_{1-\alpha/2}$ for moderate/large n and $t_{1-\alpha/2,n-1}$ when $n < 30$ (Angelopoulos et al., 2023; Efron & Tibshirani, 1993). This estimator serves as a prior-free reference that keeps the continuous score information.

Asymptotics and label efficiency. When $N_A \to \infty$ and N_H is fixed, uncertainty is dominated by the labeled correction term, motivating our Bayesian CRE which borrows strength structurally from $\mathcal{D}_A$. In finite samples, the main empirical question is whether this structural borrowing improves interval width without sacrificing coverage. *A short theoretical link showing the CRE posterior mean as a first-order shrinkage variant of the DE is provided in App. A.*

Coverage and resampling protocol. Our empirical coverage study is based on *repeated-labeling resampling* rather than a purely synthetic data-generating process. We fix an ADNI-derived prediction table with out-of-fold machine probabilities and observed clinician labels, treat the cohort prevalence

$$g_{\text{true}} = \frac{1}{N}\sum_{i=1}^{N} H_i$$

as the finite-population target, and repeatedly subsample labeled sets of size $n \in \{10, 20, 40, 80\}$ without replacement. For each label budget, we run $M = 500$ replications and report empirical coverage and mean 95% interval width. We perform this evaluation on both the full cohort and a fixed 65–70 subset. Within each replication, when comparing Uniform and Jeffreys priors, we reuse the *same labeled indices* so that differences isolate prior sensitivity rather than relabeling variability.

3.4 Operating thresholds, calibration, and propagation of threshold uncertainty

Autorater scores p are mapped to hard decisions $A = \mathbf{1}\{p \geq t\}$.

Selection of t (OOF vs. leaky). We study (i) a conventional $t = 0.5$, (ii) Youden's $t^\star_Y \in \arg\max_t \text{TPR}(t) + \text{TNR}(t) - 1$, and (iii) a cost–sensitive Bayes threshold

$$t^\star_{\text{Bayes}} = \frac{C_{10}(1-\pi)}{C_{10}(1-\pi) + C_{01}\pi}.$$

where π is prevalence and C_{01}, C_{10} are false–negative/false–positive costs (Youden, 1950; Fluss et al., 2005). To avoid optimistic bias, we choose data–driven thresholds by K–fold *out–of–fold* (OOF) selection: for each fold, fit the autorater on $K-1$ folds, compute $t_Y^\star$ on the held–out fold only, and pool OOF decisions across folds. We report the gap between OOF and leaky selection as an audit metric.

Propagation into CRE. Given a chosen t (possibly stratum–specific), we re–compute A and re–fit CRE so that posterior uncertainty reflects the $(\theta_{H|1}, \theta_{H|0})$ induced by t. To quantify uncertainty in $t_Y^\star$, we bootstrap the labeled set B times, obtain $\{t_{Y,b}^\star\}$, and summarize its spread; optionally, we integrate over t by averaging posteriors across $t_{Y,b}^\star$ (a simple Bayesian bootstrap over cut–points). At very small label budgets the sampling variability of t can dominate; we therefore report the bootstrap dispersion of $t^\star$ and interpret CRE intervals conditionally on the chosen thresholding policy.

Calibration before thresholding. If reliability diagrams and Brier scores indicate miscalibration, we apply (within–stratum) temperature scaling or isotonic regression to produce calibrated $\tilde{p}$, then threshold $\tilde{p}$ at the chosen rule. These monotone mappings preserve ranking and thus AUC, improving decision quality without affecting CRE's form (Brier, 1950; Murphy, 1973; Guo et al., 2017; Zadrozny & Elkan, 2002).

3.5 Case–study pipeline: data curation, autorater, and end–to–end flow

The medical–imaging case study uses ADNI baseline T1-weighted MRI and clinician labels (Mueller et al., 2005; Jack Jr et al., 2008). DICOM→NIfTI conversion uses `dcm2niix`, and volumes are loaded with NiBabel (Li et al., 2016; Brett et al., 2020). Preprocessing consists of nonzero bounding-box crop/pad, resampling to 64^3, and per-volume min–max normalization to $[0, 1]$. The autorater is a lightweight 3D CNN with two Conv3D–ReLU–MaxPool blocks, a dense hidden layer, and a sigmoid output, trained with Adam at learning rate 10^{-4}, batch size 2, and 5 epochs in PyTorch/`numpy` (Kingma & Ba, 2014; Paszke et al., 2019; Harris et al., 2020).

For downstream inference and threshold selection, we use *out-of-fold* (OOF) predicted probabilities on the full cohort so that operating-point analyses are not driven by in-sample fitted scores. Scores p are then mapped to hard decisions A under the selected thresholding policy, and CRE is fit on $(\mathcal{D}_A, \mathcal{D}_H)$ globally and by age strata (50–73, 74–79, 80–100). The end-to-end routine follows Section J.2 for thresholding/calibration and Section J.1 for Bayesian inference.

Robustness to dataset shift and label missingness. If labeled cases differ from the unlabeled pool (covariate shift), MCAR is violated. Practical mitigations include stratified subsampling for labeling, *propensity–overlap* diagnostics (low AUC for a labeled–vs–unlabeled classifier), propensity–weighted PPCs, and hierarchical pooling that absorbs modest shifts (Rosenbaum & Rubin, 1983). When MAR mechanisms are suspected, one can augment CRE with a logistic missingness model for H and perform joint inference; we leave this to future work and rely on design–stage controls plus PPCs in this study. When labeled and unlabeled pools are not exchangeable due to covariate shift, we outline an importance-weighted variant of CRE (IW-CRE) that replaces unweighted cell counts by density-ratio–weighted counts; see Appendix L.

4 Experiments

4.1 Evaluation design, SBC calibration, and posterior behavior

We report two complementary evaluations. First, we use simulation-based calibration (SBC) to validate the conjugate Bayesian engine under draws from the model itself. Second, we report a repeated-labeling resampling study on ADNI-derived out-of-fold prediction tables, which evaluates the finite-population prevalence target under repeated subsampling of scarce labels.

SBC calibration. For SBC, we draw $(\theta_A, \theta_{H|1}, \theta_{H|0})$ from the prior, simulate $(\mathcal{D}_A, \mathcal{D}_H)$ under the base Beta–Bernoulli model, fit the conjugate posterior, and record posterior ranks for the primitive parameters and the derived functional g. We use $M_{\mathrm{SBC}} = 500$ replications. Full rank-histogram panels and goodness-of-

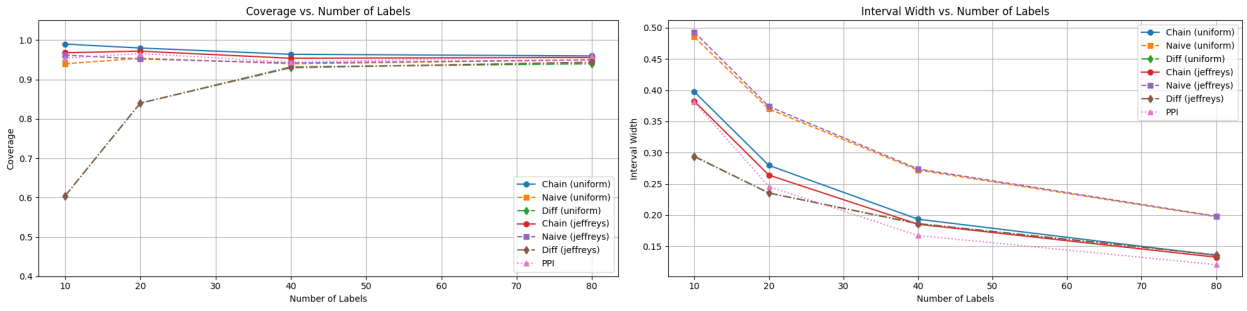
fit p-values appear in Appendix Section I. The resulting histograms are visually close to uniform, supporting calibration of the conjugate inference engine.

Repeated-labeling resampling study. For the empirical coverage study, we fix an ADNI-derived prediction table and repeatedly subsample labeled sets of size $n \in \{10, 20, 40, 80\}$. The full cohort contains $N = 2{,}116$ subjects with finite-population prevalence $g_{\text{true}} = 0.308129$, and the fixed 65–70 subset contains $N = 244$ subjects with $g_{\text{true}} = 0.225410$. For each label budget, we run $M = 500$ replications. The same labeled indices are reused across Uniform and Jeffreys prior comparisons within each replicate. Base CRE posteriors use direct Beta sampling; the labeled-only Bayesian baseline also uses closed-form Beta updates; the classical difference estimator is paired with bootstrap intervals; and the prior-free PPI baseline uses continuous probabilities together with analytic normal/t intervals.

Full cohort: DE vs. CRE. Table 1 summarizes empirical coverage (Cov.) and mean 95% interval width (W) on the full cohort. At the smallest label budget ($n = 10$), CRE attains near-nominal coverage under both priors (Uniform: 0.990; Jeffreys: 0.968) with materially smaller widths than the labeled-only Bayesian baseline, while the classical difference estimator substantially undercovers (0.604). As the label budget increases, CRE remains close to nominal and consistently narrower than the labeled-only Bayesian baseline; the difference estimator improves with more labels but remains less stable at small n. The prior-free analytic PPI baseline is competitive, but in the full cohort it does not dominate CRE on the joint coverage–width tradeoff.

Table 1: Empirical resampling study (M=500): empirical 95% coverage (Cov.) and mean interval width (W) for four estimators. PPI is prior-free; CRE/Naïve/Diff are shown under Uniform/Jeffreys.

n_{labels}	Prior	PPI		CRE		Naïve		Diff	
		Cov.	W	Cov.	W	Cov.	W	Cov.	W
10	Uniform	0.954	0.381	0.990	0.397	0.940	0.485	0.604	0.293
10	Jeffreys	0.954	0.381	0.968	0.382	0.962	0.493	0.604	0.294
20	Uniform	0.966	0.246	0.980	0.280	0.954	0.370	0.840	0.235
20	Jeffreys	0.966	0.246	0.972	0.264	0.952	0.374	0.840	0.235
40	Uniform	0.944	0.167	0.964	0.193	0.940	0.272	0.932	0.186
40	Jeffreys	0.944	0.167	0.954	0.185	0.942	0.274	0.930	0.186
80	Uniform	0.960	0.120	0.960	0.135	0.950	0.197	0.940	0.136
80	Jeffreys	0.960	0.120	0.956	0.132	0.950	0.198	0.944	0.136



(a) Coverage vs. # labels (PPI, CRE, Naïve, Diff; CRE (b) Mean 95% interval width (same four estimators) shown under Uniform/Jeffreys)

Figure 1: Full cohort repeated-labeling resampling results for $n \in \{10, 20, 40, 80\}$. PPI is prior-free; CRE uses Uniform/Jeffreys priors.

Fixed 65–70 subset. For completeness, we repeat the same repeated-labeling study on a fixed 65–70 subset; full plots and the numeric table are deferred to Appendix Section F. Here the qualitative pattern

remains similar but the gains are clearly attenuated, which is consistent with the smaller cohort size. In particular, CRE remains competitive and well calibrated, but it is no longer uniformly narrower than the labeled-only Bayesian baseline. This subset is therefore best interpreted as a smaller-cohort sensitivity analysis rather than a second primary win.

Scope of the current evaluation. The present version focuses on SBC, repeated-labeling resampling, and the ADNI case study. We do not claim a comprehensive synthetic stress-test over all prevalence regimes or dataset-shift mechanisms in this version. Those broader stress tests are natural follow-up experiments, but they are not needed to support the narrower claim made here: that the conjugate binary chain-rule estimator is a useful and well-calibrated deployment-time estimator for prevalence monitoring under scarce labels.

4.2 ADNI case study: discrimination, thresholds, calibration, and subgroup analysis

We evaluate the MRI autorater on the ADNI cohort (ages 50–100) with the preprocessing and CNN described in Section 3 (Mueller et al., 2005; Jack Jr et al., 2008; Paszke et al., 2019; Harris et al., 2020). Overall ROC and AUC (Fig. 2a) use nonparametric bootstrap (B=2000) (Hanley & McNeil, 1982; Efron & Tibshirani, 1993; Fawcett, 2006). Age-stratified ROCs (Fig. 2b) show uniformly strong discrimination with slight attenuation for 80–100; Table 3 reports stratum sizes, prevalence, AUC, and 95% CIs. Pairwise AUC comparisons via two-sided permutation tests show no significant differences; Holm-adjusted p-values are in Table 11.

Operating thresholds per stratum are selected by maximizing Youden's index; $t_Y^\star$ increases with age, trading small TPR drops for larger TNR gains (Table 4) (Youden, 1950; Fluss et al., 2005). We then refit CRE per stratum to propagate the induced $(\theta_{H|1}, \theta_{H|0})$, enabling age-aware monitoring with calibrated intervals.

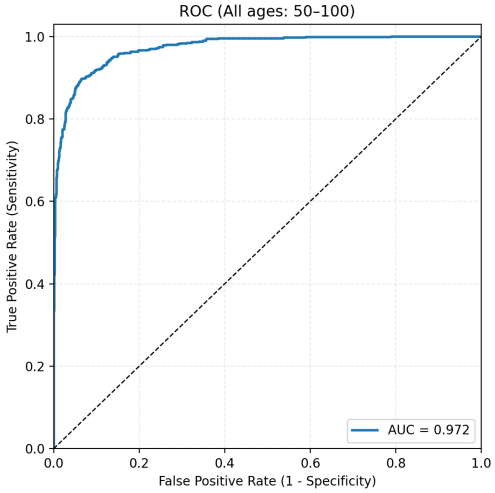
Probability reliability complements discrimination: Brier scores (50–73: 0.070; 74–79: 0.079; 80–100: 0.106) and reliability diagrams indicate mild underconfidence in the youngest stratum and overconfidence in the oldest. Post-hoc calibration (temperature scaling, isotonic regression) improves probability reliability without altering ranking; AUC remains unchanged while thresholded decisions A better align with clinical preferences (Brier, 1950; Murphy, 1973; Guo et al., 2017; Niculescu-Mizil & Caruana, 2005; Zadrozny & Elkan, 2002). Because CRE consumes only A, it integrates seamlessly with probability-level adjustments.

Prevalence estimation on real data: PPI vs. CRE. On ADNI, PPI and CRE produce closely aligned prevalence estimates overall, while age-stratified estimates show small but non-negligible differences reflecting thresholding and subgroup-specific conditional rates; see Table 2. We therefore interpret the table as evidence that CRE is practically compatible with a prior-free continuous-score baseline, while still providing a coherent Bayesian posterior for the thresholded deployment estimand.

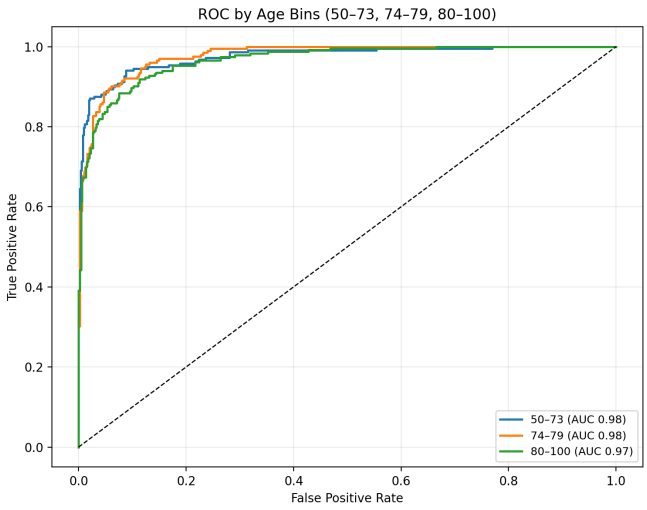
Table 2: Prevalence estimates g on ADNI (overall and by age).

Scope	N	PPI $\hat{g}$ [95%]	CRE (Uniform) [95%]	CRE (Jeffreys) [95%]
Overall	2,116	0.308 [0.283, 0.333]	0.308 [0.289, 0.329]	0.308 [0.288, 0.328]
50–73	737	0.343 [0.313, 0.372]	0.332 [0.307, 0.359]	0.332 [0.305, 0.359]
74–79	695	0.310 [0.280, 0.341]	0.306 [0.279, 0.332]	0.306 [0.280, 0.333]
80–100	684	0.268 [0.234, 0.302]	0.292 [0.266, 0.318]	0.291 [0.267, 0.318]

To probe subgroup equity, we examine dispersion of calibration errors across age strata and re-run CRE after stratum-specific calibration. Multicalibration-style checks (empirical calibration within computationally identifiable subgroups) show reduced dispersion after age-wise calibration (Hebert-Johnson et al., 2018; Pleiss et al., 2017).



(a) Overall ROC on the full cohort (ages 50–100). Shaded belt: 95% bootstrap band.



(b) Age–stratified ROC curves with 95% bootstrap bands.

Figure 2: Overall and age–stratified ROC curves with 95% bootstrap bands.

Table 3: Age–stratified discrimination: sample size (n), prevalence, AUC, and 95% bootstrap CIs.

Age bin	n	**Prevalence**	**AUC**	**95% CI**
50–73	737	0.294	0.975	[0.963, 0.986]
74–79	695	0.291	0.977	[0.967, 0.986]
80–100	684	0.341	0.967	[0.953, 0.978]

Table 4: Operating points by age: fixed t=0.5 vs. Youden's $t_Y^\star$. Entries show point estimates with 95% CIs (Wilson), and AUC with 95% bootstrap CI.

Age bin	**Threshold**	**ACC**	**TPR**	**TNR**	**AUC (95% CI)**
50–73	$t = 0.5$	0.916 [0.894, 0.934]	0.917 [0.873, 0.947]	0.915 [0.888, 0.936]	0.975 [0.963, 0.986]
50–73	$t_Y^\star = 0.484$	0.920 [0.898, 0.937]	0.940 [0.900, 0.965]	0.912 [0.884, 0.933]	0.975 [0.963, 0.986]
74–79	$t = 0.5$	0.901 [0.876, 0.921]	0.946 [0.905, 0.969]	0.882 [0.851, 0.908]	0.977 [0.967, 0.986]
74–79	$t_Y^\star = 0.586$	0.928 [0.906, 0.945]	0.901 [0.852, 0.935]	0.939 [0.914, 0.957]	0.977 [0.967, 0.986]
80–100	$t = 0.5$	0.858 [0.830, 0.882]	0.953 [0.917, 0.973]	0.809 [0.771, 0.843]	0.967 [0.953, 0.978]
80–100	$t_Y^\star = 0.682$	0.911 [0.887, 0.930]	0.884 [0.837, 0.919]	0.925 [0.896, 0.946]	0.967 [0.953, 0.978]

4.3 Computation, diagnostics, and reproducibility

For the base estimator studied in the main experiments, posterior computation is fully conjugate and requires only direct Beta sampling; no MCMC is needed. On a standard laptop CPU, repeated-labeling coverage runs with $M = 500$ complete in seconds to minutes depending on the label budget and bootstrap overhead for the non-Bayesian baselines. NUTS is reserved for non-conjugate extensions only, where standard diagnostics such as rank-normalized $\hat{R}$, ESS, divergences, and E–BFMI apply. We release scripts for data preparation, resampling, model fitting, and figure/table generation to ensure end-to-end reproducibility.

Summary. Across SBC, empirical resampling, and real-data analyses, the main conclusion is consistent: the conjugate chain-rule estimator provides a well-calibrated and practically efficient default for binary prevalence monitoring with scarce labels. Its strongest advantage appears in the full cohort and at small label budgets, where it improves substantially on the instability of the classical difference estimator while remaining narrower than the labeled-only Bayesian baseline. In smaller subsets, the gains are weaker but the estimator remains competitive and well behaved.

5 Conclusion

We studied a narrow but practically important instance of Bayesian prediction-powered inference: binary prevalence estimation through the chain-rule functional $g = \theta_A\theta_{H|1} + (1-\theta_A)\theta_{H|0}$. Our main point is not to re-introduce Bayesian PPI in full generality, but to show that this particular base case is fully conjugate, easy to implement, and already useful for deployment-time monitoring. In this base model, posterior inference requires only direct Beta sampling; MCMC is needed only for explicit non-conjugate extensions such as hierarchical pooling or richer prior structures.

Empirically, the strongest evidence comes from two places. First, simulation-based calibration supports the correctness of the conjugate posterior engine. Second, repeated-labeling resampling on ADNI-derived out-of-fold prediction tables shows that CRE achieves near-nominal coverage in the full cohort, is consistently narrower than the labeled-only Bayesian baseline, and substantially improves on the instability of the classical difference estimator at very small label budgets. In the fixed 65–70 subset, the same qualitative story remains visible but the gains are attenuated, which is exactly the kind of behavior one should expect in a smaller cohort rather than a sign of failure.

A second practical contribution is procedural rather than purely statistical. We pair the estimator with deployment-oriented checks: a prior-free analytic PPI baseline using continuous probabilities, labeled–unlabeled propensity-overlap diagnostics, and OOF versus leaky threshold selection together with bootstrap cut-point dispersion. These components make the method easier to audit and harder to overstate in retrospective evaluation.

The present version intentionally keeps its claims narrow. We do not argue that the current experiments exhaust all shift regimes or all prevalence settings, nor do we claim that the 65–70 subset provides a second independent headline result. Instead, we view this paper as establishing a clean base estimator, a careful evaluation protocol, and a realistic case study that together make Bayesian chain-rule PPI more usable in practice. Future work should expand the stress testing, strengthen the theoretical comparison to semiparametric estimators, and explore richer non-conjugate extensions in settings where the conjugate base model is no longer adequate.

Use of large language models. During manuscript preparation, we used a large language model (ChatGPT) only for minor copy-editing; all technical content and experiments were authored and checked by authors, who take full responsibility for any remaining errors.

References

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Lihua Lei, and Jing Lei. Prediction-powered inference. *Science*, 382(6669):669–674, 2023. doi: 10.1126/science.adg6862.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.

Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Matthew Brett, Michael Hanke, Christopher J. Markiewicz, Paul McCarthy, Marc-Alexandre Côté, Ben Cipollini, Matteo di Oleggio Castello, Satrajit Ghosh, Demian Wassermann, Stephan Gerhard, et al. Nibabel: Neuroimaging in Python, 2020. URL `https://zenodo.org/record/4295521`.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1): 1–3, 1950.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017. doi: 10.18637/jss.v076.i01.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

Abhishek Chakrabortty and T. Tony Cai. Efficient and adaptive linear regression in semi-supervised settings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):907–938, 2018. doi: 10.1111/rssb.12282.

William G. Cochran. *Sampling Techniques*. Wiley, 3 edition, 1977.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992. doi: 10.1080/01621459.1992.10475217.

Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993. ISBN 9780412042317.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Ronit Fluss, David Faraggi, and Benjamin Reiser. Estimation of the Youden index and its associated cutoff point. *Biometrical Journal*, 47(4):458–472, 2005.

Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A*, 182(2):389–402, 2019. doi: 10.1111/rssa.12378.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3 edition, 2013. ISBN 9781439840955.

Jessica L. Gronsbell and T. Tony Cai. Semi-supervised approaches to the estimation of the ROC curve and the area under the curve. *Biometrics*, 74(2):552–562, 2018. doi: 10.1111/biom.12753.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Yue Guo and Lihua Lei. Confidence sets from prediction-powered inference. *Journal of the American Statistical Association*, 2021. doi: 10.1080/01621459.2021.1996374. URL `https://doi.org/10.1080/01621459.2021.1996374`. Published online.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, et al. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

Ursula Hebert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Robert A. Hofer, Ryan J. Tibshirani, and Alexander N. Angelopoulos. A Bayesian formulation of prediction-powered inference. arXiv preprint arXiv:2405.06034, 2024. URL `https://arxiv.org/abs/2405.06034`.

Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014. URL `https://jmlr.org/papers/v15/hoffman14a.html`.

Clifford R. Jack Jr, Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J. Britson, Jennifer Whitwell, Chadwick Ward, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ravin Kumar, Christopher Carroll, Ari Hartikainen, and Osvaldo Martin. Arviz: A unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, 4(33):1143, 2019. doi: 10.21105/joss.01143.

Xiangrui Li, Paul S. Morgan, John Ashburner, Stephen M. Smith, and Chris Rorden. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods*, 264:47–56, 2016.

Sharon L. Lohr. *Sampling: Design and Analysis*. CRC Press, 2 edition, 2019. ISBN 9780367197117.

Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman & Hall/CRC, Boca Raton, FL, 2 edition, 2020.

Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's disease neuroimaging initiative (ADNI). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4): 595–600, 1973.

Radford M. Neal. Mcmc using hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng (eds.), *Handbook of Markov Chain Monte Carlo*, pp. 113–162. Chapman and Hall/CRC, 2011. ISBN 9781420079425.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pp. 625–632, 2005. doi: 10.1145/1102351.1102430.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Geoff Pleiss, Aditi Raghunathan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, 2017.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016. doi: 10.7717/peerj-cs.55.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer, 1992.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. In *JMLR Workshop and Conference Proceedings (PASCAL Workshop on Statistics and Optimization of Clustering Workshop)*, pp. 1–8, 2007.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv:1804.06788*, 2018.

Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011. doi: 10.1007/978-1-4419-9782-1.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017. doi: 10.1007/s11222-016-9696-4.

Aki Vehtari, Andrew Gelman, Paul-Christian Bürkner, Jonah Gabry, Daniel Simpson, and Stan Development Team. Bayesian workflow. arXiv preprint arXiv:2011.01808, 2021a.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16(2):667–718, 2021b. doi: 10.1214/20-BA1221.

W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002. doi: 10.1145/775047.775151.

A Theory snap-in: CRE posterior mean as a first-order shrinkage variant of the difference estimator

Proposition A.1 (Posterior mean as a first-order shrinkage of the difference estimator)**.** *Consider the base Beta–Bernoulli CRE with independent priors* $\theta_A \sim \mathrm{Beta}(\alpha_A, \beta_A)$, $\theta_{H|1} \sim \mathrm{Beta}(\alpha_1, \beta_1)$, $\theta_{H|0} \sim \mathrm{Beta}(\alpha_0, \beta_0)$, *and let* $\hat{\theta}_A = n_A/N_A$, $\hat{\theta}_{H|1} = n_{11}/(n_{11}+n_{10})$, $\hat{\theta}_{H|0} = n_{01}/(n_{01}+n_{00})$. *Define the classical difference estimator* $\hat{g}_{\mathrm{diff}} = \bar{A} + \overline{(H-A)}$. *Then, under exchangeable labeling and* $N_A \gg N_H$,

$$\mathbb{E}[g \mid \mathcal{D}] = \hat{g}_{\mathrm{diff}} + \lambda_A\big(\mathbb{E}[\theta_A \mid \mathcal{D}] - \hat{\theta}_A\big) + \lambda_1\big(\mathbb{E}[\theta_{H|1} \mid \mathcal{D}] - \hat{\theta}_{H|1}\big) + \lambda_0\big(\mathbb{E}[\theta_{H|0} \mid \mathcal{D}] - \hat{\theta}_{H|0}\big) + O_{\mathbb{P}}(N_A^{-1}),$$

where the weights are $\lambda_A = \hat{\theta}_{H|1} - \hat{\theta}_{H|0}$, $\lambda_1 = \hat{\theta}_A$, $\lambda_0 = 1 - \hat{\theta}_A$. *With Jeffreys priors* $\mathrm{Beta}\big(\frac{1}{2}, \frac{1}{2}\big)$, *each* $\mathbb{E}[\theta \mid \mathcal{D}] - \hat{\theta} = O(N^{-1})$, *so the correction is a small, data–adaptive shrinkage that contracts extreme cells and stabilizes interval width near boundaries (Gelman et al., 2013).*

Proof sketch. Write $g(\boldsymbol{\theta}) = \theta_A\theta_{H|1}+(1-\theta_A)\theta_{H|0}$. A first-order delta expansion about $\hat{\boldsymbol{\theta}} = (\hat{\theta}_A, \hat{\theta}_{H|1}, \hat{\theta}_{H|0})$ gives

$$\mathbb{E}[g \mid \mathcal{D}] \approx g(\hat{\boldsymbol{\theta}}) + \nabla g(\hat{\boldsymbol{\theta}})^\top \Big(\mathbb{E}[\boldsymbol{\theta} \mid \mathcal{D}] - \hat{\boldsymbol{\theta}}\Big), \quad \nabla g(\hat{\boldsymbol{\theta}}) = \big(\hat{\theta}_{H|1} - \hat{\theta}_{H|0},\ \hat{\theta}_A,\ 1 - \hat{\theta}_A\big).$$

Under $N_A \gg N_H$ and exchangeability, $g(\hat{\boldsymbol{\theta}})$ is asymptotically equivalent to the design–unbiased difference form $\hat{g}_{\text{diff}}$ up to $O_{\mathbb{P}}(N_A^{-1})$ terms (linearization and the independence of the large-N_A and labeled components), matching classical calibration/difference–estimator theory (Cochran, 1977; Särndal et al., 1992; Lohr, 2019). Beta posteriors yield $\mathbb{E}[\theta_A \mid \mathcal{D}] = (\alpha_A + n_A)/(\alpha_A + \beta_A + N_A)$ and similarly for $\theta_{H|a}$; substituting gives the stated shrinkage correction. □

Remark (stability near the boundaries). When some (2×2) cells are tiny, Jeffreys or weakly-informative priors keep posterior means away from $\{0, 1\}$, tightening the tail behavior of g and making the $(O(N^{-1}))$ correction beneficial in practice (Gelman et al., 2013).

B Prior-free PPI baseline (analytic): estimator and CIs

For the prevalence functional, the *prediction–powered* estimator using probabilities is

$$\hat{g}_{\text{PPI}} = \bar{p} + \overline{(H - p)} = \frac{1}{N_A}\sum_{i=1}^{N_A} p_i + \frac{1}{N_H}\sum_{i=1}^{N_H}(H_i - p_i).$$

with large-sample variance $\widehat{\text{Var}}(\hat{g}_{\text{PPI}}) = \widehat{\text{Var}}(p)/N_A + \widehat{\text{Var}}(H-p)/N_H$. For very small n we use a t-critical value with Satterthwaite's degrees of freedom; otherwise $z_{0.975}$ suffices. **This is distinct from the *difference estimator* based on binary** A, $\hat{g}_{\text{diff}} = \bar{A} + \overline{(H - A)}$, for which we report nonparametric bootstrap percentile CIs (see Methods).

C Exchangeability diagnostics (propensity overlap)

We train a simple "labeled vs. unlabeled" propensity model and report AUC (near 0.5 implies overlap).

Table 5: Propensity AUC and sample sizes (exchangeability check).

R	n_{labeled}	N	AUC (mean)	95% CI
50	634	2,116	0.522	[0.507, 0.543]

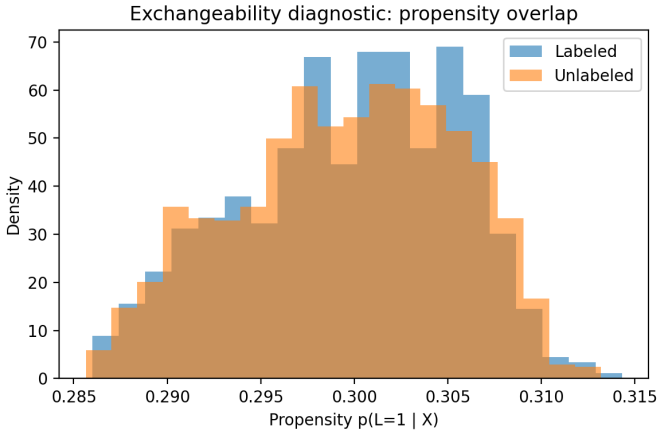


Figure 3: Propensity overlap histograms for labeled vs. unlabeled pools (near-complete overlap).

D Threshold selection: OOF vs. leaky and bootstrap dispersion

OOF vs. leaky (train=eval) selection

Table 6: Five-fold OOF thresholds and performance.

Fold	t_{train}	ACC	TPR	TNR
1	0.6026	0.9151	0.8678	0.9340
2	0.6073	0.9243	0.8790	0.9431
3	0.6026	0.9220	0.8944	0.9359
4	0.6026	0.9173	0.8872	0.9310
5	0.6122	0.9291	0.9318	0.9278

Table 7: Leaky (train=eval) selection

Fold	t_{train}	ACC	TPR	TNR
leaky_full	0.6026	0.9230	0.8972	0.9344

Bootstrap dispersion of Youden thresholds

Table 8: Bootstrap summary of $t_Y^\star$ and induced metrics.

mean(t)	t_{lo}	t_{hi}	ACC_m	ACC_ℓ	ACC_h	TPR_m	TPR_ℓ	TPR_h	TNR_m	TNR_ℓ	TNR_h
0.601	0.503	0.631	0.921	0.894	0.926	0.896	0.876	0.936	0.933	0.876	0.947

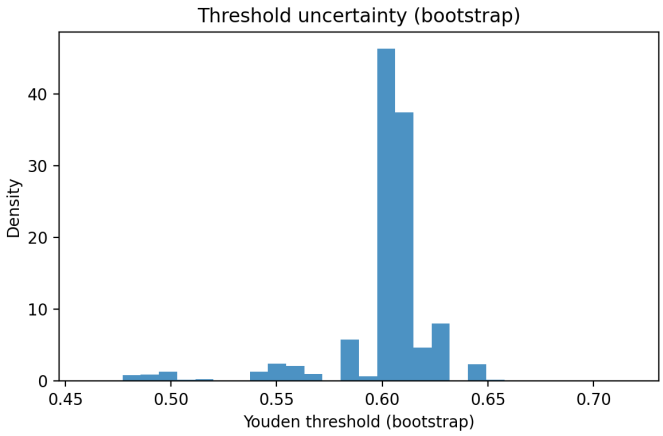
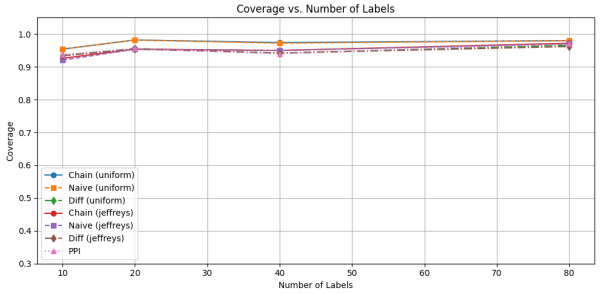


Figure 4: Bootstrap distribution of $t_Y^\star$ and operating metrics.

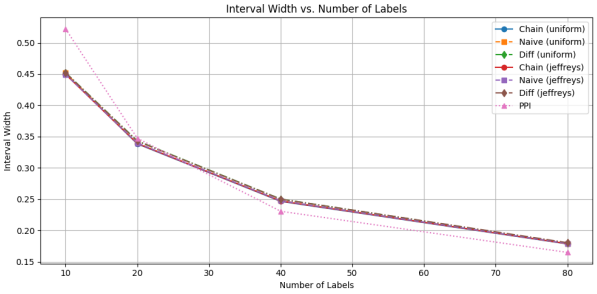
E DE and PPI baselines: additional simulation tables

Detailed DE vs. CRE numbers are reported in the main text (Table 1); extended tables and per-replication summaries are included in the repository.

F Additional results: Age 65–70 subset



(a) Coverage vs. # labels (65–70; PPI prior-free; CRE under Uniform/Jeffreys)



(b) Mean 95% interval width (same estimators)

Figure 5: 65–70 subset: coverage and width across $n \in \{10, 20, 40, 80\}$.

Table 9: 65–70 subset: empirical 95% coverage (Cov.) and mean interval width (W). PPI is prior-free and listed once. CRE/Naïve/Diff are shown under Uniform/Jeffreys.

n_{labels}	Prior	PPI (prob.)		CRE		Naïve		Diff	
		Cov.	W	Cov.	W	Cov.	W	Cov.	W
10	Uniform	0.938	0.522	0.954	0.450	0.954	0.452	0.934	0.452
20	Uniform	0.956	0.347	0.982	0.338	0.982	0.340	0.956	0.343
40	Uniform	0.940	0.231	0.974	0.246	0.972	0.248	0.942	0.250
80	Uniform	0.970	0.165	0.980	0.178	0.980	0.179	0.966	0.180
10	Jeffreys	0.938	0.522	0.926	0.449	0.920	0.449	0.934	0.452
20	Jeffreys	0.956	0.347	0.954	0.339	0.954	0.340	0.954	0.343
40	Jeffreys	0.940	0.231	0.950	0.247	0.950	0.247	0.942	0.250
80	Jeffreys	0.970	0.165	0.972	0.179	0.970	0.179	0.962	0.180

G Prevalence estimates on real data (overall and by age)

Extended versions of Table 2 (stratifications and alternative thresholds) are available in the repository; we keep a single summary table in the main text to avoid duplication.

H K-bin sensitivity (quantile binning)

We compare $K \in \{2, 4, 5\}$ with Dirichlet prior on $(P(B{=}k))$ and Beta priors on $P(H{=}1 \mid B{=}k)$. Under the label budgets considered, increasing K does not materially change $\hat{g}$; CI widths change minimally.

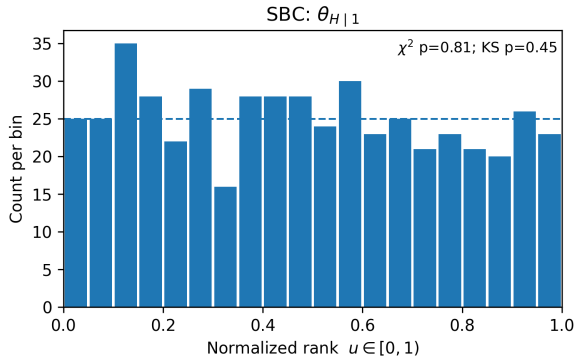
Table 10: K-bin sensitivity summary (numbers generated by repo scripts).

Prior	K	ΔMean vs $K{=}2$	Width ratio $(K/2)$
Uniform	4	-8.55×10^{-5}	1.009
Uniform	5	3.46×10^{-4}	1.001
Jeffreys	4	-7.25×10^{-5}	0.990
Jeffreys	5	1.07×10^{-4}	0.990

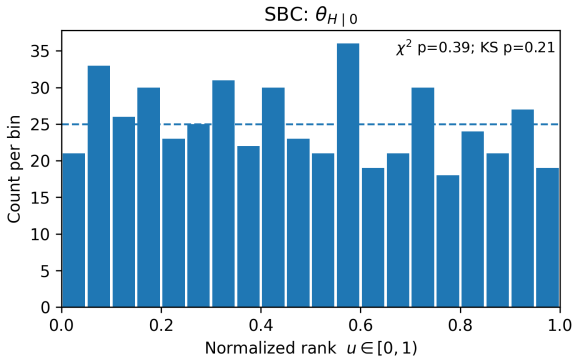
Protocol. We fix labeled indices across priors (prior-invariant comparison), use quantile binning with minimal tie drops, and report Δmean and width ratios.

I SBC and additional diagnostics

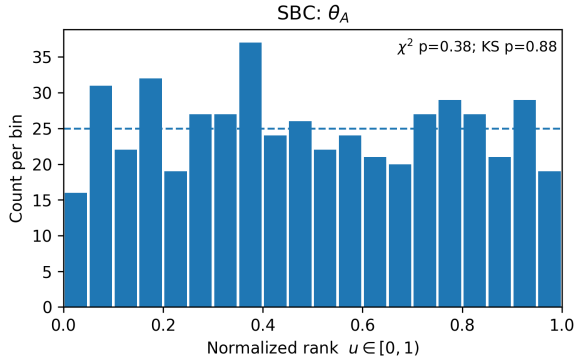
Over $M_{\text{SBC}} = 500$, posterior rank histograms for $\theta_A, \theta_{H|1}, \theta_{H|0}$ and the derived g are visually close to uniform. With $B = 20$ bins (expected count $= M/B = 25$), fluctuations fall within the nominal 95% binomial band $\approx [15, 35]$. Goodness–of–fit tests are non–significant for all parameters (per–panel p–values in captions), supporting approximate calibration.



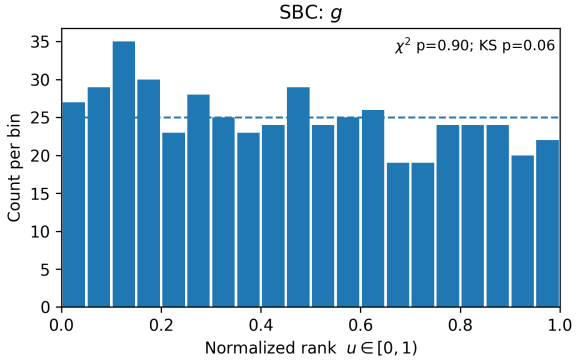
(a) $\theta_{H|1}$: $\chi^2\, p = 0.81$; KS $p = 0.45$



(b) $\theta_{H|0}$: $\chi^2\, p = 0.39$; KS $p = 0.21$

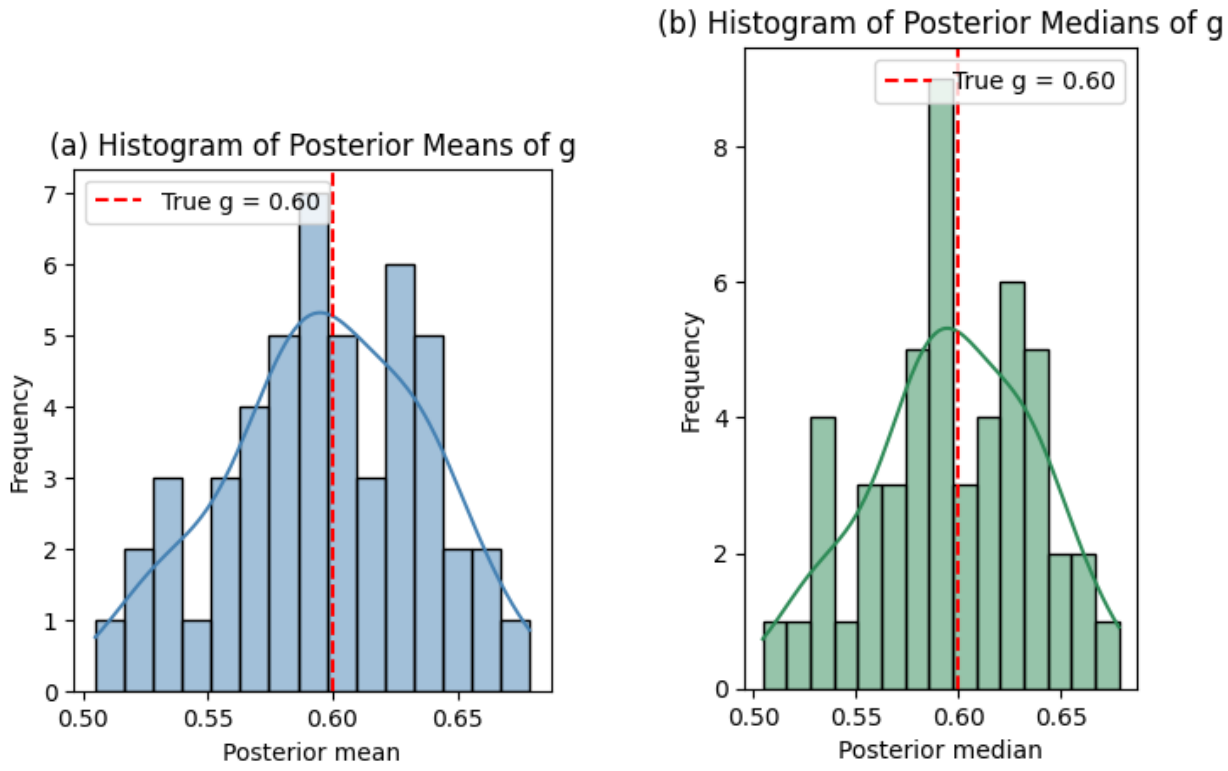


(c) θ_A: $\chi^2\, p = 0.38$; KS $p = 0.88$



(d) $g = \theta_A\theta_{H|1} + (1-\theta_A)\theta_{H|0}$: $\chi^2\, p = 0.90$; KS $p = 0.06$

Figure 6: SBC rank histograms (M=500, B=20). Dashed line marks the expected per–bin count (25). Counts fluctuate within $[15, 35]$, consistent with binomial variability; tests do not reject uniformity.



(a) Posterior means across repeated datasets.

(b) Posterior medians across repeated datasets.

Figure 7: Posterior mean/median summaries across repeated datasets (vertical line: $g_{\text{true}} = 0.60$).

J Algorithms (pseudocode)

J.1 Bayesian CRE inference

Algorithm 1 CRE (base: conjugate; extensions: NUTS)

Require: $\mathcal{D}_A = \{A_i\}_{i=1}^{N_A}$, $\mathcal{D}_H = \{(A_i, H_i)\}_{i=1}^{N_H}$, priors
1: Compute cell counts $(n_A, n_{11}, n_{10}, n_{01}, n_{00})$
2: **if** base Beta–Bernoulli **then**
3: Draw $\theta_A, \theta_{H|1}, \theta_{H|0}$ from Beta posteriors (Prop. 1)
4: Map draws to $g = \theta_A\theta_{H|1} + (1-\theta_A)\theta_{H|0}$
5: **else** ▷ hierarchy / non-Beta priors / K-bin / joint t
6: Run NUTS: `draws=2000, tune=1000, target_accept=0.95, chains=4`
7: Check $\hat{R} < 1.01$, bulk/tail ESS, divergences; then map draws to g
8: **end if**
9: PPCs on (A, H) margins and g; optionally SBC for end-to-end checks

J.2 Operating threshold selection and uncertainty propagation

Algorithm 2 Thresholding and calibration workflow

Require: Scores p_i, labels H_i (dev split), policy $\in \{t{=}0.5, t^\star_Y, t^\star_{\text{Bayes}}\}$
1: **if** $t^\star_Y$ **then**
2: Sweep t; pick t maximizing $J(t) = \text{TPR}(t) + \text{TNR}(t) - 1$
3: Bootstrap labeled set B times to obtain $\{t^\star_{Y,b}\}$ and summarize
4: **end if**
5: Calibrate (temperature scaling or isotonic) if needed; apply t to get A
6: Fit CRE on $(\mathcal{D}_A, \mathcal{D}_H)$; report posterior for g (and by strata)

K Permutation tests: age-stratum AUC comparisons

Holm-adjusted p-values (two-sided permutation with label shuffles, bin sizes fixed). No comparison is significant at $\alpha = 0.05$.

Table 11: Pairwise AUC differences across age strata.

Comparison	AUC diff	Adj. p
50–73 vs. 74–79	-0.002190	0.779
50–73 vs. 80–100	0.008512	0.654
74–79 vs. 80–100	0.010702	0.496

L Importance-weighted CRE under covariate shift

Suppose unlabeled data follow $p_{\text{pop}}(X)$ while labels come from $p_{\text{lab}}(X)$. Let stabilized importance weights $w(x) \propto p_{\text{pop}}(x)/p_{\text{lab}}(x)$, and normalized $\tilde{w}_i = w(X_i)/\big(\frac{1}{n}\sum_j w(X_j)\big)$. Replace unweighted (A, H) margins by their weighted analogs; for the unlabeled pool use $\tilde{v}_i$ analogously. Kish effective sizes $n^{(H)}_{\text{eff}} = \big(\sum_i \tilde{w}_i\big)^2/\sum_i \tilde{w}_i^2$, $N^{(A)}_{\text{eff}} = \big(\sum_i \tilde{v}_i\big)^2/\sum_i \tilde{v}_i^2$ quantify variance inflation. Conjugate updates proceed with fractional counts:

$$\theta_A \,|\, \mathcal{D} \sim \text{Beta}\big(\alpha_A{+}\tilde{n}_A,\ \beta_A{+}\tilde{N}_A{-}\tilde{n}_A\big),\ \theta_{H|1} \,|\, \mathcal{D} \sim \text{Beta}\big(\alpha_1{+}\tilde{n}_{11},\ \beta_1{+}\tilde{n}_{10}\big),\ \theta_{H|0} \,|\, \mathcal{D} \sim \text{Beta}\big(\alpha_0{+}\tilde{n}_{01},\ \beta_0{+}\tilde{n}_{00}\big).$$

Map draws to g as usual. In practice, clip extreme weights and report n_{eff}.

M Reproducibility checklist

- Seeds fixed (`numpy/PyMC/python`); versions pinned in `environment.yml`.
- Config-driven runs; figures/tables built via `scripts/` and Makefile targets.
- Diagnostics reported: rank-normalized $\hat{R}$, bulk/tail ESS, divergences, E–BFMI.
- SBC: $M_{\mathrm{SBC}} = 500$ with rank histograms.
- Data splits, bootstrap/permutation indices saved and versioned.

N Code and data availability

The full codebase, including `code/`, `scripts/`, and configuration files used for all figures and tables, is provided as an anonymized code archive attached to this submission. ADNI participant-level data cannot be redistributed under the ADNI Data Use Agreement; qualified researchers can apply at `adni.loni.usc.edu`. We release code and summary-level results only.