

Population-Aware Bayesian Fairness in AI under Complex Survey Designs

Debashis Chatterjee

`debashis.chatterjee@visva-bharati.ac.in`

Visva-Bharati University

Research Article

Keywords: Bayesian fairness, complex survey design, hierarchical modeling, population inference, pseudo-posterior calibration, equal opportunity

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11335320/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Population-Aware Bayesian Fairness in AI under Complex Survey Designs

Debashis Chatterjee^{*1}

¹Department of Statistics, Visva-Bharati University, Santiniketan, West Bengal, India

October 11, 2026

Abstract

Fairness assessments of predictive systems often conflate disparities in a collected sample with disparities in the population of interest. We develop a Bayesian population-aware fairness framework that integrates target-population true-positive-rate estimands, complex-survey weighting, hierarchical partial pooling, and uncertainty-based fairness decisions. We establish an exact composition–rate transport decomposition, a posterior certification envelope, and sharp partial-identification bounds when population strata lack support. The framework is evaluated across five finite-population sampling mechanisms, each with 250 repeated samples, and applied to the 2018 Alabama American Community Survey using an employment-status classifier trained on the 2017 release. In simulations, unadjusted weighted pseudo-posterior intervals undercover under strongly informative sampling, while scalar design-variance alignment improves calibration. In the ACS audit, the estimated group true-positive-rate gap decreases from 0.0536 without weights to 0.0212 with survey weights. Working posterior intervals are narrower than replicate-weight uncertainty, and a simultaneous fairness certificate remains inconclusive. These results demonstrate why target definitions, survey design, and uncertainty calibration must be integrated before drawing population-level fairness conclusions, without implying causal discrimination or universally valid posterior certification.

Mathematics Subject Classification (2020): 62D05, 62F15 (Primary); 62H30, 62C10, 68T05 (Secondary).

Keywords: Bayesian fairness; complex survey design; hierarchical modeling; population inference; pseudo-posterior calibration; equal opportunity.

1 Introduction

An algorithm evaluated on a dataset does not necessarily have the same performance in the population in which it will be used. This is particularly consequential for group fairness: a reported difference in true-positive rates (TPRs), for example, is a statement about a *distribution*, not a context-free characteristic of a classifier. The distribution may change because the sample was drawn unequally, because the composition of protected groups varies across locations, or because conditional outcomes change over time. Foundational discussions emphasize that definitions of fairness involve substantive choices about populations, outcomes, and interventions (Hardt et al., 2016; Mitchell et al., 2021; Mehrabi et al., 2021; Pessach and Shmueli, 2022). We address one part of this larger ethical problem: how to make the inferential target, sampling mechanism, and uncertainty operationally visible.

^{*}Corresponding author. Email: debashis.chatterjee@visva-bharati.ac.in

Survey statisticians have long distinguished finite-population inference from inferences conditional on an observed sample. Hierarchical modeling and poststratification provide tools for sparse domains, while informative-sampling methods address differences between the sample and target populations (Park et al., 2004; Gelman, 2007; Lax and Phillips, 2009; Si et al., 2015; Savitsky and Toth, 2016). Statistical fairness research, meanwhile, has developed Bayesian uncertainty measures for group-specific classification outcomes (Barrainkua et al., 2024, 2023; Panero et al., 2026) and fairness-constrained Bayesian inference (Lee et al., 2026). Recent guidance explicitly recognizes the consequences of complex surveys for machine-learning evaluation (Oh et al., 2026). Thus, none of Bayesian fairness, poststratification, or survey-aware machine learning can independently be claimed as novel here.

Our focus is a unified and auditable *target-population fairness assessment* that (i) identifies the population and the outcome-conditioned denominator of a fairness quantity, (ii) separates known design corrections from model-based extrapolation, (iii) propagates uncertainty in within-cell rates and target composition, and (iv) allows an explicit conclusion of non-identifiability or insufficient evidence. We refer to this approach as Bayesian population-aware AI (BPAI).

Contributions and scope. This article makes four contributions. First, we formalize a population-conditional equal-opportunity estimand, including the outcome-conditioned cell weights that are required for correct poststratification. Second, we establish an exact composition-rate decomposition, a simultaneous posterior fairness-certification bound, and sharp partial-identification bounds under support failure (proofs in Appendix A). Third, we investigate five sampling regimes through 250 repeated samples per regime, evaluating bias, root mean squared error, interval coverage, and erroneous low-risk fairness declarations. Fourth, we implement a survey-weighted audit using the 2018 Alabama American Community Survey (ACS), with a classifier trained on the 2017 Alabama release; we compare sample and design-weighted disparities, pseudo-posterior uncertainty, and simultaneous fairness certification. We distinguish algebraic results from empirically assessed operating characteristics and do not claim universal design calibration, causal identification, or algorithmic superiority.

Organization of the paper. Section 2 reviews the relevant fairness and survey-inference literature. Sections 3–7 introduce the population estimand, the inferential framework, the transport decomposition, the fairness-certification result, and the partial-identification analysis. Section 8 specifies the empirical design and survey-weighted estimators. Section 9 reports the finite-population experiments, followed by the ACS case study in Section 10. Sections 11–13 synthesize the findings, discuss limitations and further research, and conclude; mathematical proofs appear in Appendix A.

2 Related work and research gap

Hardt et al. (2016) formalized an equal-opportunity criterion in supervised prediction. Reviews by Mitchell et al. (2021) and Mehrabi et al. (2021) emphasize that fairness indicators reflect assumptions about data generation and populations. These contributions make it inappropriate to treat an observed confusion matrix as the complete normative object.

Bayesian uncertainty in fairness is not new. Barrainkua et al. (2024) developed an uncertainty-aware method for comparing classification performance and fairness using Bayesian group confusion matrices; Panero et al. (2026) discussed frequentist and Bayesian fairness uncertainty; and Lee et al. (2026) studied fairness through a matched Gibbs posterior. Their existence moves our proposal beyond the generic claim that posterior intervals should accompany fairness metrics: the target population, sampling design, and transport assumptions must be explicit *within* the

quantity whose posterior is reported.

The survey literature provides equally important precedents. [Gelman \(2007\)](#) and [Park et al. \(2004\)](#) analyzed the difficulties and advantages of survey-aware multilevel and poststratified estimates; [Si et al. \(2015\)](#) proposed hierarchical modeling of weights; and [Savitsky and Toth \(2016\)](#) established conditions for pseudo-posterior consistency under informative sampling. [Williams and Savitsky \(2021\)](#) showed why complex-design dependence can compromise uncertainty quantification from an unadjusted pseudo-posterior and proposed a covariance adjustment. Most directly, the survey-aware machine-learning review of [Oh et al. \(2026\)](#) addresses design-aware estimation and evaluation. BPAI builds on, rather than replaces, these lines of research by treating the *target-population fairness event*, its design-dependent uncertainty, and any resulting audit decision as linked inferential objects.

Table 1: Positioning of the BPAI framework relative to established lines of work. These distinctions concern primary emphases, not universal absences in the existing literature.

Line of research	Main established contribution	Additional BPAI question
Algorithmic fairness	Definitions, disparities, constrained decisions	Which population and denominator define the reported disparity?
Bayesian fairness	Posterior uncertainty and fairness-aware learning	How do target cell shares and sampling mechanisms enter the posterior event?
Survey Bayes / MRP	Partial pooling, weight modeling, poststratification	Which fairness functionals admit calibrated design-aware uncertainty?
Survey-aware ML	Workflow safeguards against ignoring design metadata	When can a population fairness claim be certified, deferred, or ruled non-identifiable?
Transport analysis	Distribution changes across domains	Which portion of a fairness shift is compositional rather than within-cell?

3 Target-population fairness as the primary estimand

Let $U^\star$ denote a finite target population or, when appropriate, let $\mathcal{P}^\star$ denote a target superpopulation law. For individual i , write $X_i \in \mathcal{X}$ for predictors, $Y_i \in \{0, 1\}$ for an observed outcome, $G_i \in \{A, B\}$ for a group, and S_i for geographic/time context. A trained classifier $h : \mathcal{X} \rightarrow \{0, 1\}$ is *fixed* for the fairness audit. This distinction prevents a posterior for the audit data from being silently treated as training uncertainty. When h is itself uncertain, an outer layer is required and is left for future research.

Define $Z_i = h(X_i)$, and let $I_i \in \{0, 1\}$ indicate inclusion and response under survey design $\mathcal{S}$. Provided the conditioning events have positive probability, the population equal-opportunity gap is

$$\Delta^\star(h) = \text{TPR}_A^\star(h) - \text{TPR}_B^\star(h), \quad \text{TPR}_g^\star(h) = \Pr_{\mathcal{P}^\star}\{Z = 1 \mid Y = 1, G = g\}. \quad (1)$$

By contrast, a naive sample gap conditions additionally on $I = 1$. Equality of the two requires conditions on the selection mechanism or suitable correction; it is not guaranteed by a large sample size.

For a finite population with $N_{g+} = \sum_{i \in U^\star} \mathbf{1}\{Y_i = 1, G_i = g\} > 0$, the same target has the descriptive form

$$\text{TPR}_g^\star(h) = \frac{\sum_{i \in U^\star} \mathbf{1}\{G_i = g, Y_i = 1\} h(X_i)}{\sum_{i \in U^\star} \mathbf{1}\{G_i = g, Y_i = 1\}}. \quad (2)$$

This is an association with recorded Y , not a proof of causal fairness or non-discrimination. Outcomes may reflect institutional discrimination, historical inequality, missing-not-at-random reporting, or selective labels (Mitchell et al., 2021; Rubin, 1976).

3.1 Poststratification requires the correct conditional composition

Partition the joint covariate-context space into cells $C \in \mathcal{C}$ and set

$$q_{gc}^* = \Pr_{\mathcal{P}^*}(C = c \mid Y = 1, G = g), \quad \theta_{gc}^* = \Pr_{\mathcal{P}^*}(Z = 1 \mid Y = 1, G = g, C = c). \quad (3)$$

Then the *exact identity*

$$\text{TPR}_g^* = \sum_{c \in \mathcal{C}} q_{gc}^* \theta_{gc}^* \quad (4)$$

uses cell shares among *population positives*, not shares among all people in group g . If $r_{gc}^* = \Pr_{\mathcal{P}^*}(C = c \mid G = g)$ and $\kappa_{gc}^* = \Pr_{\mathcal{P}^*}(Y = 1 \mid G = g, C = c)$, then

$$q_{gc}^* = \frac{r_{gc}^* \kappa_{gc}^*}{\sum_{d \in \mathcal{C}} r_{gd}^* \kappa_{gd}^*}, \quad \sum_d r_{gd}^* \kappa_{gd}^* > 0. \quad (5)$$

Consequently, external population totals for demographics alone do not identify target-population TPRs when group- or cell-level outcome prevalences are unknown. A joint model for (r, κ, θ) , or weighted estimators with credible design information, is needed.

4 Bayesian and survey-aware inferential framework

4.1 Hierarchical component

One interpretable component is a hierarchical logistic model for positive-outcome prevalence,

$$\begin{aligned} Y_i \mid X_i, G_i, S_i, \beta, \mathbf{u}, \mathbf{v} &\sim \text{Bernoulli}(\kappa_i), \\ \text{logit}(\kappa_i) &= \mathbf{x}_i^\top \beta + u_{S_i} + v_{G_i} + \gamma_{S_i, G_i}, \\ u_s &\sim N(0, \tau_u^2), \quad v_g \sim N(0, \tau_v^2), \quad \gamma_{sg} \sim N(0, \tau_\gamma^2). \end{aligned} \quad (6)$$

For fairness evaluation, conditional rates θ_{gc} can be modeled using a binomial or logistic component for $Z = h(X)$ among cases with $Y = 1$; incorporating measured predictors may reduce residual cell heterogeneity. Equivalently, a joint model for (X, Y, G, S) plus fixed h defines θ implicitly. Hierarchical shrinkage can stabilize small cells but can also hide genuine subgroup disparities under misspecified pooling assumptions. Priors for scale parameters, interaction structures and inclusion variables must therefore be stress-tested rather than treated as automatically benign.

4.2 Two conceptually different Bayesian routes

Route I: generative population–selection modeling. Specify a finite-population model and a sampling/response model $p(I_i \mid X_i, Y_i, G_i, S_i, \xi)$, together with known design constraints. Infer unobserved population quantities and then the functional in Eq. (1). This provides a generative interpretation when selection and nonresponse assumptions are defensible, but known first-order weights alone rarely identify the full selection process.

Route II: weighted pseudo-Bayesian inference. If a population likelihood $p_{\vartheta}(y_i, z_i \mid x_i, g_i, s_i)$ is posited, one may construct

$$\tilde{\pi}(\vartheta \mid \mathcal{D}, \mathcal{S}) \propto \pi(\vartheta) \prod_{i:I_i=1} p_{\vartheta}(y_i, z_i \mid x_i, g_i, s_i)^{\tilde{w}_i}, \quad \tilde{w}_i = \frac{nw_i}{\sum_{j:I_j=1} w_j}. \quad (7)$$

The exact normalization can be adapted to the estimation domain and documented. Here w_i represents available inclusion/nonresponse-adjusted survey weights; $\tilde{\pi}$ is a *pseudo-posterior*, not generally an ordinary likelihood-based Bayesian posterior. Consistency needs assumptions of the kind studied by [Savitsky and Toth \(2016\)](#), and the nominal spread may not reproduce complex-design variance ([Williams and Savitsky, 2021](#)). In particular, raising contributions to weights does not remove within-cluster dependence or guarantee frequentist coverage.

Design-sensitive uncertainty. Let $\hat{\vartheta}$ solve weighted estimating equations. A large-sample design-based approximation has sandwich form $H^{-1}JH^{-1}$ under suitable regularity conditions, where H is a sensitivity matrix and J incorporates the actual design covariance of the weighted score. A pragmatic sensitivity analysis transforms pseudo-posterior disparity draws so that their scalar variance agrees with a design-based variance estimate, as motivated by the broader calibration literature ([Williams and Savitsky, 2021](#)). We examine this particular adjustment in Section 9. It is specific to the reported functional: agreement on a scalar variance does not establish joint-posterior calibration, and subgroup TPR ratios may remain difficult to estimate when positive outcomes are rare.

5 An algebraic transport decomposition

Let P denote a source environment and Q a target environment. For each group g and cell c , let q_{gc}^P, q_{gc}^Q be the shares conditional on $(Y = 1, G = g)$ and let $\theta_{gc}^P, \theta_{gc}^Q \in [0, 1]$ denote the corresponding conditional TPR contributions. Define $T_g^R = \sum_c q_{gc}^R \theta_{gc}^R$ for $R \in \{P, Q\}$.

Proposition 1 (Exact composition–rate decomposition). *For each group g ,*

$$T_g^Q - T_g^P = \underbrace{\sum_c (q_{gc}^Q - q_{gc}^P) \theta_{gc}^P}_{\text{positive-outcome composition}} + \underbrace{\sum_c q_{gc}^Q (\theta_{gc}^Q - \theta_{gc}^P)}_{\text{within-cell rate change}}. \quad (8)$$

If $\text{TV}(q_g^Q, q_g^P) = \frac{1}{2} \sum_c |q_{gc}^Q - q_{gc}^P|$, then

$$|\Delta^Q - \Delta^P| \leq \sum_{g \in \{A, B\}} \left[\text{TV}(q_g^Q, q_g^P) + \sum_c q_{gc}^Q |\theta_{gc}^Q - \theta_{gc}^P| \right]. \quad (9)$$

Its proof is in Appendix A. The decomposition is descriptive and does not attribute causal effects to a survey design or demographic differences. It nonetheless prevents a material mistake: interpreting a change in a marginal fairness indicator as necessarily reflecting a change in classifier behavior within comparable cells.

5.1 A worked hypothetical reversal (not an ACS result)

Consider two cells, $c = 1, 2$, and one fixed classification rule with no within-cell shift. Suppose $(\theta_{A1}, \theta_{A2}) = (0.90, 0.50)$ and $(\theta_{B1}, \theta_{B2}) = (0.85, 0.55)$. In a hypothetical *observed positive-outcome sample*, take $q_A^P = (0.20, 0.80)$ and $q_B^P = (0.80, 0.20)$. Then $T_A^P = 0.58$, $T_B^P = 0.79$, and $\Delta^P = -0.21$. In a hypothetical *target positive-outcome population*, instead take $q_A^Q = (0.70, 0.30)$

and $q_B^Q = (0.30, 0.70)$. Then $T_A^Q = 0.78$, $T_B^Q = 0.64$, and $\Delta^Q = +0.14$. The sign reverses despite unchanged cell-specific θ . These numbers are *constructed for mathematical illustration*; they are not estimated from the Folktables or the ACS data. The example illustrates why sample composition and target composition must be distinguished before drawing a conclusion about population fairness.

6 Uncertainty-certified fairness claims

A reported posterior mean of Δ^* is insufficient for high-consequence decisions. Let $\varepsilon > 0$ be a context-dependent tolerable magnitude and consider the *violation event* $V_\varepsilon = \{|\Delta^*| > \varepsilon\}$, with posterior probability $\rho_\varepsilon = \Pi(V_\varepsilon \mid \mathcal{D}, \mathcal{S}, \mathcal{P}^*)$. Here Π denotes an explicitly declared *working uncertainty distribution*; if derived from a pseudo-posterior, its interpretation and calibration must be qualified.

Theorem 2 (Posterior fairness-certification envelope). *Fix reference positive-outcome cell compositions $\bar{q}_g \in \mathbb{S}^{|C|-1}$ and reference rates $\bar{\theta}_{gc} \in [0, 1]$. Define $\bar{\Delta} = \sum_c \bar{q}_{Ac} \bar{\theta}_{Ac} - \sum_c \bar{q}_{Bc} \bar{\theta}_{Bc}$. Suppose the uncertainty distribution assigns probability at least $1 - \alpha$ to the simultaneous event*

$$E = \bigcap_{g=A,B} \left\{ \text{TV}(q_g^*, \bar{q}_g) \leq \eta_g, \quad |\theta_{gc}^* - \bar{\theta}_{gc}| \leq e_{gc} \text{ for every } c \right\}, \quad (10)$$

where all tolerances are nonnegative. Put

$$B = \sum_{g=A,B} \left\{ \eta_g + \sum_c \bar{q}_{gc} e_{gc} \right\}. \quad (11)$$

Then

$$\Pi \left(|\Delta^* - \bar{\Delta}| \leq B \mid \mathcal{D}, \mathcal{S}, \mathcal{P}^* \right) \geq 1 - \alpha. \quad (12)$$

Consequently, (i) if $|\bar{\Delta}| + B \leq \varepsilon$, then $\rho_\varepsilon \leq \alpha$; and (ii) if $|\bar{\Delta}| - B > \varepsilon$, then $\rho_\varepsilon \geq 1 - \alpha$.

The proof appears in Appendix A. The theorem is a conditional algebraic certificate: it says what follows *if* the asserted simultaneous posterior event is valid. It does *not* establish that a particular pseudo-posterior has nominal design-based coverage. Constructing and empirically validating useful simultaneous E is an important research task. The result also clarifies an asymmetry: the existence of a large uncertain posterior interval does not imply either fairness or unfairness; it supports abstention from a strong claim.

6.1 Decision-theoretic audit, deployment and abstention

Let $a \in \{\text{approve, defer}\}$, with zero loss for approval when V_ε does not occur, loss $c_H > 0$ for approval if it does occur, and a constant review cost $c_R \in (0, c_H)$ for deferral. The conditional expected losses are $c_H \rho_\varepsilon$ and c_R . The Bayes action approves only if

$$\rho_\varepsilon \leq c_R / c_H. \quad (13)$$

The costs require institutional justification, and a posteriori decision rules cannot substitute for legal or substantive ethics. The proposed action is an *audit decision*, not an automatic license for deployment.

7 Positivity, identification, and a principled “unknown”

Transport is not identifiable merely because an analyst specifies a hierarchical model. Suppose a target cell contains group-positive individuals but no comparable source support. An extrapolated cell rate is then dependent on modeling assumptions or external information rather than identified by the sampled outcomes.

Proposition 3 (Sharp ignorance bounds without supported cells). *For $g \in \{A, B\}$, suppose the target positive-outcome composition q_{gc}^* is known and rates θ_{gc}^* are identified for cells $c \in O_g$, while for $c \in U_g = \mathcal{C} \setminus O_g$ nothing is known except $0 \leq \theta_{gc}^* \leq 1$. Define $L_g = \sum_{c \in O_g} q_{gc}^* \theta_{gc}^*$ and $m_g = \sum_{c \in U_g} q_{gc}^*$. Then the sharp identified interval is*

$$T_g^* \in [L_g, L_g + m_g], \quad \Delta^* \in [L_A - L_B - m_B, L_A + m_A - L_B]. \quad (14)$$

If both groups contain an unsupported target mass, no additional observations from *other* cells can, in general, reduce these worst-case widths. Hierarchical priors can yield a posterior within the interval, but do not make extrapolated information empirically identified. Missing-not-at-random outcomes lead to analogous, potentially more serious, nonidentification (Rubin, 1976). Appendix A supplies a proof and explains sharpness.

8 Empirical design and survey-weighted estimation

8.1 Data sources and fixed-classifier audit

The empirical application uses the American Community Survey (ACS) one-year Public Use Microdata Sample (PUMS), accessed through the `folkttables` Python package (Ding et al., 2021; Folkttables Developers, 2026; U.S. Census Bureau, 2026). Its `ACSEmployment` benchmark defines an observed employment-status outcome using the ACS variable `ESR` (positive class `ESR==1`) and includes `RAC1P` group coding. The primary audit population is Alabama in 2018. A sample-weighted logistic prediction model is fitted using Alabama 2017 and then frozen before auditing the 2018 release; hence, the audit does not reuse the model-fitting records. The contrast compares the White-alone and Black-alone categories recorded in the ACS. Other recorded race categories contribute to training and aggregate prediction summaries but not to this two-group contrast. The target is predictive performance for recorded employment status, not an effect of hiring decisions or a causal measure of discrimination.

8.2 Design-based anchor and uncertainty

Let w_i denote the released person weight of an audit record and Z_i the prediction made by the fixed classification rule. For any group with a positive weighted outcome-positive denominator, define

$$\hat{T}_{g,w} = \frac{\sum_{i \in s} w_i \mathbf{1}\{Y_i = 1, G_i = g\} Z_i}{\sum_{i \in s} w_i \mathbf{1}\{Y_i = 1, G_i = g\}}, \quad \hat{\Delta}_w = \hat{T}_{A,w} - \hat{T}_{B,w}. \quad (15)$$

A zero denominator signifies an undefined ratio and must be reported, rather than numerically imputed. The ACS release contains person weights `PWGTP` and 80 replicate person weights `PWGTP1, ..., PWGTP80` (U.S. Census Bureau, 2026). The *entire* nonlinear ratio is recomputed under every replicate weight. For the applicable ACS PUMS successive-difference-replication (SDR) convention, the design-variance estimator is

$$\widehat{\text{Var}}_{\text{SDR}}(\hat{\Delta}_w) = \frac{4}{80} \sum_{r=1}^{80} (\hat{\Delta}_{w,r} - \hat{\Delta}_w)^2. \quad (16)$$

These replicate weights quantify design variability; they are neither independent population draws nor samples from a Bayesian posterior. The SDR estimate conditions on the frozen classifier and does not include its training uncertainty.

8.3 Inferential comparators and validation criteria

The finite-population experiments compare a naive sample audit, a known-inclusion-probability design-weighted audit, independent-cell-weighted pseudo-Bayes, hierarchical-weighted pseudo-Bayes, and a scalar design-variance-aligned version of hierarchical disparity draws. We evaluate bias, root mean squared error, interval coverage, and false low-risk fairness declarations; structural support loss is studied through identified intervals. The ACS audit compares the unweighted gap, the weighted ratio and SDR interval, independent-cell and hierarchical pseudo-posterior summaries, and a variance-aligned draw distribution. Outcome-positive age–sex cell composition and cell-specific conditional rates are analyzed separately. Throughout, posterior event probabilities are interpreted conditionally on the working model, while repeated-design coverage and ACS SDR uncertainty are evaluated as different, noninterchangeable quantities. The concrete sample sizes, numerical outputs, and diagnostic plots are reported in Sections 9 and 10.

9 Finite-population simulations: estimation accuracy, coverage, and fairness certification

9.1 Simulation design and inferential comparators

The simulation examines the operational implications of the population-conditional equal-opportunity estimand in Eq. (1), the transport identity in Eq. (8), and the identification limits in Eq. (14). The target in each setting is a *fixed, finite-population* true-positive-rate (TPR) difference, $\Delta_N = T_{A,N} - T_{B,N}$, calculated before survey selection. For each of five mechanisms, we independently generate one population of $N = 30,000$ individuals and hold that population fixed over $R = 250$ repeated independent Poisson samples. The simulated group membership has $\Pr(G = A) = 0.62$, with group-specific categorical cell probabilities over six age–sex strata. Conditional on group and cell, the binary outcome and fixed binary prediction are generated from cell-specific Bernoulli distributions; among $Y = 0$, the false-positive probability is 0.12. Accordingly, the experiment evaluates uncertainty for an *already-fixed* classification rule, rather than incorporating uncertainty arising from fitting a classifier.

The five selection mechanisms are (i) near-uniform inclusion, $\pi_i = 0.12$; (ii) informative inclusion depending on (G_i, C_i, Y_i) ; (iii) strongly informative inclusion with more extreme subgroup contrasts; (iv) prediction-dependent inclusion additionally depending on Z_i ; and (v) deliberate loss of support, with $\pi_i = 0$ for $G_i = B, C_i = 6$. Apart from this structural zero, programmed propensities are constrained to $[0.015, 0.72]$. The finite-population estimands, expected sample sizes and population-positive mass in unsupported cells are recorded in Table 2. Since each mechanism is assigned its own generated finite population, small differences among the true gaps are not sampling effects of the *same* population.

We compare five procedures. *Naïve* estimates the unweighted TPR difference and uses the usual binomial normal-approximation interval; *design* uses the inverse-inclusion-probability Hájek TPR difference with a Poisson-design linearization interval based on the *known* simulated π_i ; *pseudo* uses independent beta pseudo-posteriors for weighted positive-class cell rates with Dirichlet pseudo-draws for q_{gc} ; *hierarchical* replaces the independent cell model by a Gaussian-prior hierarchical logistic specification, approximating its weighted pseudo-posterior by a Laplace expansion around the posterior mode; and *adjusted* recentres and rescales hierarchical disparity

Table 2: Simulation regimes and fixed-population targets. All scenarios have $N = 30,000$ and 250 independent Poisson-selection replications; the expected size is $\sum_i \pi_i$. The final column is the outcome-positive composition mass that is not observed with positive inclusion probability, not the proportion of all population records excluded.

Sampling mechanism	Expected n	True Δ_N	Unsupported positive mass
Near-uniform	3,600	0.0768	0.0000
Informative	4,738	0.0734	0.0000
Strongly informative	7,618	0.0764	0.0000
Prediction-dependent	6,548	0.0628	0.0000
Lack of support	4,708	0.0615	0.0914

draws to the weighted point estimate and *design* variance. Weights entering the working Bayesian cell likelihoods are normalized to have a sample mean of 1. The last procedure is an estimand-specific *variance alignment*, not a full joint design-based posterior, and the two unadjusted pseudo-posterior intervals are not guaranteed to be repeated-design calibrated (Savitsky and Toth, 2016; Williams and Savitsky, 2021).

Each Bayesian-type procedure uses 1,300 working posterior draws per replication, and each interval is an equal-tail 95% interval. For method m , the empirical bias, root mean squared error and interval coverage are computed as

$$\begin{aligned}
\widehat{\text{Bias}}_m &= R^{-1} \sum_{r=1}^R (\hat{\Delta}_{mr} - \Delta_N), \\
\widehat{\text{RMSE}}_m &= \left\{ R^{-1} \sum_{r=1}^R (\hat{\Delta}_{mr} - \Delta_N)^2 \right\}^{1/2}, \\
\widehat{\text{Cov}}_m &= R^{-1} \sum_{r=1}^R \mathbf{1}\{\Delta_N \in I_{mr}\}.
\end{aligned} \tag{17}$$

The empirical coverage standard error is $\{\widehat{\text{Cov}}_m(1 - \widehat{\text{Cov}}_m)/R\}^{1/2}$. Thus, a coverage estimate near 0.95 has a Monte Carlo standard error of about 0.014; small deviations must not be overinterpreted.

9.2 An exact population-composition reversal

We begin with a deterministic, two-cell calculation, displayed separately from the Monte Carlo simulations. Holding the group-by-cell conditional true-positive rates *fixed*, but replacing sample-conditioned positive-outcome cell shares by target-population shares, changes the group TPR difference from -0.21 to $+0.14$ (Figure 1). The difference of $+0.35$ is entirely the composition term of Eq. (8); the conditional-rate term is exactly zero. This is a constructive verification of possible *sign reversal*, not a frequency estimate from a stochastic experiment and not a claim about the ACS data.

9.3 Estimation accuracy and repeated-design uncertainty calibration

Figure 2 contrasts root mean squared errors and nominal 95% interval coverages; the exact values for all five settings and all five methods are transcribed in Table 3. Under prediction-dependent selection, the naive estimator exhibits bias -0.0401 and RMSE 0.0411, compared with design-weighted bias -0.0007 and RMSE 0.0234. Its binomial interval covers the finite-population truth in only 2.8% of replicates. In contrast, the design-based interval covers in 94.0% of replicates,

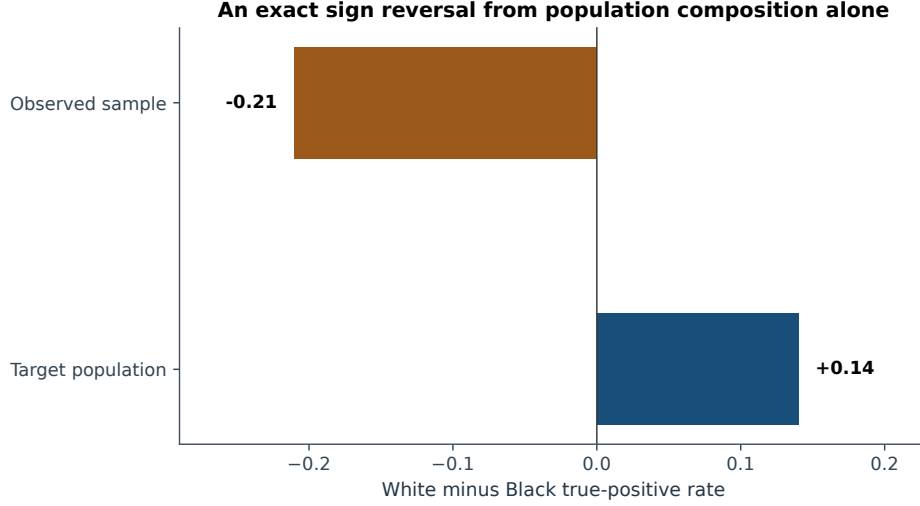


Figure 1: Deterministic two-cell verification of the transport identity. A shift in the composition of $Y = 1$ records reverses the White-minus-Black TPR difference, although within-cell TPRs are unchanged.

close to the nominal level within Monte Carlo variation. This is direct evidence, *under the stipulated known-propensity design*, that an apparently precise sample disparity can be badly misleading for the target population.

The gain does not apply uniformly to squared-error performance. Under the ordinary informative regime, the naive RMSE (0.0166) is lower than the weighted RMSE (0.0180), even though the weighted absolute bias (0.0019) is lower than the naive absolute bias (0.0044). In the strongly informative regime, the weighted estimate again improves bias (0.0013 versus 0.0105), but its RMSE (0.0260) exceeds that of the naive estimate (0.0215). These are finite-sample bias–variance trade-offs rather than evidence that unweighted inference is population-correct.

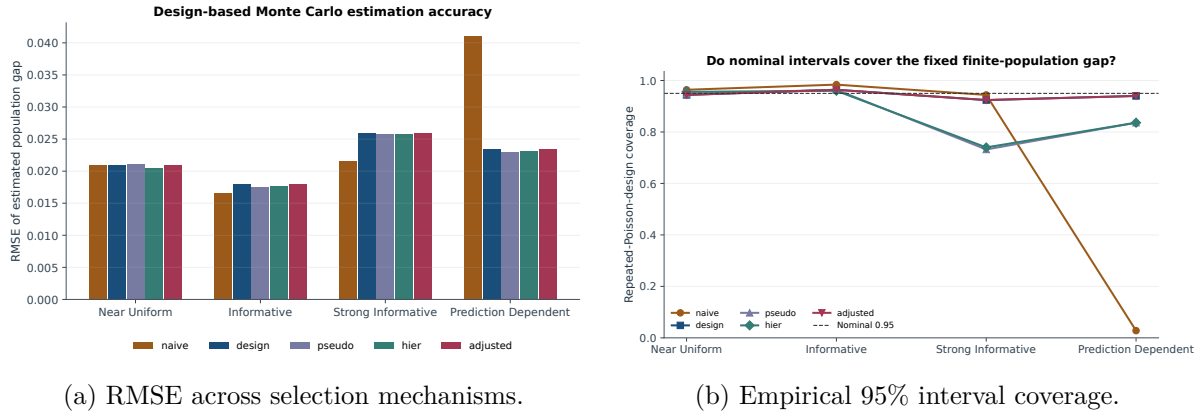


Figure 2: Monte Carlo performance for the four regimes with positive inclusion probabilities. Each plotted value is computed over 250 repeated samples from a fixed finite population. The loss-of-support regime, for which the full target is only partially identified, is examined separately in Figure 3a.

Notably, the two unadjusted working Bayesian procedures undercover with more severe selection: in the strongly informative regime, the independent-cell and hierarchical intervals attain only 0.732 and 0.740 coverage, respectively; under prediction-dependent selection, both attain 0.836. The scalar variance-alignment procedure improves these values to 0.924 and

Table 3: Estimation accuracy and interval calibration over 250 simulations per scenario. Each table panel contains numerical values transcribed from the full-run summary CSV; bold does not denote statistical significance. Method abbreviations: N, naive; D, design-weighted; P, independent-cell pseudo-Bayes; H, hierarchical pseudo-Bayes; A, scalar design-aligned hierarchical draws.

Panel A: root mean squared error					
Mechanism	N	D	P	H	A
Near-uniform	0.021	0.021	0.021	0.021	0.021
Informative	0.017	0.018	0.018	0.018	0.018
Strongly informative	0.022	0.026	0.026	0.026	0.026
Prediction-dependent	0.041	0.023	0.023	0.023	0.023
Lack of support	0.018	0.018	0.018	0.018	0.018
Panel B: nominal 95% interval coverage					
Mechanism	N	D	P	H	A
Near-uniform	0.964	0.944	0.948	0.956	0.944
Informative	0.984	0.964	0.964	0.960	0.964
Strongly informative	0.944	0.924	0.732	0.740	0.924
Prediction-dependent	0.028	0.940	0.836	0.836	0.940
Lack of support	0.976	0.952	0.956	0.960	0.952

0.940, respectively, although even its strong-informativeness result remains below 0.95 in this experiment. Because its point estimate is fixed to the design-weighted estimate, its RMSE matches the design comparator by construction. These results support the need for complex-survey calibration, but they do *not* establish a universal coverage theorem or validate the full joint posterior of (q, θ) .

9.4 Support failure, sharp identification and erroneous fairness clearance

The loss-of-support regime sets $\pi_i = 0$ for one of the six cells in group B . Consequently, the observed design supplies no information about that cell’s conditional TPR without additional restrictions. Using the exact positive-outcome composition from the simulated finite population and allowing its unsupported conditional rate to range over $[0, 1]$ produces the sharp interval $[0.0410, 0.1325]$, which contains the true disparity 0.0615 (Figure 3a). Its width, 0.0914, equals the unsupported positive-outcome mass. The observed coverage values for methods in this no-support regime are conditional numerical outcomes for the particular population and do not confer design validity for an unidentified full-population target. The singleton intervals in the other four regimes are *oracle identification intervals based on the known synthetic finite population*, not estimated 95% confidence intervals. They are used to verify the algebra of the bounds, not to imply that finite samples identify the exact disparity.

For the working posterior decision experiment, a fairness tolerance $\varepsilon = 0.05$ and illustrative acceptable violation probability $\alpha_d = 0.10$ are specified. A *false low-risk clearance* occurs if $\widehat{\Pr}(|\Delta| > \varepsilon \mid \mathcal{D}) \leq \alpha_d$ when the true finite-population gap exceeds ε . All five simulated finite populations happen to have $|\Delta_N| > 0.05$; therefore, these experiments evaluate *false clearance only*, not false high-risk alerts in truly low-disparity populations. Figure 3b and Table 4 report the observed frequencies.

In the prediction-dependent setting the hierarchical working pseudo-posterior falsely clears 7.6% of samples, compared with 4.0% after scalar variance alignment. Under strong informativeness, the corresponding frequencies are 4.4% and 1.6%. Lower false-clearance frequencies do not

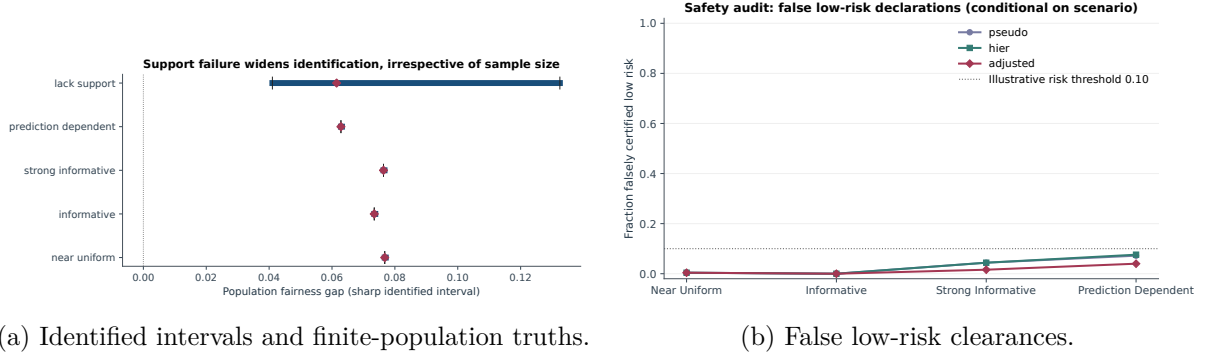


Figure 3: Two complementary limits to confident fairness assessment: structural support failure (left) and model-dependent posterior risk statements (right). The right panel plots four positive-propensity regimes; values for the unsupported regime remain in Table 4. These working risk rules differ from the simultaneous-envelope certificate of Eq. (12).

Table 4: False low-risk clearance frequencies (proportion of 250 replications), for $\varepsilon = 0.05$ and working risk cutoff 0.10. The naive approach has no associated posterior-risk declaration. In the non-support setting, these are model-dependent outcomes and *cannot* substitute for identification.

Mechanism	Design normal	Pseudo	Hierarchical	Aligned
Near-uniform	0.004	0.004	0.004	0.004
Informative	0.000	0.000	0.000	0.000
Strongly informative	0.016	0.044	0.044	0.016
Prediction-dependent	0.040	0.072	0.076	0.040
Lack of support	0.036	0.024	0.028	0.036

establish a uniformly superior policy: these settings contain no genuinely low-risk populations, and the frequency of unnecessary deferral cannot therefore be assessed. Moreover, the risk cutoff 0.10 is a *decision rule applied to pointwise posterior draws*; it does not replace the joint-event fairness certificate proposed in the theory section.

9.5 Key findings from the simulation study

The experiment yields four specific findings. *First*, the deterministic reversal shows that a sign change of a population TPR disparity can arise exclusively through shifts in outcome-positive composition, with no change to conditional rates. *Second*, informative and prediction-dependent selection can make naive sample fairness intervals severely misleading, while appropriately supplied inclusion weights correct the leading distortion in the tested settings. *Third*, unadjusted weighted pseudo-posterior credible intervals need not achieve nominal repeated-design coverage; scalar design-variance alignment improves coverage here but remains an approximation and does not necessarily improve point-estimation RMSE. *Fourth*, a structural zero in the selection mechanism imposes a nonvanishing identification interval that no increase in observed sample size can resolve without new support or additional assumptions. These conclusions are conditional on the synthetic mechanisms, fixed classification rules, known propensities and the chosen inferential approximations; they do not constitute a general performance guarantee for Bayesian population-aware AI.

10 American Community Survey application: population-aware employment prediction audit

10.1 Dataset, frozen prediction rule and target estimand

We evaluate the population-aware fairness workflow using the *American Community Survey* (ACS) one-year Public Use Microdata Sample (PUMS), as distributed by the U.S. Census Bureau and exposed through the `folktables` software (Ding et al., 2021; Folktables Developers, 2026; U.S. Census Bureau, 2026). The specific benchmark is `ACSEmployment`, with positive outcome $Y_i = \mathbf{1}\{\text{ESR}_i = 1\}$, and race coding `RAC1P`. The disparity compares people recorded as White alone (`RAC1P=1`) and Black alone (`RAC1P=2`); the remaining recorded race categories are retained in model training and overall predictive metrics but are not part of the two-group fairness contrast. Official data and documentation are available at <https://www.census.gov/programs-surveys/acs/microdata/documentation.html>; the benchmark implementation is at <https://github.com/socialfoundations/folktables>. Neither the outcome label nor the reported TPR gap identifies causal hiring discrimination; both concern recorded employment status and its predictions.

A sample-weighted logistic classifier was fitted once to the **2017 Alabama** ACS release ($n_{\text{train}} = 47,645$). Age was standardized, and categorical covariates were one-hot encoded. The predictors explicitly exclude race and sex while including age, education, marital and relationship status, disability-related and other recorded covariates; the exclusion of these two variables does not eliminate proxies or guarantee fairness. Its classification threshold is fixed at 0.50. The frozen classifier was then applied without re-fitting to the 2018 Alabama PUMS records; the different survey year avoids a within-release training-audit split, although uncertainty from model fitting is not propagated by the present audit.

Let $Z_i = \mathbf{1}\{\hat{p}_i \geq 0.50\}$. The relevant target is the Alabama 2018 person-population TPR difference $\Delta^{\text{AL},2018} = \Pr(Z = 1 \mid Y = 1, G = A) - \Pr(Z = 1 \mid Y = 1, G = B)$, rather than the raw contrast among released survey respondents. We compare the unweighted estimate with the

PUMS person-weighted ratio in Eq. (15). The ACS provides the main person weight $PWGTP$ and 80 corresponding person replicate weights $PWGTP1, \dots, PWGTP80$; each full TPR ratio is recomputed under every replicate, and the ACS successive-difference-replication (SDR) standard error is calculated by Eq. (16) (U.S. Census Bureau, 2026). The replicate weights are a survey-design variance tool, not posterior samples.

Table 5: Characteristics of the 2017 Alabama training sample and the 2018 Alabama audit sample, together with survey-weighted predictive performance. The weighted total is the sum of person weights, not an unweighted count. Predictive metrics use all released 2018 Alabama PUMS person records.

Characteristic	Value
Training release and sample size	Alabama 2017; 47,645
Audit release and sample size	Alabama 2018; 47,777
White-alone records / outcome positives	35,551 / 15,071
Black-alone records / outcome positives	10,115 / 3,714
Sum of released 2018 person weights	4,887,871
Survey-weighted outcome-positive prevalence	0.4206
Survey-weighted accuracy	0.7987
Survey-weighted Brier score	0.1404
Survey-weighted ROC AUC	0.8739

10.2 Survey-weighted fairness versus sample-based inference

Table 6 and Figure 4 summarize the numerical contrast. The unweighted sample TPR difference is 0.05364; the person-weighted difference is 0.02120, a decrease of 0.03244 (approximately 60.5% relative to the naive point estimate). Thus, the point estimate lies above the illustrative tolerance $\varepsilon = 0.05$ before weighting but below it after correction to the specified population target. This comparison is a direct empirical illustration of why the fairness *estimand* must be defined before judging a numerical disparity. It is not an effect of a fairness intervention, because the prediction rule is unchanged.

The ACS SDR standard error for the weighted gap is 0.00848, yielding the nominal 95% Wald interval [0.00458, 0.03783]. This interval excludes zero and lies below 0.05 under the frozen-model, replicate-design calculation. Such an interval is not the probability statement $\Pr(|\Delta| > 0.05 \mid \mathcal{D})$, however, and it does not account for uncertainty in classifier training. The hierarchical working pseudo-posterior gives mean 0.02143 and equal-tail 95% interval [0.00855, 0.03429]; the independent-cell pseudo-posterior gives 0.02169 and [0.00898, 0.03434]. The scalar SDR-aligned draw distribution is recentered at 0.02120 and yields [0.00447, 0.03791]. The narrower, unadjusted model-conditional credible intervals must not be presented as automatically design-valid credible intervals.

10.3 Outcome-conditioned composition and hierarchical cell rates

To connect the empirical results to the exact poststratification identity in Eq. (4), we partition the 2018 Alabama positive-outcome records by the six cross-classified age–sex cells (under 35, 35–54, and at least 55, crossed with the survey’s recorded binary sex coding). We distinguish the *composition* $q_{gc} = \Pr(C = c \mid Y = 1, G = g)$ from the *conditional prediction rate* $\theta_{gc} = \Pr(Z = 1 \mid Y = 1, G = g, C = c)$; substituting unconditional demographic shares for q_{gc} would generally produce the wrong TPR denominator. As implemented, independent weighted-Dirichlet pseudo-draws describe q_{gc} and a hierarchical weighted-logit Laplace approximation supplies

Table 6: Alabama 2018: target-population fairness gap and different forms of uncertainty. Posterior tail counts use 3,500 stored draws and the event $|\Delta| > 0.05$; a zero count represents the resolution of this finite Monte Carlo run, not a mathematical probability of zero. “SDR” denotes a design-based Wald confidence interval; the pseudo-posterior intervals condition on the stated working model.

Method	Gap mean/point	95% interval	Draws exceeding tolerance
Naive, unweighted	0.05364	—	—
Weighted ratio (SDR)	0.02120	[0.00458, 0.03783]	—
Independent-cell pseudo-Bayes	0.02169	[0.00898, 0.03434]	0/3500
Hierarchical pseudo-Bayes	0.02143	[0.00855, 0.03429]	0/3500
Scalar SDR-aligned draws	0.02120	[0.00447, 0.03791]	1/3500

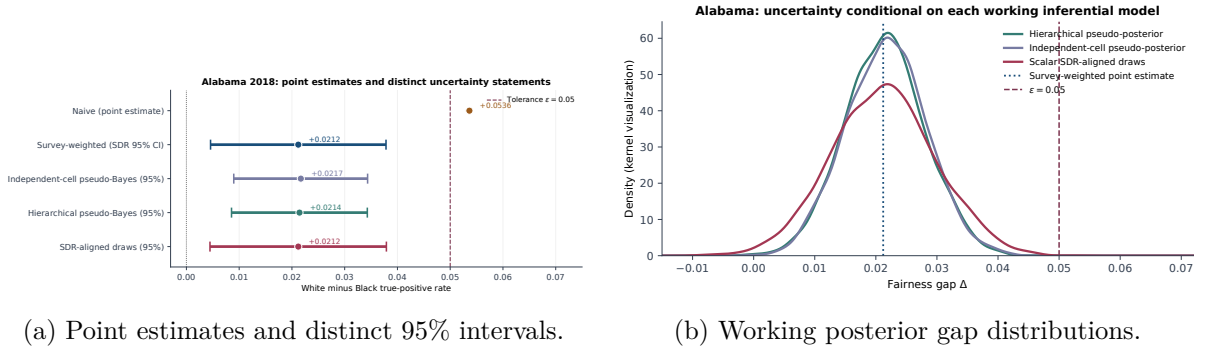


Figure 4: Alabama 2018 audit, calculated from design-weighted estimates and 3,500 stored pseudo-posterior draws. The naive sample estimate has no corresponding interval in the archived summary; therefore no naive interval is invented. The vertical threshold is $\varepsilon = 0.05$. The SDR-aligned density is a variance-adjusted draw distribution, *not* a full design-based posterior.

cell-rate draws. Their reported intervals condition on the working pseudo-model, rather than providing a fully design-adjusted analysis of every subgroup.

Table 7 manually records the posterior means from the 3,500 stored joint-composition and hierarchical-rate draws; Figure 5 shows the corresponding 95% equal-tail intervals. For example, the White-alone and Black-alone positive-outcome composition shares in the 35–54 male cell are 0.228 and 0.183, whereas the conditional rates in that cell are 0.887 and 0.776. In the under-35 female cell, the conditional rates are more similar (0.900 and 0.903), but the group cell shares differ (0.153 and 0.219). These descriptive contrasts illustrate why changes in both q and θ can affect an aggregated fairness statistic. They neither identify causal explanations for group differences nor provide a transported out-of-state decomposition.

Table 7: Alabama 2018: posterior-mean shares q_{gc} among outcome-positive records and hierarchical conditional rates θ_{gc} , computed directly from saved draws. Columns refer to recorded White-alone (W) and Black-alone (B) race groups; each group’s six q values sum to approximately one (rounding aside). Uncertainty intervals are displayed in Figure 5.

Age–sex cell	q_{Wc}	θ_{Wc}	q_{Bc}	θ_{Bc}
Under 35, male	0.176	0.901	0.188	0.875
Under 35, female	0.153	0.900	0.219	0.903
35–54, male	0.228	0.887	0.183	0.776
35–54, female	0.192	0.911	0.233	0.846
55+, male	0.140	0.582	0.083	0.537
55+, female	0.111	0.559	0.094	0.524

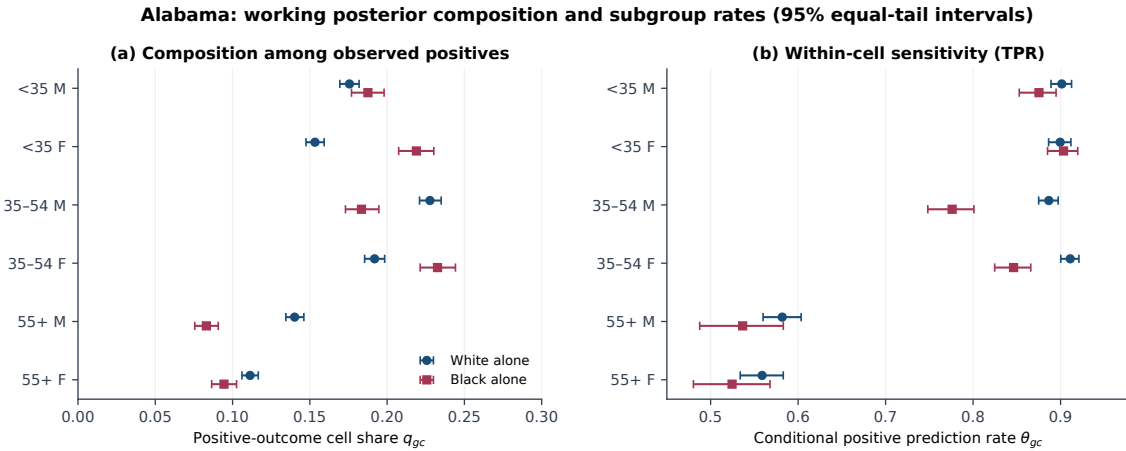


Figure 5: Alabama 2018, six-cell decomposition of the positive-outcome TPR: (a) group-conditional outcome-positive composition q_{gc} and (b) within-cell positive prediction rates θ_{gc} . Symbols show means and horizontal bars show 95% equal-tail intervals obtained from the working pseudo-posterior draws. The binary sex categories follow the historical PUMS variable, not an exhaustive taxonomy of gender identity.

Marginalizing over the stored draws yields working posterior means of 0.8167 for the White-alone TPR and 0.7952 for the Black-alone TPR, with marginal 95% intervals of [0.8103, 0.8229] and [0.7840, 0.8062], respectively. Their dependent difference is summarized in Table 6; its interval cannot generally be recovered by subtracting the endpoints of the two marginal TPR intervals. All 12 outcome-positive subgroup cells meet the software’s minimal observed-positive threshold (the smallest cell contains 409 sample positives), so the six-cell audit does not flag an *empirical* empty-cell problem in Alabama. This fact does not prove transport positivity for other places or time periods.

10.4 Design calibration, fairness certificates and inferential scope

The variance ratio between the hierarchical working pseudo-posterior and the ACS SDR estimate is 0.5926: its disparity draws have approximately 59.3% of the SDR variance. This concrete discrepancy motivates closer examination of design calibration, although a single survey-realization variance comparison is not, by itself, evidence of repeated-design undercoverage in the ACS. In the stored working draws, none of the 3,500 hierarchical or independent-cell values exceeds $|\Delta| = 0.05$, and one SDR-aligned draw does. These Monte Carlo counts are highly conditional on the modeling choices and do not provide proof of population fairness.

We also assess the *simultaneous-envelope certificate* from Eq. (12), which should not be confused with the pointwise tail-probability decision. The recorded bound is $B = 0.15583$ around a reference gap 0.02143, with empirical joint-event mass approximately 0.9497. Because $|0.02143| + 0.15583 > 0.05$ and $|0.02143| - 0.15583 < 0.05$, the predefined certificate is *inconclusive*; it establishes neither low disparity nor a definitive threshold violation at the stipulated joint credible level. The conservatism of the worst-cell simultaneous construction is especially visible compared with the narrow marginal draw distributions. Although the workflow’s illustrative action flag reads “approve for further review,” that flag is based on the *pointwise* working risk rule and is not a formal joint fairness certificate.

The empirical analysis is an audit of a fixed prediction rule for the specified Alabama 2018 target population. The empirical gap and its design-based uncertainty are not causal estimates, and the comparison of pseudo-posterior and SDR variances cannot alone establish repeated-design coverage. Conclusions about other populations or prediction systems require the corresponding target definition, sampling information, and empirical assessment.

10.5 Key findings from the ACS application

The Alabama ACS analysis provides three substantive observations. *First*, an unchanged frozen employment-status classifier yields materially different unweighted and target-survey-weighted fairness gaps (0.05364 versus 0.02120), illustrating the importance of population definition and survey correction. *Second*, working hierarchical and independent-cell pseudo-posterior estimates closely agree with the weighted point estimate, but their uncertainty spread is smaller than that suggested by the released replicate-weight variance; consequently, posterior risk statements require explicit calibration qualifications. *Third*, the six-cell composition/rate representation offers an interpretable account of how subgroup composition and conditional TPRs enter the population gap, while the rigorous simultaneous-envelope audit appropriately returns *inconclusive* despite low pointwise posterior tail counts. These results pertain to the specified Alabama 2018 population; none of the observed disparities identifies causal discrimination or establishes the behavior of an AI-based employment decision system.

11 Discussion

11.1 From sample disparity to population disparity

The mathematical framework and the two empirical investigations point to the same distinction: fairness is a property of a specified target-population distribution, whereas an unweighted audit describes the distribution induced by the observed sample. The exact decomposition in Proposition 1 clarifies why a marginal true-positive-rate gap may change even when the classification rule and conditional cell rates do not. In the deterministic construction, compositional change alone reverses the sign of the disparity (Figure 1). In the Alabama application, retaining the same fixed classifier while accounting for ACS person weights changes the estimated White-minus-Black

TPR gap from 0.05364 to 0.02120 (Table 6). The 0.03244 difference is substantial relative to the illustrative tolerance of 0.05, but it does not demonstrate that weighting changes the fairness of the classifier: it changes the population-relevant quantity being estimated.

11.2 Accuracy, calibration and inferential strength

The simulations distinguish correcting a sampling distortion from increasing estimator precision. In the prediction-dependent inclusion regime, weighting substantially reduces bias and restores interval coverage toward the nominal level. Under other informative sampling regimes, smaller bias need not imply smaller mean squared error. More importantly, the unadjusted weighted pseudo-posterior intervals can have inadequate repeated-design coverage. Hierarchical pseudo-Bayes attains coverage 0.740 under strongly informative sampling and 0.836 under prediction-dependent sampling, whereas the scalar design-aligned procedure attains 0.924 and 0.940, respectively (Table 3). The alignment is intentionally limited: it adjusts dispersion for a single estimated contrast without correcting the entire joint posterior or guaranteeing calibration under other designs.

The Alabama audit reinforces that qualification rather than resolving it. The hierarchical pseudo-posterior point estimate is close to the design-weighted ratio, but its estimated disparity variance is only 0.5926 times the ACS SDR variance. This is a discrepancy between two uncertainty estimates for a single released survey sample, not direct evidence of repeated-design undercoverage in Alabama. The joint fairness-certification envelope remains inconclusive even though the stored pointwise pseudo-posterior draws seldom exceed the selected disparity threshold. Such a contrast is informative: a small estimated posterior tail probability and a conservative simultaneous guarantee answer different questions. Reporting both avoids equating a convenient working model with definitive certification.

11.3 Identification, practical implications and limitations

Proposition 3 identifies a more fundamental boundary than posterior imprecision. When an outcome-positive target stratum lacks sampling support, its unobserved conditional prediction rate can vary over the entire unit interval unless substantive restrictions are introduced. The simulation’s sharp identification interval therefore persists even if the observed cells are measured with negligible error (Figure 3a). Hierarchical pooling can provide conditional extrapolations, but prior-driven extrapolation should not be described as empirical identification.

The scope of the numerical results must be interpreted carefully. The simulation fixes finite populations and classifiers, assumes known selection probabilities, and studies a limited set of sampling designs; the ACS study uses one geographic target and two recorded race categories, while treating the fitted classifier as fixed. Released survey weights do not supply a fully identified nonresponse or selective-label mechanism, and neither the SDR interval nor the working posterior incorporates training-stage uncertainty. Employment-status prediction is not the same as a real hiring or resource-allocation decision, and predictive disparities cannot establish causal discrimination. The chosen tolerance and error costs are policy parameters, not statistical constants. Lastly, subgroup coding, small-domain reporting, nonresponse, missing-not-at-random outcomes, and disclosure protection require ethical and substantive scrutiny alongside mathematical diagnostics (Mitchell et al., 2021; Rubin, 1976).

12 Future research

Several methodological extensions follow from the empirical limitations rather than from the assumption that the present approach has solved fairness certification.

Joint design-calibrated Bayesian inference. A priority is to investigate transformations of the full joint pseudo-posterior for outcome-conditioned cell shares and conditional TPRs, with uncertainty informed by replicate weights, within-design dependence, and nonlinear ratio functionals. Repeated-design coverage should be evaluated over additional informative-selection mechanisms, sparse positive-outcome cells, and a wider range of disparity magnitudes. In particular, designs with truly small population gaps are necessary to measure false alarms and inappropriate deferrals, which the present high-disparity simulation regimes cannot quantify.

Geographic and temporal transport. The composition–rate identity suggests evaluating population targets across different states and survey years, while keeping source and destination populations explicit. These extensions require compatible variables, defensible support assumptions, and uncertainty propagation for both outcome-positive composition and within-cell rates. Where support fails, partial-identification bounds should be reported alongside any hierarchical extrapolations rather than replaced by them.

More realistic selection and model uncertainty. The fixed-classifier assumption can be relaxed by propagating training-stage variability, model selection, and calibration uncertainty through the entire fairness audit. Additional work should address nonignorable survey nonresponse, selectively observed labels, changes in the meaning of the outcome, and fairness across intersectional or multiple groups. Sensitivity analyses are particularly important where survey weights cannot identify the selection mechanism.

Decision design and simultaneous certification. Finally, less conservative but valid simultaneous credible regions should be studied under complex sampling, along with partial-identification-aware decisions and abstention rules. Evaluation should separately quantify false fairness clearances, false alarms, review burden, and robustness to prior distributions and cost choices. The appropriate output may often be a bounded or explicitly inconclusive conclusion, rather than a single fairness score.

13 Conclusion

This study develops and evaluates a population-aware approach to Bayesian fairness assessment that integrates outcome-conditioned estimands, complex survey inference, hierarchical uncertainty modeling, and decisions about fairness claims. The exact transport decomposition explains how population composition can alter marginal disparities; the posterior-certification result specifies conditions for defensible probabilistic statements; and the partial-identification bounds expose what unsupported target strata prevent us from learning. Across five finite-population sampling regimes, design weighting corrects serious distortions arising from prediction-dependent selection, but unadjusted weighted pseudo-posteriors do not uniformly achieve nominal coverage. In the 2018 Alabama ACS audit, the unweighted TPR gap of 0.05364 decreases to 0.02120 after population weighting. The close agreement between pseudo-posterior point estimates and the weighted ratio does not eliminate their uncertainty-calibration limitations, and the simultaneous certificate remains inconclusive. Thus, the principal contribution is not a claim that a predictive

system has become fair; it is a reproducible inferential framework for making population-specific fairness statements more transparent, qualified, and empirically testable.

Declarations

Funding: To be completed by the authors before submission.

Competing interests: To be completed by the authors before submission.

Ethics and consent: This study uses publicly released ACS PUMS microdata for secondary statistical analysis; there was no direct recruitment, participant contact, or intervention. Any applicable institution-specific review and publication requirements should be confirmed by the authors.

Data availability: ACS PUMS data and documentation are available from the [U.S. Census Bureau](#); the benchmark implementation is provided by the [Folktables repository](#). Reproducibility code for the simulations and ACS audit and aggregate numerical outputs were prepared for this study.

Code Availability The Python implementations used for the simulation study and empirical analysis are publicly available under the MIT License at https://github.com/debashisdotchatterjee/BPAI_viewpoint1.

Author Contributions

Debashis Chatterjee: Conceptualization, methodology, formal analysis, software, validation, investigation, data analysis, visualization, writing—original draft, and writing — review and editing requirements.

A Proofs and additional mathematical details

A.1 Proof of Proposition 1

Proof. Insert and subtract $\sum_c q_{gc}^Q \theta_{gc}^P$:

$$\begin{aligned} T_g^Q - T_g^P &= \sum_c q_{gc}^Q \theta_{gc}^Q - \sum_c q_{gc}^P \theta_{gc}^P \\ &= \sum_c (q_{gc}^Q - q_{gc}^P) \theta_{gc}^P + \sum_c q_{gc}^Q (\theta_{gc}^Q - \theta_{gc}^P). \end{aligned}$$

For probability vectors q, q' and any $0 \leq f_c \leq 1$, $|\sum_c (q_c - q'_c) f_c| \leq \text{TV}(q, q')$; this follows by splitting $q - q'$ into positive and negative parts, which have equal total mass $\text{TV}(q, q')$. Apply this to the first component and the triangle inequality to the second. The difference of group gaps obeys $|\Delta^Q - \Delta^P| \leq |T_A^Q - T_A^P| + |T_B^Q - T_B^P|$, proving Eq. (9). $\square$

A.2 Proof of Theorem 2

Proof. On E , for each g ,

$$\begin{aligned} \left| \sum_c q_{gc}^* \theta_{gc}^* - \sum_c \bar{q}_{gc} \bar{\theta}_{gc} \right| &\leq \left| \sum_c (q_{gc}^* - \bar{q}_{gc}) \theta_{gc}^* \right| + \sum_c \bar{q}_{gc} |\theta_{gc}^* - \bar{\theta}_{gc}| \\ &\leq \text{TV}(q_g^*, \bar{q}_g) + \sum_c \bar{q}_{gc} e_{gc} \\ &\leq \eta_g + \sum_c \bar{q}_{gc} e_{gc}. \end{aligned}$$

Subtract the two groups and use the triangle inequality to obtain $|\Delta^* - \bar{\Delta}| \leq B$ on E . Since $\Pi(E \mid \mathcal{D}, \mathcal{S}, \mathcal{P}^*) \geq 1 - \alpha$, the posterior probability of that inequality is at least $1 - \alpha$. If $|\bar{\Delta}| + B \leq \varepsilon$, then on E , $|\Delta^*| \leq \varepsilon$, whence $V_\varepsilon \subseteq E^c$, giving $\rho_\varepsilon \leq \alpha$. If $|\bar{\Delta}| - B > \varepsilon$, then on E , $|\Delta^*| > \varepsilon$, whence $E \subseteq V_\varepsilon$, giving $\rho_\varepsilon \geq 1 - \alpha$. $\square$

Remark 1. *A simultaneous credible event is essential. Independent marginal $1 - \alpha$ statements for every rate do not, in general, imply that their intersection has probability $1 - \alpha$. One may construct a joint credible region from posterior draws, or use a conservatively allocated union bound; either approach must be empirically calibrated when survey pseudo-posteriors are involved.*

A.3 Proof of Proposition 3

Proof. For each group g , split $T_g^* = \sum_{c \in O_g} q_{gc}^* \theta_{gc}^* + \sum_{c \in U_g} q_{gc}^* \theta_{gc}^* = L_g + R_g$. Since each unobserved rate belongs to $[0, 1]$, $0 \leq R_g \leq m_g$. Both endpoints are attained by setting all unsupported θ_{gc}^* equal to 0 or 1, respectively. Every intermediate value can be attained by setting all unsupported rates equal to the same $t \in [0, 1]$ (when $m_g > 0$), so the interval is sharp. The feasible range of $T_A^* - T_B^*$ follows by choosing independently the endpoints of the two group intervals, giving $[L_A - (L_B + m_B), (L_A + m_A) - L_B]$, also sharp under the stated absence of cross-cell constraints. $\square$

A.4 Why total-population poststratification weights can fail for TPR

By Bayes' rule, assuming $\Pr(Y = 1 \mid G = g) > 0$,

$$\Pr(C = c \mid Y = 1, G = g) = \frac{\Pr(Y = 1 \mid C = c, G = g) \Pr(C = c \mid G = g)}{\sum_d \Pr(Y = 1 \mid C = d, G = g) \Pr(C = d \mid G = g)}.$$

Thus, population shares among all members of a group generally differ from shares among true positives, even if demographic margins are perfectly known. Replacing q_{gc}^* by r_{gc}^* in Eq. (4) estimates a different functional. This does not require a probabilistic theorem beyond Bayes' rule, but it is a potentially important source of empirical target mismatch.

A.5 Bayes decision threshold derivation

From the proposed loss specification, the posterior expected loss of *approve* is $c_H \Pi(V_\varepsilon \mid \mathcal{D}, \mathcal{S}, \mathcal{P}^*)$ and that of *defer* is c_R . Comparing the two quantities gives Eq. (13). The result does not state that the posterior event probability is well calibrated; it only identifies the Bayes action conditional on a given uncertainty distribution and loss function.

References

- Ainhize Barrainkua, Paula Gordaliza, Jose A. Lozano, and Novi Quadrianto. Uncertainty in fairness assessment: Maintaining stable conclusions despite fluctuations, 2023. URL <https://doi.org/10.48550/arXiv.2302.01079>. Preprint related to the 2024 AISTATS publication.
- Ainhize Barrainkua, Paula Gordaliza, Jose A. Lozano, and Novi Quadrianto. Uncertainty matters: Stable conclusions under unstable assessment of fairness results. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1198–1206, 2024. URL <https://proceedings.mlr.press/v238/barrainkua24a.html>. Official PMLR proceedings record; no proceedings DOI independently confirmed.
- Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning, 2021. URL <https://doi.org/10.48550/arXiv.2108.04884>. NeurIPS 2021; DOI points to the corresponding arXiv preprint.
- Folktables Developers. Folktables: A benchmarking toolkit for u.s. census data. Software repository, 2026. URL <https://github.com/socialfoundations/folktables>. Accessed 11 October 2026; software repository, no DOI.
- Andrew Gelman. Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2):153–164, 2007. DOI: [10.1214/088342306000000691](https://doi.org/10.1214/088342306000000691). URL <https://doi.org/10.1214/088342306000000691>.
- Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016. URL <https://doi.org/10.48550/arXiv.1610.02413>. NeurIPS 2016; DOI links to the arXiv version.
- Jeffrey R. Lax and Justin H. Phillips. How should we estimate public opinion in the states? *American Journal of Political Science*, 53(1):107–121, 2009. DOI: [10.1111/j.1540-5907.2008.00360.x](https://doi.org/10.1111/j.1540-5907.2008.00360.x). URL <https://doi.org/10.1111/j.1540-5907.2008.00360.x>.
- Jihu Lee, Kunwoong Kim, Sehyun Park, Insung Kong, Dongyoon Yang, and Yongdai Kim. A fair bayesian inference through matched gibbs posterior. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/61c2975281d60d3b1ce4cefc157d99df-Abstract-Conference.html. Official ICLR conference record; no DOI independently confirmed.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021. DOI: [10.1145/3457607](https://doi.org/10.1145/3457607). URL <https://doi.org/10.1145/3457607>.
- Shira Mitchell, Eric Potash, Solon Barocas, Alexander D’Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021. DOI: [10.1146/annurev-statistics-042720-125902](https://doi.org/10.1146/annurev-statistics-042720-125902). URL <https://doi.org/10.1146/annurev-statistics-042720-125902>.
- YongKyung Oh, Henry W. Zheng, Jeffrey Feng, and Alex A. T. Bui. Survey-aware machine learning: A guideline for valid population health inference based on scoping review. In *Proceedings of the 7th Conference on Health, Inference, and Learning*, volume 333 of *Proceedings of Machine Learning Research*, pages 198–220, 2026. DOI: [10.48550/arXiv.2605.08963](https://doi.org/10.48550/arXiv.2605.08963). URL <https://proceedings.mlr.press/v333/oh26a.html>. DOI identifies the corresponding arXiv preprint, not the PMLR proceedings.

- Francesca Panero, Ernst C. Wit, and Marco Scutari. The role of uncertainty in assessing the fairness of machine learning models, 2026. URL <https://doi.org/10.48550/arXiv.2609.07959>. arXiv preprint, September 2026.
- David K. Park, Andrew Gelman, and Joseph Bafumi. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12(4):375–385, 2004. DOI: [10.1093/pan/12.4.375](https://doi.org/10.1093/pan/12.4.375). URL <https://doi.org/10.1093/pan/12.4.375>.
- Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys*, 55(3):1–44, 2022. DOI: [10.1145/3494672](https://doi.org/10.1145/3494672). URL <https://doi.org/10.1145/3494672>.
- Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. DOI: [10.1093/biomet/63.3.581](https://doi.org/10.1093/biomet/63.3.581). URL <https://doi.org/10.1093/biomet/63.3.581>.
- Terrance D. Savitsky and Daniell Toth. Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10(1):1677–1708, 2016. DOI: [10.1214/16-EJS1153](https://doi.org/10.1214/16-EJS1153). URL <https://doi.org/10.1214/16-EJS1153>.
- Yajuan Si, Natesh S. Pillai, and Andrew Gelman. Bayesian nonparametric weighted sampling inference. *Bayesian Analysis*, 10(3):605–625, 2015. DOI: [10.1214/14-BA924](https://doi.org/10.1214/14-BA924). URL <https://doi.org/10.1214/14-BA924>.
- U.S. Census Bureau. American community survey public use microdata sample (pums): Documentation and replicate-weight guidance. Official online documentation, 2026. URL <https://www.census.gov/programs-surveys/acs/microdata/documentation.html>. Accessed 11 October 2026; online documentation, no DOI.
- Matthew R. Williams and Terrance D. Savitsky. Uncertainty estimation for pseudo-bayesian inference under complex sampling. *International Statistical Review*, 89(1):72–107, 2021. DOI: [10.1111/insr.12376](https://doi.org/10.1111/insr.12376). URL <https://doi.org/10.1111/insr.12376>.