# ShiftBench: A Benchmark for Per-Cohort Certify-or-Abstain Decisions on Positive Predictive Value Under Covariate Shift

**Ananya Salian** [1]

## Abstract

Many machine learning benchmarks emphasize aggregate accuracy on a single dataset, while formal statistical confidence in each cohort's precision under deployment shift is rarely the primary unit of reporting. Standard evaluation tools typically assume independently and identically distributed sampling, an assumption that has been observed to fail in a range of real deployment settings. As such, we present **ShiftBench**, a benchmarking protocol that, given a classifier, a calibration set, and a target distribution, returns a per-cohort certify-or-abstain decision on precision with tunable controls over the family-wise false-certification rate.

Across the 38 real datasets and 9 candidate methods we evaluated, ShiftBench observed no false certifications in 19,600 boundary-null trials. We prove that the weighted-conformal lower-bound formula, when used directly as a one-sided hypothesis test on PPV, permits closed-form limits strictly greater than our given $\tau$ under the null. In our benchmark, this manifests as certification-under-null rates of 65 to 81% against $\alpha = 5\%$. As part of a methodological audit, we provide a ranking of methods by reliability and statistical power and further identify alternative bounds that appear to retain validity under the protocol.

## 1. Introduction

Machine learning models are increasingly deployed across populations that differ from the calibration sets they were tuned on, yet practitioners often need per-cohort statistical confidence in metrics like PPV (positive predictive value, precision), rather than simply aggregating accuracy. Existing approaches, including conformal prediction, importance weighting, and distributionally robust optimization (DRO), each address parts of this problem, but there is currently no unified protocol that returns formal certify-or-abstain decisions on PPV under covariate shift with multiple-testing control. To address this gap, we present **ShiftBench**, a benchmarking protocol that takes a trained classifier, a calibration set, and a target distribution and returns a per-cohort certify-or-abstain decision on PPV with control over the family-wise error rate (FWER) across resulting (cohort, $\tau$) hypotheses. Across 38 real datasets and 9 candidate methods we evaluated, ShiftBench observes no false certifications across 19,600 boundary-null trials, and quantifies which methods certify, when, and why on this benchmark.

This deployment shift problem arises across a wide range of settings. Fairness audits evaluate model behavior across demographic subgroups, where per-group calibration and risk control are first-class concerns [1, 2, 3, 4, 5]. Drug-discovery pipelines screen candidate molecules across novel chemical scaffolds, where the deployment distribution is by construction, disjoint from training. Furthermore, large-scale benchmark suites have brought distribution shift forward as a first-class evaluation concern, with empirical demonstrations of shift across geographies, time periods, chemical scaffolds, and demographic subgroups in WILDS and DomainBed, across reproduced test sets in image classification, and fine-grained shift taxonomies [6, 7, 8, 9, 10]. In each setting, the practitioner often cares less about the model's average performance on the original test set than about its performance in each cohort they intend to deploy on. For example, a model reporting 90% PPV on average may achieve 95% on one cohort and 40% on another, with only a per-cohort breakdown being able to indicate whether this model is statistically reliable.

Two challenges make this hard. First, the calibration data rarely matches the deployment distribution, a problem that motivates much of current covariate-shift literature [11, 12, 13, 14, 15]. Second, even when data is well-matched, statistical confidence in per-cohort estimates requires careful accounting. As the number of cohorts increases, the probability that at least one cohort is incorrectly certified by chance increases alongside this, a familiar instance of the multiplicity problem, hence a Holm or Bonferroni-style correction is often needed in such cases

[1]Independent. Correspondence to: Ananya Salian <contact@ananyasalian.com>.

*Accepted to the 1st Workshop on Combining Theory and Benchmarks*, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

[16, 17, 18, 19]. Standard machine learning evaluation tools typically assume independently and identically distributed (IID) samples and have been documented to underreport variance and to omit formal multiple testing correction in benchmark practices, specifically in NLP power-analysis audits and benchmark-variance studies, as seen in broader critiques of benchmark selection [20, 21, 22, 23].

The community has developed three broad responses to these challenges.

**Conformal prediction** provides distribution-free, finite-sample coverage guarantees for prediction sets, and has become a strong uncertainty quantification tool due to its model agnostic nature and sole assumption of exchangeability [25, 26, 27]. The classical form assumes calibration and target data are drawn IID from the same distribution, an assumption that frequently fails under distribution shifts. Weighted and split variants of this restore marginal coverage on the target when a known importance ratio is available [13, 24]. However, two properties of this guarantee leave it misaligned for certification. First, this guarantee concerns prediction-set coverage rather than a one-sided test on a population metric. Second, the coverage it controls is marginal, holding the average over the entire target population, rather than on a per-cohort basis. As we show in (Theorem 1), using its weighted-conformal lower bound directly as a per-cohort PPV test does not inherit the coverage guarantee, and is anti-conservative under the null, meaning that this tool is mismatched for certification.

**Importance weighting** adapts to shift by reweighting calibration samples by the density ratio between target and source. This technique underlies most shift-aware estimators. When the density ratio is smooth and accurately estimated, density-ratio methods such as KLIEP, uLSIF, KMM, and BBSE recover accurate target-domain estimates [11, 12, 28, 29, 30]. Its central weakness, however, is variance. As the distributions diverge, weights grow heavy-tailed and the effective sample size falls sharply, at which point, the concentration inequalities used to form confidence bounds, become either overconservative or, in some cases, systematically misleading [31, 32, 33].

**Distributionally robust optimization** (DRO) intervenes on the model during training rather than at evaluation, instead optimizing the model against the worst-case or worst-group risk within an uncertainty set at training time. Group DRO has been reported to improve worst-group accuracy on subpopulation shift benchmarks [34]. It directly addresses the weakest cohort, therefore requiring no density-ratio estimation, simultaneously avoiding the weight-instability failure mode described above. However, this yields a more robust trained model, rather than a post-hoc, per-cohort certify-or-abstain decision with finite-sample error control, and in practice, can be conservative, requiring group membership to be known in advance.

These separate responses have begun to merge in shift-aware extensions, most notably with weighted conformal prediction (WCP), which adapts conformal quantile estimates using importance weights, alongside adjacent work on distribution-free risk control and selective inference [13, 35].

Across our benchmark we report three findings, in order of practical impact:

**Stability diagnostics** were the dominant factor in whether the protocol could issue valid certifications under covariate shift in our experiments. The choice of importance-weight estimator was determined to be second-order.

**Statistical bound family** was the strongest determinant of how often the protocol could certify. The two families we compared were binomial-exact bounds (a count of correctly classified calibration samples, with inverted binomial distribution to obtain this bound) and concentration bounds (bounds how far the empirical mean can deviate from the population mean using sample variance). In this paper, we observed that concentration bounds remain valid but tend to become conservative at moderate effective sample size, whereas binomial-exact bounds tend to remain powerful in this regime.

**Density-ratio estimator choice** (uLSIF, KLIEP, BBSE, rULSIF) changed the result by less than a percentage point across our datasets within a given bound family.

**Concrete deployment example**

Consider a health system deploying a triage model across eight hospitals, each requiring $\mathrm{PPV} \geq 0.8$ on its predicted-positive cohort across five thresholds, for a total of 40 hypotheses controlled jointly by Holm's procedure at $\alpha = 0.05$. The per-hospital importance weights satisfy the PSIS tail-index $\hat{k}$ and ESS (effective sample size) weight-reliability gates defined in Section 3, as target cases lie within the support of the calibration set.

At hospitals with $n_{\text{eff}} \approx 40$, this protocol certifies only the strongest cohorts, via the binomial bound, abstaining on the remainder (Table 2), rather than issue an over-confident certification that a $n_{\text{eff}}$-blind procedure may permit. At hospitals with $n_{\text{eff}} \geq 200$, the two bound families agree, and the concentration bound is preferred when the target-domain importance-weighted PPV is the metric of interest. Subgroups with $n_{\text{eff}} < 30$ are returned as NO-GUARANTEE before any false certification can be issued, prompting upstream sample collection rather than a misleading guarantee. WCP, despite its formal coverage guarantee, is not used here as a PPV test, for the reasons discussed in Section 5.

The benchmark and per-experiment breakdowns are released as a resource (see Appendix).

## 2. Related Work

### 2.1. Distribution shift, conformal prediction, and weight diagnostics

Distribution shift refers to the mismatch between the joint distributions of inputs and labels at training versus deployment time. This takes several forms in current literature: covariate shift (input distribution changes while conditional label distribution does not), label shift (marginal label distribution changes), and concept shift (conditional itself changes). Empirical benchmarks have curated datasets and algorithm comparisons across geographies, time periods, chemical scaffolds, and demographic subgroups, with WILDS and DomainBed as widely cited references [6, 7]. These benchmarks largely report aggregate or worst-cohort accuracy, rather than ranking methods by per-cohort statistical robustness.

Conformal prediction provides distribution-free coverage guarantees for prediction sets. The classical version assumes exchangeability between calibration and target data, and weighted variants reweight the calibration distribution by a known importance ratio to recover this coverage on the target distribution [13, 24, 36, 37].

Related extensions adapt conformal scores to selective inference and false-discovery control under shift [38, 39]. However, the question of whether the WCP lower-bound formula retains validity when used directly as a per-cohort PPV hypothesis test under covariate shift does not appear to have been discussed in current literature. A practical concern with importance weighting is heavy-tailed or near-degenerate weights.

Kish's effective sample size and Pareto-smoothed importance sampling diagnostics have appeared in the survey-sampling and importance-sampling literature, though their use as explicit gating conditions for refusing a per-cohort statistical guarantee remains uncommon in shift-aware machine learning [31, 32].

### 2.2. Density-ratio estimation

Importance-weighted shift correction relies on the importance ratio $w(x) = p_{\text{target}}(x)/p_{\text{cal}}(x)$ and density-ratio estimation aims to estimate $w(x)$ directly without estimating the two densities separately. Prevalent approaches differ in what they match: moment-matching (KMM), KL-divergence minimization (KLIEP), least-squares fitting (uLSIF & the relative variant rULSIF), and label-shift estimation via confusion matrices (BBSE), where the latter applies in cases where only the label marginal $p(y)$ shifts while the conditional $p(x \mid y)$ remains unchanged [11, 12, 28, 29, 30]. Group DRO [34] sits alongside this body of work and addresses worst-cohort performance during training rather than estimating $w(x)$ directly.

### 2.3. Multiple testing and certification benchmarks

When several statistical claims are tested simultaneously, the chance that at least one is incorrectly accepted by chance grows with the number of hypotheses. The multiple-testing literature provides many methods, including closed testing and the Holm step-down rule, that aim to control the family-wise error rate while preserving more power on the strongest hypotheses than a flat Bonferroni split would [16, 17]. For the single-hypothesis lower bound that sits inside each test, two complementary tools are widely used: empirical-Bernstein (EB) bounds, which use observed variance to tighten Hoeffding-style inequalities, and more recently, betting-martingale confidence sequences, which provide anytime-valid bounds despite the absence of fixed sample sizes [40, 41]. The algorithmic fairness literature has separately developed formal bounds on subgroup performance metrics, including PPV [1, 4].

However, no published benchmark currently issues a per-cohort three-way decision (certify, abstain, no-guarantee) under FWER control across many domains and methods. Furthermore, the interaction between these single-hypothesis bounds and multi-cohort FWER control under covariate shift, in the specific context of PPV certification, has not been studied as a unified protocol in known literature.

## 3. Method

### 3.1. Protocol

A trained classifier serves predictions across multiple cohorts (for example, age groups, hospitals, or chemical-scaffold classes). For each cohort, we require a statistical guarantee that its PPV clears a given threshold $\tau$ at confidence level $1 - \alpha$. The challenge is that the calibration data the model was trained on typically differs from the deployment population, an instance of covariate shift. To correct for this, we re-weight calibration samples by the importance ratio $w(x) = p_{\text{target}}(x)/p_{\text{cal}}(x)$ and compute a confidence lower bound on the weighted PPV. If this bound clears $\tau$, we CERTIFY, otherwise we ABSTAIN. If weight-reliability diagnostics flag instability (heavy tails, low effective sample size, or weight concentration), we return NO-GUARANTEE rather than risk a vacuous claim. The family-wise error rate (FWER) is controlled over all (cohort, $\tau$) hypotheses per dataset using Holm's step-down procedure [17].

**Setup and assumptions** A model $f : \mathcal{X} \to \{0, 1\}$ has been trained on a source population. A calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i, \hat{y}_i)\}_{i=1}^{n}$ is drawn from $P_{\text{cal}}$, and a test set $\mathcal{D}_{\text{test}} = \{x_j\}$ (labels withheld) is drawn from $P_{\text{target}} \neq P_{\text{cal}}$.

The validity guarantees that we report rest on three assump-

tions:

(A1): Covariate shift
$P_{\text{cal}}(Y \mid X) = P_{\text{target}}(Y \mid X)$, so the source and target distributions differ only in $X$. Concept shift currently remains out of scope.

(A2): Common support
$p_{\text{target}}(x) > 0 \Rightarrow p_{\text{cal}}(x) > 0$, so $w(x)$ is finite almost everywhere on the target support.

(A3): Within-cohort IID sampling
For each cohort $g$ the calibration samples and test points are IID from their respective distributions, while cross-cohort independence is not required as cohorts are disjoint by construction.

A fourth assumption, (A4): Bounded importance weights, $0 < w_{\min} \leq W_i \leq w_{\max} < \infty$, is used only by the WCP analysis in Section 5 and is introduced there.

**Confidence bound and FWER control.** For a cohort $g$ and threshold $\tau \in (0, 1)$, we certify when the importance-weighted PPV lower bound clears $\tau$ at confidence $1 - \alpha$. The primary bound is the Maurer-Pontil empirical-Bernstein (EB) bound [40]:

$$\mathrm{LB}(g, \tau) = \hat{\mu}_g - \sqrt{\frac{2\,\hat{V}_g \log(2/\alpha_k)}{n_{\text{eff}} - 1}} - \frac{7 \log(2/\alpha_k)}{3(n_{\text{eff}} - 1)}, \quad (1)$$

where $\hat{\mu}_g = \sum_i w_i y_i / \sum_i w_i$ is the importance-weighted PPV estimate over the predicted-positive subset ($\hat{y}_i = 1$), $\hat{V}_g$ is the importance-weighted variance, $n_{\text{eff}} = (\sum_i w_i)^2 / \sum_i w_i^2$ is the Kish effective sample size (a conservative plug-in substitution for the original sample size $n$ in the Maurer-Pontil bound), and $\alpha_k$ is the Holm-adjusted level for hypothesis $k$. For $G$ cohorts and $|\mathcal{T}| = 5$ thresholds $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, the family contains $G \cdot |\mathcal{T}|$ hypotheses; we use $\alpha = 0.05$ throughout. The empirical-Bernstein $p$-values obtained by inverting the bound below are super-uniform under the null $\mathrm{PPV} = \tau$ (a corollary of the one-sided Maurer-Pontil bound [40], with the Kish $n_{\text{eff}}$ substituted for $n$, which is conservative since $n_{\text{eff}} \leq n$), and Clopper-Pearson $p$-values are super-uniform by construction; both are therefore valid inputs to Holm step-down, which controls FWER under arbitrary dependence [17].

### 3.2. Datasets

ShiftBench draws on 38 real datasets across three domains: 15 molecular (Murcko scaffold cohorts; RDKit 208-feature descriptors), 12 tabular (demographic and temporal cohorts), and 11 text (temporal, geographic, and category cohorts; TF-IDF with 5,000 features). Datasets range from 60-25,520 calibration samples (total dataset sizes from 303-93,087) and from 3-7,636 cohorts after filtering cohorts with zero positive calibration samples (Table 1). The 23 tabular and text datasets exercise the active-certification regime in our experiments, while the 15 molecular datasets primarily exercise a structural-abstention regime: Murcko scaffold partitioning produces hundreds of novel-scaffold cohorts where models predict zero positives, leaving PPV undefined and the protocol abstaining due to insufficient predicted positives.

*Table 1.* Domain summary (uLSIF, single-method baseline). Effective sample sizes per domain are in Table 2.

| Domain | $n$ datasets | Mean cert% | Cohort definition |
|---|---|---|---|
| Molecular | 15 | 0.03% | Murcko scaffold |
| Tabular | 12 | 7.51% | Demographic bins |
| Text | 11 | 33.90% | Temporal or geographic |

### 3.3. Methods evaluated

We evaluate 9 candidate methods within a unified evaluation harness. The methods are uLSIF, KLIEP, KMM, rULSIF, BBSE (applicable under label shift), split conformal, CV+, group DRO, and weighted conformal prediction (WCP) [28, 12, 11, 29, 30, 24, 34, 13]. The first eight emit hypothesis-test $p$-values (empirical-Bernstein or Clopper-Pearson) that are valid inputs to Holm step-down. WCP, which we include for comparability, was originally designed to provide a marginal per-cohort coverage guarantee on prediction sets [13] rather than a hypothesis-test rejection on PPV; we report its certify rates separately and discuss its behavior as a PPV test in Section 5.

### 3.4. Stability diagnostics and weight behavior on real data

The protocol returns NO-GUARANTEE when the importance weights themselves look unreliable, rather than risk a vacuous PPV claim. Every evaluation therefore tracks three weight-reliability diagnostics drawn from the survey-sampling and importance-sampling literatures. The first is the PSIS tail index $\hat{k}$, a Pareto-smoothed-importance-sampling diagnostic that estimates the heaviness of the upper tail of the weights from the top 20% of samples; we use a gate threshold of 0.7, above which plug-in concentration bounds become unreliable. The second is the Kish effective sample size fraction ($n_{\text{eff}}/n$), a coarse summary of how concentrated the weight mass is, gated at 0.3. The third is the clip mass, the share of total weight concentrated in the top 1% of samples, gated at 0.1. A NO-GUARANTEE decision is returned if any of the three gates fails.

The gate thresholds (PSIS $\hat{k}$, ESS fraction, clip mass) are conservative safety margins fitted on the 38 datasets reported here, not validity prerequisites for the empirical-Bernstein bound. The EB bound itself remains valid for

any importance-weighted estimator with bounded variance regardless of gate setting; failed gates trigger abstention as a conservative practice, not because the bound becomes invalid. See Theorem 1 for the WCP-specific asymptotic limit, which is independent of any gate setting.

To check that these diagnostics are not vacuously triggered (or vacuously satisfied) on the 38 datasets we evaluate, we record their observed ranges. Across all datasets under uLSIF weights, ESS fractions span roughly 0.95 to 1.00 and $\hat{k}$ is approximately 0 in every dataset, suggesting that real-world covariate shift in this suite tends to be mild compared with the gate thresholds. For numerical stability the $\hat{k}$ estimator applies a near-uniformity guard that returns $\hat{k} = 0$ when the maximum-to-median weight ratio falls below 3, a cheap regime check before fitting the Pareto tail; this guard triggered on all 38 datasets, so none of them exercised the gates at certification time. The diagnostics are therefore reported as a post-hoc audit rather than active abstention triggers on this benchmark. The experiments at lognormal $\sigma \in \{1.0, 1.5\}$ discussed in Section 4.1 deliberately exceed naturally observed shift in order to probe protocol behavior in regimes where the gates do begin to trigger.

## 4. Results

### 4.1. Validity calibration

Before any method comparison, we check that the protocol controls FWER at the claimed 5% level on this benchmark. We ran 19,600 boundary-null trials, composed of 10,000 synthetic trials (lognormal weight stress at $\sigma \in [0.5, 3]$, oracle, naive, and cross-fit weight learners, gate-threshold sensitivity sweeps, and estimand-gap stress tests) and 9,600 trials on real fairness data (Adult, COMPAS, and German Credit). Across all 19,600 trials we observed zero false certifications (type I errors). In a separate weighted-null calibration study (B1, Appendix C), the empirical-Bernstein FWER stayed at 0.0% throughout, while the Clopper-Pearson FWER spanned roughly 1.9 to 4.1%, near the nominal 5% level. Without $n_{\text{eff}}$ gating, FWER inflated to roughly 12% at $\sigma = 1.5$, a realistic-stress condition, and to roughly 42% at $\sigma = 3$, an adversarial extreme. Per-experiment tables, Wilson 95% confidence intervals, and per-domain breakdowns are reproduced in Appendix C.

### 4.2. Effective sample size primarily accounts for certification rates

We find that the effective sample size, more than domain identity, accounts for cross-domain variation in certification rates on our benchmark. To see this, we first hold the density-ratio method fixed (uLSIF) and report per-cohort certification rates by domain in Table 2. Under this single-method baseline, the fraction of (cohort, $\tau$) pairs that the protocol certifies spans about three orders of magnitude across the three domains: about 0.03% on molecular, 7.51% on tabular, and 33.90% on text. The domain hierarchy tracks the mean effective sample size in each domain closely.

*Table 2.* Cross-domain certification landscape under a single density-ratio method (uLSIF). Each row aggregates the (cohort, $\tau$) pairs in one domain. Mean $n_{\text{eff}}$ is the average Kish effective sample size across the cohorts in that domain. Cert% is the fraction of pairs the protocol certifies under Holm step-down at $\alpha = 0.05$.

| Domain | Number of datasets | Cert% | Mean $n_{\text{eff}}$ |
|---|---|---|---|
| Molecular | 15 | 0.03% | 0.15 |
| Tabular | 12 | 7.51% | 33.6 |
| Text | 11 | 33.90% | 437.7 |

Because correlation between domain and $n_{\text{eff}}$ does not on its own demonstrate which factor accounts for the certification gap, we ran a controlled subsampling intervention. We subsampled the IMDB text dataset down to molecular-level effective sample sizes, holding all else fixed (uniform weights, $\tau = 0.8$, 50 replicates per level, $n \in \{3, 10, 25, 50, 100, 250, 500, 1000\}$). The certification rate sat at 0% for $n_{\text{eff}} \leq 50$, jumped to roughly 60% at $n_{\text{eff}} \approx 98$, and remained stable for larger $n_{\text{eff}}$. No other variable changed across the sweep. The pattern matches the concentration-inequality power curve and is consistent with $n_{\text{eff}}$ driving certification difficulty on our benchmark through a sample-size mechanism rather than a domain-intrinsic property. A 27-dataset regression supports this reading: $R^2$ for $\log(n_{\text{eff}})$ alone is about 0.65, and the partial $R^2$ for $n_{\text{eff}}$ given domain is about 0.41 versus about 0.11 for domain given $n_{\text{eff}}$. The full regression is in Appendix D.

### 4.3. Method ranking

Table 3 reports the per-method certification rate, broken down by the three domains, for each of the 9 candidate methods under Holm step-down at $\alpha = 0.05$. The eight hypothesis-test methods appear in the upper block of the table; weighted conformal prediction (WCP) is reported separately at the bottom as a coverage method, since its conformal coverage guarantee does not translate directly into a PPV hypothesis-test rejection in our setup (Section 5; Appendix F).

On tabular data, split conformal (SC) certifies roughly 16.5% of (cohort, $\tau$) pairs, ahead of the next-best fully-covered empirical-Bernstein method (BBSE, roughly 8.1%) by about 8 percentage points. On text the ordering compresses, with split conformal at roughly 38.1% and BBSE at roughly 33.9%, a gap of about 4 percentage points. (We take the best among the fully-covered empirical-Bernstein methods as the comparator; group DRO is excluded here for its partial dataset coverage, although on the text datasets it ran it certifies a slightly higher fraction.) Combining tabular and text,

the split-conformal advantage is in the range of roughly 1 to 11 percentage points: the point estimate is about 7 percentage points and a dataset-level bootstrap (10,000 iterations) gives a 95% confidence interval of $[+0.8, +11.2]$ percentage points on the combined gap. The lower end of this interval is small, and we treat the precise size of the gap as uncertain on this benchmark; the qualitative ordering (split conformal $\geq$ each empirical-Bernstein method on every active pair we observed) is the more robust finding. The Holm correction we apply controls the family-wise error rate *within* each method's set of hypotheses, not *across* the nine methods; the method-to-method comparisons in Table 3 are therefore descriptive summaries of per-method certification rates rather than multiplicity-corrected tests of a between-method difference, and we read them as such.

### 4.4. Why split conformal leads

We next ask why split conformal certifies more (cohort, $\tau$) pairs than the empirical-Bernstein methods on this benchmark. Two explanations are plausible. The first is that split conformal targets a different estimand (the unweighted source-domain PPV) and so benefits from a power gain that does not extend to the target-domain quantity. The second is that split conformal and the empirical-Bernstein methods target the same estimand, in practice, but use $p$-values that differ in tightness. We find that the second explanation is more consistent with our benchmark.

To distinguish the two, we compare the Clopper-Pearson and Maurer-Pontil $p$-values on the same data. In the regime where most active pairs sit (effective sample size in the rough range 30 to 100, margin in the rough range 0.2 to 0.4), the exact binomial $p$-value tends to be roughly two to three orders of magnitude smaller than the Maurer-Pontil empirical-Bernstein $p$-value in our experiments. The mean $\log_{10}$ ratio $\log_{10}(p_{\mathrm{EB}}/p_{\mathrm{SC}})$ is roughly $+2.87$ in the $n_{\mathrm{eff}} \in [30, 75]$ bin and drops to roughly $+0.45$ for $n_{\mathrm{eff}} > 500$, where the two families largely agree. We observed zero EB-only certifications relative to split conformal across the 351 active (cohort, $\tau$) pairs in our benchmark,[1] which means every pair that any empirical-Bernstein method certifies, split conformal also certifies.

The estimand-scope explanation is harder to reconcile with the data. Of the 1,325 pre-Holm SC-only (method, cohort, $\tau$) quadruples in our benchmark (counted before the Holm correction, so not directly comparable to the active-pair count above), 1,322 (roughly 99.8%) have $\hat{\mu}_{\mathrm{EB}} \geq \tau$. In other words, the empirical-Bernstein point estimate typically does clear the threshold; what fails is the $p$-value, not the estimate. We read this as evidence for a $p$-value-power advantage rather than an estimand-scope artifact, although either reading is compatible with this benchmark and stronger evidence is left for future work. The full $p$-value-by-bin analysis, robustness checks (cap sensitivity, $\alpha$ robustness, and per-$\tau$ Holm), and a bounded-weight conformal prediction dominance conjecture are reproduced in Appendix E.

[1] We call a (cohort, $\tau$) pair *active* for a comparison when at least one of the compared methods certifies it; here that is the 351 pairs certified by split conformal or some empirical-Bernstein method (of which 228 are also certified by an empirical-Bernstein method and 123 by split conformal alone). This post-Holm head-to-head count is distinct from the pre-Holm SC-only quadruple count discussed in the next paragraph.

### 4.5. Density-ratio estimator choice is second-order

The four density-ratio estimators we compared (uLSIF, KLIEP, rULSIF, and BBSE) agreed on roughly 99% of active (cohort, $\tau$) decisions in our benchmark and differed by less than 0.6 percentage points on tabular certification rates. On text, the four are identical to within rounding. KLIEP and uLSIF disagreed on 2 of 225 active pairs (those where uLSIF or KLIEP certifies), both boundary-local (lower-bound margin roughly 0.09 to 0.15 above $\tau$), with KLIEP certifying and uLSIF abstaining. We observed no EB-only certifications. Within our benchmark, uLSIF is a reasonable default; its closed-form solution is typically faster than KLIEP's iterative optimization (median roughly $2\times$ across our datasets, ranging from comparable to about $17\times$).

### 4.6. Abstention occurs in sparse cohorts

In our 15 molecular datasets, roughly 99.5% of the 55,670 (cohort, $\tau$) pairs have $n_{\mathrm{pos}}(g) = 0$, meaning the model predicts no positives in the scaffold cohort and PPV is undefined, and roughly 99.8% have $n_{\mathrm{eff}} < 10$. The protocol returns NO-GUARANTEE on these pairs by design: a system that certified at $n_{\mathrm{pos}}(g) = 0$ would necessarily inflate FWER well above 5%. We read this as the protocol working as intended on cohorts where the metric itself is undefined. Aggregating to task-level cohorts with at least 50 positives is one practical fix in our analysis; classification reaches roughly 0.02% CERTIFY across all molecular pairs and regression reaches roughly 1.29%, since positives are less structurally rare for regression targets.

## 5. Discussion

### 5.1. WCP as a PPV test: a guard rail

WCP is widely used for prediction-set coverage under shift, and a practitioner who wants to test a per-cohort PPV claim might be tempted to repurpose its lower-bound formula directly as a hypothesis-test statistic. We show analytically that this repurposing is structurally anti-conservative.

**Setup and algorithm.** Fix one cohort, one threshold $\tau \in (0, 1)$, and the boundary null $H_0 : \Pr(Y = 1 \mid \hat{Y} = 1) = \tau$. Let $D_+ = \{(X_i, Y_i, W_i)\}_{i=1}^{n_+}$ be the calibration units

*Table 3.* Per-method CERTIFY rate (%) under Holm step-down at $\alpha = 0.05$. Each cell is the fraction of (cohort, $\tau$) pairs in the named domain that the protocol certifies. The protocol returns three possible outcomes per pair (CERTIFY, ABSTAIN, NO-GUARANTEE); we report only the CERTIFY rate, with the remainder split across the other two outcomes. The eight hypothesis-test methods in the upper block are valid Holm inputs. Weighted conformal prediction (bottom row) is reported as a coverage method and is not FWER-controlled; its rate is not directly comparable with the upper block. KMM$^\dagger$ is skipped on datasets with $n_{\text{cal}} > 1000$ (ran on 14/38: 9 molecular, 5 tabular, 0 text); group DRO$^\dagger$ skipped at $n_{\text{cal}} > 5000$ (ran on 24/38: 12 molecular, 9 tabular, 3 text). Their cells aggregate over only the datasets where each ran; pair denominators therefore differ from the 55,670 / 905 / 410 baseline used by the remaining seven methods. The combined tabular+text SC vs best-EB gap is 7.07 percentage points (dataset-level bootstrap 95% CI $[0.8, 11.2]$). The 38 datasets break down into 15 molecular, 12 tabular, and 11 text.

| Method | Molecular | Tabular | Text |
|---|---|---|---|
| split conformal | 0.08 | **16.46** | **38.05** |
| BBSE | 0.03 | 8.07 | 33.90 |
| KLIEP | 0.03 | 7.73 | 33.90 |
| rULSIF | 0.03 | 7.51 | 33.90 |
| uLSIF | 0.03 | 7.51 | 33.90 |
| CV+ | 0.00 | 7.29 | 30.98 |
| group DRO$^\dagger$ | 0.27 | 5.66 | 37.78 |
| KMM$^\dagger$ | 0.02 | 0.43 | n/a |
| weighted conformal (coverage) | 0.39 | 41.66 | 49.02 |

predicted positive, with importance weights $W_i = \omega(X_i) > 0$ and (A4) $0 < w_{\min} \le W_i \le w_{\max} < \infty$ in addition to the §3 assumptions; we further assume the weights are label-independent within the predicted-positive cohort, $W_i \perp Y_i$ (a condition stronger than covariate shift, since $W_i = \omega(X_i)$ and $Y$ depends on $X$). The WCP lower bound is $\mathrm{lb}(\alpha) = \sum_{i \in S_\alpha} W_{(i)} Y_{(i)} / \sum_{i \in S_\alpha} W_{(i)}$, where the trim index set $S_\alpha$ is obtained by sorting $D_+$ on $Y$ (ascending), then removing the bottom-$\alpha$ fraction of cumulative weight. WCP certifies $H_1 \colon \mathrm{PPV} > \tau$ iff $\mathrm{lb}(\alpha) \ge \tau$.

**Theorem 1** (WCP PPV lower bound is anti-conservative)**.** *Under (A1) to (A4), the label-independence condition $W_i \perp Y_i$, and $\alpha \in (0, 1-\tau)$, define $q^\star_\alpha = \tau/(1-\alpha) > \tau$. As $n_+ \to \infty$: **(i)** $\mathrm{lb}(\alpha) \xrightarrow{P} q^\star_\alpha$; **(ii)** $\Pr(\mathrm{lb}(\alpha) \ge \tau) \to 1$; **(iii)** for every fixed $\alpha_m \in (0, 1-\tau)$ (e.g. $\alpha_m = \alpha/m$), $q^\star_{\alpha_m} = \tau/(1-\alpha_m) > \tau$ and $\Pr(\mathrm{lb}(\alpha_m) \ge \tau) \to 1$, so no Bonferroni-style reduction of $\alpha$ restores validity.*

*Proof sketch; full proof in Appendix F.* By the strong law and (A1)+(A4), the total weights of the zero- and one-labeled units satisfy $T_0/n_+ \to (1-\tau)\mu_W$ and $T_1/n_+ \to \tau\mu_W$ a.s., where $\mu_W = \mathbb{E}[W_1]$ is the mean weight. For $\alpha < 1-\tau$, the trim cutoff $\hat{K}_\alpha$ falls strictly inside the zero-block w.p. $\to 1$ (a uniform law of large numbers for the cumulative-weight partial-sum process, plus the label-independence condition $W \perp Y$). Hence the trim removes only zero-labeled units, so the remaining $S_\alpha$ satisfies $\sum W_{(i)} Y_{(i)} = T_1$ while $\sum W_{(i)} = T - R_\alpha$ with $R_\alpha/n_+ \to \alpha\mu_W$. Slutsky gives $\mathrm{lb}(\alpha) \xrightarrow{P} \tau\mu_W/[(1-\alpha)\mu_W] = \tau/(1-\alpha) = q^\star_\alpha > \tau$, proving (i); (ii) is immediate. For (iii), every fixed $\alpha_m \in (0, 1-\tau)$ has $q^\star_{\alpha_m} = \tau/(1-\alpha_m) > \tau$, so (ii) gives $\Pr(\mathrm{lb}(\alpha_m) \ge \tau) \to 1$ at each Bonferroni level $\alpha_m = \alpha/m$; shrinking $\alpha$ moves $q^\star_{\alpha_m}$ toward $\tau^+$ but never to or below $\tau$ (full proof in Appendix F). □

**Finite-sample sanity.** At $\alpha = 0.05$ the lb asymptote is $q^\star_{0.05} \in \{0.526, 0.737, 0.947\}$ for $\tau \in \{0.5, 0.7, 0.9\}$ (claim (i)). The certify-probability limit in claim (ii) is 1, not $q^\star$; at finite $n_+$ the two trajectories sit apart. B2 (2,000 trials, $n_+ \in [30, 80]$, weights unshifted at $\sigma = 0$) reports certify rates $\Pr(\mathrm{lb} \ge \tau \mid H_0) \in \{0.651, 0.669, 0.811\}$ for $\tau \in \{0.5, 0.7, 0.9\}$, far exceeding the nominal $\alpha = 0.05$. This is the upward direction predicted by claim (ii); the finite-$n_+$ values themselves do not pin down the limit, only its sign relative to $\alpha$. A non-asymptotic Berry-Esseen bound on the rate of convergence is left as an open question.

**Why Bonferroni does not rescue.** Theorem 1(iii): at every Bonferroni level $\alpha_m = \alpha/m$, $q^\star_{\alpha_m} \to \tau^+$ stays strictly above $\tau$, so the certify probability under $H_0$ tends to 1, still $\gg \alpha_m$. Under the bounded-weight assumptions of Theorem 1 this bias is a property of the bound's asymptote rather than a multiple-testing artifact; whether it persists under genuinely heavy-tailed weights is left open (Section 6).

WCP's original prediction-set coverage guarantee [13] is not in question; the lower bound it produces is not a calibrated hypothesis-test statistic for population-parameter testing. The algorithmic detail and full proof are reproduced in Appendix F.

### 5.2. When the protocol can certify

A protocol that controls FWER but never certifies is of limited practical value, and the cross-domain results in Section 4.2 indicate that whether the protocol certifies depends strongly on the effective sample size of the cohort. We therefore characterize the regime in which certification be-

*Table 4.* Estimated certification power (split conformal with Clopper-Pearson under Holm step-down) as a function of effective sample size $n_{\text{eff}}$ and margin (true PPV minus $\tau$). At margin 0.1, $n_{\text{eff}}$ near 200 is the interpolated 50% crossing.

| $n_{\text{eff}}$ | margin $= 0.1$ | margin $= 0.2$ | margin $= 0.3$ |
|---|---|---|---|
| 25 | 0.0% | 0.0% | ≈0% |
| 50 | 0.0% | 0.0% | 0.0% |
| 100 | 5.8% | ≈10% | ≈20% |
| 150 | 25.8% | ≈40% | ≈50% |
| 200 | 49.6% | ≈70% | ≈80% |
| 500 | 99.8% | ≈100% | ≈100% |

comes feasible by sweeping the effective sample size at a fixed margin $\text{true_PPV} - \tau$. The resulting power curve in Table 4 suggests roughly the following thresholds in our setting: $n_{\text{eff}}$ near 200 is needed for 50% power at margin 0.1 (the interpolated 50% crossing is $n_{\text{eff}} = 200 \rightarrow 49.6\%$ and $n_{\text{eff}} = 210 \rightarrow 55.5\%$); $n_{\text{eff}} \geq 100$ for 50% power at margin 0.2; and $n_{\text{eff}} \geq 75$ at margin 0.3. Below $n_{\text{eff}}$ near 30 the empirical-Bernstein penalty in Eq. 1 exceeds any realistic margin even at the least-conservative Holm level, so no method we tested has power in that regime. Molecular scaffold cohorts in our benchmark have median $n_{\text{eff}}$ between roughly 1 and 6, tabular demographic cohorts between 25 and 40, and text temporal and geographic cohorts between 200 and 500. This range is consistent with the wide cross-domain certification gap reported in Section 4.2.

### 5.3. Beyond classifiers

Our benchmark uses small-to-medium classifiers; validity was empirically audited only at that scale. In principle the same FWER-control requirement and the same WCP-as-PPV-test caveat apply to any per-cohort metric evaluation that reports (cohort, $\tau$) families, including foundation-model evaluations, though whether the diagnostic gate thresholds calibrated on our 38 datasets transfer to foundation-model regimes (where importance-weight tails can be much heavier) remains open. As a concrete illustration, the artifact repository includes a Colab notebook applying this protocol to a fine-tuned LLM classifier on the CivilComments toxicity dataset with per-demographic-group cohorts; a systematic sweep across multiple open LLMs is staged for the extended Datasets and Benchmarks version of this work.

## 6. Limitations

Despite our work's potential utility for per-cohort certification under shift, our approach has several limitations that we address in this section. Our protocol addresses the covariate-shift family of distribution shift, where the input distribution changes between calibration and target while the conditional label distribution is preserved. Label shift, conditional drift, and concept shift are out of scope and not covered by the validity guarantee; expected behavior under mild covariate-shift violations is discussed in the next paragraph.

**Behavior under mild covariate-shift violations.** When (A1) holds only approximately (small conditional drift), the empirical-Bernstein lower bound remains a valid one-sided confidence bound on the source-domain weighted PPV, but that weighted quantity no longer coincides with the target-domain PPV, so the bound no longer tracks the deployment metric without bias. We expect the certify rate to drop (more abstentions) rather than spike (false certifications): residual conditional drift biases the lower bound relative to the target-side PPV, while in heavier-tailed regimes we would expect the PSIS $\hat{k}$ and clip-mass values to rise toward the NO-GUARANTEE thresholds (though on the 38 benchmark datasets the gates never triggered, $\hat{k} \approx 0$ throughout). A formal sensitivity result under bounded conditional drift is left as future work.

**Gate machinery under adversarial regimes remain minimally tested.** Our realism bridge (Section 4.1) finds all 38 datasets near-uniform under uLSIF density-ratio estimation (PSIS $\hat{k} \approx 0$, ESS $\geq 0.95$). The empirical FWER guarantees we report are therefore validated under realistic shift but not under the adversarial heavy-tail regime where weighted methods are most prone to fail. Two consequences follow. First, Theorem 1 assumes bounded weights (A4); this assumption is satisfied trivially on this benchmark because empirical weights are near-uniform, so whether the closed-form anti-conservativity result transfers to deployment regimes with genuinely heavy-tailed weights is left to future work. Second, the diagnostic gates never triggered on real data in our benchmark; their calibration against synthetic lognormal stress (Section 4.1) is the strongest test of the gate machinery we run.

We also focus on positive predictive value because PPV is, in our experience, the hardest case for sparse cohorts: it conditions on predicted positives, which amplifies weight concentration into near-zero effective sample size. The same harness extends to recall, AUC, and calibration metrics once the relevant estimand and its null are defined, and PPV serves as our stress test rather than a fundamental limit of the approach.

Real deployment weights are rarely fully characterized, and we therefore treat the protocol as a benchmarked baseline rather than a deployment-ready certificate; users who plan to act on a CERTIFY decision should validate the diagnostics on their own deployment data. Two of the methods we evaluate are subject to scalability caps in our runs: KMM is skipped on datasets where $n_{\text{cal}} > 1{,}000$ (ran on 14/38) and group DRO on $n_{\text{cal}} > 5{,}000$ (ran on 24/38). Their cells in Table 3 aggregate over the datasets where each ran, with smaller pair denominators than the seven fully-

covered methods; method-ranking conclusions involving KMM or group DRO are not directly comparable to the fully-covered methods. Results for the seven uncapped methods are unaffected. Finally, the default gate thresholds (PSIS tail index 0.7 and ESS fraction 0.3) were selected to balance abstention against power on these 38 datasets, and the abstention and power tradeoff at these thresholds may differ on datasets with substantially different positive rates or label distributions not represented in this benchmark. Deriving principled gate thresholds analytically from the Maurer-Pontil concentration constants is left as future work.

## 7. Conclusion

ShiftBench is a benchmarking protocol for shift-aware certify-or-abstain decisions on per-cohort PPV under covariate shift, with Holm-based family-wise error control across the (cohort, $\tau$) hypotheses of each dataset. Across 38 real datasets and 9 candidate methods, the protocol issued no false certifications in 19,600 boundary-null trials, in which each cohort's true PPV is placed exactly at its threshold. Among the family-wise-error-controlled methods, split conformal certified the largest fraction of cohorts on tabular and text data, by a modest margin whose precise size is uncertain on this benchmark; the qualitative ordering is the more robust finding, and the choice of density-ratio estimator was second-order. Using the weighted-conformal lower bound directly as a one-sided PPV test, by contrast, is anti-conservative under the null and is not a valid basis for certification.

The protocol is metric-agnostic and model-agnostic, and its two central lessons extend beyond the present setting: family-wise error control is required whenever a metric is certified separately across many cohorts, and a coverage guarantee such as weighted conformal prediction must not be read as a hypothesis test on a population quantity. The harness accommodates recall, AUC, and calibration once the relevant estimand and its null are specified, and applying it to foundation-model evaluation, where per-subgroup reporting is already standard, is a natural next step. The principal open questions are whether the weight-reliability gates calibrated here transfer to the heavier-tailed importance weights that larger models and richer features can induce, how to derive the gate thresholds analytically rather than empirically, and the non-asymptotic rate at which the weighted-conformal anti-conservativity established in Theorem 1 emerges at finite sample sizes.

## References


[1] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153-163, 2017. https://doi.org/10.1089/big.2016.0047.

[2] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, 2012. https://doi.org/10.1145/2090236.2090255.

[3] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Innovations in Theoretical Computer Science (ITCS)*, 2017. https://arxiv.org/abs/1609.05807.

[4] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. https://arxiv.org/abs/1610.02413.

[5] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness (FairWare)*, 2018. https://doi.org/10.1145/3194770.3194776.

[6] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021. https://arxiv.org/abs/2012.07421.

[7] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*, 2021. https://arxiv.org/abs/2007.01434.

[8] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1902.10811.

[9] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations (ICLR)*, 2022. https://arxiv.org/abs/2110.11328.

[10] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009. https://mitpress.mit.edu/9780262170055/dataset-shift-in-machine-learning/.

[11] Jiayuan Huang, Alexander J Smola, Arthur Gretton, Karsten M Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007. https://papers.nips.cc/paper_files/paper/2006/hash/a2186aa7c086b46ad4e8bf81e2a3a19b-Abstract.html.

[12] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699-746, 2008. https://doi.org/10.1007/s10463-008-0197-x.

[13] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1904.06019.

[14] Amos Storkey. When training and test sets are different: Characterizing learning transfer. *Dataset Shift in Machine Learning (MIT Press)*, pages 3-28, 2009. https://mitpress.mit.edu/9780262170055/dataset-shift-in-machine-learning/.

[15] Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press, 2012. https://mitpress.mit.edu/9780262017091/.

[16] Ruth Marcus, Eric Peritz, and K. Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655-660, 1976. https://doi.org/10.1093/biomet/63.3.655.

[17] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65-70, 1979. https://www.jstor.org/stable/4615733.

[18] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289-300, 1995. https://www.jstor.org/stable/2346101.

[19] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29 (4):1165-1188, 2001. https://doi.org/10.1214/aos/1013699998.

[20] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2021. https://arxiv.org/abs/2103.03098.

[21] Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. https://arxiv.org/abs/2010.06595.

[22] Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021. https://arxiv.org/abs/2107.07002.

[23] Sara Hooker. Moving beyond "algorithmic bias is a data problem". In *Patterns*, 2021. https://arxiv.org/abs/2105.00031.

[24] Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *Annals of Statistics*, 49(1): 486-507, 2021. https://doi.org/10.1214/20-AOS1965.

[25] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. https://doi.org/10.1007/b106715.

[26] Yaniv Romano, Evan Patterson, and Emmanuel J Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1905.03222.

[27] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B*, 76(1):71-96, 2014. https://doi.org/10.1111/rssb.12021.

[28] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391-1445, 2009. https://jmlr.org/papers/v10/kanamori09a.html.

[29] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution

comparison. *Neural Computation*, 25(5):1324-1370, 2013. https://doi.org/10.1162/NECO_a_00442.

[30] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning (ICML)*, 2018. https://arxiv.org/abs/1802.03916.

[31] Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965. https://archive.org/details/surveysampling0000kish.

[32] 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. https://doi.org/10.1007/s11222-016-9696-4.

[33] Aditya Krishna Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. *International Conference on Machine Learning (ICML)*, 2016. https://arxiv.org/abs/1605.07720.

[34] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020. https://arxiv.org/abs/1911.08731.

[35] Stephen Bates, Anastasios N Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021. https://arxiv.org/abs/2101.02703.

[36] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094-1111, 2018. https://doi.org/10.1080/01621459.2017.1307116.

[37] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494-591, 2023. https://doi.org/10.1561/2200000101.

[38] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021. https://arxiv.org/abs/2110.01052.

[39] Ying Jin and Emmanuel J. Candès. Model-free selective inference under covariate shift via weighted conformal P-values. *arXiv preprint arXiv:2307.09291*, 2023. https://arxiv.org/abs/2307.09291.

[40] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, 2009. https://arxiv.org/abs/0907.3740.

[41] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B*, 86 (1):1-27, 2024. https://doi.org/10.1093/jrsssb/qkad009.

## Reproducibility Statement

All code, preprocessed data, result tables, and figures are committed to the anonymous repository at https://anonymous.4open.science/r/shiftbench-B098. The per-dataset artifacts (features, labels, cohort assignments, and calibration/test splits) are committed directly, so the benchmark reproduces without rerunning the one-time featurization step; the featurization scripts are included for completeness but are not on the runtime path. The full cross-domain sweep (38 datasets $\times$ 9 methods $\times$ 5 thresholds) runs on commodity CPU hardware with no GPU required, taking roughly 20 minutes of wall-clock time on a single commodity machine, and all randomness is seeded for every method so the result tables regenerate bit-for-bit. The statistical core (empirical-Bernstein and Clopper-Pearson bounds, $p$-value inversion, and the Holm step-down) is unit-tested against pinned golden values.

## A. Density Ratio Agreement: Full Tables (H1)

KLIEP vs. uLSIF agreement from the H1 ablation study (6 datasets, 30 trials each, multiple $\tau$ and $\alpha$ settings; note: a separate multi-trial ablation from the full C1 benchmark) is reported in Table 5:

*Table 5.* H1 ablation: KLIEP vs. uLSIF agreement on certify/abstain decisions (6 datasets, 30 trials per dataset).

| Dataset | Total pairs | Disagree | Agree% |
|---|---|---|---|
| Adult | 750 | 0 | 100% |
| BACE | 750 | 0 | 100% |
| BBBP | 276 | 32 | 88.4% |
| COMPAS | 750 | 0 | 100% |
| IMDB | 175 | 0 | 100% |
| Yelp | 600 | 0 | 100% |
| **Overall** | **3,301** | **32** | **99.0%** |

In the ablation, all 32 disagreements fall in the $[\tau + 0.1, \tau + 0.2]$ LB-margin bin on BBBP; zero disagreements below $\tau$ or above $\tau + 0.2$. **Full C1 benchmark** (38 datasets, 56,985 cohort$\times\tau$ pairs): KLIEP and uLSIF produce exactly 2 disagreements, on `adult` (White_Male_55+, $\tau$=0.5) and `credit_default` (mid_edu3, $\tau$=0.7); BBBP shows zero disagreements in the single-run C1 data. The 1% figure quoted in §4.5 is computed over the 225 active C1 pairs.

## B. Method Coverage Table

Seven of nine methods ran on all 38 datasets. KMM and group_dro are skipped on datasets that exceed their per-method $n_{\text{cal}}$ thresholds, so their coverage is partial: KMM ran on 14/38; group DRO on 24/38. Their cells in Table 3 aggregate only over the datasets where each ran; Table 6 lists the per-method coverage in full.

*Table 6.* Method coverage across the 38 datasets (15 molecular, 12 tabular, 11 text). Seven of the nine methods run on every dataset; KMM and group DRO are skipped on datasets exceeding their $n_{\text{cal}}$ caps.

| Method | Molecular (/15) | Tabular (/12) | Text (/11) | Coverage |
|---|---|---|---|---|
| `split_conformal` | 15 | 12 | 11 | 100% |
| `bbse` | 15 | 12 | 11 | 100% |
| `ulsif` | 15 | 12 | 11 | 100% |
| `kliep` | 15 | 12 | 11 | 100% |
| `rulsif` | 15 | 12 | 11 | 100% |
| `cvplus` | 15 | 12 | 11 | 100% |
| `weighted_conformal` | 15 | 12 | 11 | 100% |
| `group_dro`$^\dagger$ | 12 | 9 | 3 | 63% (24/38) |
| `kmm`$^\dagger$ | 9 | 5 | 0 | 37% (14/38) |

†: Skipped on datasets where $n_{\text{cal}} > 5{,}000$ (group_dro) or $n_{\text{cal}} > 1{,}000$ (KMM); pair denominators in Table 3 therefore differ from the 55,670 / 905 / 410 used by the remaining seven methods. Method-ranking conclusions involving KMM or group DRO are not directly comparable to the fully-covered methods.

## C. Validity Calibration: Per-Experiment Detail

Per-experiment FWER tables (B1: CP-vs-EB under lognormal weights, 2,000 trials per condition $\times 4\sigma \times 2$ domains; T1: oracle / naive / cross-fit weight learners, 1,000 trials each; T1-ext: Adult / COMPAS / German Credit fairness datasets, 72,000 trials (a separate, larger real-data FWER study, distinct from the 9,600 real-data boundary-null trials counted in Section 4.1); T2: gate-threshold sensitivity, 13,500 trials; T3: realism bridge under uLSIF; T4: estimand-gap stress test) and the corresponding Wilson 95% confidence intervals are reproduced in `results/holm_corrected/cert_rate_summary_C1.csv` and `results/numbers.json`. The headline figures cited in Section 4.1 (zero false certifications across the 19,600 boundary-null trials, empirical-Bernstein FWER 0.0% throughout, Clopper-Pearson FWER spanning roughly 1.9 to 4.1%, roughly 12% inflation at $\sigma = 1.5$ without $n_{\text{eff}}$ gating, and roughly 42% at $\sigma = 3$) all derive from those files.

## D. Effective Sample Size: Regression Detail

OLS regression of (cohort, $\tau$) certify rates on $\log(n_{\text{eff}})$, domain, and shift magnitude across 27 datasets in our benchmark gives $R^2 \approx 0.65$ for $\log(n_{\text{eff}})$ alone, $R^2 \approx 0.50$ for domain alone, $R^2 \approx 0.71$ for the joint model, and partial $R^2(n_{\text{eff}} \mid \text{domain}) \approx 0.41$ versus $R^2(\text{domain} \mid n_{\text{eff}} + \text{shift}) \approx 0.11$, consistent with $n_{\text{eff}}$ being the dominant explanatory variable in this regression. The 27-dataset regression mixes methods (uLSIF for molecular and BBSE for tabular/text); the single-method table in Section 4.2 confirms the cross-domain hierarchy is preserved when one method is held fixed.

## E. P-Value Family: Mechanism Detail

Mean $\log_{10}(p_{\text{EB}}/p_{\text{SC}})$ across 295,765 (dataset, cohort, $\tau$, EB-method) quadruples, binned by effective sample size: $\approx +1.00$ at $n_{\text{eff}} < 30$ ($\approx 10\times$ ratio), $\approx +2.87$ at $n_{\text{eff}} \in [30, 75]$ ($\approx 740\times$), $\approx +2.28$ at $[75, 150]$, $\approx +1.98$ at $[150, 500]$, and $\approx +0.45$ at $n_{\text{eff}} > 500$ ($\approx 3\times$; methods largely agree). Of 1,325 pre-Holm SC-only quadruples in our benchmark, 1,322 ($\approx 99.8\%$) have $\hat{\mu}_{\text{EB}} \geq \tau$, which is consistent with the gap being driven by $p$-value family rather than estimand. Robustness checks: the SC versus EB gap is preserved at $\alpha \in \{0.01, 0.05, 0.10\}$ (roughly 8.2 to 8.4 pp on tabular); under per-$\tau$ Holm the tabular gap widens to roughly 8.84 pp. A bounded-weight conjecture, namely $p_{\text{CP}} \leq p_{\text{EB}}$ when weights are bounded and labels are Bernoulli, is empirically supported by zero EB-only certifications across the 351 active pairs in our benchmark but is not formally proved; whether tighter betting-martingale EB sequences [41] recover EB-only certifications in the $n_{\text{eff}} \in [30, 150]$ regime remains open.

## F. P1.1: Canonical WCP Sanity (Tibshirani 2019 vs. PPV Lower Bound)

**Setup.** Fix one cohort and threshold $\tau \in (0, 1)$. Let $D_+ = \{(X_i, Y_i, W_i)\}_{i=1}^{n_+}$ be the calibration units predicted positive in this cohort, with importance weights $W_i = \omega(X_i) > 0$, and assume covariate shift, common support, within-cohort IID, and bounded weights $0 < w_{\min} \leq W_i \leq w_{\max} < \infty$. Under the boundary null $H_0$: PPV $= \tau$ identically, the WCP PPV lower bound is $\text{lb}(\alpha) = \sum_{i \in S_\alpha} W_{(i)} Y_{(i)} / \sum_{i \in S_\alpha} W_{(i)}$, after sorting positives by $Y_i$ and trimming according to the importance-weighted bottom-$\alpha$ rule.

**Empirical limit.** On 2,000 simulated trials per condition with $n_+ \in [30, 80]$ and $\alpha = 0.05$, the certify probability under $H_0$ takes values roughly $\{0.65, 0.67, 0.81\}$ for $\tau \in \{0.5, 0.7, 0.9\}$. These values are well above the nominal $\alpha$ and increase with $\tau$, which is consistent with the lower bound being biased upward in the limit through the importance-weighted trim, with the asymptote $q_\alpha^\star = \tau/(1-\alpha) > \tau$ established in Theorem 1; only the non-asymptotic Berry-Esseen rate of convergence to that asymptote remains open. The point relevant to a practitioner is that this empirical certify rate is well above $\alpha$ and does not shrink to $\alpha$ under any Bonferroni-style adjustment that we tested.

**Full proof of Theorem 1.** Throughout, (A1) to (A4) are the covariate-shift, common-support, within-cohort IID, and bounded-weight assumptions stated above; we additionally use the label-independence condition $W_i \perp Y_i$ within the predicted-positive cohort (stated with the theorem setup, and stronger than covariate shift since $W_i = \omega(X_i)$ and $Y$ depends on $X$), and write $\mu_W = \mathbb{E}[W_1]$, $\sigma_W^2 = \text{Var}(W_1) < \infty$.

*Proof idea, in words.* Trimming the lowest-$\alpha$ share of weight removes only zero-labeled units (they sort first), so the surviving set is artificially enriched with true positives. Its weighted PPV then settles not at $\tau$ but at $\tau/(1-\alpha)$, which is always above $\tau$; the bound therefore certifies the null with probability tending to one. Steps 1 to 4 make this precise, and Step 5 shows that shrinking $\alpha$ (the Bonferroni fix) does not remove the bias.

*Proof. Step 1: reduction to a weighted-mean ratio.* Let $T_0 = \sum_{i:Y_i=0} W_i$ and $T_1 = \sum_{i:Y_i=1} W_i$, so $T = T_0 + T_1$. By the strong law, the label-independence condition $W_i \perp Y_i$, and (A4),

$$T_0/n_+ \xrightarrow{\text{a.s.}} (1-\tau)\mu_W, \quad T_1/n_+ \xrightarrow{\text{a.s.}} \tau\mu_W, \quad T/n_+ \xrightarrow{\text{a.s.}} \mu_W.$$

After the $y$-sort, the first $n_0 := |\{i : Y_i = 0\}|$ entries are exactly the zero-labeled units, and the remaining $n_+ - n_0 = n_1$ are the one-labeled units. Set $\hat{\pi}_0 = T_0/T$. By Slutsky, $\hat{\pi}_0 \xrightarrow{P} 1-\tau$.

*Step 2: behavior of the trim cutoff $\hat{K}_\alpha$.* The cumulative-weight CDF $G_{n_+}(k) = T^{-1}\sum_{i\le k} W_{(i)}$ is piecewise constant with jumps $W_{(i)}/T$. Because $\alpha < 1-\tau$ and $\hat{\pi}_0 \to 1-\tau$, eventually $\alpha < \hat{\pi}_0$ a.s., so the cutoff falls strictly inside the zero-block: $\hat{K}_\alpha \le n_0$. Conditional on $\{i \le n_0\}$, the $W_{(i)}$ are IID from the conditional law of $W_1$ given $Y_1 = 0$. By the label-independence condition $W_i \perp Y_i$ this conditional law equals the marginal law of $W_1$, so the within-zero-block weights are IID copies of $W_1$. The empirical mean process indexed by $t \in [0,1]$, $G_{n_+}(\lfloor tn_0 \rfloor) = T^{-1}\sum_{i\le\lfloor tn_0\rfloor} W_{(i)}$, converges uniformly in $t$ to $(1-\tau)\,t \cdot (\mu_W/\mu_W) = (1-\tau)t$ by a uniform law of large numbers for the partial-sum process of IID bounded weights, followed by continuity of the inverse map (Vervaat's lemma) to convert convergence of the cumulative-weight process into convergence of the quantile. Hence

$$\hat{K}_\alpha/n_0 \xrightarrow{P} \alpha/(1-\tau).$$

*Step 3: numerator and denominator of* $\mathrm{lb}(\alpha)$. Let $R_\alpha = \sum_{i\le\hat{K}_\alpha} W_{(i)}$ be the removed weight and $\bar{T}_\alpha = T - R_\alpha$ the remainder. By Step 2 and the LLN, $R_\alpha/n_+ \xrightarrow{P} \alpha\mu_W$ and $\bar{T}_\alpha/n_+ \xrightarrow{P} (1-\alpha)\mu_W$. Because $\hat{K}_\alpha \le n_0$ eventually, the trim removes only zero-labeled units, so $\sum_{S_\alpha} W_{(i)}Y_{(i)} = T_1$. Therefore

$$\mathrm{lb}(\alpha) = \frac{T_1}{T - R_\alpha} \xrightarrow{P} \frac{\tau\,\mu_W}{(\tau + (1-\tau) - \alpha)\,\mu_W} = \frac{\tau}{1-\alpha} =: q^\star_\alpha.$$

Since $\alpha \in (0, 1-\tau) \subset (0,1)$ implies $1-\alpha < 1$, $q^\star_\alpha = \tau/(1-\alpha) > \tau$ strictly, proving (i).

*Step 4: certify probability.* By Step 3, $\mathrm{lb}(\alpha) - \tau \xrightarrow{P} q^\star_\alpha - \tau > 0$. For any $\varepsilon \in (0, q^\star_\alpha - \tau)$, $\Pr(\mathrm{lb}(\alpha) - \tau < \varepsilon) \to 0$, hence $\Pr(\mathrm{lb}(\alpha) \ge \tau) \to 1$, proving (ii).

*Step 5: Bonferroni regime.* Claims (i) and (ii) hold for every fixed $\alpha_m \in (0, 1-\tau)$. In particular, at any Bonferroni level $\alpha_m = \alpha/m$ the asymptote $q^\star_{\alpha_m} = \tau/(1-\alpha_m) > \tau$, so Step 4 gives $\Pr(\mathrm{lb}(\alpha_m) \ge \tau) \to 1$. As $m$ grows, $q^\star_{\alpha_m} \to \tau^+$ from above but never reaches $\tau$, so reducing $\alpha$ shifts the asymptotic bias toward zero without changing its sign and the certify probability stays at 1 for each fixed level; no Bonferroni-style correction restores validity, proving (iii). The coupled regime $\sqrt{n_+}\,\alpha_m = O(1)$, in which the rejection probability would have a non-degenerate limit between $1/2$ and 1, depends on the convergence rate flagged in Remark 3 and is left open. □

*Remark* 2 (Heavy-tailed weights). (A4) imposes bounded weights. In our B2 simulation we use lognormal $W_i \sim \mathrm{LN}(0,\sigma)$ for $\sigma \in \{0, 0.5, 1.0, 1.5\}$, which have finite moments but are unbounded above for every $\sigma > 0$ and therefore fall outside (A4), so Theorem 1 does not formally cover them. Heavier tails inject more variance into $\mathrm{lb}(\alpha)$ and pull more trials below $\tau$ for fixed $n_+$; we expect (but do not prove for unbounded weights) the asymptotic limit $q^\star_\alpha$ to be unchanged, as it depends on $\mu_W$ only through a ratio in which $\mu_W$ cancels. Empirically (B2 sanity, $\tau = 0.7$, $n = 100$) the certify rate moves from $0.712$ ($\sigma = 0$) through $0.729$ ($\sigma = 0.5$) to $0.658$ ($\sigma = 1.0$) and $0.621$ ($\sigma = 1.5$), heavier weights delay convergence to anti-conservatism but do not reverse it. The unbounded-weight regime (e.g. Pareto with shape $< 2$) is not covered by Theorem 1 and is left open.

*Remark* 3 (Finite-sample rate is open). The result above is asymptotic: it says where the bound lands as $n_+$ grows, not how fast it gets there. A finite-sample rate (a Berry-Esseen bound of order $1/\sqrt{n_+}$) is harder because the random trim cutoff $\hat{K}_\alpha$ makes the constant depend on the weight tail; we leave it open.

**Task.** We contrast: (A) Tibshirani 2019 conformal coverage for individual predictions; (B) PPV lower-bound type I error.

**Setup.** Binary labels $y \sim \mathrm{Bernoulli}(0.7)$; conformity score $s = 1 - y$; importance weights $w \sim \mathrm{LN}(0,\sigma)$. Table 7 reports both quantities side by side.

*Table 7.* P1.1 sanity check: WCP coverage holds for individual predictions (Tibshirani 2019 setting) but the same WCP lower bound badly inflates type I error when re-purposed as a PPV test ($n_{\mathrm{cal}} = 100$, 2,000 trials per row).

| $\sigma$ | Tibshirani coverage | PPV-lb type I error | Required |
|---|---|---|---|
| 0.0 | **1.000** $\geq 0.95$ ✓ | **0.712** $\gg 0.05$ × | Cov. $\geq 0.95$; TI $\leq 0.05$ |
| 0.5 | **1.000** $\geq 0.95$ ✓ | **0.729** $\gg 0.05$ × | |
| 1.0 | **1.000** $\geq 0.95$ ✓ | **0.658** $\gg 0.05$ × | |
| 1.5 | **1.000** $\geq 0.95$ ✓ | **0.621** $\gg 0.05$ × | |

2,000 trials per condition; $n_{\mathrm{cal}} = 100$.

Tibshirani's coverage guarantee holds; the PPV lower-bound formula exhibits between 62 and 73% type I error here, consistent with the asymptotic upward bias $\mathrm{lb} \to \tau/(1-\alpha) > \tau$ established in Theorem 1. This bias is not constrained by the original coverage guarantee, which constrains a different property (prediction set coverage, not population parameter testing).

**W2a: Bonferroni does not rescue WCP.** Even reducing $\alpha$ via Bonferroni correction, WCP type I error stays between 57 and 83% across the $\alpha' < 1 - \tau$ values we tested, consistent with Theorem 1(iii): the asymptotic bias shrinks but does not change sign as $\alpha' \to 0$, so it is not removed by a smaller significance threshold (Table 8).

*Table 8.* W2a: Bonferroni cannot rescue WCP ($m = 20$ tests, 2,000 trials).

| Correction | $\alpha'$ ($m = 20$) | $P(\text{certify} \mid H_0)$ | FWER ($m = 20$) |
|---|---|---|---|
| None ($\alpha = 0.05$) | 0.0500 | 82.5% | 100% |
| Bonferroni | 0.0025 | 56.9% | 100% |
| Diagonal (valid) | 0.0025 | $\leq 0.25\%$ | $\leq 5\%$ |