Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Auditing Conformal Prediction under Distribution Shift: A Detectability Boundary, Exact Label-Budget Design, and Repair

Maha Moussa

maha.moussa@usu.edu

Utah State University

Research Article

Keywords: conformal prediction, covariate shift, concept shift, uncertainty quantification, effective sample size, deployment monitoring

Posted Date: August 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10396495/v1

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

Additional Declarations: No competing interests reported.

Auditing Conformal Prediction under Distribution Shift: A Detectability Boundary, Exact Label-Budget Design, and Repair

Maha Moussa[1,2*]

[1*]Department of Mathematics and Statistics, Utah State University, Logan, Utah, USA.
[2]Department of Statistics, Mathematics and Insurance, Faculty of Commerce, Benha University, Benha, Egypt.

Corresponding author(s). E-mail(s): maha.moussa@usu.edu;

Abstract

Conformal prediction quantifies uncertainty without parametric outcome assumptions, but standard coverage relies on exchangeability. Importance weighting can restore validity under covariate shift when density ratios are known. In deployment, practitioners often have labeled source data and only unlabeled target covariates, raising a basic question: can those data determine whether coverage remains credible? We introduce DriftGuard, a two-stage audit. Its label-free stage reports domain discriminability, effective sample size of estimated density ratios, weight stabilization, and visible abstention. We establish exact and approximate detectability bounds, a minimax testing consequence, and a sharp $[\mathbf{0}, \mathbf{1}]$ identified set for target coverage under unrestricted conditional shift. DriftGuard-L crosses this boundary with a probability sample of target labels, using an exact one-sided coverage test, prospective label-budget calculation, and finite-sample conformal recalibration. Under strong Gaussian covariate shift, ordinary conformal coverage fell to 0.789, while oracle and cross-fitted weighted conformal prediction attained 0.926 and 0.923. A published doubly robust comparator was robust when either nuisance was oracle, whereas both-misspecified coverage fell to 0.812. In an untouched chronological deployment of 8,734 hourly Bike Sharing observations, source conformal, estimated weighting, and estimated doubly robust calibration undercovered (0.568–0.594); 2,176 delayed target labels raised subsequent coverage from 0.558 to 0.884. Under pure concept shift, an overlap score remained near 0.96 while coverage fell from 0.90 to 0.40. Exact label budgets quantified power, and target recalibration restored synthetic evaluation coverage. The framework separates overlap diagnosis, visible abstention, target-label falsification, and repair.

Keywords: conformal prediction, covariate shift, concept shift, uncertainty quantification, effective sample size, deployment monitoring

1 Introduction

Conformal prediction (CP) converts almost any predictive algorithm into prediction intervals or sets with finite-sample marginal coverage under exchangeability [1–3]. That guarantee is attractive precisely because it avoids specifying a full outcome distribution. Deployment data, however, are rarely exchangeable with development data. Geographic transfer, changing case mix, sensor replacement, policy changes, and temporal drift can alter the

covariate distribution P_X, the conditional outcome law $P_{Y|X}$, or both [4].

Weighted conformal prediction (WCP) extends validity to covariate shift by weighting calibration scores with the density ratio $w(x) = dQ_X(x)/dP_X(x)$ [5]. Related work treats time-varying distributions, nonexchangeable sequences, auditing, and multiple sources [6–9]. These methods establish what can be achieved under stated shift assumptions. A complementary operational problem has received less direct attention: what can an analyst know when target labels are absent?

This paper makes four contributions. First, we characterize label-free auditing as an identification problem: observational equivalence yields a minimax testing lower bound and, absent restrictions on $Q_{Y|X}$, a sharp $[0, 1]$ identified set for coverage. Second, DriftGuard reports discriminability and effective sample size (ESS) separately, optionally summarizes them with an explicitly nonprobabilistic overlap score, and makes abstention visible. Third, DriftGuard-L supplies an exact marginal-coverage test, exact prospective label-budget design, and target recalibration with finite-sample validity. Fourth, we evaluate detectable and deliberately undetectable failures, a published doubly robust comparator, biased audit sampling, 45 rolling temporal deployments, and a held-out year of hourly transportation demand while reporting power, width, and abstention jointly. Concurrent work uses small target-labeled samples to learn auxiliary audit models for conditional classification coverage [10]; our objective differs by exactly testing marginal undercoverage, quantifying label budgets, and covering regression as well as classification.

The Capital Bikeshare deployment is a central validation rather than a supplementary illustration: source CP, estimated WCP, and estimated doubly robust calibration are evaluated on an untouched future year, followed by disjoint delayed-label repair and nine prequential monthly deployments. This design tests the framework under temporal dependence and outcome evolution that deliberately exceed the assumptions of the clean Gaussian experiments.

2 Methods

2.1 Setting and conformal baselines

Let $D_s = \{(X_i, Y_i)\}_{i=1}^n$ be labeled source observations and $U_t = \{X_j^t\}_{j=1}^m$ unlabeled target covariates. A fitted predictor $\hat{f}$ and calibration nonconformity scores R_i yield a split-conformal threshold from the empirical $(1-\alpha)$ quantile. For regression, $R_i = |Y_i - \hat{f}(X_i)|$. For classification, $R_i = 1 - \hat{p}_{Y_i}(X_i)$ and the prediction set contains classes with $\hat{p}_k(x) \geq 1 - q$.

WCP replaces the empirical distribution of R_i with a weighted distribution using $w_i = dQ_X/dP_X(X_i)$. This targets a changed covariate distribution under the identifying condition $Q_{Y|X} = P_{Y|X}$. When ratios are estimated, finite-sample validity generally needs additional conditions or doubly robust constructions [11]; our estimated-weight results are therefore empirical rather than advertised as exact guarantees. Proposition 1 isolates the population price of ratio error without treating estimated weights as exact.

For exact test-point weighting, each target covariate x receives weight $w(x)$. The weighted empirical distribution assigns masses

$$\frac{w(X_i)}{\sum_{j=1}^n w(X_j) + w(x)} \text{ at } R_i,$$
$$\frac{w(x)}{\sum_{j=1}^n w(X_j) + w(x)} \text{ at } +\infty.$$

The $(1-\alpha)$ quantile is target-specific and may be infinite when calibration support is inadequate. We compare oracle Gaussian ratios, five-fold cross-fitted logistic ratio estimates, and estimates clipped to $[1/20, 20]$. Infinite intervals are reported rather than silently removed.

Proposition 1 (Estimated-ratio validity gap)**.** *Assume covariate shift. Let Q have density $w(X)$ relative to the source joint law P, and let $\widetilde{Q}$ have density $\widetilde{w}(X)$, with both ratios nonnegative and integrating to one. For any fixed prediction set C, if $\widetilde{Q}\{Y \in C(X)\} \geq 1 - \alpha$, then*

$$Q\{Y \in C(X)\} \geq 1 - \alpha - \tfrac{1}{2} E_P |w(X) - \widetilde{w}(X)|.$$

Proof. Coverage is the probability of the event $\{(X, Y) : Y \in C(X)\}$. The difference in this event probability under Q and $\widetilde{Q}$ is bounded by

their total variation distance, which here equals $\frac{1}{2}E_P|w - \widetilde{w}|$. □

The bound is diagnostic rather than directly estimable without assumptions on ratio error. It makes the separation precise: oracle test-point WCP has its stated finite-sample guarantee; estimated WCP inherits an approximation term and is not relabeled "exact."

2.2 Primary component diagnostics and optional overlap summary

DriftGuard fits a probabilistic classifier to distinguish calibration covariates from target covariates. Let A be its held-out area under the ROC curve and let $\hat{w}_i$ be the corresponding prior-corrected odds ratio, clipped to $[0.02, 50]$. Define

$$n_{\text{eff}} = \frac{(\sum_i \hat{w}_i)^2}{\sum_i \hat{w}_i^2}, \tag{1}$$

$$R = \left[\max\{0, 2(1-A)\} \min\left\{1, \frac{n_{\text{eff}}}{n}\right\}\right]^{1/2}. \tag{2}$$

The primary diagnostics are A, n_{eff}/n, the ratio range, and the clipping or stabilization rule. The optional overlap score $R \in [0, 1]$ merely compresses two observable requirements—covariate overlap and nonconcentration of importance weights—for dashboards or triage. It is not a coverage probability, inferential quantity, or statistical certificate, and no scientific conclusion depends on it. When $n_{\text{eff}} < 0.35n$, weights are shrunk toward one, $\tilde{w}_i = 1 + \lambda(\hat{w}_i - 1)$, using the largest $\lambda \in [0, 1]$ satisfying the ESS floor. A score cutoff of 0.45 is examined only as an illustrative operational rule; users should instead predeclare actions from the component diagnostics and application costs. Sensitivity over 0.25–0.75 is supplied to demonstrate the rule's nonuniversality.

2.3 A label-free impossibility result

Proposition 2 (Detectability boundary)**.** *Let T be any audit statistic measurable with respect to (D_s, U_t). Consider two target laws Q_0 and Q_1 with the same covariate marginal $Q_{0,X} = Q_{1,X}$ but potentially different conditional laws $Q_{0,Y|X}$ and $Q_{1,Y|X}$. Then T has the same distribution under Q_0 and Q_1. In particular, no such label-free audit can uniformly certify target coverage over unrestricted changes in $Q_{Y|X}$.*

Proof. The observed target component U_t has the same law under Q_0 and Q_1, while D_s is unchanged. Therefore the joint law of all arguments of T is identical. Yet $Q_{1,Y|X}$ can place target outcomes outside any fixed prediction set with arbitrarily high probability. Hence target coverage may differ while the distribution of T cannot. □

Corollary 1 (Minimax indistinguishability)**.** *For any binary label-free audit $\phi(D_s, U_t)$ and any observationally equivalent pair (Q_0, Q_1) in Proposition 2, its type-I and type-II errors satisfy*

$$Q_0(\phi = 1) + Q_1(\phi = 0) = 1.$$

Consequently, their maximum is at least $1/2$; no label-free decision rule performs uniformly better than guessing on this pair.

Proof. Proposition 2 gives $Q_0(\phi = 1) = Q_1(\phi = 1)$. Substitute $Q_1(\phi = 0) = 1 - Q_1(\phi = 1)$. □

Proposition 3 (Approximate indistinguishability)**.** *Let P_0^{obs} and P_1^{obs} denote the laws of all data visible to a label-free audit. For every binary audit ϕ,*

$$P_0^{\text{obs}}(\phi = 1) + P_1^{\text{obs}}(\phi = 0) \geq 1 - \text{TV}(P_0^{\text{obs}}, P_1^{\text{obs}}).$$

If the source law is unchanged and the audit observes m independent target covariates, then

$$P_0^{\text{obs}}(\phi = 1) + P_1^{\text{obs}}(\phi = 0) \geq 1 - \sqrt{\tfrac{m}{2}\,\text{KL}(Q_{0,X} \| Q_{1,X})},$$

with the lower bound truncated at zero.

Proof. The first inequality is Le Cam's two-point testing bound [12]. The unchanged source component contributes zero divergence, product structure gives $\text{KL}(P_0^{\text{obs}} \| P_1^{\text{obs}}) = m\,\text{KL}(Q_{0,X} \| Q_{1,X})$, and Pinsker's inequality completes the result. □

Proposition 3 extends the exact boundary: even when target covariate marginals differ slightly, a finite unlabeled sample can have little power to distinguish deployment regimes.

Proposition 4 (Sharp coverage identified set)**.** *Fix Q_X and a prediction set $C(x)$. Suppose that, for Q_X-almost every x, the outcome space contains measurable selections both inside and outside $C(x)$.*

Over unrestricted target conditionals $Q_{Y|X}$ compatible with the observed Q_X, the sharp identified set for target coverage is $[0, 1]$.

Proof. Choose the inside selection with probability γ and the outside selection with probability $1 - \gamma$, conditionally at every x. The resulting target law has coverage γ and the same observed Q_X. Every $\gamma \in [0, 1]$ is therefore attainable, and probabilities cannot lie outside $[0, 1]$. □

2.4 DriftGuard-L: minimal-label falsification and repair

When the label-free audit cannot exclude conditional shift, draw a simple random audit sample of b labeled target observations, independent of model fitting and source calibration. For fixed source-calibrated prediction sets, let K_b be the number of covered audit outcomes. DriftGuard-L tests

$$H_0 : c_t \geq 1 - \alpha \quad \text{versus} \quad H_1 : c_t < 1 - \alpha$$

using the exact lower-tail p-value $p_b = \Pr\{\text{Binomial}(b, 1 - \alpha) \leq K_b\}$. If undercoverage is detected, the target audit sample can replace the source calibration sample: target nonconformity scores are recalibrated using order statistic $\lceil (b + 1)(1 - \alpha) \rceil$ for subsequent target predictions.

Proposition 5 (Audit and repair validity)**.** *If target audit and subsequent evaluation observations are i.i.d., and the original prediction rule is fixed independently of audit outcomes, the lower-tail test has size at most η for every $c_t \geq 1 - \alpha$. If target recalibration uses the split-conformal order statistic above, a subsequent target outcome is covered with probability at least $1 - \alpha$.*

Proof. Conditional on the fixed prediction rule, $K_b \sim \text{Binomial}(b, c_t)$. Its lower tail is stochastically largest over the null at $c_t = 1 - \alpha$, giving size at most η. The repair claim follows from exchangeability of the b target calibration scores and the subsequent target score and the standard conformal rank argument. □

For prospective design, let $k_\eta(b)$ be the largest integer satisfying $\Pr\{\text{Binomial}(b, 1-\alpha) \leq k_\eta(b)\} \leq \eta$. At a scientifically relevant alternative $c_1 < 1-\alpha$, exact power is

$$\pi(b; c_1) = \Pr\{\text{Binomial}(b, c_1) \leq k_\eta(b)\}. \quad (3)$$

The smallest budget attaining power $1-\beta$ is $b_{\min} = \min\{b : \pi(b; c_1) \geq 1 - \beta\}$. Thus "a small label sample" becomes a prespecified design choice.

Operational decision protocol.

DriftGuard executes four predeclared steps: (i) report domain-classifier AUC, ESS fraction, ratio range, and clipping; (ii) use oracle WCP only when ratios are known and label estimated WCP as approximate; (iii) abstain or acquire labels when overlap is inadequate, while treating acceptable overlap as silent about concept shift; and (iv) choose b from Eq. (3), audit a probability sample, then recalibrate on target labels if undercoverage is detected. The contribution is therefore a formally specified decision problem with distinct information regimes, not a claim that a composite score is a new coverage theorem.

Table 1 makes the relationship with doubly robust calibration explicit. Yang et al.'s construction protects asymptotically against misspecification of either the ratio or conditional-score nuisance under covariate shift. DriftGuard-L answers a different question: after representative target outcomes are acquired, it supplies a finite-sample test and recalibration guarantee without requiring either nuisance model, but only under target exchangeability and a fixed prediction rule. These guarantees are complementary rather than competing.

2.5 Experiments

We used four public scikit-learn datasets: diabetes regression ($n = 442$), breast-cancer classification ($n = 569$), wine classification ($n = 178$), and digits classification ($n = 1{,}797$). A nonlinear synthetic regression dataset ($n = 2{,}200$) provided controlled ground truth. Predictions averaged random forests, extremely randomized trees, and histogram gradient boosting. Each condition used 20 deterministic repetitions and nominal coverage $1 - \alpha = 0.90$.

Covariate shift was generated by exponential tilting along a randomly selected standardized feature direction, at severities $0, 0.75, 1.5$, and 2.5. Concept shift retained the target covariate law but added $s\{1.5X_1 + X_3^2 - 1\}$ to target regression outcomes. Mixed shift applied both mechanisms. Labels were used only for retrospective evaluation. We compared ordinary split CP, WCP, and ESS-stabilized WCP with the DriftGuard audit.

Table 1 Guarantee regimes and their noninterchangeable assumptions.

Procedure	Information and assumptions	Guarantee	Principal failure mode
Oracle test-point WCP	Source labels, target covariates, known ratio, covariate shift	Finite-sample marginal coverage	Conditional shift or poor support
Estimated WCP	Estimated ratio and covariate-shift model	Approximate; ratio error matters	Ratio misspecification or overlap
Doubly robust calibration	Estimated ratio and conditional score law; one nuisance consistently estimated	Asymptotic target calibration under covariate shift	Both nuisances wrong or conditional shift
DriftGuard-L	Representative target labels, fixed rule, exchangeable target calibration/evaluation	Exact audit size and finite-sample recalibration	Biased acquisition or dependence

Outcomes were coverage, interval width or set size, shift-classifier AUC, ESS, reliability, and abstention.

For DriftGuard-L, independent synthetic target populations were audited using label budgets of 25, 50, 100, and 200. We used 80 repetitions per shift type, severity, and budget. The one-sided audit used level 0.05. Evaluation always used a disjoint fixed holdout of 1,300 target observations, so power and repaired coverage were not assessed on reused labels.

The exact-WCP study used equal-covariance Gaussian source and target covariates, making the oracle ratio known analytically. It included 1,200 training, 800 calibration, and 1,000 target observations; mean shifts ranged from 0 to 2.0. Each covariate, concept, and mixed-shift condition used 80 repetitions. A separate ratio-robustness experiment used 40 repetitions per shift to compare oracle ratios, correctly specified cross-fitted logistic ratios, nonlinear boosted ratios, and clipped logistic ratios. It recorded normalized empirical L_1 ratio error, coverage, finite width, and infinite-interval rate. A 20,000-repetition binomial study verified audit size and analytical power. Label-selection bias was examined in a two-stratum population where hard and easy cases had coverages 0.55 and 0.95 and true prevalences 0.20 and 0.80.

Two naturally occurring transfer designs supplemented the controlled studies. Monthly adjusted stock prices distributed with Matplotlib [13] were transformed to log returns. In addition to one fixed chronological split, a predeclared rolling design used five assets, 120-month training, 48-month calibration, and disjoint 24-month deployment windows, yielding 45 asset–window deployments. Uncertainty used 4,000 bootstrap resamples of complete asset–window clusters. A demographic transfer used the two sex-coded groups in the public diabetes dataset; 100 source-group resamples were evaluated in each direction. These cases diagnose deployment behavior but are not used to claim i.i.d. guarantees for dependent returns.

We additionally implemented the split-sample efficient-influence-function calibration equation of Yang et al. [11]. Its two nuisances are the source-to-target density ratio and the conditional score distribution. A Gaussian experiment crossed oracle and deliberately misspecified versions of both nuisances, using 3,000 source and 2,000 target observations, 20 repetitions, and shifts 0, 1, and 2. Estimated nuisance fits used logistic domain odds and gradient-boosted conditional score classifiers. These comparisons concern asymptotic double robustness, not unconditional exact finite-sample validity.

For an external chronological validation, we used all 17,379 hourly observations in the UCI Bike Sharing data [14]. The point model used January–June 2011; source calibration used July–December 2011; and all of 2012 was untouched deployment data (8,734 observations). The year indicator and the casual and registered counts were excluded to prevent temporal identity and outcome-component leakage. We compared source CP, estimated test-point WCP, and estimated

doubly robust calibration. Confidence intervals resampled complete days 4,000 times. Finally, January–March 2012 labels were revealed for delayed-label recalibration and April–December remained a disjoint evaluation period. A second predeclared design evaluated nine monthly deployments from April through December; at each anchor, only the immediately preceding 7, 14, 30, 60, or 90 days of labels calibrated the following month. Deployment-cluster bootstraps quantified repeated-transfer uncertainty. Because hourly observations are dependent, these intervals and delayed repairs are descriptive external validation, not i.i.d. exact guarantees.

3 Results

3.1 Public-data and synthetic benchmarks

Across the benchmark suite, all methods were generally close to nominal coverage under absent or moderate covariate shift (Figure 1). WCP sometimes improved coverage at stronger shifts, but its variability increased as overlap declined. Stabilization produced similar means with reduced exposure to extreme weights. The wine and breast-cancer datasets showed greater Monte Carlo variability because their calibration and target samples were small.

The overlap score declined as shifts became more separable and density-ratio ESS fell. Across individual runs, its association with absolute coverage error was weak (Spearman $\rho = -0.10$, $p = 0.043$), confirming that it is not a calibrated probability of coverage. Threshold sensitivity reinforced the point: across cutoffs from 0.25 to 0.75, every pure-concept-shift deployment was accepted even though 75% had coverage below 0.85. For covariate shift, cutoffs at or above 0.35 rejected the single undercovered run. The score is therefore retained only as a compact overlap warning; AUC and ESS are the primary reported quantities.

3.2 Exact test-point weighting

The exact construction clarified both the benefit and price of covariate-shift correction (Figure 2). At Gaussian shift 2.0, ordinary split CP covered 0.789, whereas oracle and cross-fitted estimated WCP covered 0.926 and 0.923. Their infinite-interval rates were 0.106 and 0.098. Clipping eliminated infinite intervals but reduced coverage to 0.873. At shift 1.5, oracle and estimated WCP covered approximately 0.913 and 0.912, compared with 0.838 for ordinary CP. Thus, clipping trades visible abstention for hidden undercoverage when overlap is poor.

Across all 1,200 paired simulation conditions, estimated test-point WCP increased coverage by 0.0395 (95% CI 0.0356–0.0434) and reduced absolute deviation from 0.90 by 0.0307 (95% CI 0.0270–0.0343) relative to ordinary CP. Oracle and estimated results were nearly indistinguishable in this controlled setting.

The dedicated ratio-estimation stress test showed when oracle and estimated weighting diverge (Figure 3). At mean shift 2.0, oracle, logistic, boosted, and clipped-logistic coverage was 0.919, 0.915, 0.888, and 0.874, respectively. Their infinite-interval rates were 0.136, 0.120, 0.052, and 0.000. Mean normalized ratio L_1 error was 0.108 for logistic estimation, 0.334 for boosting, and 0.187 after clipping. Thus, a flexible ratio estimator was not automatically safer, and clipping again exchanged visible support failure for undercoverage. These findings empirically instantiate Proposition 1: estimated weighting is an approximation whose quality depends on ratio accuracy.

3.3 Doubly robust comparator and external deployment

The doubly robust stress test supplied a strong comparator rather than a new validity claim (Figure 4, left). At Gaussian shift 1, coverage was 0.907 with both nuisances oracle, 0.904 with an oracle ratio and misspecified outcome nuisance, and 0.911 with an oracle outcome nuisance and misspecified ratio. With both nuisances misspecified, coverage fell to 0.882. At severe shift 2, the corresponding coverages were 0.903, 0.886, 0.912, and 0.812; estimated doubly robust calibration attained 0.882 and estimated WCP 0.895. Thus, the oracle-outcome regime retained the nominal target despite a wrong ratio, while finite-sample instability and overlap remained visible for ratio-driven and fully estimated fits.

The untouched Bike deployment provided a materially harder external test (Figure 4, right). During 2012, source CP covered 0.568 (day-bootstrap 95% interval 0.552–0.583), estimated

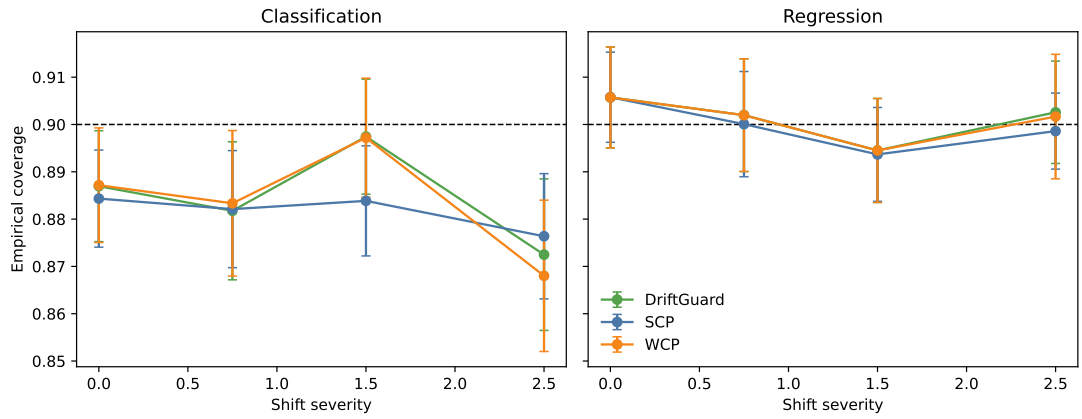


Fig. 1 Empirical coverage across public and synthetic benchmarks. Points are means across datasets and 20 repetitions; bars show 95% normal-approximation intervals for the mean. The dashed line is 0.90.

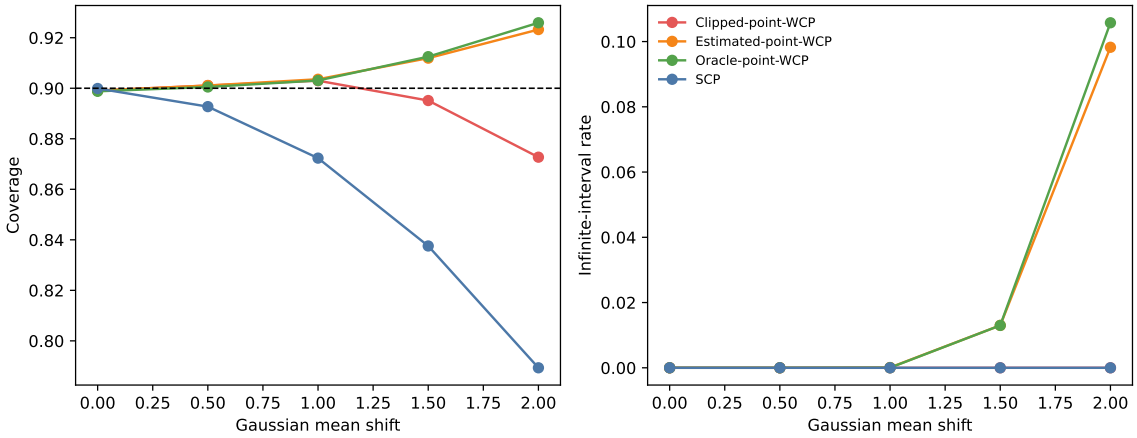


Fig. 2 Exact test-point weighted conformal prediction under Gaussian covariate shift. Infinite intervals reveal lack of target support.

WCP covered 0.594 (0.575–0.612), and estimated doubly robust calibration covered 0.589 (0.574–0.605). All three intervals exclude 0.90 by a wide margin. This is not a failed comparison to conceal: it shows that neither covariate weighting nor double robustness under a covariate-shift model repairs temporal changes outside that model.

After chronologically accumulating 2,176 labels from January–March 2012, delayed-label recalibration increased coverage on the disjoint April–December period from 0.558 (0.541–0.576) to 0.884 (0.874–0.894), with the interval threshold increasing from 81.4 to 210.1. This calendar-defined label count was not selected by the exact budget calculation. Residual undercoverage is consistent with continued temporal evolution and dependence; accordingly, this is descriptive external evidence rather than an invocation of Proposition 5. The external study therefore reproduces the framework's full logic: label-free correction fails under an unverified conditional-stability assumption, target outcomes falsify it, and delayed labels provide a large but not magical repair.

Rolling deployment removed dependence on that single calendar split (Figure 5). Across nine monthly evaluations, fixed source calibration averaged 0.556 coverage (deployment-bootstrap 95% interval 0.524–0.589). Recalibration using only prior target labels averaged 0.918 with 7 days, 0.917 with 14 days, 0.913 with 30 days, 0.907 with 60 days, and 0.900 with 90 days. The result was therefore repeatable across deployment anchors

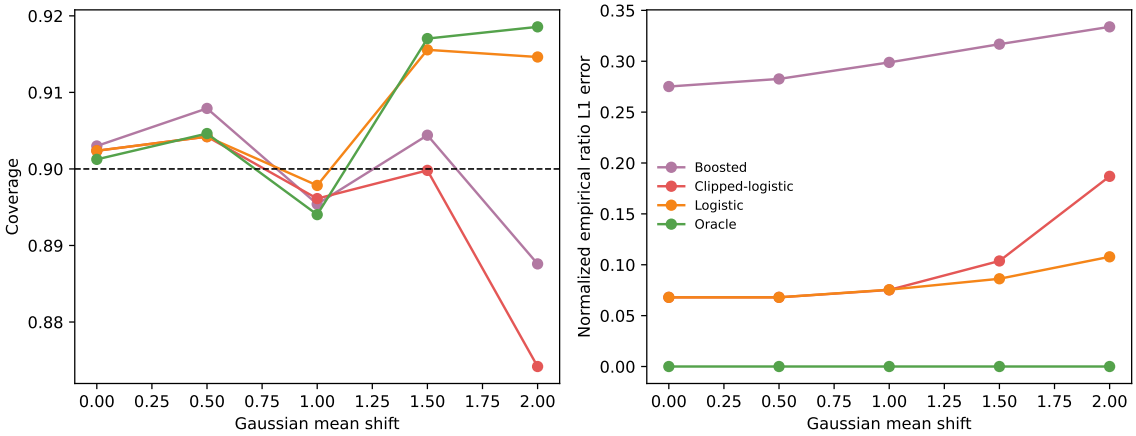


Fig. 3 Density-ratio robustness with known Gaussian oracle ratios. The left panel reports coverage; the right reports normalized empirical L_1 ratio error.

and robust to the labeling-history choice. Short histories adapted most quickly but produced wider, more variable intervals; longer histories traded responsiveness for efficiency.

3.4 Natural temporal and demographic transfers

The observational cases illustrated two qualitatively different regimes (Figure 6). Across 45 rolling stock asset–window deployments, ordinary CP coverage was 0.941 (window-bootstrap 95% interval 0.921–0.957), and estimated WCP coverage was 0.919 (0.896–0.940). The paired difference was −0.022 (−0.032 to −0.013). Mean finite width decreased for every asset, but weighting did not improve coverage. The earlier fixed chronological split was also conservative. These results are an observed stress test, not evidence of universal weighting superiority.

In diabetes sex-group transfer, ordinary CP averaged 0.913 coverage from the lower-coded to higher-coded group and 0.885 in the reverse direction. Exact estimated WCP returned infinite intervals for every target observation in both directions. Because sex itself identifies the domain and the calibration set contains only the source group, the density ratio correctly exposes a complete support failure. Reporting 100% coverage without the 100% infinite-interval rate would be misleading.

3.5 Detectability boundary

The controlled experiment produced the paper's main result (Figure 7). Under pure covariate shift, DriftGuard coverage was 0.890, 0.906, 0.908, and 0.904 as severity increased; the overlap score simultaneously declined from 0.964 to 0.375, flagging 95% of the most severe deployments. Under pure concept shift, the overlap score stayed between 0.964 and 0.976 and accepted every deployment, while coverage fell to 0.774, 0.593, and 0.404. Mixed shift combined both behaviors: the audit detected worsening overlap but could not attribute additional failure to the conditional change.

3.6 How many target labels cross the boundary?

The exact audit converted invisible conditional shift into a power–budget relationship (Figure 8). At concept-shift severity 1.0, source-calibrated evaluation coverage was 0.715. Detection power was 0.763 with 25 labels, 0.963 with 50, and 1.000 with 100 or 200. For the milder severity 0.5, where coverage was 0.839, power increased from 0.125 to 0.850 as the budget rose from 25 to 200. Mixed-shift results were similar or stronger.

Exact prospective calculations sharpened this interpretation. At level 0.05, detecting coverage 0.85 with 80% power requires 270 labels, whereas detecting coverage 0.80, 0.75, or 0.70 requires 78, 40, or 25 labels. For 90% power the corresponding budgets are 368, 109, 55, and 33. Small audits therefore detect severe failures efficiently

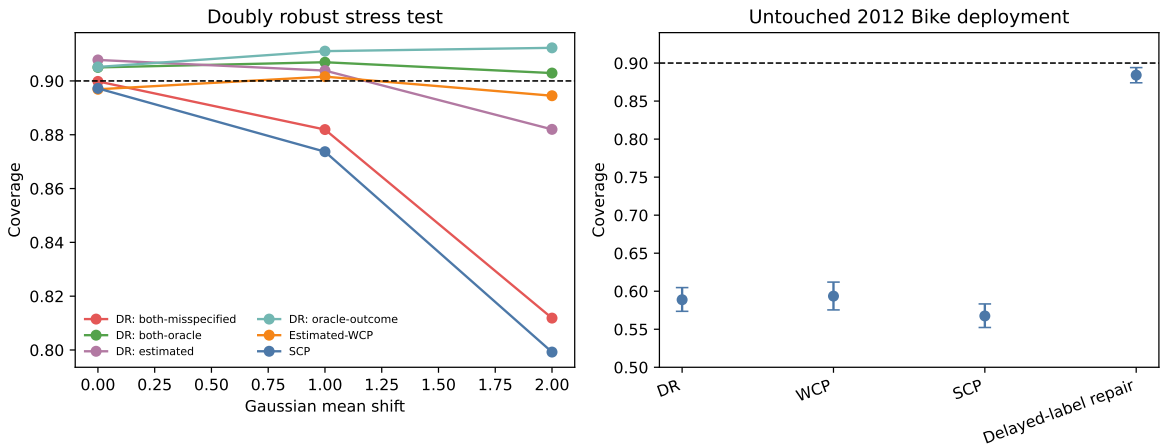


Fig. 4 Published doubly robust comparator and external validation. Left: crossed oracle and misspecified nuisances under Gaussian covariate shift. Right: day-bootstrap intervals in the untouched 2012 Bike deployment and disjoint delayed-label repair.

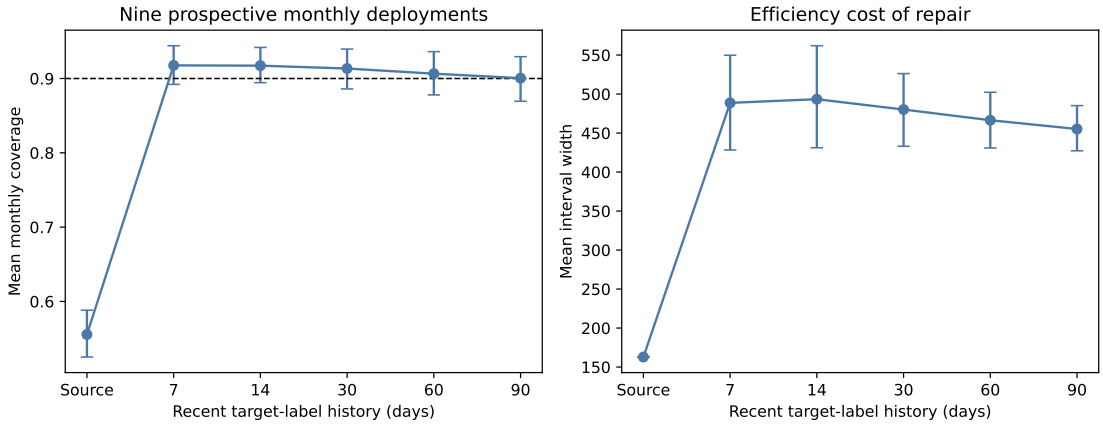


Fig. 5 Prequential external repair across nine monthly Bike deployments. Each evaluation month uses only labels observed during the immediately preceding history window; bars are deployment-bootstrap 95% intervals.

but should not be advertised as well powered for five-percentage-point undercoverage.

Target recalibration restored independent evaluation coverage, with an explicit efficiency cost (Figure 9). Under concept-shift severity 1.0, mean coverage after recalibration was 0.957, 0.922, 0.907, and 0.901 for budgets 25, 50, 100, and 200. Small budgets were conservative because conformal quantiles are coarse; larger budgets brought coverage and width toward stable values. At severity 1.5, mean width increased from about 5.0 under invalid source calibration to 11.2 with 200 target labels, making the cost of honest uncertainty visible.

3.7 Audit size, power, and sampling bias

The exact audit was conservative at small budgets and approached level 0.05 as discreteness decreased (Figure 10). When true coverage was 0.90, empirical rejection probabilities were 0.034, 0.021, 0.042, 0.042, and 0.046 for budgets 25, 50, 100, 200, and 500, closely matching exact analytical values. With true coverage 0.80, power increased from 0.381 at 25 labels to 0.873 at 100 and 0.989 at 200.

Representativeness was essential. In the two-stratum population, true marginal coverage was 0.87. With 100 audit labels sampled at the correct 20% hard-case prevalence, rejection probability

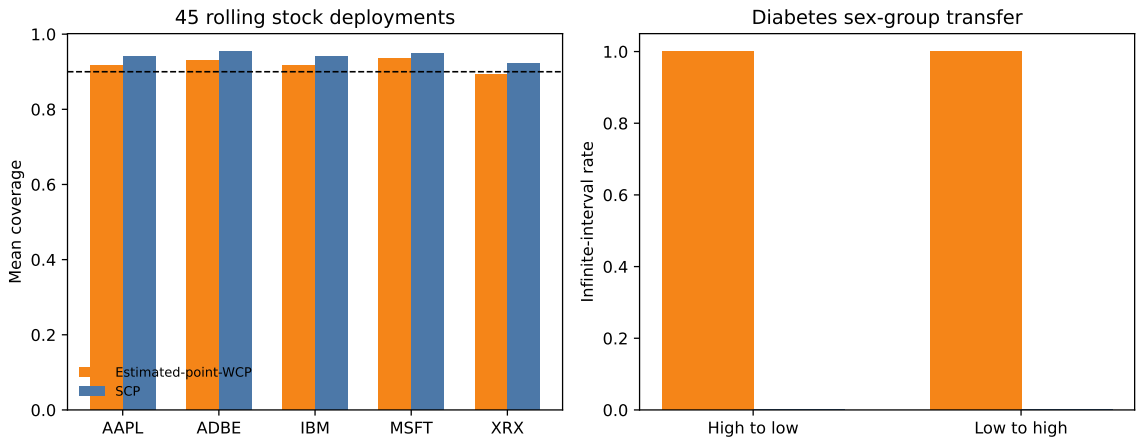


Fig. 6 Observed transfer cases. Bars in the temporal panel summarize five assets over 45 rolling deployment windows; sex-group transfer deliberately tests nonoverlapping demographic support.

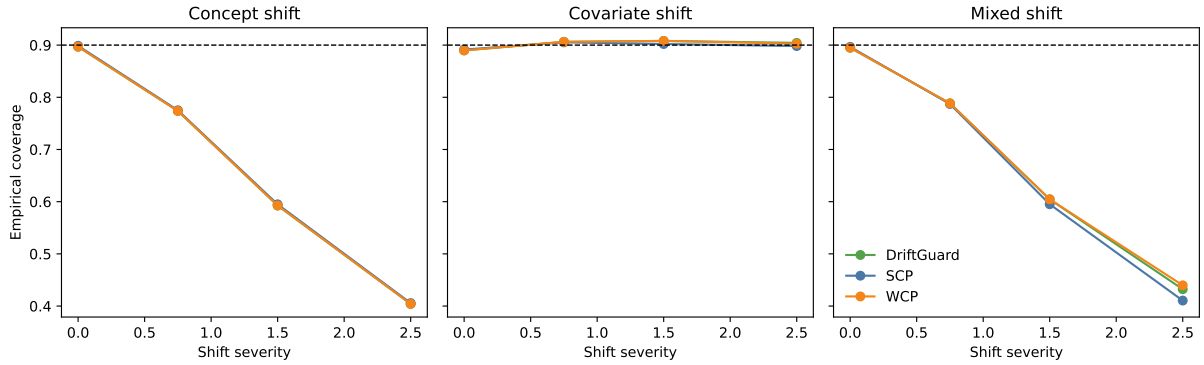


Fig. 7 Detectability boundary. Covariate shift changes observable P_X; concept shift changes only $P_{Y|X}$ and is therefore invisible without target labels.

was 0.229. Undersampling hard cases at 5% biased estimated coverage upward by 0.059 and reduced rejection to 0.002; oversampling them at 40% biased coverage downward by 0.080 and raised rejection to 0.913. The test is exact for a probability sample, not for an arbitrary convenience sample.

In practice, representativeness should be designed rather than assumed. When a simple random sample is infeasible, predeclared stratified probability sampling can deliberately oversample rare or high-risk cases while retaining known inclusion probabilities; stratum-specific coverage estimates are then aggregated using target-population stratum proportions. With unequal known inclusion probabilities, design-weighted Horvitz–Thompson or Hájek estimators and survey-sampling intervals target marginal coverage, but the simple exact binomial test no longer applies [15]. If selection probabilities are estimated from observed variables, inverse-probability or augmented outcome–selection estimators can reduce observed selection bias, while inference becomes model-assisted rather than finite-sample exact. If labeling depends on unobserved outcomes even after conditioning, neither correction identifies target coverage without additional assumptions; a sensitivity analysis over plausible selection odds or worst-case coverage of unlabeled cases should then replace certification [16]. DriftGuard-L therefore requires the report to state the sampling frame, inclusion mechanism, response rate, adjustment model, and whether inference is exact, design-based, model-assisted, or sensitivity-based.

Table 2 Controlled detectability experiment for DriftGuard.

Shift	Severity	Coverage	Overlap score	Acceptance
Covariate	0.00	0.890	0.964	1.00
Covariate	1.50	0.908	0.567	0.95
Covariate	2.50	0.904	0.375	0.05
Concept	0.00	0.897	0.976	1.00
Concept	1.50	0.593	0.965	1.00
Concept	2.50	0.404	0.969	1.00
Mixed	1.50	0.605	0.567	1.00
Mixed	2.50	0.432	0.374	0.05

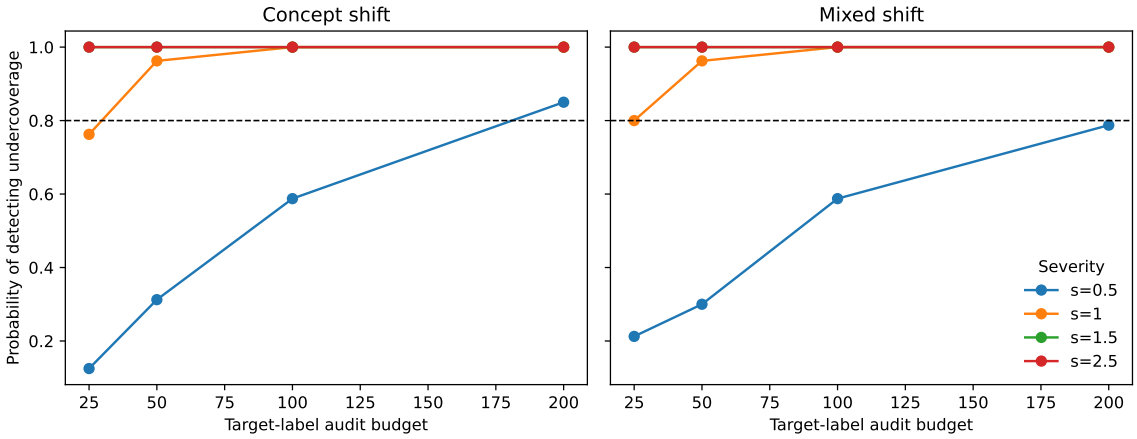


Fig. 8 Power of the exact one-sided target-coverage audit. The horizontal line marks 80% power.

Deployment acquisition blueprint.

Before auditing, an organization should record six auditable fields: (1) the deployment population and sampling frame; (2) the undercoverage alternative and power used to choose the label budget; (3) the inclusion mechanism and probabilities; (4) the separation of audit, recalibration, and evaluation outcomes; (5) the applicable guarantee class from Table 1; and (6) coverage, width, abstention, response rate, and selection-sensitivity results. This checklist turns label acquisition into a prospective monitoring design rather than an opportunistic retrospective analysis and is intended as a reusable deployment blueprint for clinical, industrial, and public-sector ML systems.

4 Discussion

The results clarify a distinction that is often blurred in applied uncertainty quantification. WCP can address changes in P_X when density ratios are estimable and $P_{Y|X}$ is stable. A target-domain classifier and ESS can reveal whether those observable requirements are strained. They cannot test the conditional-invariance assumption without target outcomes.

This boundary changes how deployment reports should be written. A low diagnostic score is actionable evidence of weak overlap, unstable weights, or extrapolation. A high score means only that target covariates resemble the source and weighting is numerically supported; it is not evidence that the outcome mechanism is unchanged. DriftGuard-L makes target-label acquisition operational: it attaches exact error control to an audit and shows the budget required for useful power. Recalibration does not repair the underlying predictive model and can produce wide intervals under severe shift; it repairs marginal uncertainty coverage, a narrower but verifiable objective.

The exact-WCP results also change the interpretation of interval efficiency. An infinite interval

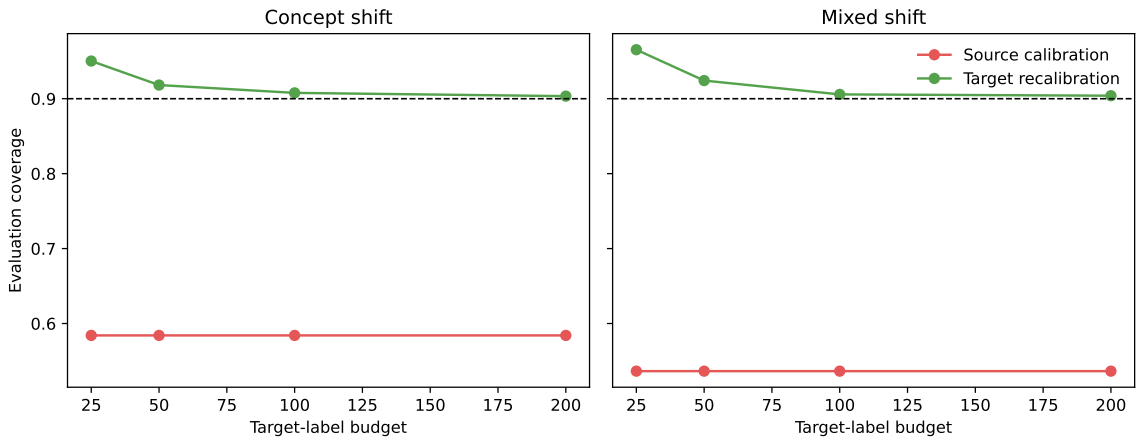


Fig. 9 Independent evaluation coverage before and after target recalibration at shift severity 1.5.

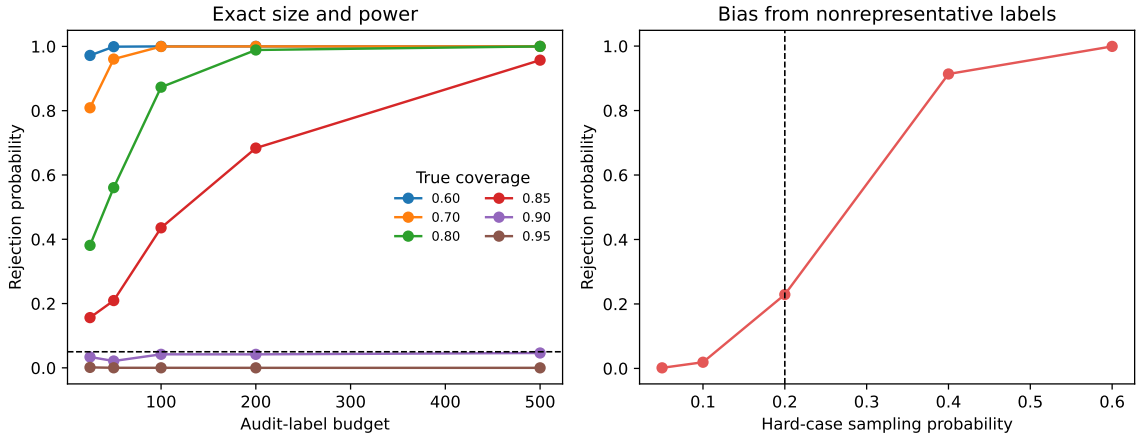


Fig. 10 Exact audit operating characteristics and the effect of nonrepresentative target-label sampling.

is not a numerical accident; it is the valid procedure's explicit declaration that the calibration sample provides insufficient weighted support for that target point. Weight clipping produces more convenient finite intervals, but the severe-shift experiment shows that this convenience can be purchased with undercoverage. Deployment reports should include the infinite-set or abstention rate alongside coverage and finite-set width.

The natural transfers reinforce this point. Rolling temporal transfer remained broadly conservative, but estimated weighting slightly reduced coverage while narrowing intervals; demographic transfer produced complete lack of support even though ordinary intervals appeared numerically usable. DriftGuard is therefore most valuable as a reporting protocol: it forces coverage, finite width, approximation status, and abstention to be interpreted jointly.

The doubly robust and Bike studies sharpen the positioning. DriftGuard is not proposed as a replacement for doubly robust calibration: the oracle-nuisance experiment confirms that method's intended asymptotic advantage when either nuisance is consistently estimated. DriftGuard-L instead provides finite-sample audit and repair after representative labels cross the information boundary. The external deployment shows why both layers are necessary: a sophisticated comparator can fail alongside weighting when the outcome mechanism evolves, whereas delayed labels reveal the failure and enable repeatable prequential repair.

The Capital Bikeshare result is therefore the empirical counterpart of the identification theory, not merely an application example.

The study has limitations. The financial and transportation studies are retrospective and dependent, not prospective multi-site trials; day- and window-cluster bootstraps do not create i.i.d. conformal validity. Public benchmark datasets are moderate in size. The doubly robust implementation follows a published estimating equation, but its fully estimated performance remains sensitive to finite samples, nuisance fitting, and overlap and is not unconditionally exact. No finite experiment spans every misspecification. Proposition 1 quantifies why estimated ratios remain approximate rather than granting them finite-sample exactness. The optional overlap-score threshold is an operational display choice; component diagnostics and the full threshold sweep are reported, and no validity claim rests on the score. The exact label audit assumes a probability sample. Known-probability designs permit design-based adjustment, but estimated or nonignorable selection sacrifices exactness and may require sensitivity analysis. The framework repairs marginal coverage, not conditional calibration or the underlying predictive model. A prospective multi-site validation remains the most important next step. These limitations constrain empirical generalization but do not alter the identification, testing, or synthetic exactness results.

5 Conclusion

Label-free conformal auditing has a useful but sharply bounded role. It can quantify overlap and identify weight degeneracy, but cannot certify coverage against an unobserved change in $P_{Y|X}$. DriftGuard-L crosses this boundary with a small probability sample of target labels, an exact undercoverage test, and valid target recalibration. The combined workflow reports observable warnings, unverifiable assumptions, the power purchased by a labeling budget, and the interval-width cost of restoring honest marginal coverage.

Statements and Declarations.

Funding. No external funding was received specifically for this study.

Competing interests. The author declares no financial or non-financial competing interests.

Author contributions. Maha Moussa conceived the study, developed the methodology, implemented the analyses, interpreted the results, and wrote and approved the manuscript.

Data availability. All data are public datasets distributed with scikit-learn, Matplotlib, or the UCI Machine Learning Repository, or are generated by the included scripts. Exact copies of the Bike Sharing and stock inputs are included in the reproducibility package with SHA-256 checksums.

Code availability. The accompanying reproducibility package contains fixed seeds, an exact Python dependency lock, a one-command workflow, raw replicate-level outputs, numerical checkpoints, summaries, figures, and per-script runtimes. A complete clean-order execution of all 13 stages passed the archived structural and numerical checks.

Ethics approval and consent to participate. Not applicable. The study used public, simulated, or non-identifiable data and involved no human participants or animals.

References

[1] Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)

[2] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. Journal of the American Statistical Association **113**(523), 1094–1111 (2018)

[3] Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. Foundations and Trends in Machine Learning **16**(4), 494–591 (2023)

[4] Quiñonero-Candela, J., et al.: Dataset shift in machine learning. MIT Press (2009)

[5] Tibshirani, R.J., Foygel Barber, R., Candès, E.J., Ramdas, A.: Conformal prediction under covariate shift. Advances in Neural Information Processing Systems **32** (2019)

[6] Gibbs, I., Candès, E.: Adaptive conformal inference under distribution shift. In: Advances in Neural Information Processing Systems, vol. 34, pp. 1660–1672 (2021)

[7] Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J.: Conformal prediction beyond exchangeability. The Annals of Statistics **51**(2), 816–845 (2023)

[8] Prinster, D., Liu, A., Saria, S.: Jaws: Auditing predictive uncertainty under covariate shift. In: Advances in Neural Information Processing Systems, vol. 35, pp. 35907–35920 (2022)

[9] Liu, Y., *et al.*: Multi-source conformal inference under distribution shift. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235 (2024)

[10] Zhou, Y., Fathony, R., Nguyen, N.H., Sesia, M.: Audited conformal prediction for classification under unknown distribution shift. arXiv preprint arXiv:2606.14909 (2026)

[11] Yang, Y., Kuchibhotla, A.K., Tchetgen Tchetgen, E.J.: Doubly robust calibration of prediction sets under covariate shift. Journal of the Royal Statistical Society Series B **86**(4), 943–965 (2024)

[12] Tsybakov, A.B.: Introduction to Nonparametric Estimation. Springer, New York (2009)

[13] Hunter, J.D.: Matplotlib: A 2d graphics environment. Computing in Science & Engineering **9**(3), 90–95 (2007)

[14] Fanaee-T, H., Gama, J.a.: Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence **2**, 113–127 (2014) https://doi.org/10.1007/s13748-013-0040-3

[15] Lohr, S.L.: Sampling: Design and Analysis, 3rd edn. Chapman and Hall/CRC, Boca Raton (2021)

[16] Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 3rd edn. Wiley, Hoboken (2019)