All Actuaries Summit 2026
25 – 27 May 2026, Melbourne

Fairness Testing for Insurance Pricing: A Statistical Inference Framework

Prepared by Fei Huang and Giles Hooker

Presented to the Actuaries Institute
2026 All-Actuaries Summit
25-27 May 2026

Actuaries Institute
ABN 69 000 423 656
Level 34, Australia Square, 264 George Street, Sydney NSW 2000
P +61 (0) 2 9239 6100 | actuaries.asn.au

Fairness Testing for Insurance Pricing: A Statistical Inference Framework

Fei Huang*[1] and Giles Hooker[2]

[1] *UNSW Sydney, School of Risk and Actuarial Studies*

[2] *University of Pennsylvania, Wharton School, Department of Statistics and Data Science*

Abstract

Ensuring fairness in insurance pricing has become a central concern for regulators, insurers, and the public. Existing approaches often rely on descriptive disparity measures or significance testing without explicit reference to fairness concepts or acceptable tolerances, leading to fragmented practices and ambiguous conclusions. We develop a unified framework that grounds fairness testing in classical statistical inference, providing clear definitions, reproducible methods, and auditable protocols.

Our framework advances fairness testing along four key dimensions. First, we formalise widely discussed fairness criteria as hypotheses on identifiable estimands, making fairness notions statistically precise. Second, we introduce decision rules that incorporate explicit tolerance thresholds, reflecting regulatory standards in practice. Third, we propose inference procedures that bridge actuarial practice with fairness guarantees. Fourth, we design a quote-audit protocol that specifies how to collect, test, and validate fairness claims in a manner that is transparent and replicable.

By combining inferential rigour with regulatory practicality, our approach delivers a coherent methodology that both regulators and insurers can implement. The framework integrates uncertainty quantification and audit design into a single pipeline, ensuring that fairness tests are interpretable, reproducible, and statistically robust. Beyond pricing, the principles extend to approval outcomes and process measures, providing a flexible foundation for responsible insurance analytics. The result is a regulator- and insurer-ready framework that aligns fairness testing with statistical guarantees and supports the broader goal of accountable and transparent insurance practices.

*Correspondence: Fei Huang, feihuang@unsw.edu.au.

1 Introduction

The premium a household pays for home, auto, or life insurance has direct consequences for financial security and social equity. Pricing algorithms that embed disparate outcomes, whether by design or as an unintended by-product of model complexity, can systematically disadvantage protected groups, eroding trust in insurance markets and creating legal and reputational risk for insurers.

In Australia, fairness in insurance pricing sits at the intersection of several regulatory frameworks. The *Insurance Contracts Act 1984* (Commonwealth of Australia, 1984) establishes the central legal framework for insurance contracts, while four federal anti-discrimination statutes, the *Racial Discrimination Act 1975*, *Sex Discrimination Act 1984*, *Age Discrimination Act 2004*, and *Disability Discrimination Act 1992*, constrain permissible rating factors. Each state and territory also maintains its own anti-discrimination legislation, such as the *Equal Opportunity Act 2010* (Victoria) and the *Anti-Discrimination Act 1977* (New South Wales), both of which cover the provision of insurance services and may impose obligations beyond those available under federal law, adding further complexity for insurers operating across jurisdictions.

The Australian Human Rights Commission and the Actuaries Institute have jointly published a Guidance Resource on AI and discrimination in insurance pricing and underwriting (Australian Human Rights Commission & Actuaries Institute, 2022), which provides practical guidance on complying with these statutes when AI is used in pricing and underwriting decisions and is the primary Australian interpretive reference for the framework developed here. APRA's prudential framework and ASIC's focus on consumer outcomes under the *Australian Securities and Investments Commission Act 2001* add further regulatory expectations. The Insurance Council of Australia's General Insurance and Life Insurance Codes of Practice reinforce the expectation that pricing systems are fair, transparent, and explainable to customers.

At the same time, the rapid uptake of machine learning and algorithmic pricing has outpaced the development of practical testing standards. Actuaries and data scientists are being asked to demonstrate fairness without agreed-upon methods for what constitutes a fair outcome, how precisely it must be measured, or how much uncertainty is acceptable. Similar regulatory developments are under way internationally. The United States has seen detailed fairness testing proposals from the Colorado Division of Insurance (Colorado Division of Insurance, 2023) and the New York Department of Financial Services (New York State Department of Financial Services, 2024), and the American Academy of Actuaries has issued methodological guidance (American Academy of Actuaries, 2023). The European Union's AI Act creates analogous obligations for high-risk AI systems, including insurance underwriting. These international frameworks provide useful benchmarks even where Australian regulation differs in specifics.

Existing regulatory proposals and actuarial guidance share three gaps. First, and most fundamentally, they prescribe testing steps without connecting those steps to a named fairness criterion. A practitioner instructed to compare group mean premiums is not told whether this operationalises demographic parity, conditional demographic parity, or something else entirely. This matters as each fairness criterion embodies a distinct ethical commitment about what equality in insurance means, and different criteria can point in opposite directions for the same dataset. Understanding the criterion is also what gives the audit its ethical justification. It situates the statistical exercise within a normative claim about fairness that can be debated, refined, and held to account. Second, existing guidance specifies *what* to compute but not how to move from a computed statistic to a defensible pass-or-fail conclusion. Measured disparities are reported without reference to sampling uncertainty, so a large dataset may flag trivially small gaps, while a small dataset may miss large ones. Third, none of the existing regulatory guidance specifies a complete audit process. Pre-audit design, tolerance margins, standard error estimation, multiplicity correction, and remediation are treated separately or not at all, leaving audits fragmented and conclusions that cannot be reproduced or independently verified.

Contributions. Our framework advances fairness testing along four dimensions. First, we formalise each fairness criterion as a testable hypothesis on an identified estimand, connecting the ethical concept to a precise statistical object and making fairness notions reproducible and auditable. Second, we treat fairness testing as a statistical inference problem, using the two one-sided tests (TOST) procedure to operationalise explicit tolerance margins that correspond directly to regulatory thresholds such as a 5% price gap or a 0.80 adverse impact ratio. Third, we provide inference procedures that bridge actuarial practice with fairness guarantees. Fourth, we design a complete quote-audit protocol that specifies how to collect, test, and validate fairness claims in a transparent and replicable manner.

Who this paper is for. The framework is designed to be directly usable by practising actuaries, data scientists, compliance teams, and regulatory examiners. The statistical methods draw on tools already familiar to actuarial practitioners, such as regression, confidence intervals, and hypothesis testing. The audit protocol provides a checklist-style workflow that a team can follow without needing to reconstruct the theoretical foundations. Readers who wish to engage with the underlying statistical theory will find it in Sections 3 and 4. Readers primarily interested in implementation can go directly to the audit protocol (Section 4) and the empirical illustration (Section 5).

Scope. We focus on pricing outcomes (P) and a single binary protected attribute (A), which covers the most common regulatory use cases in insurance. Although the paper is grounded in insurance contexts, the underlying statistical framework is general and applies to any setting where an algorithmic or model-based decision is audited for fairness across a protected group, including credit scoring, lending, and hiring decisions.

Section 2 reviews the regulatory and actuarial landscape and maps existing proposals to their underlying fairness criteria. Section 3 formalises each criterion as an inference problem and introduces the TOST framing. Section 4 presents the audit protocol. Section 5 applies the protocol to an international case study, and Section 6 concludes the paper with future directions.

2 Fairness Criteria and Existing Testing Proposals

This section interprets current regulatory and actuarial guidance through the lens of four fairness criteria. There is no formal regulatory framework for fairness testing in Australia. Internationally, the Colorado Division of Insurance (Colorado Division of Insurance, 2023) and New York Department of Financial Services (New York State Department of Financial Services, 2024) have published detailed fairness testing requirements, and the American Academy of Actuaries has issued methodological guidance (American Academy of Actuaries, 2023). These international proposals provide a concrete operational benchmark, and we draw on them throughout this section.

2.1 Disproportionate impact

Underlying criterion: a rating variable should not produce materially different predicted prices for policyholders in different protected classes, after controlling for other observable risk characteristics. The test assesses the variable-level impact on protected groups.

The American Academy of Actuaries (2023) proposes diagnosing disproportionate impact through a matched-comparison design, which involves constructing pairs of otherwise-identical policy profiles that a rating variable should not disproportionately affect predicted prices across protected classes, after controlling for other observable characteristics. The procedure involves four steps.

1. Create a matched dataset using nonparametric matching (Ho, Imai, King, & Stuart, 2007), pairing insureds who are similar in all characteristics except the protected attribute and the variable under examination. The `MatchIt` package in `R` (Ho, Imai, King, & Stuart, 2011) provides a practical implementation.
2. Fit two models on the matched dataset: one including the variable under examination, and one excluding it.
3. For each protected class, compute the average predicted outcome under each model.
4. If the inclusion of the variable leads to materially different predicted outcomes for a given protected class, the variable is deemed to exert disproportionate impact on that class.

The matched-comparison design approximates a causal effect of the variable under examination V on each protected class, but only under a selection-on-observables assumption. The matching must balance all characteristics that jointly determine group membership and predicted price. If relevant confounders are unobserved, or if matching variables are themselves affected by the protected attribute, the estimated impact may be biased. In particular, a variable may appear to have no disproportionate impact after matching while still producing disparate outcomes through unobserved pathways. The DI test is therefore best viewed as a variable-level diagnostic rather than a definitive causal assessment, and a pass verdict should be interpreted as absence of detectable disproportionate impact given the observed controls, not as proof of causal fairness.

2.2 Proxy discrimination (omitted-variable bias)

Underlying criterion: observed rating variables should not serve as statistical substitutes for a protected attribute. When a variable correlates with A and absorbs some of A's predictive power for P, removing A from the model does not eliminate its influence on prices. Instead, that influence passes through the correlated variable.

Proxy discrimination has a straightforward econometric structure. It can be viewed as a special case of omitted-variable bias in which the omitted variable is the protected characteristic A. Let P denote the pricing outcome, X_l a set of legitimate rating factors, and W a set of additional (potentially non-traditional) variables used in the model. When A is excluded from a regression of P on X_l and W, the estimated coefficients $\hat{\phi}$ on W will be biased to the extent that W is correlated with A and A has a causal effect on P. Re-introducing A into the regression and observing a shift in $\hat{\phi}$ therefore provides direct evidence that W is acting as a proxy for the protected attribute (Lindholm, Richman, Tsanakas, & Wüthrich, 2022; Pope & Sydnor, 2011).

Draft regulation by the Colorado Division of Insurance (2023, §8) (CO DOI) operationalises this insight, though without naming its connection to omitted-variable bias. The procedure compares two regression models:

1. Fit two models for the outcome of interest. The first regresses the outcome on (a) traditional underwriting factors, (b) non-traditional variables used in the pricing or approval decision, and (c) race/ethnicity indicator variables. The second regresses the outcome on sets (a) and (b) only, omitting (c). CO DOI recommend logistic regression for binary approval outcomes and linear regression for premium rates per $1,000 of face amount. Race/ethnicity is estimated via BIFSG (Bayesian Improved First Name Surname Geocoding, a method that infers the probability of an individual's race or ethnicity from their first name, surname, and residential geography) (Voicu, 2018) when not directly observed.

2. Examine the coefficients on the non-traditional variables (set b) across the two models. Under CO DOI's draft regulation, any variable whose coefficient shifts between the two models is taken as evidence that the variable may contribute to unfair discrimination.

The same logic was applied by the Federal Trade Commission (2007) to investigate whether credit-based insurance scores proxy for race/ethnicity and neighbourhood income in automobile claims-cost models, giving the CO DOI approach a precedent in federal regulatory practice.

A complementary approach, proposed by du Preez et al. (2024), builds a classifier that predicts A from the non-protected variables. Strong predictive accuracy implies that those variables can substitute for A, and thus that proxy discrimination is plausible. This approach is less model-dependent than the coefficient-shift method but yields a weaker conclusion: it establishes potential for proxy discrimination rather than measuring its extent in a specific pricing model.

The New York State Department of Financial Services (2024) (NY DFS) proposes a related concept under the label 'drivers of disparity'. NY DFS requires insurers to identify variables that "cause differences in outcomes for protected classes relative to control groups" (§17.vi) and to demonstrate that observed characteristics do not "serve as a proxy for any protected classes that may result in unfair or unlawful discrimination" (§11). We interpret this as aligning with the proxy discrimination framework described above. In particular, the requirement to assess whether variables act as proxies can be operationalised through a coefficient-shift analysis, in which changes in estimated effects after controlling for the protected attribute provide evidence of proxy behaviour. Under this interpretation, the CO DOI and NY DFS approaches are substantively aligned despite their different terminology.

2.3 Conditional Demographic Parity

Underlying criterion: after accounting for legitimate risk differences, members of different groups should face the same distribution of prices. CDP is the conditional analogue of demographic parity (DP or independence criterion). It permits price differences explained by X_l but not residual differences attributable to group membership alone (Xin & Huang, 2023).

Formally, CDP holds for a pricing outcome P if

$$\Pr(P = p \mid X_l = x_l, A = a) = \Pr(P = p \mid X_l = x_l, A = b)$$

for all p and all values x_l of the legitimate rating factors X_l.

In its strict form, CDP requires equality of the entire conditional distribution of prices across groups. In practice, and throughout this paper, we operationalise a mean-based relaxation: we

test whether the conditional expectation $\mathbb{E}[P = p \mid X_l = x_l, A = a]$ equals $\mathbb{E}[P = p \mid X_l = x_l, A = b]$, which is the quantity identified by the regression coefficient β. This is the formulation adopted in the major regulatory proposals.

A ratio relaxation, relaxed conditional demographic parity (RCDP) at tolerance level τ, permits a bounded disparity:

$$\tau \leq \frac{\Pr(P = p \mid X_l = x_l, A = a)}{\Pr(P = p \mid X_l = x_l, A = b)} \leq \frac{1}{\tau}$$

The τ threshold corresponds directly to the adverse impact ratio used in employment discrimination law, so RCDP is a natural operationalisation for regulators already working within that framework. When the conditioning on legitimate rating factors X_l is removed, the criterion reduces to demographic parity (DP), which requires equality in the marginal distribution of outcomes across groups, without adjusting for differences in underlying risk.

In its strict form, RCDP requires this ratio to lie within $(\tau, 1/\tau)$ for all p and all values x_l of the legitimate rating factors, which is a pointwise condition that is generally untestable from finite data. In practice, the regression imposes the assumption that the conditional mean gap is constant across all values of X_l. Under this assumption, the coefficient estimates a single gap that applies uniformly after controlling for X_l, rather than a pointwise condition at each x_l.

The CO DOI draft regulation (Colorado Division of Insurance, 2023, §§ 6–7) encodes a two-step test that we interpret as operationalising CDP and RCDP, respectively, though the regulation does not use this terminology.[1]

1. Regress the outcome variable on a set of race/ethnicity indicator variables and, optionally, a limited set of approved control variables. CO DOI recommend logistic regression for binary approval outcomes and linear regression for premium rates per $1,000 of face amount; unobserved race/ethnicity is estimated using BIFSG.

2. **First test (CDP):** assess whether the race/ethnicity indicators are jointly or individually significant at the 5% level. A model passes this test if none of the race/ethnicity coefficients is statistically significant.

3. **Second test (RCDP):** for any significant race/ethnicity indicator, assess whether its estimated effect exceeds the regulatory tolerance. CO DOI set this at 5 percentage points for approval rates and 5% of the mean premium for price outcomes. A model passes this test if all significant effects are below these thresholds.

When no control variables are included in the regression, the same two tests instead probe the unconditional criteria, demographic independence and the raw demographic impact ratio,

[1] The mapping from the CO DOI procedure to CDP and RCDP is our own interpretation, based on the algebraic structure of the proposed tests.

respectively. Including controls shifts the comparison to the conditional (CDP/RCDP) setting.

A fundamental limitation of the CO DOI procedure is that its pass/fail rule conflates statistical significance with practical importance in a way that is sensitive to sample size. In large datasets, even negligible disparities will be flagged as statistically significant, while in small datasets, substantively large disparities may go undetected. The TOST framework in Section 3.5 addresses this limitation by treating the tolerance margin as the primary inferential target.

Several additional metrics proposed by the NY DFS (New York State Department of Financial Services, 2024) are consistent with testing for independence (demographic parity) or CDP; these include the Adverse Impact Ratio, Denials Odds Ratio, Standardised Mean Differences, and z/t-tests (Table 5 in Appendix A).

Each metric can be computed against observed outcomes (corresponding to DP) or against regression residuals that control for X_l (corresponding to CDP). The NY DFS guidance does not resolve which mode is required, leaving the choice to the insurer. Our framework addresses this by tying the conditioning set to the pre-specified list of legitimate rating factors X_l (see Section 4.2, Step 3).

2.4 Sufficiency

Underlying criterion: the model's predictions should be calibrated consistently across groups. Sufficiency requires that knowing a policyholder's group membership, in addition to the model's predicted price, provides no additional information about the true underlying risk. When this condition fails, the model assigns the same predicted risk to different actual risks depending on group membership, a form of systematic miscalibration.

For quantitative outcomes, sufficiency requires

$$\Pr\left(P \leq p \mid \widehat{P}, A\right) = \Pr\left(P \leq p \mid \widehat{P}\right)$$

for all $p \in \mathbb{R}$: conditional on the prediction, the distribution of actual outcomes is the same for all groups. For binary classifiers, the criterion takes a more familiar form,

$$\Pr\left(P = 1 \mid \widehat{P} = 1, A = a\right) = \Pr\left(P = 1 \mid \widehat{P} = 1, A = b\right),$$

which is precisely parity of the positive predictive value (PPV, also called precision) across groups (Barocas, Hardt, & Narayanan, 2023, Ch. 3). An analogous condition on the negative predictive value (NPV) covers the $P = 0$ case.

In actuarial terms, a sufficiency violation means that policyholders with the same predicted risk have different expected loss rates across groups. That is, the mapping from predicted risk

to actual outcomes depends on group membership, which is inconsistent with sound ratemaking as well as with actuarial fairness.

Baumann and Loi (2023) propose a concrete hypothesis test for sufficiency when the outcome is continuous, by discretising $\widehat{P}$ into bins. Within each bin, the null hypothesis is that the conditional expected outcome is the same for both groups.

The American Council of Life Insurers (2024, ACLI), in a proposed revision to the CO DOI draft regulation, describe a procedure for checking a conditional relaxation of sufficiency. We interpret this as a test for conditional sufficiency, though the ACLI document does not use this terminology. The procedure involves four steps.

1. Obtain: data on some actual outcome of interest for a population of applicants or insureds; corresponding model output obtained from some algorithm or predictive model; data on a set of approved control variables; a race/ethnicity factor variable inferred from other characteristics. ACLI proposes using BIFSG to infer race/ethnicity from other characteristics, or any other approved method with demonstrably greater accuracy than BIFSG.
2. Create a 'test' regression or regression-like model relating the outcome of interest jointly to a) the corresponding 'model output' variable, b) approved control variables, and c) inferred race/ethnicity indicator variables.
3. Create a 'reference' model relating the outcome jointly to (a) and (b) only.
4. Conclude adherence to conditional sufficiency if and only if all three of the characteristics of the estimated models are true:
    - 4.1. The estimated coefficient on the model output variable is statistically significant in both the test and reference models. ACLI propose evaluating statistical significance against a 5% threshold.
    - 4.2. Confidence intervals for the coefficients on the model output variable in the test and reference models overlap materially.
    - 4.3. The estimated coefficients on the model output variable from the test and reference models are not statistically different to each other. ACLI propose using a z-test to compare coefficient values, if appropriate.

When no control variables are included in steps 2 and 3, the procedure reduces to a test of unconditional sufficiency. Including approved controls shifts it to the conditional relaxation, which is the more practically relevant setting when legitimate risk factors explain part of the outcome variation.

3 Fairness Testing via Statistical Inference

Section 2 showed that current regulatory proposals implicitly embed four distinct statistical criteria. This section gives each criterion a precise inferential formulation, specifying an estimand, a statistical model, and explicit hypotheses with a corresponding decision rule. The four criteria presented in this paper are intended as representative examples. Other fairness notions may be appropriate depending on the application context and stakeholder objectives (Xin & Huang, 2023).

Throughout this section, P denotes the response variable, the test target, which may be the quoted premium, pure premium, loss ratio, or another quantity as specified in the audit plan (Section 4.2, Step 4). A denotes the binary protected attribute with groups a and b (e.g., minority and majority racial/ethnic groups), and X_l denotes the vector of approved legitimate rating factors. Unless stated otherwise, we treat A as observed. Section 3.5 addresses the case where it must be inferred. We have used separate notation for the regression model used in each disparity criterion in order to emphasise that the coefficients in each model have meanings and interpretation that are contingent on the other terms in the model.

3.1 Disproportionate Impact

Estimand. Let V denote the variable under examination. Using a matched dataset that balances X_l across groups, define the disproportionate impact of V on protected class a as

$$\Delta_V^{\mathrm{DI}}(a) = \mathbb{E}[P^{(1)} \mid A = a] - \mathbb{E}[P^{(0)} \mid A = a],$$

where $P^{(1)}$ and $P^{(0)}$ denote predicted prices from models with and without V, respectively. The variable V is deemed to have disproportionate impact on group a if the within-group difference $\Delta_V^{\mathrm{DI}}(a)$ is statistically significant for that group.

Model. Construct a matched sample on X_l and estimate two models on that sample: one including V and one excluding it. Predictions from these models estimate $P^{(1)}$ and $P^{(0)}$. The test is applied separately within each protected class.

Hypotheses.

$$H_0 : \Delta_V^{\mathrm{DI}}(a) = 0 \quad \text{vs.} \quad H_A : \Delta_V^{\mathrm{DI}}(a) \neq 0,$$

Equivalence testing. For a pre-specified margin $\delta > 0$, we test

$$H_0 : |\Delta_V^{\mathrm{DI}}(a)| \geq \delta \quad \text{vs.} \quad H_A : |\Delta_V^{\mathrm{DI}}(a)| < \delta.$$

Decision rule. The variable V is deemed not to have disproportionate impact on group a if the confidence interval for $\Delta_V^{\mathrm{DI}}(a)$ lies entirely within $(-\delta, \delta)$. The test is applied for each

protected class and each variable under examination, with multiplicity correction applied across the resulting family of tests.

3.2 Conditional Demographic Parity

Estimand (level gap).

$$\Delta_\mu = \mathbb{E}[P \mid X_l, A = a] - \mathbb{E}[P \mid X_l, A = b].$$

Estimand (ratio).

$$R_\mu = \frac{\mathbb{E}[P \mid X_l, A = a]}{\mathbb{E}[P \mid X_l, A = b]}.$$

Model. We estimate the regression

$$P = \alpha + \beta \mathbf{1}\{A = a\} + \gamma^\top X_l + \varepsilon,$$

where β captures the conditional difference in expected price between groups.

Under the log-linear specification, the regression coefficient β has a direct population interpretation:

$$\beta = \log \frac{\mathbb{E}[P \mid X_l, A = a]}{\mathbb{E}[P \mid X_l, A = b]} = \log R_\mu,$$

so $\hat{\beta}$ is a consistent estimator of the log price ratio $\log R_\mu$, and $\exp\Big(\hat{\beta}\Big)$ estimates the ratio R_μ directly. The TOST is therefore applied to the confidence interval for $\hat{\beta}$, with the ratio tolerance band $(\log \tau,\ -\log \tau)$. The implied dollar gap at the portfolio mean premium $\bar{P}$ is

$$\widehat{\Delta}_\mu = \bar{P}\,(\exp\Big(\hat{\beta}\Big) - 1),$$

and this quantity is checked against the level tolerance $(-\delta, +\delta)$.

Hypotheses.

$$H_0 : \beta = 0 \quad \text{vs.} \quad H_A : \beta \neq 0.$$

Equivalence testing. To align inference with regulatory tolerance thresholds, we adopt an equivalence testing framework. For a pre-specified margin $\delta > 0$, we test

$$H_0 : |\Delta_\mu| \geq \delta \quad \text{vs.} \quad H_A : |\Delta_\mu| < \delta.$$

For ratio-based criteria, with tolerance $\tau \in (0, 1)$, we test

$$H_0 : |\log R_\mu| \geq \log(1/\tau) \quad \text{vs.} \quad H_A : |\log R_\mu| < \log(1/\tau).$$

Decision rule. The model is deemed to satisfy conditional demographic parity if the confidence interval for Δ_μ (or $\log R_\mu$) lies entirely within the corresponding tolerance band. Otherwise, the criterion is violated.

3.3 Sufficiency

In the sufficiency context, $\hat{P}$ denotes the model's predicted price or loss cost, and P denotes the corresponding actual outcome, such as the observed premium.

Estimand. Differences in conditional expectations across groups:

$$\Delta_{\mathrm{SU}}(r) = \mathbb{E}[P \mid \hat{P} = r, A = a] - \mathbb{E}[P \mid \hat{P} = r, A = b],$$

Model. For continuous outcomes, estimate the pair of models

$$P = \zeta + \kappa\, \hat{P} + \xi^\top X_l + \varepsilon,$$

$$P = \zeta' + \kappa'\, \hat{P} + \xi'^\top X_l + \lambda\, \mathbf{1}\{A\} + \varepsilon'.$$

Under the linear specification, $\hat{\lambda}$ is the natural estimator of the level gap in the estimand.

Hypotheses.

$$H_0 : \lambda = 0 \quad \text{vs.} \quad H_A : \lambda \neq 0.$$

Equivalence testing. For a pre-specified tolerance $\delta > 0$, we test

$$H_0 : |\lambda| \geq \delta \quad \text{vs.} \quad H_A : |\lambda| < \delta.$$

A company passes sufficiency if the confidence interval for $\hat{\lambda}$ lies entirely within $(-\delta, +\delta)$.

Decision rule. A company fails sufficiency if the confidence interval for $\hat{\lambda}$ lies outside $(-\delta, +\delta)$, indicating that the model charges more (or less) per unit of actual risk for one group than the other in a substantively meaningful way. When sufficiency fails, the slope stability test $H_0 : \kappa = \kappa'$ serves as a useful diagnostic. A significant shift in $\hat{\kappa}$ indicates that the predicted price carries different information about actual risk for the two groups.

3.4 Proxy Discrimination

Estimand. For each variable W_j in the set of candidate proxy variables W, the component-wise coefficient shift is

$$\Delta_{\mathrm{PD},j} = \phi_j - \phi'_j,$$

where ϕ_j and ϕ'_j denote the coefficient on W_j in the restricted and extended models, respectively.

Model. Consider the pair of regressions:

$$P = \alpha + \theta^\top X_l + \phi^\top W + \varepsilon,$$

$$P = \alpha' + \theta'^\top X_l + \phi'^\top W + \lambda \mathbf{1}\{A\} + \varepsilon'.$$

A shift in $\hat{\phi}_j$ when A is added reveals that W_j was absorbing part of the influence of A in the restricted model.

Hypotheses. For each W_j, we test whether the coefficient shift is statistically distinguishable from zero:

$$H_0 : \Delta_{\mathrm{PD},j} = 0 \quad \text{vs.} \quad H_A : \Delta_{\mathrm{PD},j} \neq 0.$$

Minimum effect size test. Unlike the TOST procedure applied to other criteria, where the goal is to confirm that a gap is small enough to fall within a pre-specified tolerance band, proxy discrimination is a detection problem. The question is not whether the shift is small enough to be negligible, but whether it is large enough to be meaningful. We therefore apply the mirror image of TOST. For a pre-specified minimum shift $\rho > 0$, we test

$$H_0 : |\Delta_{\mathrm{PD},j}| \leq \rho \quad \text{vs.} \quad H_A : |\Delta_{\mathrm{PD},j}| > \rho.$$

Under TOST, a model passes fairness if the confidence interval lies entirely *within* the tolerance band. Here, a variable is flagged as a proxy if the confidence interval lies entirely *outside* and above ρ, providing positive evidence that the shift is substantively meaningful rather than merely statistically detectable. As with TOST, this approach is robust to sample size. A large dataset will not flag a negligible shift simply because it is precisely estimated, and a small dataset will not clear the threshold unless the shift is genuinely large.

Decision rule. A variable W_j is flagged as a proxy for A if the confidence interval for $\hat{\Delta}_{\mathrm{PD},j}$ lies entirely above ρ in absolute value. The threshold ρ should be pre-specified in the audit plan as a minimum relative shift (e.g. $\rho = 0.05$, corresponding to a 5% relative change in the coefficient), and corresponds directly to the role played by the tolerance margins δ and τ in the TOST decision rules for CDP and Sufficiency.

3.5 Equivalence Testing and Tolerance Margins

The subsections above pair each criterion with a conventional two-sided significance test. That framing has a practical limitation. Under conventional significance testing, the null hypothesis is that the model is fair, $H_0 : \beta = 0$, and the audit must find evidence strong enough

to overturn that presumption. A pass verdict means only that the data failed to detect unfairness, not that the model is genuinely fair. This is an *innocent until proven guilty* framing. The model passes by default unless the data prove otherwise. In small datasets, this creates a serious problem, because an insurer with a genuinely unfair model but few observations will pass simply because the test lacks statistical power to detect the disparity. In large datasets the problem runs in the opposite direction. With enough observations, even a gap of a few dollars will be flagged as statistically significant, regardless of whether it exceeds any regulatory tolerance.

Equivalence testing reverses the burden of proof. The null hypothesis becomes unfairness, $H_0 : |\beta| \geq \delta$, and the model passes only if the data provide positive evidence that the gap is smaller than the pre-specified tolerance. This is a *guilty until proven innocent* framing. The insurer must demonstrate compliance, not merely avoid detection. Large samples now help rather than hurt, because more data means a more precise estimate and a tighter confidence interval that is more likely to fit inside the tolerance band if the model is truly fair. The consequences of this false negative risk depend on the audit's purpose. A false negative here means a genuinely fair model receives a fail verdict because the confidence interval is too wide to fit inside the tolerance band, not because the true disparity is large. When the audit is purely diagnostic, a fail verdict directs further investigation rather than triggering sanctions, reducing the cost of being wrong—an approach we recommend. When the audit carries direct regulatory consequences, an adequate sample size is essential. The pre-audit power calculation should be designed to ensure the confidence interval is narrow enough to confirm fairness when the true disparity lies within the tolerance band. Without sufficient data, the audit is structurally biased toward fail verdicts even for genuinely fair models.

The right question for a fairness audit is not "is there any disparity at all?", the answer is almost always yes for a large enough dataset. The right question is: "is the disparity large enough to matter?" This is precisely what equivalence (tolerance-based) testing asks. Instead of testing whether the gap is zero, we test whether the gap is small enough to fall within a pre-specified tolerance band. A model passes if, and only if, the data provide positive evidence that the disparity is below the threshold, not merely that we failed to detect it.

The equivalence testing approach is formalised by Schuirmann (1987) as the two one-sided tests (TOST) procedure. The regulatory tolerance margins already present in the literature, the CO DOI 5% threshold, the standard 0.80 adverse impact ratio, correspond directly to the margins δ and τ required to specify a TOST. The appropriate margin for a specific product line and protected attribute should be agreed between the insurer and regulator in the pre-audit setup (Section 4.2).

Level gaps. Declare fairness with respect to margin $\delta > 0$ if the confidence interval for Δ_μ lies entirely within $(-\delta, +\delta)$:

$$H_0 : |\Delta_\mu| \geq \delta \quad \text{vs.} \quad H_A : |\Delta_\mu| < \delta.$$

Ratios. For an adverse impact tolerance $\tau \in (0, 1)$ (e.g. $\tau = 0.8$), declare fairness if the confidence interval for $\log R_\mu$ lies within $(\log \tau, -\log \tau)$:

$$H_0 : |\log R_\mu| \geq \log(1/\tau) \quad \text{vs.} \quad H_A : |\log R_\mu| < \log(1/\tau).$$

Declare the model fair with respect to a criterion if the relevant confidence interval lies entirely within the tolerance band after multiplicity correction.

Multiplicity, power, and design. When multiple tests are conducted across different groups, segments, or risk tiers, there is an increased chance of a false rejection purely by chance. Standard multiple-testing adjustments should be applied: Holm–Bonferroni for strict family-wise error control, or Benjamini–Hochberg for false discovery rate control.

Power is defined as the probability of rejecting the null hypothesis when it is false. Under TOST, the null is unfairness ($H_0 : |\Delta_\mu| \geq \delta$), so the null is false when the model is genuinely fair. Power is therefore the probability that a fair model correctly receives a pass verdict. Low power means that a fair model may fail the audit simply because the sample is too small to estimate the disparity precisely enough, producing a confidence interval too wide to fit inside the tolerance band. The pre-audit sample size calculation should therefore be designed to achieve sufficient power to pass a truly fair model, not merely to detect a disparity.

For the CDP level-gap test fitted on the original scale of P, the minimum sample size to achieve power $1 - \beta$ is approximately

$$n \;\geq\; \frac{2\,\sigma_\varepsilon^2 (z_{1-\alpha} + z_{1-\beta})^2}{(\delta - |\Delta_\mu^*|)^2},$$

where σ_ε^2 is the residual variance of P after controlling for X_l, and Δ_μ^* is the assumed true disparity. Setting $\Delta_\mu^* = 0$ gives a conservative lower bound for a model that is exactly fair. When a log-linear model is used, as in the empirical illustration, the same formula applies with $\log(1/\tau)$ replacing δ and the residual variance of $\log P$ replacing σ_ε^2.

Misclassification of protected attributes. When the protected attribute A is not directly observed and must be inferred (for example, using surname- and geography-based methods such as Bayesian Improved First Name Surname Geocoding (BIFSG) or related approaches such as Bayesian Improved Surname Geocoding (BISG)), the resulting labels may be imper-

fect. This should be treated as a measurement error problem. Misclassification of the protected attribute can bias estimated disparities, potentially understating true disparities or generating spurious ones.

Such methods are widely used in regulatory settings. For example, US regulators including the New York State Department of Financial Services and the Colorado Division of Insurance permit or recommend the use of BIFSG-style approaches to infer race or ethnicity when direct data are unavailable, particularly in fairness testing and disparate impact analysis.

In practice, proxy-based race inference should be used with caution in fairness audits. Recent research (Xin, Hooker, & Huang, 2026) shows that such methods can systematically distort regression-based disparity estimates, potentially masking meaningful differences or generating spurious signals of unfairness depending on how proxy errors interact with model structure. As a result, proxy-based testing should not be treated as a neutral substitute for observed protected attributes. Where possible, insurers should complement proxy-based analyses with alternative approaches, such as improved proxy models, sensitivity analyses, or audit designs that avoid proxy inference altogether. Proxy-based testing should therefore be viewed as a diagnostic tool rather than a definitive measure of discrimination.

Practitioner summary: what TOST does and why it matters. Table 1 summarises the key difference between conventional significance testing and the TOST approach recommended in this framework.

Table 1: Conventional significance testing versus equivalence testing (TOST) for insurance fairness audits.

Conventional significance test	**TOST equivalence test**
Null hypothesis: gap $= 0$	Null hypothesis: gap $\geq \delta$
Pass = fail to detect any gap	Pass = positive evidence gap is small
Fails with large samples (trivial gaps flagged)	Robust to sample size
No link to regulatory tolerance	Tolerance δ set before testing
Ambiguous: "not significant" $\neq$ "fair"	Clear: pass means gap is within tolerance

4 Testing Framework for Prices under Mix-Based Audits

4.1 Overview of the audit protocol

Figure 1 summarises the audit workflow from design to decision. The role of this section is operational; that is given a selected fairness criterion, outcome variable, and tolerance bands, it specifies how to collect data, estimate the relevant quantities, and reach a defensible pass/fail

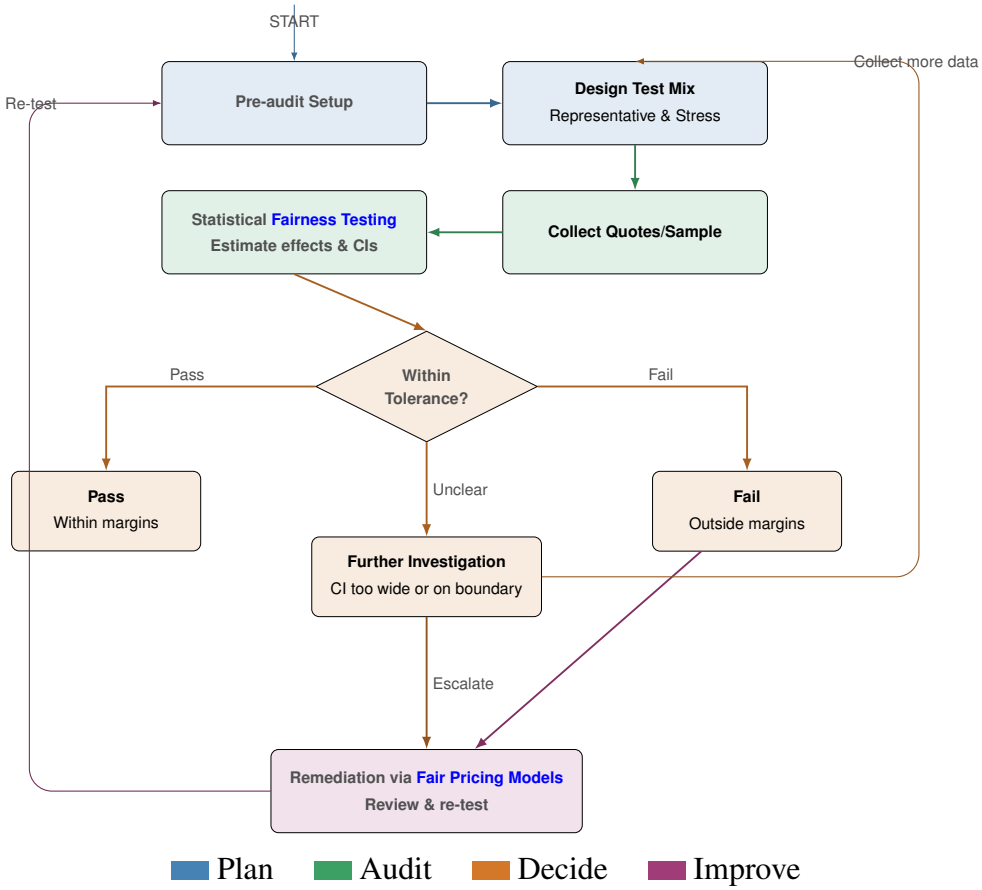

Figure 1: End-to-end flow for fairness auditing.

conclusion. Formal definitions of the criteria and their associated statistical tests are provided in Section 3.

4.2 Pre-audit setup

All design choices must be fixed prior to data collection. Each step should be recorded in an audit plan document retained for regulatory review.

1. **Select a fairness criterion.** Select a fairness criterion (e.g., Disproportionate Impact (DI), Proxy Discrimination (PD), Conditional Demographic Parity (CDP), or Sufficiency (SU)) based on regulatory requirements and the intended scope of the audit. Document the selection and the regulatory basis. There is no one-size-fits-all choice of fairness criterion. The appropriate selection depends on stakeholder preferences and the specific application context (Krafcheck, Balnozan, & Huang, 2026).

2. **Specify the protected attribute.** Define the protected attribute A (e.g., race/ethnicity, sex, age group) based on regulatory requirements. If A is not directly observed, document the inference method used and its expected accuracy in the audit population (see Section 3.5).

3. **Specify legitimate rating factors.** Define the set of permitted variables X_l that may

legitimately influence pricing. This set determines which differences are controlled for in conditional tests (CDP, SU) versus tested directly (DI).

4. **Define the response variable.** Select the outcome P to be audited. Options include quoted premium, pure premium, and loss ratio. The choice determines the scale of tolerance margins and the interpretation of results.

5. **Set tolerance margins.** Pre-specify the level-gap margin δ (e.g., a percentage of the mean outcome), the ratio margin τ (e.g., 0.80 for adverse impact), and the significance level α for confidence intervals.

6. **Specify standard error estimation.** The choice of standard error estimator should align with the audit design and the nature of the outcome. In pricing applications, model outputs are typically deterministic, so residual variation primarily reflects approximation error of the audit model rather than sampling variation in the outcome. This makes heteroskedasticity-consistent estimators (e.g., HC3 (MacKinnon & White, 1985)) a natural choice. For stochastic outcomes such as realised losses, robustness to heteroskedasticity remains important.

7. **Define test scope and multiplicity rule.** Specify whether the audit applies to a single population or multiple groups, segments, or outcomes. Pre-specify the multiplicity correction method: Holm–Bonferroni for family-wise error control, or Benjamini–Hochberg for false discovery rate control. No adjustment is required when the audit consists of a single pre-specified test.

8. **Design the quote mix.** Construct a representative mix reflecting the portfolio distribution and test requirements.

4.3 Data collection

1. **Submit requests.** Collect quotes within a short time window to minimise the risk that the pricing model is updated between submissions.

2. **Record data.** For each observation, record $(P,\ A,\ X_l,\ t,$ model version, id$)$.

3. **Log missing responses.** Record all failures or missing outputs. If missing rates differ across groups, report this as a potential process-level disparity.

4. **Ensure reproducibility.** Store the full dataset and a complete record of all audit requests, together with a hash (digital fingerprint) of the data, to ensure its integrity and enable independent verification.

4.4 Testing and decision

Apply the statistical test corresponding to the selected fairness criterion, as defined in Section 3.

The final decision rule is:

- **Pass**: the relevant confidence interval lies entirely within the tolerance margin after any required multiplicity correction.
- **Insufficient information**: the confidence interval is too wide to determine whether the true disparity lies within or outside the tolerance band, typically because the sample is too small, or the interval straddles a tolerance boundary in a way that requires further investigation. This outcome does not constitute a pass and should prompt either additional data collection or further investigation.
- **Fail**: any confidence interval lies outside the tolerance margin after correction, or multiple tests show a systematic pattern of deviations.

Prepare an audit report documenting the pre-audit plan, all test statistics and confidence intervals, the multiplicity correction applied (if any), and, where the result is a conditional pass or fail, the diagnostic findings and planned remediation.

4.5 Diagnostics and remediation

If the model fails or receives a conditional pass, further analysis is required to understand the source of the disparity and identify appropriate remediation.

1. **Diagnose sources of disparity.** Analyse potential drivers of the detected disparity, including the role of specific variables, model structure, and data composition. This may involve re-estimating the model across subpopulations, examining sensitivity to key inputs, or assessing whether disparities are concentrated in particular segments.
2. **Assess proxy and dependence effects.** Evaluate whether observed variables act as proxies for the protected attribute or whether residual dependencies remain after controlling for legitimate factors. This helps distinguish between direct, indirect, and spurious sources of disparity.
3. **Evaluate robustness.** Conduct sensitivity or robustness analyses to determine whether the results are stable under alternative model specifications, data subsets, or measurement assumptions.
4. **Remediation.** Based on the diagnostic findings, apply appropriate fair modelling approaches (Charpentier, 2024). These may include pre-processing methods (e.g., data transformation or reweighting), in-processing methods (e.g., incorporating fairness con-

straints during model training), or post-processing methods (e.g., adjusting outputs to satisfy fairness criteria). The choice of method should align with the identified source of disparity and the regulatory context.

5. **Documentation and follow-up.** Document all diagnostic analyses, remediation actions, and their rationale. Where necessary, implement ongoing monitoring and specify conditions under which the model should be re-audited.

5 Case Study

To demonstrate the framework in practice, we apply the full audit protocol to a publicly available dataset from the US auto insurance market. While the context is American, the fairness testing framework is broadly applicable to other contexts.

5.1 Data

The empirical analysis uses the ProPublica Illinois auto insurance dataset assembled by Larson, Angwin, Kirchner, and Mattu (2017) for their 2017 investigation into racial disparities in US auto insurance pricing. The dataset brings together three independently collected sources, which we describe in turn. The ProPublica analysis was contested by the insurance industry at the time of publication, with critics arguing that the aggregate loss cost variable is an imperfect control for each insurer's individual risk assessment (Lynch, 2017). We use this dataset as a publicly available worked example to demonstrate the audit framework, not to draw definitive conclusions about discriminatory intent or practice by any insurer.

Premium quotes (Quadrant Information Services). Premium data were obtained from Quadrant Information Services, a commercial provider of insurance rate data (Quadrant Information Services, 2017). Quadrant supplied approximately 30 million liability premium quotes by zip code from the leading insurance companies operating in each state (Larson, Kirchner, & Angwin, 2017). To hold driver characteristics constant across zip codes and companies, ProPublica restricted the analysis to a single standardised driver profile, a 30-year-old female schoolteacher with a bachelor's degree, excellent credit, no accidents or moving violations, purchasing standard coverage for the first time. The vehicle is a 2016 Toyota Camry. Annual mileage is approximately 13,000 miles and the coverage is $100,000 property damage and $100,000/$300,000 bodily injury. Quotes were collected in early 2017. For each zip code and company, the data record the annual bodily injury (BI) premium, the annual property damage (PD) premium, and their sum (`combined_premium`), which is the total annual liability premium. Six companies filed BI coverage only. For these, the combined premium equals the bodily injury premium.

Loss cost data (Illinois Department of Insurance). Aggregate risk data were obtained via public records requests to state insurance commissioners. Illinois provided zip-code-level data on paid losses (claims settled in the calendar year) for the period 2012–2014 (Illinois Department of Insurance, 2015), following a change in state data collection requirements in 2011 (Larson, Kirchner, & Angwin, 2017). The variable `state_risk` represents the average annual bodily injury plus property damage payout per insured vehicle across all Illinois insurers in a given zip code. This is essentially the aggregate pure premium for the zip code as a whole, the actuarial benchmark against which individual insurer quotes are compared. The state risk variable is insurer-agnostic, reflecting the average claims experience of all carriers writing in the zip code, not the experience of any individual company. This is a material limitation for the proxy discrimination and sufficiency tests, as discussed in Section 5.4.

Demographics (American Community Survey). Zip-code-level demographic data are drawn from the US Census Bureau's American Community Survey (ACS) 5-year estimates covering 2010–2014 (Larson, Kirchner, & Angwin, 2017; United States Census Bureau, 2015). The variable `white_non_hisp_pct` records the percentage of the zip code's population identifying as non-Hispanic white. ProPublica defined majority-minority zip codes as those with a non-white population share exceeding 50%, consistent with the lower threshold applied in Illinois and Missouri (where minority zip codes are less numerous than in California or Texas) to ensure adequate sample sizes.

The merged dataset contains 31,382 observations at the zip-code–company level, covering 923 Illinois zip codes each observed for each of 34 individual insurance companies (entities) belonging to 15 insurance groups. All 34 companies have complete premium data for all 923 zip codes. Table 2 provides summary statistics for the key variables.

Table 2: Summary statistics, Illinois auto insurance dataset ($n = 31{,}382$ zip-company observations). Premium and state risk are in dollars per year. Pct minority is the share of the zip population identifying as non-white (ACS 5-year estimates, 2010–2014).

Variable	Mean	SD	Min	Median	Max
Combined premium ($)	370.2	147.7	154	334	1,345
State risk ($)	162.7	50.9	52	167	298
Pct minority (%)	19.2	24.2	0	8	99
Chicago indicator	0.053	—	0	—	1
Minority flag ($\geq$50%)	0.113	—	0	—	1

Of the 923 zip codes, 104 (11.3%) are classified as majority-minority (non-white share $\geq 50\%$), and 49 (5.3%) are within the city of Chicago. The median minority share across Illinois zip codes is 8.1%, indicating that most zip codes are predominantly white. This reflects the state's highly segregated residential geography, in which minority populations are concentrated in a relatively small number of urban zip codes.

Table 3 presents unconditional comparisons of premiums and loss costs across majority-minority and majority-white zip codes.

Table 3: Unconditional premium and loss cost by zip-code minority status. Excess = combined premium − state risk. Ratio = minority mean / white mean.

Variable	White zips	Minority zips	Ratio
Mean combined premium ($)	355.9	482.2	1.355
Mean state risk ($)	160.6	179.2	1.116
Mean excess premium ($)	195.3	303.0	1.551
Zip codes (n)	819	104	—

Majority-minority zip codes face mean premiums 35.5% higher than majority-white zip codes, while their mean state risk is only 11.6% higher. The mean excess premium, the gap between quoted price and the actuarial benchmark, is 55.1% larger in minority zip codes than in white zip codes. These unconditional comparisons do not control for other rating factors and are presented here for descriptive purposes only. The conditional analysis in Section 5 uses regression to partial out the contribution of state risk and the Chicago indicator.

The dataset covers liability insurance only (the legally required coverage in Illinois), purchased by a single standardised driver profile. It does not include collision, comprehensive, or other optional coverages, which may exhibit their own disparity patterns. The 34 companies in the dataset represent the largest insurers by Illinois market share. Smaller regional carriers are excluded. The Quadrant quotes reflect the rates that would be charged to the standardised profile under each company's then-current rate filing, and may not reflect the rates actually charged to the population of Illinois policyholders, who vary in age, driving history, vehicle, and other underwriting characteristics.

The loss cost variable (`state_risk`) is aggregated across all insurers in the state, not company-specific. An insurer whose risk profile differs from the statewide average, for example, because it selects more or less risky policyholders than average, will have a company-specific risk experience that may diverge from `state_risk`. ProPublica (Larson, Kirchner, & Angwin, 2017) validated the use of the aggregate variable using a 2015 Nationwide rate filing in California that disclosed internal loss data, finding a statistically significant correlation between company-specific and statewide risk at the zip-code level (Pearson $r = 0.96$).

Finally, the unit of observation throughout is the zip code, not the individual policy. Zip-level premium comparisons are therefore subject to ecological inference bias (King, 1997). Within-zip heterogeneity in driver risk and demographics means that zip-level disparities provide an indication of, but do not directly measure, individual-level disparities. Individual policy records, which insurers hold but do not publicly release, would be required for a complete audit. The empirical analysis in Section 5 takes this constraint as given and explicitly flags it where the interpretation is affected. Table 6 in Appendix B defines all variables used in the regression models.

5.2 Pre-audit setup

Following Section 4.2, the following decisions are recorded before examining the data.

1. **Criteria.** We select Conditional Demographic Parity as the fairness criterion.
2. **Protected attribute.** A = minority flag: zip codes in which at least 50% of the population identify as non-white, following ProPublica's Illinois threshold. There are 104 majority-minority zip codes out of 923. Race/ethnicity is taken from ACS data directly, so BIFSG imputation is not required.
3. **Legitimate rating factors X_l.** Log state risk (the Illinois DOI aggregate loss cost per insured vehicle, the primary actuarial input) and the Chicago indicator (which captures urban territory relativities common in Illinois rate filings).
4. **Response variable.** Annual combined liability premium (bodily injury + property damage), log-transformed.
5. **Tolerance margins.** Following the CO DOI proposal: $\delta = 5\%$ of mean premium $= \$18.51$ and $\tau = 0.80$. Significance level $\alpha = 0.10$ (90% CIs).
6. **Test family and multiplicity.** The family comprises 34 company-level CDP tests. We apply Holm–Bonferroni correction.

5.3 Fairness Testing for Conditional Demographic Parity

For each company k we estimate

$$\log P_{iz} = \alpha_k + \beta_k \cdot \mathbf{1}\{A_z = 1\} + \gamma_k \log(\text{StateRisk}_z) + \delta_k \cdot \text{Chicago}_z + \varepsilon_{iz}, \tag{5.1}$$

where z indexes zip codes and $A_z = 1$ for majority-minority zips. Standard errors are HC3. The estimand of primary interest is β_k: the conditional log-premium gap attributable to group membership after controlling for loss cost and geography.

Table 4 reports results for all 34 companies, sorted by the implied price ratio. We declare fairness if the 90% CI for β_k lies entirely within $(\log 0.80, -\log 0.80) = (-0.223, +0.223)$ *and* the implied dollar gap lies within $(\pm\$18.51)$.

All 34 companies fail the CDP test. Price ratios range from 1.09 (Garrison Property & Casualty, a USAA subsidiary) to 1.43 (Metropolitan Property & Casualty), implying annual premiums in majority-minority zip codes that are \$34 to \$158 higher on average than in comparable-risk white zip codes. None of the 90% confidence intervals approach the $\log(0.80)$ lower boundary of the tolerance band; every failure is substantial rather than borderline. This matches and formally confirms the core finding of Larson, Angwin, et al. (2017). After controlling for

Table 4: T3 (CDP) results: companies at the top and bottom of the disparity distribution. Full results for all 34 companies are reported in Table 7 (Appendix B). β = conditional log-premium gap (HC3 SEs); Ratio = e^{β}; Gap = implied dollar difference at the mean premium ($370). TOST: $\delta = 5\%$ of mean premium, $\tau = 0.80$, $\alpha = 0.10$.

Company	Gap ($)	Ratio	90% CI for β	Dec.
Metropolitan Prop & Cas Ins Co	$158	1.427	$[+0.315, +0.397]$	FAIL
Allstate Ind Co	$138	1.374	$[+0.287, +0.349]$	FAIL
. . . 29 companies (all FAIL, ratios 1.10–1.37) . . .				
USAA Cas Ins Co	$ 35	1.095	$[+0.079, +0.102]$	FAIL
Garrison Prop & Cas Ins Co	$ 34	1.091	$[+0.074, +0.100]$	FAIL

risk differences, 33 of the 34 Illinois insurers they examined showed disparities exceeding 10%.

After Holm–Bonferroni correction across the family of 34 company-level tests, all 34 companies still reject $H_0 : \beta = 0$ at $\alpha = 0.10$. The correction makes no practical difference here because every gap is large enough that all 34 p-values are negligible even against the strictest adjusted threshold. The multiplicity adjustment is nonetheless required by the pre-specified protocol. In this dataset the standard significance test and the TOST criterion agree, because the disparities are large relative to the tolerance margins.

5.4 Limitations of the case study

Several data limitations bear on the interpretation of these results.

Aggregate-level data. The analysis operates at the zip code level. Both the minority flag and the loss cost variable are neighbourhood averages rather than individual or company-specific measures. A disparity detected at the zip code level indicates that minority neighbourhoods pay more than comparable-risk non-minority neighbourhoods on average, but does not establish that any individual policyholder was priced differently on the basis of their personal characteristics. The loss cost variable in particular represents the Illinois DOI average across all drivers and all insurers in each zip code, rather than each insurer's own risk experience for the standardised profile. Individual-level policy data and company-specific loss data would be required to make stronger claims about individual pricing and insurer-specific risk assessment.

Omitted rating variables. Critics of the original ProPublica analysis noted that the comparison between a standardised single-profile premium and a population-average loss cost may not accurately reflect each insurer's individual risk assessment for that profile. In this case study, some rating factors are effectively held fixed by the standardised driver profile, including gender, vehicle type, credit score, age, and driving history. Other approved rating variables, such as territory relativities and tier assignments, are not observed in the data. A disparity that survives these limited controls may be partially or fully explained by such unobserved approved

variables. A regulator receiving a fail verdict on this basis should request the full rate filing and company-specific loss data before drawing conclusions about the source of the gap.

Loss cost timing and aggregation. The state risk variable is an Illinois DOI aggregate from 2012–2014, while premiums are from 2017. Any company that updated its risk model over that interval will have residuals that partly reflect the timing mismatch rather than discriminatory pricing. The variable is also insurer-agnostic: it does not capture each company's own loss experience or modelling choices.

Quote-based premiums. The Quadrant data are standardised quotes for a single driver profile, not actual premiums paid by policyholders. They capture the rate filing but not the effect of individual discounts, surcharges, or coverage elections. A full audit would use actual policy-level data.

Single driver profile. All quotes use one standardised profile, a 30-year-old female schoolteacher with good credit. Disparities may differ for other risk profiles, and the results should not be generalised to the full portfolio without further testing across a representative mix of profiles.

Binary minority flag. The 50% non-white threshold treats all zip codes above the cutoff as equivalent. A continuous specification of minority share, or individual-level demographic inference via BIFSG, would allow more granular analysis at the cost of additional modelling assumptions.

6 Conclusion

Fairness testing in insurance pricing requires more than computing a disparity statistic. It requires knowing which fairness criterion the statistic operationalises, how to move from a point estimate to a defensible pass or fail decision, and how to assemble those steps into a reproducible audit process. Existing regulatory guidance does not fully address these three requirements in a unified or comprehensive manner. This paper fills the gap by connecting four fairness criteria to formal statistical estimands, adopting equivalence testing as the decision framework, and specifying a complete audit protocol from pre-audit design through to remediation.

The empirical application to 34 Illinois auto insurers illustrates what the framework delivers in practice. All 34 companies fail the conditional demographic parity test. The majority-minority zip codes are charged \$34–\$158 more per year than comparable-risk white zip codes, with price ratios ranging from 1.09 to 1.43. Every failure is substantial rather than borderline.

Several directions remain open for future research. The tolerance margins δ and τ are taken as regulatory inputs. However, a welfare-theoretic basis for setting them would strengthen the normative foundations of the framework. The case study is limited to a single driver profile

and two control variables. Applying the protocol to actual policy-level data with the full set of approved rating factors would provide a more complete picture of how Australian insurers could implement the audit in practice. Extending the framework to multi-category protected attributes, panel data settings, and real-time monitoring of deployed pricing models are further directions for future work.

Acknowledgements

The authors gratefully acknowledge Igor Balnozan for excellent research assistance, and Chris Dolman and Xi Xin for helpful comments and suggestions.

References

American Academy of Actuaries. (2023). *Approaches to identify and/or mitigate bias in property and casualty insurance.* https://www.actuary.org/sites/default/files/2023-02/CPCdataBiasIB.2.23_0.pdf. (Accessed: 24 June 2024)

American Council of Life Insurers. (2024). *Concerning quantitative testing of external consumer data and information sources, algorithms, and predictive models used for life insurance underwriting for unfairly discriminatory outcomes.* https://drive.google.com/file/d/1SCk4w7vPFqdt9BXhCLvsnowTwHI0CZwU/view. (Accessed: 12 June 2024)

Australian Human Rights Commission, & Actuaries Institute. (2022, December). *Guidance resource: Artificial intelligence and discrimination in insurance pricing and underwriting.* Australian Human Rights Commission. Retrieved from https://humanrights.gov.au/our-work/technology-and-human-rights/publications/guidance-resource-ai-and-discrimination-insurance

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press. Retrieved from https://fairmlbook.org/

Baumann, J., & Loi, M. (2023). Fairness and risk: an ethical argument for a group fairness definition insurers can use. *Philosophy & Technology*, *36*(3), 45.

Charpentier, A. (2024). *Insurance, biases, discrimination and fairness* (1st ed. 2024. ed.). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-49783-4

Colorado Division of Insurance. (2023). *Concerning quantitative testing of external consumer data and information sources, algorithms, and predictive models used for life insurance underwriting for unfairly discriminatory outcomes.* https://drive.google.com/file/d/1BMFuRKbh39Q7YckPqrhrCRuWp29vJ44O/view. (Accessed: 12 June 2024)

Commonwealth of Australia. (1984). *Insurance contracts act 1984*. Federal Register of Legislation. Retrieved from https://www.legislation.gov.au/Details/C2023C00307 (As amended)

du Preez, V., Bennet, S., Byrne, M., Couloumy, A., Das, A., Dessain, J., ... others (2024). From bias to black boxes: understanding and managing the risks of ai–an actuarial perspective. *British Actuarial Journal*, *29*, e6.

Federal Trade Commission. (2007). *Credit-based insurance scores: Impacts on consumers of automobile insurance*. https://www.ftc.gov/sites/default/files/documents/reports/credit-based-insurance-scores-impacts-consumers-automobile-insurance-report-congress-federal-trade/p044804facta_report_credit-based_insurance_scores.pdf. (Accessed: 24 June 2024)

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199–236.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28.

Illinois Department of Insurance. (2015). *Private passenger automobile statistical data, 2012–2014*. Public records request response, provided to ProPublica. (Aggregate loss cost data by ZIP code; described in Larson, Kirchner, and Angwin (2017))

King, G. (1997). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton, NJ: Princeton University Press.

Krafcheck, E., Balnozan, I., & Huang, F. (2026, February). *Fairness metrics for life insurance* (Tech. Rep.). Society of Actuaries Research Institute. Retrieved from https://www.soa.org/resources/research-reports/2026/fairness-metrics-life-insurance/ (Research Report)

Larson, J., Angwin, J., Kirchner, L., & Mattu, S. (2017). *Minority neighborhoods pay higher car insurance premiums than white areas with the same risk*. ProPublica and Consumer Reports. Retrieved from https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-white-areas-same-risk (Accessed: April 2026)

Larson, J., Kirchner, L., & Angwin, J. (2017). *How we examined racial discrimination in auto insurance prices*. ProPublica. Retrieved from https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-methodology (Accessed: April 2026)

Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2022). Discrimination-free insurance pricing. *ASTIN Bulletin: The Journal of the IAA*, *52*(1), 55–89.

Lynch, J. (2017, April). *Why ProPublica's auto insurance report is inaccurate, unfair and irresponsible*. Insurance Information Institute. Retrieved from https://

www.iii.org/article/why-propublicas-auto-insurance-report-is-inaccurate-unfair-and-irresponsible

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*(3), 305–325.

New York State Department of Financial Services. (2024). *Proposed insurance circular letter: January 17, 2024.* https://www.dfs.ny.gov/industry_guidance/circular_letters/cl2024_nn_proposed. (Accessed: 29 June 2024)

Pope, D. G., & Sydnor, J. R. (2011). Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy*, *3*(3), 206-231.

Quadrant Information Services. (2017). *Auto insurance premium quotes by ZIP code.* Commercial data, provided to ProPublica. (Data obtained by ProPublica; described in Larson, Kirchner, and Angwin (2017))

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680.

United States Census Bureau. (2015). *American community survey 5-year estimates, 2010–2014.* US Census Bureau. Retrieved from https://www.census.gov/programs-surveys/acs (Accessed via American FactFinder)

Voicu, I. (2018). Using first name information to improve race and ethnicity classification. *Statistics and Public Policy*, *5*(1), 1–13.

Xin, X., Hooker, G., & Huang, F. (2026). *How proxy race distorts regression-based fairness audits.* Retrieved from https://arxiv.org/abs/2603.17106

Xin, X., & Huang, F. (2023). Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. *North American Actuarial Journal*, 1–35.

A Supplementary Regulatory Material

Table 5: NY DFS proposed metrics for testing independence or CDP (New York State Department of Financial Services, 2024). Each metric can be computed against observed outcomes (independence criterion – demographic parity) or regression residuals controlling for X_l (CDP criterion).

Metric/Test	Description	Source
Adverse Impact Ratio	Ratio of favourable-outcome rates between protected and control groups.	New York State Department of Financial Services (2024, §17.i)

Table 5 continued.

Metric/Test	Description	Source
Denials Odds Ratio	Odds of adverse decisions for protected classes relative to control groups.	New York State Department of Financial Services (2024, §17.ii)
Standardised Mean Difference	Difference in mean outcomes between protected and control groups, standardised by pooled SD.	New York State Department of Financial Services (2024, §17.iv)
z- or t-test	Statistical test of whether outcome differences between groups are significant.	New York State Department of Financial Services (2024, §17.v)

B Full Empirical Results

This appendix reports variable descriptions of the case study and complete coefficient-level results for all 34 Illinois insurance companies for the Conditional Demographic Parity tests. Results are as described in Section 5; pre-audit parameters are $\delta = 5\%$ of mean premium, $\tau = 0.80$, $\alpha = 0.10$.

Table 7: Full T3 results: Conditional Demographic Parity, Illinois, all 34 companies. β = conditional log-premium gap (HC3 SEs); Ratio $= e^{\beta}$; Gap = implied dollar difference at mean premium ($370). TOST: $\delta = 5\%$ of mean, $\tau = 0.80$, $\alpha = 0.10$.

Company	Gap ($)	Ratio	90% CI for β	Dec.
Metropolitan Prop & Cas Ins Co	$ 158	1.427	$[+0.315, +0.397]$	FAIL
Allstate Ind Co	$ 138	1.374	$[+0.287, +0.349]$	FAIL
Owners Ins Co	$ 137	1.369	$[+0.282, +0.346]$	FAIL
Economy Preferred Ins Co	$ 134	1.361	$[+0.269, +0.346]$	FAIL
Metropolitan Cas Ins Co	$ 134	1.361	$[+0.271, +0.345]$	FAIL

Table 7 continued.

Company	**Gap ($)**	**Ratio**	**90% CI for β**	**Dec.**
Farmers Automobile Ins Assoc	$ 127	1.342	$[+0.243,\ +0.345]$	FAIL
Metropolitan Grp Prop & Cas Ins Co	$ 120	1.323	$[+0.249,\ +0.311]$	FAIL
Country Mut Ins Co	$ 110	1.297	$[+0.234,\ +0.287]$	FAIL
Country Pref Ins Co	$ 109	1.295	$[+0.233,\ +0.284]$	FAIL
Erie Ins Exch	$ 92	1.248	$[+0.200,\ +0.242]$	FAIL
State Farm Fire & Cas Co	$ 92	1.248	$[+0.205,\ +0.238]$	FAIL
State Farm Mut Auto Ins Co	$ 92	1.248	$[+0.205,\ +0.238]$	FAIL
Erie Ins Co	$ 91	1.247	$[+0.200,\ +0.242]$	FAIL
Allstate Fire & Cas Ins Co	$ 89	1.240	$[+0.191,\ +0.239]$	FAIL
Progressive Northern Ins Co	$ 88	1.238	$[+0.194,\ +0.232]$	FAIL
Travelers Home & Marine Ins Co	$ 86	1.233	$[+0.184,\ +0.235]$	FAIL
Travelers Commercial Ins Co	$ 86	1.232	$[+0.183,\ +0.234]$	FAIL
Liberty Mut Fire Ins Co	$ 74	1.199	$[+0.164,\ +0.200]$	FAIL
First Liberty Ins Corp	$ 74	1.199	$[+0.164,\ +0.199]$	FAIL
Geico Ind Co	$ 70	1.188	$[+0.154,\ +0.190]$	FAIL
Illinois Farmers Ins Co	$ 69	1.186	$[+0.149,\ +0.193]$	FAIL
Geico Gen Ins Co	$ 66	1.179	$[+0.147,\ +0.183]$	FAIL
Government Employees Ins Co	$ 66	1.179	$[+0.147,\ +0.183]$	FAIL
Geico Cas Co	$ 64	1.173	$[+0.148,\ +0.171]$	FAIL
Progressive Direct Ins Co	$ 63	1.170	$[+0.143,\ +0.171]$	FAIL
American Family Mut Ins Co	$ 60	1.163	$[+0.137,\ +0.164]$	FAIL
Progressive Universal Ins Co	$ 60	1.162	$[+0.136,\ +0.163]$	FAIL
Safeco Ins Co Of IL	$ 56	1.152	$[+0.128,\ +0.154]$	FAIL
American Standard Ins Co of WI	$ 54	1.147	$[+0.124,\ +0.150]$	FAIL
Trumbull Ins Co	$ 51	1.138	$[+0.119,\ +0.141]$	FAIL
USAA Gen Ind Co	$ 44	1.120	$[+0.102,\ +0.126]$	FAIL
United Serv Automobile Assn	$ 37	1.099	$[+0.082,\ +0.107]$	FAIL
USAA Cas Ins Co	$ 35	1.095	$[+0.079,\ +0.102]$	FAIL
Garrison Prop & Cas Ins Co	$ 34	1.091	$[+0.074,\ +0.100]$	FAIL

Table 6: Variable definitions for the case study.

Variable	Definition
`combined_premium`	Annual BI + PD liability premium ($) for the standardised driver profile in zip code z under company k.
`log_premium`	$\log($`combined_premium`$)$. Response variable in CDP.
`state_risk`	Average annual BI + PD payout per insured vehicle in zip code z, across all Illinois insurers, 2012–2014 (IL DOI).
`log_risk`	$\log($`state_risk`$)$. Primary legitimate rating factor X_l in the regression models.
`chicago`	Indicator equal to 1 if zip code z is within the city of Chicago. Captures urban territory loading.
`white_non_hisp_pct`	Percentage of zip code population identifying as non-Hispanic white (ACS 5-year estimates, 2010–2014).
`minority_flag`	Indicator equal to 1 if $100 -$ `white_non_hisp_pct` ≥ 50. Protected attribute A in the regression models.
`loss_ratio`	`state_risk`/`combined_premium`. Response variable in T4 (Sufficiency).