CAS RESEARCH PAPER
SERIES ON BIAS AND INSURANCE

A SCALABLE TOOLBOX FOR EXPOSING INDIRECT DISCRIMINATION IN INSURANCE RATES

Olivier Côté, MSc, ACAS; Marie-Pier Côté, PhD, FSA, ACIA; and Arthur Charpentier, PhD

CASUALTY ACTUARIAL SOCIETY

Commissioned by the Canada Race and Insurance Pricing Task Force

The CAS Research Paper Series on Bias and Insurance examines potential bias in insurance pricing and related insurance practices. Through this research, the Casualty Actuarial Society (CAS) aims to inspire and generate discussions about potential bias across insurance pricing and other insurance operations and equip actuaries to participate in conversations with regulators, insurers, and other stakeholders. Earlier papers in the series established foundational knowledge and introduced quantitative approaches for examining fairness in insurance pricing, including statistical methods used to evaluate potential bias. Originally launched in 2022 as the CAS Research Paper Series on Race and Insurance Pricing, the series has expanded to address bias and fairness more broadly across insurance operations and jurisdictions.

The Casualty Actuarial Society (CAS) is a leading international organization for credentialing, professional education and research. Founded in 1914, the CAS is the world's only actuarial organization focused exclusively on property-casualty risks and serves over 11,000 members worldwide. CAS members are sought after globally for their insights and ability to apply analytics to solve insurance and risk management problems. As the world's premier P&C actuarial research organization, the CAS reaches practicing actuaries across the globe with thought-leading concepts and solutions. The CAS has been conducting research since its inception. Today, the CAS provides thousands of open-source research papers, including its prestigious publication, *Variance* – all of which advance actuarial science and enhance the P&C insurance industry. Learn more at casact.org.

CAS RESEARCH PAPER
SERIES ON BIAS AND INSURANCE

A SCALABLE TOOLBOX FOR EXPOSING INDIRECT DISCRIMINATION IN INSURANCE RATES

Olivier Côté, M.Sc., ACAS; Marie-Pier Côté, Ph.D., FSA, ACIA; and Arthur Charpentier, Ph.D.

Casualty Actuarial Society
4350 North Fairfax Drive, Suite 250
Arlington, VA 22203
www.casact.org
(703) 276-3100

Contents

A Scalable Toolbox for Exposing Indirect Discrimination in Insurance Rates

Olivier Côté, M.Sc., ACAS;*[1] Marie-Pier Côté, Ph.D., FSA, ACIA;[1]
and Arthur Charpentier, Ph.D.[2]

Executive Summary

According to actuarial standards of practice, insurance pricing relies on grouping policyholders by risk to set adequate premiums. Modern predictive models, especially machine learning, excel at detecting statistical associations to differentiate risks, but they can learn spurious or undesired correlations. This raises concerns when socioeconomic or demographic factors may (intentionally or inadvertently) affect the fairness of insurance pricing.

Fairness in insurance is difficult to operationalize due to its ambiguity. Fairness metrics from the machine learning literature lack the segment-specific relevance actuaries require and are expressed in abstract units that obscure real-world consequences. For actuaries to intervene, proxy effects and unfair biases must be quantified in insurance-relevant terms: dollars and people.

In this paper, we focus on fairness in actuarial pricing. We study the situation wherein insurance rates should be fair with respect to a categorical (or discretized) sensitive variable, such as race or economic status, and the latter is fully observed (despite the possible privacy challenges). Our main contributions are listed below.

- We argue that actuarial fairness, solidarity, and causality form the three core dimensions of fairness in insurance pricing:
  - **Actuarial fairness** aligns premiums with expected losses, mitigating cross-subsidies.
  - **Solidarity** aligns premiums across protected groups, mitigating disparities.
  - **Causality** ensures models capture only risk effects for which proxy effects have been mitigated.
- We translate these dimensions into a five-point spectrum of premiums:
  - The **best-estimate** premium is the most accurate predictor of losses using all available information, including the sensitive variable.

[1]École d'actuariat, Université Laval.
[2]Département de Mathématiques, Université du Québec à Montréal.
*Corresponding author: Olivier Côté, olivier.cote.12@ulaval.ca.

  - The **unaware** premium is the most accurate predictor of losses using all information except the sensitive variable.
  - The **aware** premium is the most accurate predictor of losses when controlling for the sensitive variable.
  - The **corrective** premium is the most accurate predictor that enforces similar premium distributions across levels of the sensitive variable.
  - The **hyperaware** premium is the most accurate approximation of the corrective premium that does not directly discriminate on the sensitive variable.
- We define actuarially relevant local metrics that quantify the potential monetary impact of unfairness at the policyholder level. Proxy vulnerability is the difference between unaware and aware premiums. It locally measures how much the allowed variables pick up the signal of a missing sensitive variable.
- We define postpricing local metrics to evaluate the fairness of any pricing structure relative to the estimated spectrum.
- We integrate these components into a fairness assessment framework that partitions the policyholders, pinpoints segments most affected by unfairness, and evaluates local metrics to diagnose unfairness and guide intervention.
- We illustrate our approach with a large case study inspired by industry practice. The analysis relies on a real dataset of approximately 768,000 vehicles insured in Québec (2016–2017) that covers at-fault material damage claims. We examine the fairness of a pseudo commercial price with respect to discretized credit score: low (vulnerable group) versus high. Bender, Dill, et al. (2022) discuss how credit score is highly correlated with other demographic and economic factors such as race, gender, and income.
  - Proxy vulnerability is both material and skewed: while most policyholders may receive a modest rebate, a vulnerable minority of them could face 15%–30% overpricing if the regulation required that only the sensitive variable be omitted.
  - Our integrated framework (see Figure 14) illustrates that fairness in insurance pricing can be assessed efficiently, with minimal analyst effort. The framework provides simultaneous diagnostics from the three fairness dimensions, translates unfairness into dollar terms at the individual level, and highlights disparities across population segments.
- We provide additional information and the complete code illustrated on a comprehensive simulated data example in the online supplementary material.

Designed for routine portfolio monitoring, our toolbox delivers valuable insights whether or not the sensitive attribute is included in pricing, provided it is available for assessment. The toolbox's scalability, across large datasets and rich covariate sets, makes fairness operationalizable for actuaries – it is intuitive, practical, and encompassing the three fairness dimensions.

1. Introduction

Machine learning is now central to actuarial science due to its predictive power (Embrechts and Wüthrich 2022). It scales to large datasets, captures complex interactions, and excels at finding associations helpful to the predictive task (Frees et al. 2016). However, some associations are spurious; others reflect sensitive, though predictive, features. This raises ongoing debates about the legitimacy and fairness of relying on such associations in pricing.

Ratemaking datasets tend to subtly encode sensitive traits: Geographic location may hint at ethnicity, and occupation often reflects gender (Bender, Dill, et al. 2022). Combined with inequalities, e.g., racial wealth gaps or credit access disparities (Bender, Dillon, et al. 2022), nonsensitive variables associated with sensitive attributes can act as proxies, indirectly targeting protected subpopulations. Even without explicit use of protected features in ratemaking models, such proxy effects can perpetuate disparities.

The abundance of data amplifies the likelihood that combinations of input covariates inform about protected traits. This can be problematic when socioeconomic or demographic factors may (intentionally or inadvertently) affect insurance pricing.

Following Actuarial Standards of Practice No. 12 and No. 53 (Actuarial Standards Board 2005, 2017), insurance pricing requires grouping policyholders by risk to set adequate and financially sound premiums. Actuarial fairness ensures solvency; in contrast, other types of fairness may erode competitiveness, deterring their adoption without regulatory pressure.

Still, actuaries are expected to test for proxy effects in their models. A recent survey (Cavanaugh et al. 2024) indicates that US regulators broadly agree that “insurers should test to ensure that their models do not use data and information that act as proxies for disallowed rating variables.” The result is a tension: Actuaries must balance risk-based pricing with ill-defined fairness notions. Even the Actuarial Standards Board acknowledges in Actuarial Standard of Practice No. 12 that there is “no general agreement on what constitutes an ‘equitable’ classification system or ‘fair’ discrimination” (Actuarial Standards Board 2005).

Fairness in insurance lacks an operationalizable definition and meaningful metrics. Its meaning is ambiguous, and debates hinge on speculation about variable behavior in black-box models. Standard group fairness metrics offer little segment-level insight and are difficult to translate in dollars or affected policyholders. Fairness is discussed in theory, but unfairness unfolds in practice. Empirical studies remain scarce, and Fahrenwaldt et al. (2024) call for high-quality datasets to move the field forward.

To address these challenges, we develop a framework that maps fairness debates onto actuarially meaningful benchmarks. We focus on fairness in actuarial pricing, acknowledging its unique challenges. Our contributions are both conceptual and applied: Our methodology enables detection and quantification of unfairness using industry data. We illustrate on a large-scale industry dataset how potential unfairness manifests in practice.

The remainder of this paper is structured as follows. We first set the notation and scope in Section 2. In Section 3, we present the three dimensions of fairness in insurance pricing: actuarial fairness, solidarity, and causality. We then translate these dimensions as five ratemaking benchmarks in Section 4, covering the *spectrum* of fairness viewpoints.

From the estimated spectrum, we define in Section 5 actuarially relevant local metrics prior to pricing (risk spread, proxy vulnerability, fairness range, and parity cost) and postpricing (commercial loading, commercial burden, implied propensity and excess lift). In Section 6, we propose a method to detect systematic disparities by partitioning the portfolio to reveal vulnerable subpopulations. We then present two use cases: prepricing detection of proxy-vulnerable individuals (Section 6.1) and postpricing monitoring of commercial loadings (Section 6.2). We illustrate the tool's practicality with a case study using a large-scale Canadian auto insurance dataset and provide the code illustrated on a simulated data example in the online supplementary material.

2. Scope, Notation, and Setup

In this article, we focus on a one-period fairness goal that is independent of the notion of *intent*. Fairness is assessed from an output-based perspective. We assume the absence of unobserved confounders and selection bias, though these issues warrant discussion (see Côté et al. 2024; Côté et al. 2025).

The random variable Y represents the loss cost in a property and casualty insurance coverage. Variables available for pricing this coverage are denoted by the random vector $\mathbf{X}$, which we assume is measured without error.

Fairness is always relative to some prespecified prohibited (or sensitive) variable D, which here is taken to be a single, categorical, and fully observed random variable. Examples of sensitive attributes include gender in Europe, race in Texas, religion in California, and credit score in Ontario. Even though this might create privacy concerns, we assume that the insurer collects D, so this variable is fully observed in our dataset. We further refer to "protected groups" as the subpopulations formed by the different levels of D and to "vulnerable groups" as those historically disadvantaged among the protected groups.

Suppose we have a portfolio of n policyholders $(\mathbf{x}_i, d_i, e_i, Y_i)_{i=1,\ldots,n}$, where e_i is exposure to risk, measured in vehicle years. Let $\pi(\mathbf{x}, d)$ be the yearly commercial price for a policy with characteristics $\mathbf{x}$ and d, including all loads, adjustments, and profit margins.

We introduce below the setup of the real data case study used for illustrating our method.

Case study. We study fairness regarding a policyholder's credit score (high/low) in auto insurance premiums for at-fault material damage in the province of Québec, Canada.

The data, obtained through a partnership with an insurer, include over 768,000 vehicles insured from 2016 to 2017. The vector **X** comprises 16 explanatory variables,[1] including driver information, vehicle characteristics, and territorial information.

The response variable Y is the claim amount for at-fault accidents. It is highly zero inflated, with approximately 97% of observations not filing any claim. The annual average claim is around $\sum Y_i / \sum e_i \approx \$190$, where exposure $e_i \in (0, 1]$.

We take credit score as D given its link to demographic and socioeconomic factors (Bender, Dill, et al. 2022). Furthermore, Prince and Schwarcz (2019) argue in favor of considering credit score sensitive. They explain that insurers' reliance on credit data disproportionately affects low-income and ethnic minority policyholders, characteristics generally protected against discrimination. We construct a binary variable D, where $D = 1$ indicates high credit risk and represents about 40% of the sample.

We analyze a modified version of the insurer's pricing function. The modifications include the following:

1. Retrofitting the prices from available covariates (**X**, D),
2. Integrating observations unavailable at the time of developing the prices,
3. Rescaling to match average loss (preserve confidentiality of premium scale), and
4. Making additional adjustments to protect segmentation confidentiality.

We refer to this adjusted tariff as the *pseudoprice*, denoted PseudoPrice($\mathbf{x}$,d), which serves as the focal point for our fairness analysis throughout this paper.[2]

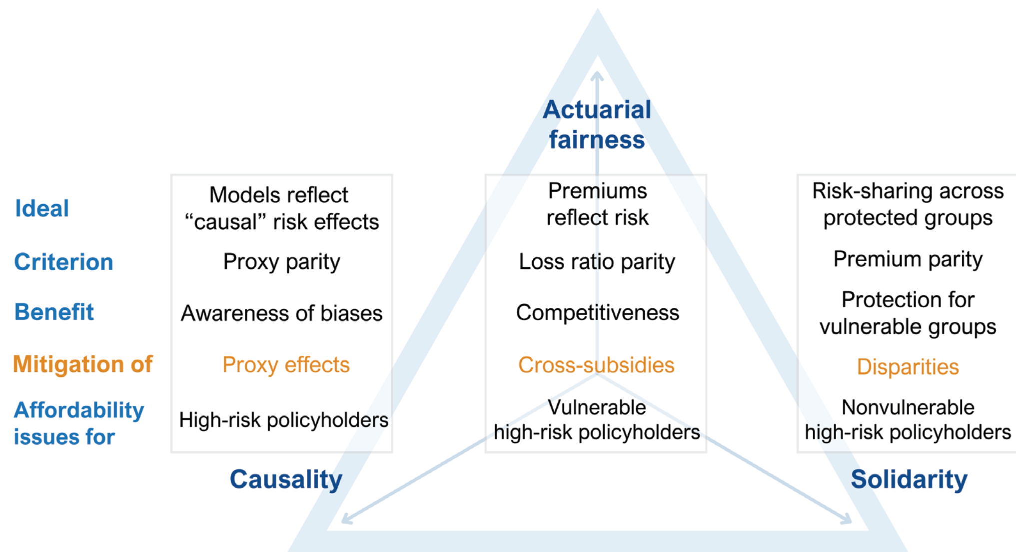
3. The Dimensions of Fairness in Actuarial Pricing

In our framework, three dimensions are needed to evaluate fairness relative to a sensitive D in actuarial pricing: *actuarial fairness* (alignment with expected loss), *solidarity* (redistribution across protected groups beyond pure risk), and *causality* (justifiable use of information). Together, the dimensions aim to cover all facets of fairness in actuarial pricing, each representing a distinct angle. Any fairness-related critique reduces to a breach of one dimension or to an explicit trade-off among them. The three dimensions of fairness are summarized in Figure 1 and presented in Sections 3.1–3.3.

[1] Strict anonymization and confidentiality measures were applied. A preselection, performed under senior actuarial oversight, reduced the size of the dataset from more than 30 candidate variables to just 16, balancing multicollinearity control with the inclusion of essential risk factors for the models.

[2] While the pseudoprice is constructed to maintain realism within the case study, its nonequivalence to the actual pricing function precludes any conclusion regarding the fairness of the partner insurer's pricing.

Figure 1. The three dimensions of fairness in actuarial pricing: ideal, corresponding criterion, and benefit. We clarify what each dimension aims at mitigating and which policyholder group might face affordability issues if this dimension were prioritized.



3.1. Actuarial Fairness

Actuarial fairness underpins viable insurance pricing. Originating in economic theory (Arrow 1963), it requires policyholders to contribute to the insurance pool in proportion to their own risk. Specifically, a premium is actuarially fair if it is "an unbiased estimate of the expected value of all future costs associated with the risk transfer" (Casualty Actuarial Society 1988).

A model is actuarially fair if it captures all risk differences and is locally balanced, keeping risk estimates unbiased and aligned with observed losses at all portfolio scales (Denuit et al. 2024). Each group, including each protected one, must have self-sustaining loss ratios. This avoids cross-subsidies and aims at a constant expected profit margin across policyholders. Actuarial fairness is about aligning premiums with expected losses.

Actuarial fairness reflects the predictive performance of the pure premium model and the lack of non-risk-based commercial adjustments. The criteria of loss ratio parity (see, e.g., Bender et al. 2025) and sufficiency (see, e.g., Mosley and Wenman 2022) align with this dimension.

3.2. Solidarity

Solidarity, in the form of risk sharing, is the foundation of insurance. In modeling with variables ($\mathbf{X}$, D), the solidarity dimension of fairness pertains to prohibited variables D, allowing risk differentiation based on $\mathbf{X}$ as long as premium distributions are similar across D. We intentionally create cross-subsidies if risk differs across groups of D, with the intent to promote societal welfare. Solidarity aligns with premium parity, as explained

Complement 1 – The Shrinking Homogeneous Pool

Initially, all policyholders were pooled in a collective fight against risk. By relying on data, insurers were able to create smaller, "homogeneous" pools, segmenting the original solidarity to better capture risk heterogeneity. With big data, the pools shrank further; Barry (2020) discusses a shift toward fairness rooted in individualized pricing. In an unrealistic extreme, oracle insurers – capable of perfectly predicting both the amount and timing of individual claims – might charge each policyholder precisely their discounted future claim amount, questioning the very concept of insurance risk transfer. Increasingly granular risk factors widen the separation between actuarial fairness and solidarity.

in Bender et al. (2025), and with demographic parity[3] of premiums: equal average premiums (weak parity) or identical premium distributions (strong parity) across protected groups. This is also referred to as the independence fairness criterion in Mosley and Wenman (2022).

3.3. Causality

Causal inference is the art and science of reflecting the causal impact of some variable on a given target, stripped of proxy effects. In a causal pricing model, each rating variable's influence corresponds to its causal effect on the claim risk Y and does not include any indirect link through sensitive attributes D.

For causal pricing, only allowed risk factors causally linked to Y belong in the model, and their premium influence must match their actual risk contribution. While controllable risk factors are valuable for prevention, noncontrollable ones, such as age, are admissible when their causal link to risk is well supported.

Proxy and causality are about effects, not specific variables. A variable's use in a model, not the variable itself, determines its role as a proxy. Talking about proxy effects rather than "proxies" emphasizes this distinction. Even valid risk factors may induce proxy effects. Causal inference tools can isolate the "nonproxy" component of a rating factor.

Common causal inference strategies include using control variables to mitigate bias during training, ensuring the model is uncontaminated by D. The causality dimension of fairness aligns with the proxy-free fairness in Charpentier (2024), the proxy parity fairness criterion of Côté et al. (2024), and the proxy discrimination metric of Lindholm et al. (2026).

[3] See definitions in Chapter 8 of Charpentier (2024, Definition 8.5, Definition 8.6, Proposition 8.1).

Complement 2 – Aligned Tools: Causal Thinking and Actuarial Judgment

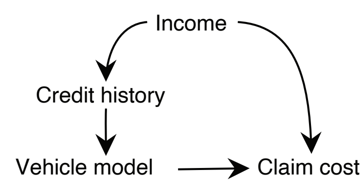
Statistical models capture associations – how claims vary with age or vehicle type. Causal models ask what happens when a variable changes. All causal models are statistical, but not all statistical models support causal claims.

A variable "causes" another if intervening on it shifts the distribution of what follows. Causation reflects a consequential distributional shift, not a deterministic outcome.

Causal modeling depends on assumptions – some testable, some not – often made explicit in causal graphs. These diagrams clarify which variables influence others (see, e.g., Moodie and Stephens 2022).

Figure 2 encodes one such structure: Vehicle model directly affects claim cost; credit history correlates with claim cost through income, an unobserved common cause. Estimating the effect of vehicle model requires controlling for income or credit history, the latter being feasible via credit score data.

Figure 2. Illustrative causal graph for insurance pricing.



Causal graphs reveal valid signals and potential biases. **Actuaries know that rating factor choice is never solely about fit.** Causal thinking formalizes that intuition. Many causal assumptions mirror those already made, explicitly or not, in practice.

Causality matters for fairness. Proxy effects hide in variables that appear overly predictive but reflect something sensitive. Discrimination lies in data, not models (Charpentier 2024). Causal reasoning helps separate valid from spurious signals, which is key to detecting unfairness along the causality dimension: **proxy effects**.

4. The Spectrum of Fairness

Côté et al. (2025) describe disparate impact as the association between premiums and sensitive attributes, lying between two extremes: **solidarity**, which aligns premiums across protected groups, and **actuarial fairness**, which aligns premiums with risk. Solidarity levels premiums; actuarial fairness levels profits. This tension is the core of fairness debates.

Nuances exist. Premium disparities relative to D are not uniformly problematic. They range from no association (**solidarity**), to justified association (**causality**), to association inflicted by proxy effects, and to direct exploitation of D for maximal predictive accuracy (**actuarial fairness**). The challenge is in pinpointing where the disparate impact falls along this spectrum and whether it crosses the blurry line between fair and unfair.

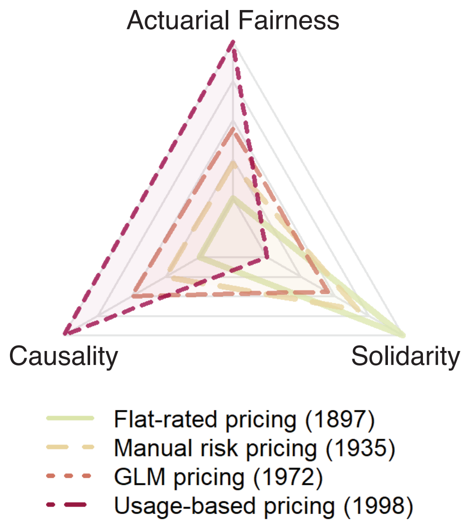
The five premium families of Côté et al. (2025) span this spectrum: best estimate, unaware, aware, hyperaware, and corrective. These families represent how fairness considerations regarding D influence a pricing structure and how permissible variables $\mathbf{X}$ are leveraged

Complement 3 – Revisiting the History of Risk Classification

Flat-rated pricing (first private auto in 1887; Insurance Information Institute 2023) maximized solidarity but ignored risk heterogeneity and causality. Manual pricing tables (Kormes 1935) added experience-based risk tables, improving actuarial fairness and likely causality. Multivariate analysis, such as minimum bias procedure (Bailey and Simon 1960), refined segmentation and risk factor use based on insurer-specific experience.

Usage-based insurance (first policy sold in 1998 by Progressive; Heller and Delong 2021) aligned premiums with behavior. Although they dilute the predictive power of protected attributes (Boucher and Pigeon 2024), behavioral data may still encode disparities, undermining solidarity as pictured in Figure 3.

Figure 3. The dimensions of fairness with increasing segmentation capacity.



Risk differences across protected groups may exist due to historical and socioeconomic factors. As data granularity increases, so does the potential for actuarial justification in perpetuating these disparities.

in relation to D. Each family offers distinct trade-offs between actuarial fairness, causality, and solidarity. We use this spectrum to investigate potential unfairness.

This section is structured as follows: we present the five benchmark premiums, one for each family, in Section 4.1, and we explain how we estimate them in Section 4.2. In Section 4.3, we reveal trade-offs between fairness dimensions in our case study.

4.1. The Five Families as Five Benchmarks

We present one model per family to obtain five benchmarks. We focus on the expectation, in line with the "**expected** value of all future costs" of Casualty Actuarial Society (1988).

First, the **best-estimate premium** $\mu^B(\mathbf{x}, d)$ aligns with **actuarial fairness**, grouping risks following $\mathbf{X}$ and D to set premiums. One example is $\mu^B(\mathbf{x}, d) = E(Y|\mathbf{X} = \mathbf{x}, D = d)$.

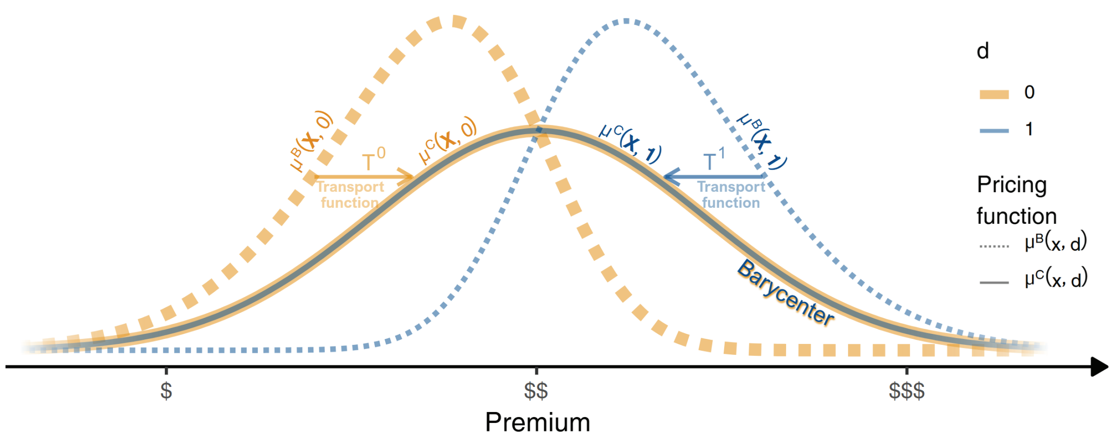
The **corrective premium** $\mu^C(\mathbf{x}, d)$ leverages both $\mathbf{X}$ and D to satisfy **solidarity**, that is, equal premium distribution per protected group. One example is $\mu^C(\mathbf{x}, d) = T^d\{\mu^B(\mathbf{x}, d)\}$, where T^d is a transport function detailed in Complement 4 and the online material.

The **unaware premium** $\mu^U(\mathbf{x})$ is the best nondirectly discriminatory approximation of the best-estimate premium, an example of unaware premium being $\mu^U(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$.

Complement 4 – From Best Estimate to Corrective: Optimal Transport

Optimal transport shifts one distribution onto another. The Wasserstein distance measures how far the $D = 1$ premium distribution is from that of $D = 0$. Solidarity holds when this distance is zero (Charpentier et al. 2023). Transport functions T^d push best-estimate premiums to a barycenter, yielding corrective premiums (see Figure 4).

Figure 4. Density of best-estimate and corrective premiums per *D*. Annotation illustrates how best-estimate premiums are transported to corrective premiums.



The **hyperaware premium** $\mu^H(\mathbf{x})$ is the best nondirectly discriminatory approximation of the corrective premium. One example is $\mu^H(\mathbf{x}) = E\{\mu^C(\mathbf{X}, D)|\mathbf{X} = \mathbf{x}\}$.

Finally, the **aware premium** $\mu^A(\mathbf{x})$ captures the effect of **X** on Y when controlling for D. An example is $\mu^A(\mathbf{x}) = E_D\{\mu^B(\mathbf{x}, D)\}$, a discrimination-free price of Lindholm et al. (2022).

The best-estimate premium sets the baseline with both **X** and D in a "paying for your own risk" approach. The unaware, aware, and hyperaware premiums omit D but handle proxies differently: The first reflects proxy effects, the second resists them, and the third uses them toward premium parity. The hyperaware and corrective explicitly pursue solidarity. The best estimate and unaware are solely risk focused. The aware permits parity shifts when causally justified by **X**. Together, these benchmarks reveal the trade-offs in fair pricing regarding D across fairness dimensions. We summarize them in Table 1.

4.2. Estimation of the Five Premiums

We give an example of procedure to estimate the spectrum of fairness. For additional details and examples, see the online supplement.

Table 1. Properties of the five fair premiums from Section 4.1.

Premium	Best estimate	Unaware	Aware	Hyperaware	Corrective
Notation	$\mu^B(\mathbf{x}, d)$	$\mu^U(\mathbf{x})$	$\mu^A(\mathbf{x})$	$\mu^H(\mathbf{x})$	$\mu^C(\mathbf{x}, d)$
Formula	$\mathrm{E}(Y \mid \mathbf{X}=\mathbf{x}, D=d)$	$\mathrm{E}(Y \mid \mathbf{X}=\mathbf{x})$	$\mathrm{E}_D\{\mu^B(\mathbf{x}, D)\}$	$\mathrm{E}\{\mu^C(\mathbf{x}, D) \mid \mathbf{X}=\mathbf{x}\}$	$T^d\{\mu^B(\mathbf{x}, d)\}$
Direct discrimination	✓	×	×	×	✓
Proxy discrimination	–	✓	×	✓	–
Demographic disparities	✓	✓	✓	×	×
Dimension prioritized	**Actuarial fairness**	**Actuarial fairness**	**Causality**	**Solidarity**	**Solidarity**

1. **Best estimate:** Fit a `lightgbm` (Ke et al. 2017) that includes ($\mathbf{X}$, D) to predict Y using an appropriate distribution (e.g., Tweedie):

$$\hat{\mu}^B(\mathbf{x}, d) = \hat{\mathrm{E}}\left(Y \mid \mathbf{X}=\mathbf{x}, D=d\right).$$

Alternatively, one can rely on a (directly discriminating) technical price.

Complement 5 – Technical or Data-Driven Best-Estimate Premium?

Insurers typically start pricing by estimating **indicated rates**, the actuary's best estimate of risk-based prices. As discussed in Section 4.1, the last four fairness families are derived from $\hat{\mu}^B$. The choice of anchor for $\hat{\mu}^B$ shapes the fairness assessment:

a) **Indicated rates as $\hat{\mu}^B$:** Indicated rates can be used as $\hat{\mu}^B$ if the sensitive variable D is included as a rating variable. This approach benefits from actuarial oversight but ties fairness benchmarking to actuarial choices. Consequently, biases within technical pricing may remain undetected.

b) **Data-driven $\hat{\mu}^B$:** To guard against institutional or analyst-induced bias, actuaries may estimate $\hat{\mu}^B$ directly from data using flexible algorithms like `lightgbm` (Ke et al. 2017), detached from technical or commercial pricing.

2. **Unaware:** Train a second `lightgbm` to approximate $\hat{\mu}^B(\mathbf{X}, D)$ using only $\mathbf{X}$:

$$\hat{\mu}^U(\mathbf{x}) = \hat{\mathrm{E}}\left\{\hat{\mu}^B(\mathbf{X}, D) \mid \mathbf{X}=\mathbf{x}\right\}.$$

3. **Aware:** Compute the empirical proportions of D in the training set. For each $\mathbf{x}$, average group-specific best-estimate premiums weighted by empirical frequencies

$$\hat{\mu}^A(\mathbf{x}) = \sum_d \hat{\mu}^B(\mathbf{x}, d)\,\widehat{\Pr}(D = d).$$

4. **Corrective:** Train the optimal transport function $\hat{T}^d$ of best-estimate premiums using `Equipy` (Fernandes Machado et al. 2025). The corrective premium is then

$$\hat{\mu}^C(\mathbf{x}, d) = \hat{T}^d\left\{\hat{\mu}^B(\mathbf{x}, d)\right\}.$$

5. **Hyperaware:** Train a last `lightGBM` model to regress $\hat{\mu}^C$ on $\mathbf{X}$ only

$$\hat{\mu}^H(\mathbf{x}) = \hat{E}\left\{\hat{\mu}^C(\mathbf{X}, D) | \mathbf{X} = \mathbf{x}\right\},$$

removing any direct discrimination on D while partly preserving parity corrections.

Complement 6 – Implementation Tips

- If D was discretized, store bin definitions for future use.
- Scale each of the five premiums by a constant to align with revenue targets.

All models are estimated on the same data, with the same features and overall target profit. Aside from natural estimation variability, differences reflect only the fairness goal.

4.3. Deviations Within the Spectrum

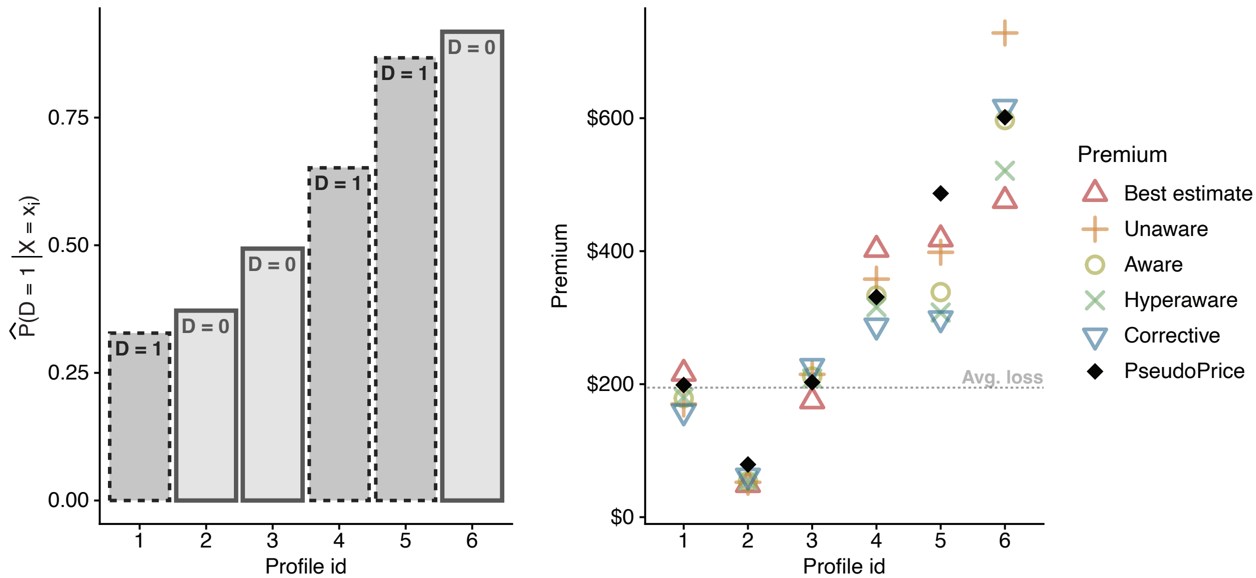
Deviations of a commercial price from a fairness benchmark suggest either misalignment with its intent or potential for predictive gain without fairness sacrifice. The meaning of the spectrum emerges only through the lens of the three fairness dimensions (Section 3), explored next.

Case study (Cont'd)**.** We estimate the spectrum of fairness following the methodology recommended in Section 4.2. We obtain our best-estimate premium $\hat{\mu}^B$ purely from data.

In the left panel of Figure 5, we display the `lightgbm` estimated propensity $\widehat{\Pr}(D = 1|\mathbf{X} = \mathbf{x})$, with dashed lines indicating observed high credit risk $D = 1$ for individuals 1, 4, and 5. In the right panel of Figure 5, we show the estimated spectrum and the pseudoprice for six individuals, partly described in Table 2. These graphs offer initial intuition on fairness:

- For individual 5, the pseudoprice lies outside the fairness spectrum. While methodology and commercial strategy (e.g., marketing or customer experience) may explain this, any deviation from the spectrum warrants close attention.

Figure 5. Propensity to observe $D = 1$ (left) and estimated spectrum of premiums along with the pseudoprice (right) for six individuals in the case study.



- For individuals with $D = 1$, the best-estimate premium (red) is the highest of the spectrum. The corrective premium $\hat{\mu}^C(\mathbf{x}, d)$ (blue) exceeds $\hat{\mu}^B(\mathbf{x}, d)$ only if $D = 0$.
- Fair premium ranges vary. For individual 2, all premiums closely align, suggesting fairness adjustments matter little for some policyholders. The intuition that higher risk appears linked to wider premium range will be exemplified in Figure 14.
- Individual $i = 5$ shows higher $\widehat{\Pr}(D = 1|\mathbf{X} = \mathbf{x}_i)$ than individual $i = 4$, pulling the unaware premium (orange cross) closer to the best estimate (red triangle), illustrating proxy discrimination. For outlier cases like $i = 6$, with $D = 0$ despite high propensity $\widehat{\Pr}(D = 1|\mathbf{X} = \mathbf{x}_6)$, the unaware premium is far from the best estimate.
- Higher propensity for $D = 1$ in individuals {4, 5, 6} versus {1, 2, 3} aligns with higher risk estimates, reinforcing the core motivation for fairness. Vulnerable individuals (high credit risk $D = 1$) tend to present allowed covariates $\mathbf{X}$ associated with higher claim propensity, driving up premiums via both protected and allowed variables.[4]

To formalize these observations, we assess each premium's alignment with the fairness dimensions introduced in Section 3. With a binary sensitive variable, Wasserstein distance measures distributional differences between the two protected groups ($D = 0$ and $D = 1$).

1. For **actuarial fairness,** we assess loss ratio parity by measuring the distance between the distributions of loss ratios across groups. Because every value of D yields a single loss ratio, we partition the data into 100 random subsamples to allow estimation of distributional differences. This sampling is used only for this dimension.

[4] In practice, biased covariates (e.g., uneven law enforcement of traffic violations) may further inflate premiums for vulnerable groups (Leong et al. 2024).

Table 2. Partial description of six profiles for the analysis in the case study.

id	GenderMainDriver	DrivExp (year)	DriverAge (year)	YearlyMileage (kilometers)	Location (from ZipCode)	OccType	hasPropertyIns	Credit score *D* (1 indicates low)
1	Female	19	35	10,000	Island of Montreal	Full-time	Yes	1
2	Male	25	42	10,000	Capitale-Nationale	None	Yes	0
3	Female	56	80	5,000	Laurentians	None	Yes	0
4	Male	8	24	20,000	Island of Montreal	Full-time	Yes	1
5	Male	0	42	15,000	Island of Montreal	Full-time	No	1
6	Male	3	19	15,000	Centre-du-Québec	Full-time	No	0

Table 3. Wasserstein distance between distributions for *D* = 0 and *D* = 1 of loss ratios (actuarial fairness), deviations from the aware family (causality), and premiums (solidarity) in the case study.

	Actuarial Fairness	Causality	Solidarity
Best estimate	0.039	67.054	116.938
Unaware	0.282	12.800	62.684
Aware	0.343	0.000	49.884
Hyperaware	0.392	10.071	49.840
Corrective	0.604	50.167	0.877
PseudoPrice	0.098	52.533	102.417

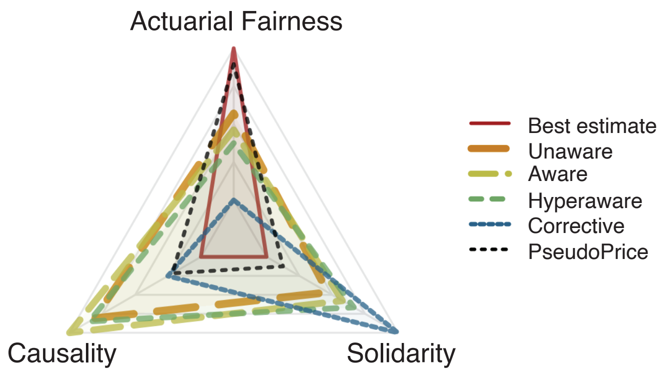
2. For **causality**, we assess proxy parity. Following Lindholm et al. (2026) and Côté et al. (2024), we treat deviations from the aware premium as proxy effects and compare their distributions across groups.
3. For **solidarity**, we compare premium distributions between protected groups.

These metrics are illustrative; they can be adapted or refined depending on the context.

Table 3 presents the Wasserstein distances for all premiums and the pseudoprice for at-fault material damage coverage. A Wasserstein distance of zero implies the corresponding fairness criterion is satisfied. This provides context for the pseudoprice's position within the fairness spectrum.

Figure 6 displays a radar plot that ranks premiums by their alignment with each fairness dimension. The closer a premium is to a triangle's vertex, the stronger is its adherence to that principle. As expected, the best-estimate, aware, and corrective premiums strongly adhere to actuarial fairness, causality, and solidarity, respectively. The three nondirectly discriminatory premiums (unaware, aware, hyperaware) cluster together, with the unaware leaning toward actuarial fairness and the hyperaware tilting toward solidarity, as expected.

Figure 6. Alignment of premiums with fairness dimensions in the case study.



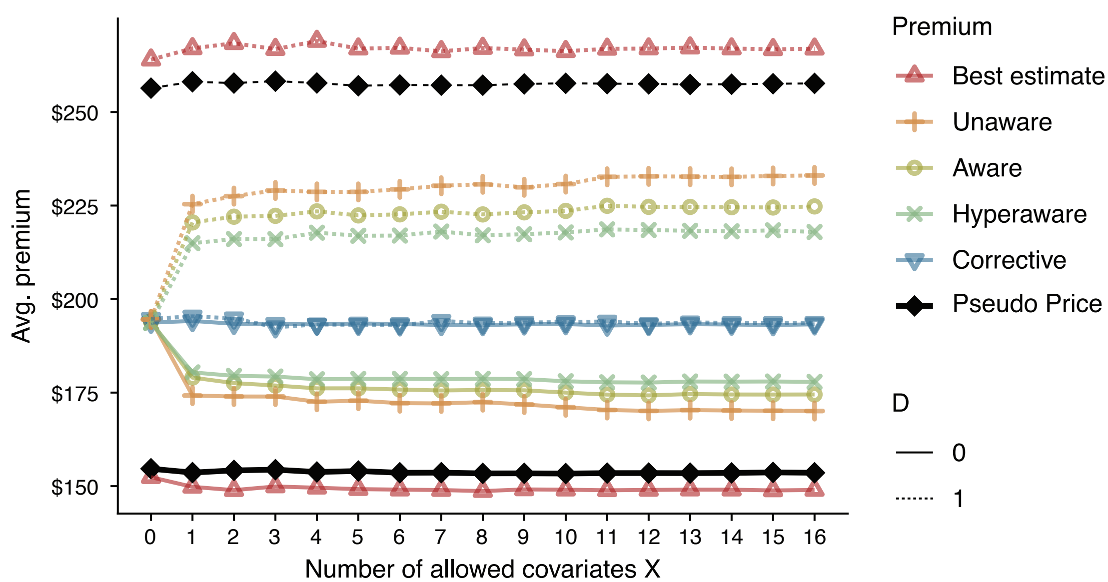
The pseudoprice aligns with a best-estimate strategy, which is unsurprising given the sensitive variable's inclusion in industrywide rates during the study period. When a variable is not deemed sensitive, insurers legitimately prioritize actuarial fairness.

We illustrate in Figure 7 how average premiums per protected group evolve when covariates are sequentially introduced. The covariates in **X** are added in order of variable importance in the best estimate `lightgbm`; the first is the main driver's experience in years `DrivExp`, which is strongly correlated with credit score. Both the best-estimate and corrective premiums directly discriminate on *D* but with opposing intents: The best estimate reflects all risk differences in *D*, whereas the corrective offsets group disparities through corrective direct discrimination. Even with abundant proxy effects, the unaware premium's ability to capture risk differences from the prohibited variable *D* is limited, as evidenced by the persistent gap between its average (orange) and the average of the best-estimate premium (red). The ordering of premium families aligns with intuition.

Figure 6 reminds us that fairness dimensions cannot be satisfied simultaneously. Classical fairness criteria (e.g., independence, proxy parity, and sufficiency), each tied to a distinct fairness dimension in Section 3, are fundamentally incompatible: extensive results, e.g., by Kleinberg et al. (2016), Charpentier (2024), and Lindholm et al. (2024) demonstrate that satisfying one typically violates another. In insurance setups, Bender et al. (2025) illustrate this impossibility with actuarial examples.

Universal fairness breaks on one truth: disparities *do exist* to begin with as seen, e.g., in Figure 7. No premium can be deemed universally fair; fairness is – and will remain – an elusive ideal, requiring ongoing governance of trade-offs across the three dimensions. All dimensions align only in the trivial case where the sensitive variable is entirely unrelated to the rest of the dataset (see, e.g., Côté et al. 2025).

Figure 7. Evolution of the mean premium for *D* = 0 (solid) or 1 (dashed) by family as a function of how many covariates are allowed in the vector X in the case study. The best estimate, corrective, and pseudoprice directly discriminate on *D*.



5. Actuarially Relevant Local Fairness Metrics

The case study suggests that deviations from the fair premium spectrum provide meaningful insights at the individual level. In this section, we first interpret key deviations within the spectrum (prepricing) and then examine deviations from a given tariff $\pi(\mathbf{x}, d)$ to the benchmarks (postpricing).

5.1. Prepricing Local Metrics

First, the **risk spread** measures the range of best estimates for different sensitive-attribute values. Building on proxy effects of Lindholm et al. (2026), we interpret the **proxy vulnerability** as the deviation between the unaware and aware benchmark premiums. Next, the **fairness range** reflects the overall range of the spectrum. Last, we define the **parity cost**, the overcharge experienced when going from loss ratio parity to premium parity.

5.1.1. Risk Spread

For a segment **x**, the **risk spread**, denoted $\Delta_{\text{risk}}(\mathbf{x})$, measures the range of data-driven risk estimates across different values of the sensitive attribute D:

$$\Delta_{\text{risk}}(\mathbf{x}) = \max\left\{\mu^B(\mathbf{x},1), \mu^B(\mathbf{x},0)\right\} - \min\left\{\mu^B(\mathbf{x},1), \mu^B(\mathbf{x},0)\right\} = \left|\mu^B(\mathbf{x},1) - \mu^B(\mathbf{x},0)\right|,$$

the within-segment premium gap between $D = 1$ and $D = 0$. It captures how much the best-estimate premium attributes risk differences to D within a given segment **x**.

The risk spread represents the model's incentive to capture risk differences driven by D for the segment **x**. The risk spread is positive; a larger value indicates a greater potential for disparate treatment across protected groups should pricing differentiate on D.

5.1.2. Proxy Vulnerability

For a segment **x**, **proxy vulnerability** quantifies the unintended price shift between a model that ignores D and one that controls for it. We define it as

$$\Delta_{\text{proxy}}(\mathbf{x}) = \mu^U(\mathbf{x}) - \mu^A(\mathbf{x}). \tag{1}$$

A large proxy vulnerability indicates that a segment **x** is prone to proxy effects, where seemingly neutral variables in **x** serve as proxies for D. This occurs when a significant risk spread exists and **X** informs on D. A positive value means the unawareness model overcharges the segment, while a negative value results from underpricing due to proxy. Studying proxy vulnerability highlights segments that are most exposed to potential proxy discrimination, providing a best guess regarding its monetary magnitude and direction.

Because the proxy phenomenon captures risk differences across groups of D indirectly, it is bounded by the distance between the aware premium and surrounding best estimates:

$$\min\left\{\mu^B(\mathbf{x},1),\mu^B(\mathbf{x},0)\right\}-\mu^A(\mathbf{x})\le\Delta_{\text{proxy}}(\mathbf{x})\le\max\left\{\mu^B(\mathbf{x},1),\mu^B(\mathbf{x},0)\right\}-\mu^A(\mathbf{x}).$$

Proxy vulnerability arises from the interplay between risk spread (potential direct discrimination on D) and propensity (ability to exploit it when using only $\mathbf{x}$).

Case study (Cont'd)**.** The top left panel of Figure 8 plots proxy vulnerability $\hat{\Delta}_{\text{proxy}}(\mathbf{x}_i)$ against the estimated propensity $\widehat{\Pr}(D=1|\mathbf{X}=\mathbf{x}_i)$ for individuals i in the test set. Black triangles relate to observed low credit score ($d_i=1$) and purple circles to high credit score ($d_i=0$). The upward spline trend indicates that proxy vulnerability increases with the likelihood of belonging to the $D=1$ group. Colors suggest that the model for $D|X$ (horizontal axis) is good, with more black triangles (true $d=1$) on the right.

The top right panel of Figure 8 shows box plots of the proxy vulnerability, expressed as a percentage of the aware premium. Proxy vulnerability is higher for the $D=1$ subpopulation and often exceeds 10%. Low-credit-score individuals ($d_i=1$) have increased claim risk, and the unaware premium μ^U captures this even when D is unobserved or excluded. This outcome is unavoidable when predictive covariates correlate with protected traits.

The middle row of Figure 8 plots the relationship between propensity and aware premium, using both original (left) and normalized rank (right) scales. Though the aware premium is constructed to be invariant to the propensity to observe $D=1$ (Section 4.1), a dependence persists: Individuals with higher propensity for $D=1$ tend to have higher aware premiums. In this case study, vulnerable individuals ($D=1$) have higher claim risk and are more likely to exhibit values of allowed covariates $\mathbf{x}$ linked to higher claim risk. The normalized rank plot confirms this: The upper tail is mainly populated by $D=1$ individuals.

The bottom line of Figure 8 depicts proxy vulnerability (color scale, low in purple, high in black) as a function of risk spread (x-axis) and propensity for $D=1$ (y-axis). Patterns reveal that proxy vulnerability arises when risk differs by protected group (large risk spread) and when $\mathbf{x}$ allows indirect inference of D (propensity near 0 or 1).

We depict in Figure 9 a map of Québec aggregated by forward-sortation area, our chosen geographic unit. Each area is colored by the Tail Value-at-Risk $TVaR_{0.95}$ of proxy vulnerability $\hat{\Delta}_{\text{proxy}}(\mathbf{x})$, computed as the average of the top 5% of values within that unit. This highlights regions where the most vulnerable individuals are concentrated, with inset maps detailing Québec City and Montréal. We see that some regions, such as Alma, Montréal-Nord, and St-Georges, exhibit high proxy vulnerability (darker purple). Representing proxy vulnerability on the map and supporting the analysis with census data may help to pinpoint sensitive demographics in these specific regions, e.g., the large proportion (42%) of immigrants in Montréal-Nord (Ville de Montréal 2018).

This is a compelling illustration of the materiality of proxy effects and why they warrant scrutiny. Even without explicit use of a prohibited attribute, its statistical imprint propagates through associated covariates, sustaining disparities in ways that evade direct detection. Proxy effects are material, and their potential impact is not evenly distributed. ▲

Figure 8. Dashboard of proxy vulnerability for the case study: proxy vulnerability (in CAN$) as a function of the propensity to observe *D* = 1 (top left), proxy vulnerability in percentage of aware premium per protected group (top right), propensity as a function of aware premiums on their original scale (middle left) and on the scale of their normalized ranks (middle right), and estimated propensity to observe *D* = 1 in terms of risk spread and colored by proxy vulnerability interval (bottom). vuln. = vulnerability.

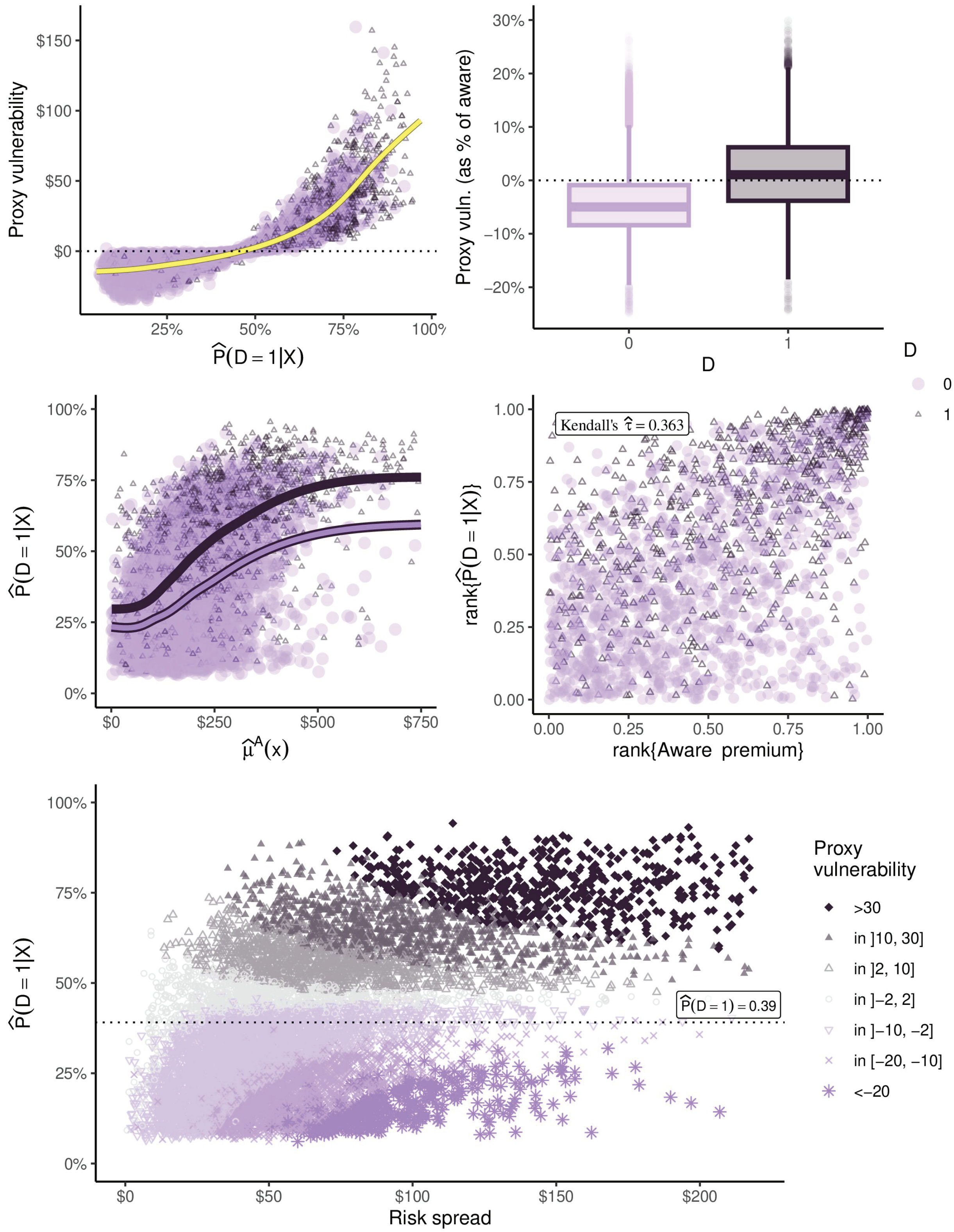
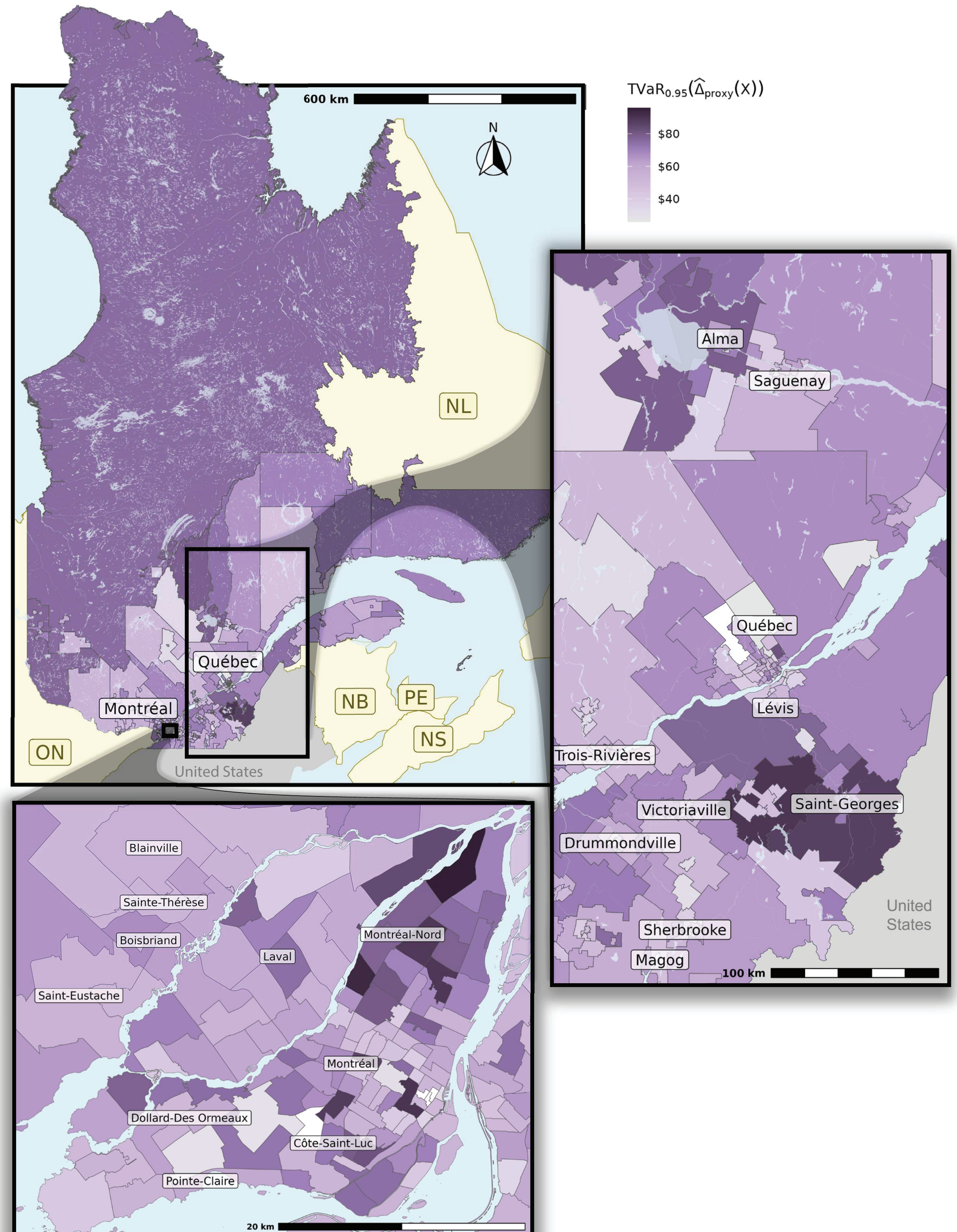


Figure 9. Geographic distribution of the empirical 95% TVaR of proxy vulnerability, assessed at the forward-sortation area level, in the case study. The top left map shows the entire province, and the inset maps present the central region and the island of Montréal.



5.1.3. Fairness Range

For a specific segment $(\mathbf{x}, d)$, the **fairness range**, denoted $\Delta_{\text{fair}}(\mathbf{x}, d)$, is defined as

$$\Delta_{\text{fair}}(\mathbf{x},d)=\max\left\{\mu^B(\mathbf{x},d),\mu^U(\mathbf{x}),\mu^A(\mathbf{x}),\mu^H(\mathbf{x}),\mu^C(\mathbf{x},d)\right\}$$
$$-\min\left\{\mu^B(\mathbf{x},d),\mu^U(\mathbf{x}),\mu^A(\mathbf{x}),\mu^H(\mathbf{x}),\mu^C(\mathbf{x},d)\right\}.$$

The fairness range measures how much premiums vary across fairness methods. A large value indicates that pricing is sensitive to fairness considerations for the segment.

5.1.4. Parity Cost

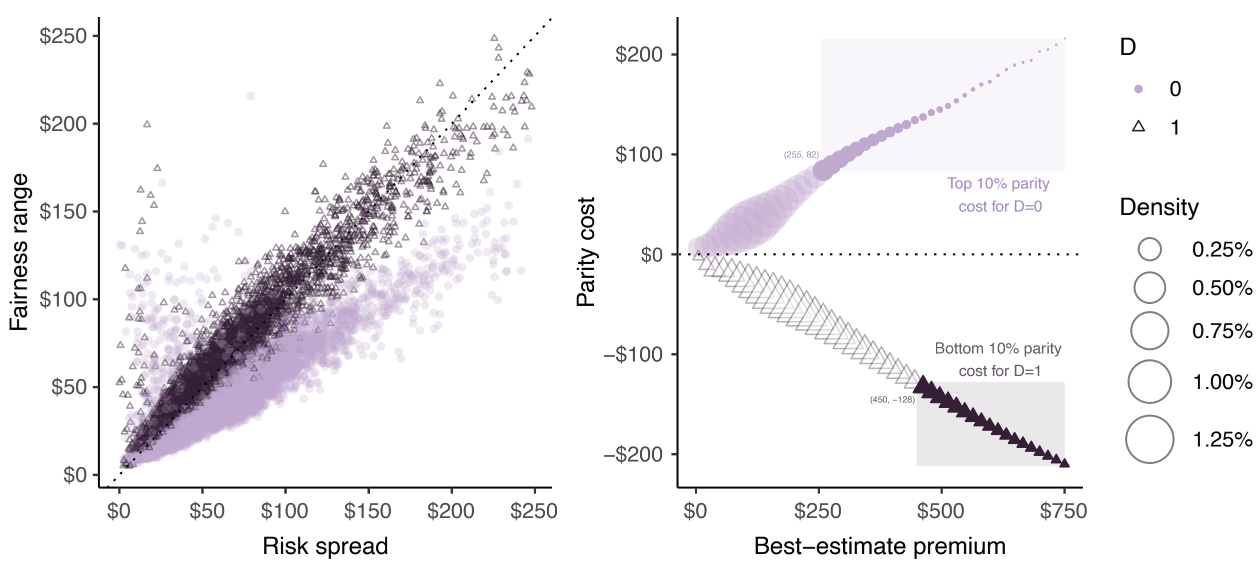
The **parity cost** is the (monetary) cost for a policyholder of enforcing premium parity compared to a "pay for your own risk" approach. For a segment $(\mathbf{x}, d)$, it is defined as

$$\Delta_{\text{parity}}(\mathbf{x},d)=\mu^C(\mathbf{x},d)-\mu^B(\mathbf{x},d).$$

A higher parity cost signals that larger adjustments are needed to enforce demographic parity. It quantifies how solidarity objectives conflict with actuarial fairness for an individual.

Case study (Cont'd). Figure 10 shows fairness range and parity cost. On the left, fairness range tracks risk spread, revealing that sensitivity to fairness adjustments follows potential for disparate treatment. Conditional on risk spread, the $D=1$ group shows greater sensitivity to fairness adjustments. On the right, parity cost reflects discounts for high credit risk ($D=1$)

Figure 10. Fairness range as a function of risk spread (left) and parity cost as a function of best-estimate premiums (right) for the case study. Colors denote protected group.



Complement 7 – Subsidizing Fairness Without Cross-Subsidies

A public scheme could fund fairness by reimbursing insurers the **parity cost**: the gap between corrective and best-estimate premiums. Vulnerable individuals ($D = 1$) pay the reduced corrective rate; others ($D = 0$) retain their actuarially fair price. This lowers prices for protected groups without burdening others, enabling fairness without cross-subsidies or insurer loss.

and surcharges for low credit risk ($D = 0$) individuals. Point size reflects sample density. The distribution for $D = 1$ centers on large discounts; for $D = 0$, on small surcharges. Thus, achieving premium parity involves imposing modest levies on the nonvulnerable group ($D = 0$) to fund substantial rebates for the vulnerable group ($D = 1$).

This captures the redistributive effect of the corrective premium relative to its risk-based counterpart. Critics of premium parity often highlight cross-subsidies as problematic; the parity cost explicitly quantifies them. ▲

5.2. Postpricing Local Metrics

Commercial pricing includes adjustments unrelated to risk or fairness. Bender et al. (2025) advise assessing final prices by their "actual impact on policyholders." We introduce local fairness metrics to evaluate a given commercial price $\pi(\mathbf{X}, D)$ relative to the spectrum.

5.2.1. Commercial Loading and Commercial Burden

For a segment $(\mathbf{x}, d)$, the **commercial loading** is defined as

$$\Delta_{\text{load}}(\mathbf{x}, d; \pi, \mu) = \pi(\mathbf{x}, d) - \mu(\mathbf{x}, d),$$

where μ denotes a reference premium intended to best represent indicated rates. If direct use of D is allowed, set $\mu = \mu^B$ to assess actuarial fairness, aligning, e.g., with the "deviation from indicated rates" metric of the Financial Services Regulatory Authority of Ontario (2024). If not, or to monitor proxy effects, use $\mu = \mu^A$ to examine causality.

The **commercial burden**, denoted $\rho_{\text{burden}}(\mathbf{x}, d; \pi, \mu)$, is the commercial loading as a percentage of μ. High burdens may raise affordability concerns for low-income policyholders:

$$\rho_{\text{burden}}(\mathbf{x}, d; \pi, \mu) = \frac{\pi(\mathbf{x}, d)}{\mu(\mathbf{x}, d)} - 1 = \frac{\Delta_{\text{load}}(\mathbf{x}, d; \pi, \mu)}{\mu(\mathbf{x}, d)}.$$

If the reference premium is the best-estimate μ^B, the commercial burden equals the markup over the claim costs, which is inversely proportional to the expected loss ratio.

5.2.2. Implied Propensity

The **implied propensity**, denoted $\tilde{P}_D(\mathbf{x}; \pi)$, is the implicit weight of $D = 1$ for segment $\mathbf{x}$ when expressing a (nondirectly discriminatory) π as a linear combination of best-estimate premiums across values of D:

$$\pi(\mathbf{x}) = \mu^B(\mathbf{x}, 1)\tilde{P}_D(\mathbf{x}; \pi) + \left\{1 - \tilde{P}_D(\mathbf{x}; \pi)\right\}\mu^B(\mathbf{x}, 0).$$

Because π is unconstrained, $\tilde{P}_D(\mathbf{x}; \pi)$ may lie outside [0, 1]. Solving for $\tilde{P}_D$ yields

$$\tilde{P}_D(\mathbf{x}; \pi) = \frac{\pi(\mathbf{x}) - \mu^B(\mathbf{x}, 0)}{\mu^B(\mathbf{x}, 1) - \mu^B(\mathbf{x}, 0)}.$$

The implied propensity is well defined when $\mu^B(\mathbf{x}, 1) \neq \mu^B(\mathbf{x}, 0)$. Values outside [0, 1] reveal targeting of a protected group. Naturally, $\tilde{P}_D(\mathbf{x}; \mu^U) = \Pr(D = 1|\mathbf{X} = \mathbf{x})$. An implied propensity aligned with $\Pr(D = 1|\mathbf{X} = \mathbf{x})$ or $1 - \Pr(D = 1|\mathbf{X} = \mathbf{x})$ reflects proxy effects or solidarity, respectively.

5.2.3. Excess Lift

For directly discriminatory pricing function, we define the excess lift for segment $\mathbf{x}$ as

$$\Delta_{\text{excess}}(\mathbf{x}; \pi) = \left|\pi(\mathbf{x}, 1) - \pi(\mathbf{x}, 0)\right| - \left|\mu^B(\mathbf{x}, 1) - \mu^B(\mathbf{x}, 0)\right| = \left|\pi(\mathbf{x}, 1) - \pi(\mathbf{x}, 0)\right| - \Delta_{\text{risk}}(\mathbf{x}).$$

The excess lift quantifies how “excessively” a pricing function π differentiates on D for a segment relative to the “pure risk” best-estimate premium. By construction, $\Delta_{\text{excess}}(\mathbf{x}; \mu^B) = 0$ for every segment $\mathbf{x}$. Strictly positive values signal overdifferentiation on D; negative values signal underdifferentiation between $D = 0$ and $D = 1$ (e.g., from smoothing, regulatory caps, or solidarity efforts) and imply cross-subsidization within that segment.

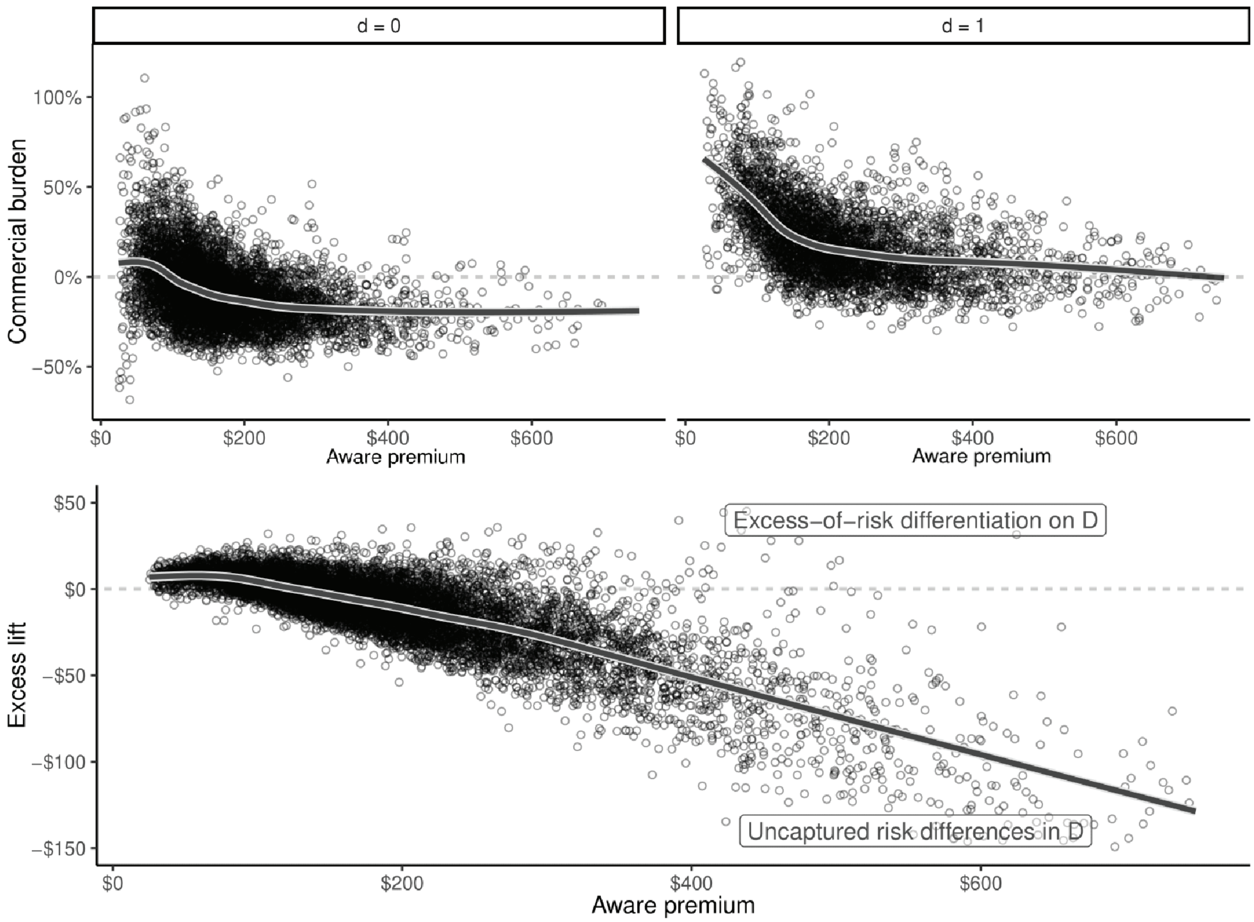
Case study (Cont’d). Figure 11 shows commercial burden $\hat{\rho}(\mathbf{x}, d; \text{PseudoPrice}, \hat{\mu}^A)$ (top) and excess lift $\hat{\Delta}_{\text{excess}}(\mathbf{x}; \text{PseudoPrice})$ (bottom) plotted against the aware premium. Though the best-estimate μ^B reflects industry norms during the study period, we use μ^A as the reference premium to assess targeting of D, whether directly or via proxy effects.

In the top panel, both groups show high variability. Commercial burden is on average higher for vulnerable populations than for others. In the bottom panel, excess lift declines with $\hat{\mu}^A$. Excess lift on D is positive for low $\hat{\mu}^A$ but negative for large premiums, possibly due to fixed costs, solidarity efforts, or rigid models (e.g., models without interactions). ▲

6. Exposing Systematic Disparities Through Partitioning

Fairness metrics on broad groups can mislead. A model may seem fair with respect to gender while masking disparities for smaller vulnerable groups, such as single mothers. Tailored fairness assessments for specific subpopulations help to target corrections. To detect

Figure 11. Commercial burden $\hat{\rho}(x, d; \text{PseudoPrice}, \hat{\mu}^A)$ by protected group (top) and excess lift $\hat{\Delta}_{\text{excess}}(x; \text{PseudoPrice})$ (bottom) plotted against aware premiums for the case study.



systematic fairness disparities, we propose in this section a simple methodology to partition policyholders following a relevant fairness metric.

To create these supervised partitions, we use decision trees (see, e.g., Loh 2014) for simplicity and interpretability. Five essential components guide this process: **data**, **feature space**, **response variable**, **loss function**, and **algorithm**, further detailed in the online supplement. The complexity of partitioning should align with the analyst's need for fairness granularity rather than be dictated by loss minimization alone.

We leverage the partitioning algorithm with relevant local fairness quantities to differentiate segments depending on a given fairness quantity: We identify segments with high proxy vulnerability upfront in Section 6.1, and we detect commercial loading in rates in Section 6.2.

6.1. Prepricing Policyholder Partitioning by Proxy Vulnerability

We identify segments most exposed to potential proxy discrimination by partitioning policyholders based on proxy vulnerability (defined in Section 5.1.2).

Case study (Cont'd). We apply a regression tree to predicted proxy vulnerability $\hat{\Delta}_{\text{proxy}}(\mathbf{x})$, using all allowed variables **X** for partitioning. The resulting regularized tree,[5] depicted in Figure 12, reveals the following:

- The highest proxy vulnerability leaves (numbered 1–5) are split by `hasPropertyIns` (property insurance indicator), `DrivExp` (driving experience), `NbTraffViolation` (count of traffic violations), and `OccType` (type of occupation).
- The indicator `hasPropertyIns` likely captures property ownership (clearly associated with credit score), `DrivExp` correlates with age or financial constraints, and a high value of `NbTraffViolation` may reflect low risk aversion.
- Inexperienced drivers, such as `DrivExp` < 5 and `hasPropertyIns == 'N'` for nodes 1 and 2, stand out as a group that is vulnerable to high proxy effects. Their causal risk may be inflated by implicit inference of their credit score.
- Proxy vulnerability averages to zero but hides wide disparities. In leaf node 1, the median overcharge is $70 on mean losses of $540 (13%). Elsewhere, the proxy rebate is at most $23. Proxy effects are both material and asymmetric.

Leaves with high predictions highlight segments wherein proxy effects are most likely to be exploited by a model. ▲

6.2. Postpricing Policyholder Partitioning by Commercial Loading

Building a regression tree on commercial loading (Section 5.2.1) reveals rating patterns.

Case study (Cont'd). We fit a regression tree to $\hat{\Delta}_{\text{load}}(\mathbf{x}; \hat{\mu}^A)$ and depict it in Figure 13. The leaves with highest commercial loading (numbered 1–6) are split by `DrivExp` (driving experience of the main driver), `MaritalStatus` (marital status), `GenderMainDriver` (gender of the main driver), `DriverAge` (age of the main driver), and `OccType` (type of occupation). Highest loadings occur among inexperienced married drivers (leaves 1, 2, 4, and 5) and single males with little driving experience (leaves 3 and 6).

Unlike proxy vulnerability, commercial loading stems from more than just proxy effects: differences in technical pricing methodology; lag in historical loss modeling for *Y*; commercial strategies; new business discounts; capping; and most important, direct discrimination on *D*. Pricing decisions, when compounded, may produce unintended disparities, disadvantaging groups beyond the insurer's intent and/or awareness.

Combined with the prepricing partition on proxy vulnerability (Section 6.1), the two partitions may help track model behavior across flagged subgroups. We integrate these components into a structured fairness assessment framework, depicted in Figure 14, combining the partitions illustrated in Figures 12 and 13 with the actuarial metrics defined in Section 5.

[5] Our partitioning with `evtree` uses a subsample of 50,000 observations and evaluates a population of 150 candidate evolutionary trees; the final model is the single best tree (no ensemble). See Grubinger et al. (2014) for more details.

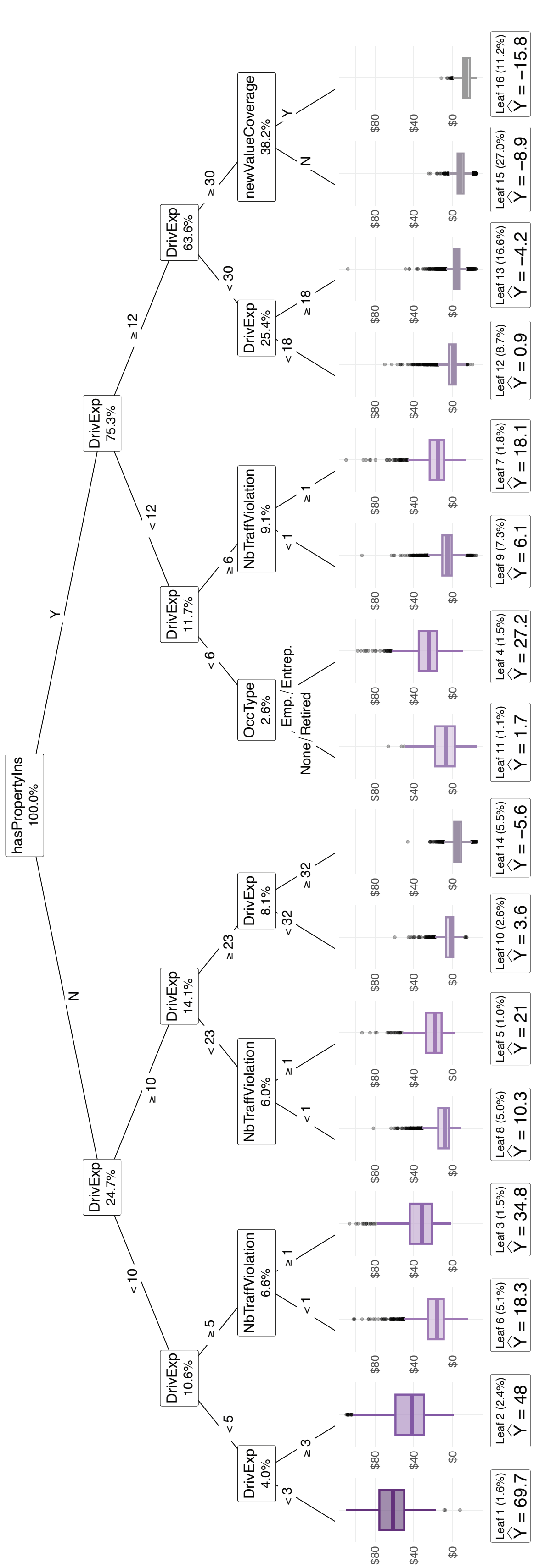


Figure 12. Optimal partitioning of $\hat{\Delta}_{proxy}(x)$ for the case study.

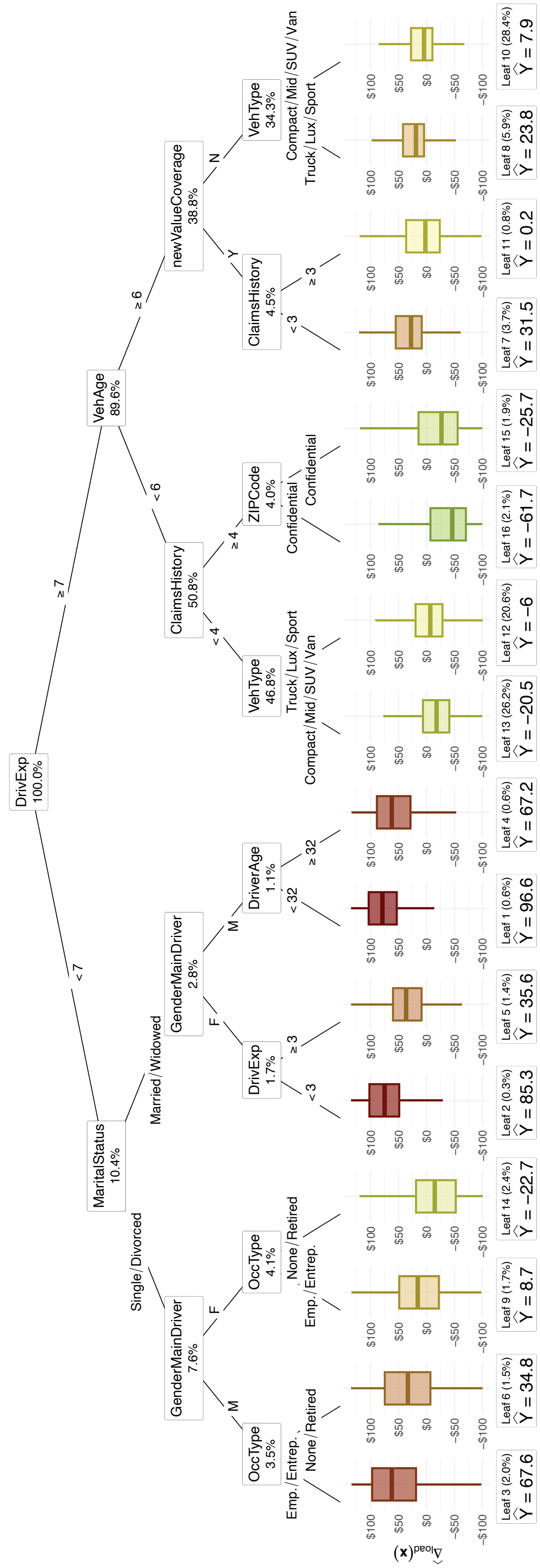


Figure 13. Optimal partitioning of $\hat{\Delta}_{load}(x)$ for the case study.

Figure 14. Fairness monitoring table summarizing demographic, pricing, and fairness metrics for partitioned subpopulations in the case study. Masking symbols ("#") preserve the partner insurer's confidentiality. Subpop = Subpopulation; Est. = Estimated; Wass. Dist. = Wasserstein distance; Best-est. = Best Estimate.

					Proxy vulnerable groups							Commercially loaded groups						
Section	Variable	Statistic	Subpop	All Data	1	2	3	4-13	14	15	16	1	2	3	4-13	14	15	16
Demographics	Exposure	Sum	All	164,064	2,405	3,882	2,295	79,797	9,694	46,763	19,227	852	419	2,964	149,637	3,877	3,114	3,201
	Credrisk (lvl %)	Level %	Credrisk = 1	37.5%	81.6%	74.8%	70.1%	48.1%	33.6%	21.3%	18.3%	76.7%	73.3%	76.1%	35.9%	59.1%	36.2%	38.0%
	DrivExp	Mean	All	#######	1.44	3.53	6.82	#####	42.46	42.29	42.24	4.40	1.31	4.23	#####	3.70	30.87	31.18
	VehAge	Mean	All	#######	5.22	4.93	4.31	#####	5.99	7.57	1.79	4.31	4.71	4.62	#####	5.52	2.33	2.33
	DrivAge	Mean	All	#######	21.89	22.30	24.36	#####	62.33	62.65	62.16	24.23	30.69	22.57	#####	20.52	50.08	51.42
	Zip Code (first character only)	Level %	J	#######	####	####	####	#####	####	####	####	####	####	####	#####	####	####	####
		Level %	G	#######	####	####	####	#####	####	####	####	####	####	####	#####	####	####	####
		Level %	H	#######	####	####	####	#####	####	####	####	####	####	####	#####	####	####	####
	OccType	Level %	Employed	60.6%	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
		Level %	Retired	18.7%	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
		Level %	None	10.3%	58.2%	50.7%	26.1%	11.8%	5.3%	4.9%	3.4%	14.9%	25.0%	0.0%	8.4%	100.0%	2.8%	4.2%
		Level %	Other	10.5%	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
	HasPropertyIns	Level %	Y	#######				#####		100%	100%	59%	56%	22%	#####	24%	83%	83%
		Level %	lvl % : N	#######	100%	100%	100%	#####	100%			41%	44%	78%	#####	76%	17%	17%
Loss and pseudo price	Loss	% of 0	All	97.2%	94.1%	95.1%	96.2%	97.1%	97.5%	97.8%	96.8%	96.1%	94.8%	94.9%	97.4%	95.8%	96.1%	94.2%
		Mean	All	189.25	539.73	414.37	322.26	202.52	134.01	108.18	187.61	198.57	315.87	74.841	177.73	256.44	403.98	441.82
	Severity	Mean	All	3743	4686	4544	4351	3948	3289	3046	3593	4126	4233	5441	3689	3809	2973	3672
		TVaR 0.95	All	19093	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
	Pseudoprice	VaR 0.05	All	68.40	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
		Mean	All	191.25	509.87	416.36	355.28	210.82	151.21	125.10	186.18	362.93	399.29	470.30	176.60	347.86	257.60	290.33
		VaR 0.95	All	401.59	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
	Performance metrics	MAE	All	272.15	712.19	586.57	483.54	296.82	214.04	175.85	267.70	466.07	545.53	666.83	251.25	469.10	334.85	433.41
		Avg. Dev. Twee.	All	137.20	226.26	200.34	192.22	150.27	117.06	106.12	138.29	221.77	233.75	233.01	132.23	174.53	157.82	180.45
Loss and price per group	Price	Mean	D=0	152.74	416.44	325.45	290.66	171.48	135.32	114.67	175.43	301.75	326.45	369.51	145.13	274.23	221.33	252.51
		Mean	D=1	255.31	530.92	447.06	382.84	253.27	182.59	163.65	234.15	381.53	425.76	502.00	232.75	398.71	321.53	351.98
		VaR 0.95	D=0	288.64	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
		VaR 0.95	D=1	497.85	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
	Loss Ratio	-	D=0	87.4%	86.1%	95.4%	102.4%	88.6%	84.7%	80.5%	93.9%	76.3%	113.6%	84.9%	86.5%	102.1%	74.6%	111.6%
		-	D=1	102.3%	109.4%	100.5%	86.9%	101.5%	94.3%	101.9%	123.9%	87.6%	110.6%	117.3%	102.3%	77.3%	84.2%	137.8%
	Est. P(D=1|X)	Mean	All	0.49	0.82	0.76	0.73	0.54	0.38	0.26	0.21	0.77	0.80	0.77	0.47	0.65	0.45	0.47
	Wass. Dist.	Prem(d = 0) vs Prem(d = 1)		103.99	113.22	122.78	94.74	82.14	48.01	49.88	59.27	82.84	99.28	136.92	88.42	123.19	101.78	101.34
Fairness Spectrum	Best est.	Mean	All	191.16	537.97	437.34	399.48	213.28	147.83	116.51	184.78	300.44	347.84	454.87	174.57	392.08	277.52	345.69
	Unaware	Mean	All	191.16	535.91	435.05	401.29	213.52	147.58	116.52	184.57	300.39	350.52	454.15	174.58	392.45	277.18	346.05
	Aware	Mean	All	191.16	468.35	389.01	366.31	209.20	153.09	125.43	200.63	271.32	317.39	400.77	176.35	368.23	283.84	351.65
	Hyperaware	Mean	All	191.16	429.22	360.74	339.62	204.61	157.22	132.10	216.09	247.98	288.18	370.75	177.49	350.26	290.47	357.14
	Corrective	Mean	All	191.16	426.82	359.46	341.01	204.73	157.32	131.99	215.51	247.65	290.23	370.30	177.40	349.98	290.57	357.76
Prepricing local metrics	Proxy vulnerability	Mean	All	-0.08	67.56	46.03	34.98	4.32	-5.51	-8.91	-16.07	29.07	33.13	53.38	-1.77	24.23	-6.66	-5.59
		Mean	D=1	7.85	68.59	48.37	36.34	7.66	-3.54	-8.02	-14.82	30.13	35.67	56.35	4.59	32.08	0.84	4.04
		Mean	D=0	-4.84	63.01	39.13	31.81	1.23	-6.51	-9.15	-16.35	25.59	26.14	43.94	-5.33	12.86	-10.92	-11.51
		TVaR 0.95	All	57.97	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
		TVaR 0.95	D=1	76.71	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
		TVaR 0.95	D=0	34.31	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
	Risk spread	Mean	All	65.90	178.05	150.62	130.62	71.93	51.53	42.80	65.42	87.67	101.95	157.52	60.30	138.71	91.90	118.66
	Fairness spread	Mean	All	55.38	161.03	131.06	118.66	61.22	42.49	33.43	54.97	86.73	101.53	137.47	50.23	116.34	83.26	104.59
	Parity cost	Mean	All	0.12	-111.15	-77.88	-58.47	-8.55	9.48	15.48	30.73	-52.79	-57.62	-84.57	2.83	-42.10	13.05	12.07
Postpricing local metrics	Commercial loading	Mean	All	0.09	-28.10	-20.99	-44.20	-2.46	3.37	8.59	1.39	62.50	51.45	15.43	2.03	-44.22	-19.92	-55.36
		Mean	D=1	-8.66	-42.25	-34.20	-56.53	-9.62	3.02	10.49	-1.60	57.83	43.62	1.80	-5.88	-68.30	-23.91	-73.51
		Mean	D=0	5.35	34.74	18.12	-15.30	4.17	3.55	8.08	2.06	77.86	72.99	58.74	6.47	-9.37	-17.66	-44.23
		TVaR 0.95	All	72.32	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
		TVaR 0.95	D=1	77.57	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
		TVaR 0.95	D=0	68.71	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
	Comm. burden	Mean	All	10.7%	0.0%	-1.4%	-7.8%	5.3%	16.3%	24.0%	4.1%	24.5%	17.6%	9.3%	12.0%	-9.0%	-5.9%	-14.6%
	Excess lift	Mean	All	-11.53	-76.78	-57.63	-43.98	-14.43	-4.43	-0.16	-9.38	-19.51	-26.73	-56.78	-8.98	-51.46	-14.26	-33.75
Extra summary statistics	Commercial burden	Prop > 0%	All	56.5%	36.5%	37.0%	24.4%	50.8%	61.1%	68.8%	58.6%	93.8%	86.0%	64.6%	58.6%	19.8%	29.6%	8.7%
		Prop > 0%	D=1	48.3%	29.2%	28.5%	18.6%	44.8%	60.8%	69.9%	54.1%	92.0%	81.4%	55.3%	50.3%	8.4%	29.5%	5.9%
		Prop > 0%	D=0	61.5%	69.0%	62.1%	37.9%	56.3%	61.2%	68.5%	59.7%	99.6%	98.4%	94.1%	63.3%	36.4%	29.7%	10.5%
		Prop > 9.2%	All	38.2%	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
		Prop > 9.2%	D=1	30.3%	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
		Prop > 9.2%	D=0	43.0%	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
		Prop > 16.9%	All	25.9%	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
		Prop > 16.9%	D=1	19.2%	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
		Prop > 16.9%	D=0	29.9%	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####	#####
				Raw decision rule from partitioning	hasPropertyIns is N AND DrivExp < 3	hasPropertyIns is N AND 3 <= DrivExp < 5	hasPropertyIns is N AND 5 <= DrivExp < 10 AND NbTraffViolation >= 1		hasPropertyIns is N AND 32 <= DrivExp	hasPropertyIns is Y AND 30 <= DrivExp AND newValueCoverage is N	hasPropertyIns is Y AND 30 <= DrivExp AND newValueCoverage is Y	DriverExp < 7 AND MaritalStatus is Married/Widowed AND	DriverExp < 7 AND MaritalStatus is Married/Widowed AND	DriverExp < 7 AND MaritalStatus is Single/Divorced AND		DriverExp < 7 AND MaritalStatus is Single/Divorced AND	DriverExp >= 7 AND VehAge < 6 AND ClaimsHistory >= 4 AND ZIP Code is [Confidential]	DriverExp >= 7 AND VehAge < 6 AND ClaimsHistory >= 4 AND ZIP Code is [Confidential]

Our toolbox guides actuaries in pinpointing which segments (partitioning columns) may warrant premium adjustment and which fairness indicators (rows) should be evaluated.

The tool assembles key actuarial components for each subgroup:

- Demographic summaries on selected covariates
- Classical pricing diagnostics: expected losses, premiums, predictive performance
- Basic fairness assessment: group disparities in premiums and losses
- Pre- and postpricing local fairness indicators
- The partitioning rule identifying the subgroup at the bottom
- Optional enrichment from external data (e.g., census)

Analysis reveals a high proxy vulnerability among groups with elevated commercial loading. The alignment between prepricing proxy vulnerability and postpricing commercial loading suggests that proxy vulnerability is captured by the pseudoprice – a predictable outcome given the sensitive attribute's acceptability during the study period. Both partitions point to inexperienced drivers (low `DrivExp`) as a critical group, offering a clear path to intervene in the pseudoprice ratemaking algorithm or to apply postprocessing adjustments for mitigating unfairness with respect to credit score.

Integrating the three dimensions of fairness in model assessment may form part of future actuarial standards. See the online supplement for code applied to simulated data. ▲

Bender et al. (2025) group biases as systemic, statistical, and human. Assuming *D* is not a causal factor, risk spread flags segment-level systemic bias (measurement, sampling, label), and parity cost is the dollar cost to undo it. Proxy vulnerability estimates statistical bias from omitting *D*. With a data-driven spectrum,[6] prepricing metrics capture systemic and (potential) statistical bias; prior-pricing metrics can reflect all three to guide mitigation.

7. Discussion

Our case study grounds fairness in a large-scale, realistic setting. The fairness spectrum translates dimensions into pricing benchmarks. It provides context to judge any given (commercial) price. It also supports intuitive metrics, such as proxy vulnerability by group or the share of policyholders facing high commercial burden, expressed in dollars and policyholders. By making fairness practical, we hope actuaries engage. The collaboration that made this case study possible reflects insurers' interest in understanding fairness: its meaning, materiality, and actuarial relevance.

In our case study, prepricing analysis shows proxy vulnerability is material and skewed. While many receive small rebates, some face 15%–30% overpricing. The pseudoprice

[6] A spectrum anchored on a data-driven best-estimate premium (Option 1 of Complement 5).

burdens the vulnerable group $D = 1$ more than the $D = 0$ group, but to a lesser extent than risk alone would justify, suggesting efforts toward solidarity.

Partitioning before and after pricing (Section 6) extends fairness analysis beyond large protected groups. Combined with local metrics (Section 5), it supports tools like Figure 14 to pinpoint the need for fairness adjustments within specific subgroups. In our case study, the pseudoprice appears to capture proxy vulnerability because commercially loaded groups also exhibit high proxy vulnerability. This surfaced concerns about fairness, which were not initially flagged, in specific policyholder segments such as inexperienced drivers.

In this paper, we progress from fairness principles to detection under assumptions that warrant scrutiny. In the current state of research, the three dimensions of fairness presented in Section 3 are necessary, but whether they are exhaustive remains an open question. Also, fairness dimensions are general – multiple premiums may reflect the same dimension. Both a corrective and a flat-rate price satisfy solidarity, suggesting a broader range of models. We ignored uncertainty in estimating benchmarks. How should we account for it?

7.1. Topics for Further Research

This study had a specific scope. Advancing fairness requires expanding it:

1. Fairness often assumes **access to protected attributes**, which may be unavailable. Can we assess fairness without it? Prediction of D (e.g., with *Bayesian Improved First Name Surname Geocoding* as in Voicu 2018) and census data help, but they are not substitutes for direct access.
2. **Market dynamics** are ignored; portfolio fairness may conflict with market fairness (Côté et al. 2024). Can insurers contribute to market fairness using their own data?
3. Fairness typically is studied as a one-year objective, but its **long-term welfare** effects remain unclear (Shimao et al. 2025). Which fairness approaches perpetuate, mitigate, or reverse disparities over time?
4. Seemingly neutral variables can mediate the link between protected traits and losses. **Behavioral data** may attenuate the impact of protected attributes on premiums by enriching $\mathbf{X}$ and detailing the causal risk chain (Boucher and Pigeon 2024). This offers actuarial justification for disparities, but does it resolve proxy issues?
5. Insurance operates between law and statistics: One demands fairness case by case; the other defends differentiation at scale. Applying antidiscrimination **regulations** meets resistance where actuarial justification holds authority. Can regulations reconcile these perspectives to fairly serve insurers, regulators, and policyholders?

Statement on the Use of Generative AI

We used generative AI tools to refine wording and syntax, draft or refactor small code snippets, and accelerate literature discovery (query formulation and complementing of other bibliographic search tools). All AI outputs (text, code, and references) were independently

Table 4. CAS Canada Race and Insurance Task Force members.

Craig Sloss, *chair*	Kexin Cai	Aliya Nathoo
Mallika Bender, *staff chair*	Renxia Huang	Andrew Nguyen
Carlos Araiza-Iturria	Shengli Huang	Yiyin (Jasmine) Pang
Frederick Au	Jordan Krasowski	Shayan Sen
Elizabeth Bellefleur-McCaul	Jingwen Li	Saul Warhaft
Davina Bhandari	Hannan Madamidola	Stephanie Zeng

reviewed, verified, and edited before inclusion; no analysis, modeling choices, or conclusions were delegated to these tools. We take full responsibility for the accuracy and integrity of all content in this publication.

Acknowledgments

We are grateful for the contributions of the CAS Canada Race and Insurance Research Task Force in overseeing, reviewing, and editing this paper. The members of the task force are listed in Table 4.

We thank Marouane Il Idrissi for discussions about projections, Fei Huang for discussions about encompassing frameworks for commercial prices, Agathe Fernandes Machados for assistance with Equipy, and Ewen Gallic for guidance on economic aspects. We thank our industrial partner for its meaningful involvement, which greatly enhanced the maturity and practicability of our work.

The first author is thankful for financial support from the Society of Actuaries' Hickman Scholars Program, Desjardins General Insurance Group, Mitacs, Fonds de recherche du Québec, the Chaire d'actuariat of Université Laval, Leadership Université Laval, and the Viger Family. The second author acknowledges the National Sciences and Engineering Research Council (NSERC) for funding (RGPIN-2019-04190). The third author acknowledges the SCOR Foundation for Science and NSERC for funding (RGPIN-2019-07077).

8. References

Actuarial Standards Board. 2005, December. *Risk Classification (for All Practice Areas)*. ASOP 12. https://www.actuarialstandardsboard.org/asops/risk-classification-practice-areas/.

Actuarial Standards Board. 2017, December. *Estimating Future Costs for Prospective Property/Casualty Risk Transfer and Risk Retention*. ASOP 53. https://www.actuarialstandardsboard.org/asops/estimating-future-costs-prospective-propertycasualty-risk-transfer-risk-retention/.

Arrow, K. J. 1963. "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review* 53 (5): 941–73.

Bailey, R. A., and L. J. Simon. 1960. "Two Studies in Automobile Insurance Ratemaking." *ASTIN Bulletin* 1 (4): 192–217.

Barry, L. 2020. "Insurance, Big Data and Changing Conceptions of Fairness." *European Journal of Sociology* 61 (2): 159–84.

Bender, M., C. Dill, M. Hurlbert, C. Lindberg, and S. Mott. 2022. "Understanding Potential Influences of Racial Bias on P&C Insurance: Four Rating Factors Explored." Casualty Actuarial Society Research Paper Series on Race and Insurance Pricing. https://www.casact.org/publications-research/research/research-paper-series-race-and-insurance-pricing.

Bender, M., D. L. Dillon, R. A. Harbage, K. Hurta, and B. J. Mullen. 2022. "Approaches to Address Racial Bias in Financial Services: Lessons for the Insurance Industry." Casualty Actuarial Society Research Paper Series on Race and Insurance Pricing. https://www.casact.org/publications-research/research/research-paper-series-race-and-insurance-pricing.

Bender, M., B. Margaret, E. Krafcheck, C. Sloss, G. Wang, and M. Woods. 2025. "Practical Application of Bias Measurement and Mitigation Techniques in Insurance Pricing." Casualty Actuarial Society Research Paper Series on Race and Insurance Pricing, Part 1 and 2. https://www.casact.org/publications-research/research/research-paper-series-race-and-insurance-pricing.

Boucher, J.-P., and M. Pigeon. 2024. "Balancing Risk Assessment and Social Fairness: An Auto Telematics Case Study." Casualty Actuarial Society Research Paper Series on Race and Insurance Pricing. https://www.casact.org/publications-research/research/research-paper-series-race-and-insurance-pricing.

Casualty Actuarial Society. 1988. "Statement of Principles Regarding Property and Casualty Insurance Ratemaking." https://www.casact.org/statement-principles-regarding-property-and-casualty-insurance-ratemaking. Originally published in 1988, rescinded in 2020, and reinstated in 2021 for reference.

Cavanaugh, L., S. Merkord, T. Davis, and D. Heppen. 2024. "Regulatory Perspectives on Algorithmic Bias and Unfair Discrimination." Casualty Actuarial Society Research Paper Series on Race and Insurance Pricing. https://www.casact.org/publications-research/research/research-paper-series-race-and-insurance-pricing.

Charpentier, A. 2024. *Insurance, Biases, Discrimination and Fairness*. Springer.

Charpentier, A., F. Hu, and P. Ratz. 2023. "Mitigating Discrimination in Insurance with Wasserstein Barycenters." Paper presented at BIAS 2023, 3rd Workshop on Bias and Fairness in AI, International Workshop of ECML-PKDD, Torino, Italy, September 22.

Côté, M.-P., O. Côté, and A. Charpentier. 2024, November 12. "Selection Bias in Insurance: Why Portfolio-Specific Fairness Fails to Extend Market-Wide." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.5018749.

Côté, O., M.-P. Côté, and A. Charpentier. 2025. "A Fair Price to Pay: Exploiting Causal Graphs for Fairness in Insurance." *Journal of Risk and Insurance* 92: 33–75.

Denuit, M., J. Huyghe, J. Trufin, and T. Verdebout. 2024. "Testing for Auto-Calibration with Lorenz and Concentration Curves." *Insurance: Mathematics and Economics* 117: 130–39.

Embrechts, P., and M. V. Wüthrich. 2022. "Recent Challenges in Actuarial Science." *Annual Review of Statistics and Its Application* 9 (1): 119–40.

Fahrenwaldt, M., C. Furrer, M. E. Hiabu, et al. 2024. "Fairness: Plurality, Causality, and Insurability." *European Actuarial Journal* 14: 317–28.

Fernandes Machado, A., S. Grondin, P. Ratz, A. Charpentier, and F. Hu. 2025, March 12. "Equipy: Sequential Fairness Using Optimal Transport in Python." Preprint, arXiv. https://arxiv.org/abs/2503.09866.

Financial Services Regulatory Authority of Ontario. 2024. *Proposed Guidance: Automobile Insurance Rating and Underwriting Supervision*. No. AU0142INT. https://www.fsrao.ca/industry/auto-insurance/regulatory-framework/guidance-auto-insurance/proposed-guidance-automobile-insurance-rating-and-underwriting-supervision-guidance.

Frees, E., G. Meyers, and R. Derrig. 2016. *Predictive Modeling Applications in Actuarial Science*. Case Studies in Insurance 2. Cambridge University Press.

Grubinger, T., A. Zeileis, and K.-P. Pfeiffer. 2014. "evtree: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R." *Journal of Statistical Software* 61 (1): 1–29.

Heller, Douglas, and Michael Delong. 2021. “Watch Where You’re Going: What’s Needed to Make Auto Insurance Telematics Work for Consumers.” Consumer Federation of America. https://consumerfed.org/reports/watch-where-youre-going/. Accessed: December 23, 2025.

Insurance Information Institute. 2023. “Insurance 101: The History of Insurance.” https://www.iii.org/article/insurance-101.

Ke, G., Q. Meng, T. Finley, et al. 2017. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree.” *Advances in Neural Information Processing Systems* 30: 3146–54.

Kleinberg, J., S. Mullainathan, and M. Raghavan. 2016, September 19. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” Preprint, arXiv. https://arxiv.org/abs/1609.05807.

Kormes, M. 1935. “The Experience Rating Plan as Applied to Workmen’s Compensation Risks – Part II.” *Proceedings of the Casualty Actuarial Society* 22: 81–108.

Leong, J., R. Moncher, and K. Jordan. 2024. “A Practical Guide to Navigating Fairness in Insurance Pricing.” Casualty Actuarial Society Research Paper Series on Race and Insurance Pricing. https://www.casact.org/publications-research/research/research-paper-series-race-and-insurance-pricing.

Lindholm, M., R. Richman, A. Tsanakas, and M. V. Wüthrich. 2022. “Discrimination-Free Insurance Pricing.” *ASTIN Bulletin* 52 (1): 55–89.

Lindholm, M., R. Richman, A. Tsanakas, and M. V. Wüthrich. 2026. “Sensitivity-Based Measures of Discrimination in Insurance Pricing.” *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2026.01.021.

Lindholm, M., R. Richman, A. Tsanakas, and M. V. Wüthrich. 2024. “What Is Fair? Proxy Discrimination vs. Demographic Disparities in Insurance Pricing.” *Scandinavian Actuarial Journal* 2024 (9): 935–70.

Loh, W.-Y. 2014. “Fifty Years of Classification and Regression Trees.” *International Statistical Review* 82 (3): 329–48.

Moodie, E. E. M., and D. A. Stephens. 2022. “Causal Inference: Critical Developments, Past and Future.” *Canadian Journal of Statistics* 50 (4): 1299–1320.

Mosley, R., and R. Wenman. 2022. “Methods for Quantifying Discriminatory Effects on Protected Classes in Insurance.” Casualty Actuarial Society Research Paper Series on Race and Insurance Pricing. https://www.casact.org/publications-research/research/research-paper-series-race-and-insurance-pricing.

Prince, A. E., and D. Schwarcz. 2019. “Proxy Discrimination in the Age of Artificial Intelligence and Big Data.” *Iowa Law Review* 105: 1257.

Shimao, H., F. Huang, and W. Khern-am-nuai. 2025, January 27. “Welfare Implications of Fairness Regulations in Insurance Cost Modeling: A Multi-Method Study.” Preprint, SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5112616.

Ville de Montréal. 2018. Profil sociodémographique: Recensement 2016 – arrondissement de montréal-nord. ville.montreal.qc.ca/portal. Accessed December 23, 2025.

Voicu, I. 2018. “Using First Name Information to Improve Race and Ethnicity Classification.” *Statistics and Public Policy* 5 (1): 1–13.

Appendix A. Brief Summary of the Paper

Let Y denote the claim cost for property-casualty coverage. Let $\mathbf{X}$ denote the vector of pricing covariates. Fairness is defined relative to a prespecified (sensitive) variable D, taken here to be a single, binary, fully observed random variable. See Section 2 for details.

protected groups. Levels of the sensitive attribute D.

vulnerable groups. Subset of protected groups showing historic disadvantage.

fairness. An elusive, contested ideal; a three-dimensional trade-off to manage (Section 3):

- **actuarial fairness.** Aligns premiums with expected losses, mitigating cross-subsidies.
- **solidarity.** Aligns premiums across protected groups, mitigating disparities.
- **causality.** Ensures models capture only proxy-free risk effects.

spectrum of fair premiums. The five fair premiums of the spectrum (Section 4) are:

Premium	Best estimate	Unaware	Aware	Hyperaware	Corrective
Notation	$\mu^B(\mathbf{x}, d)$	$\mu^U(\mathbf{x})$	$\mu^A(\mathbf{x})$	$\mu^H(\mathbf{x})$	$\mu^C(\mathbf{x}, d)$
Direct discrimination	✓	×	×	×	✓
Dimension prioritized	**Actuarial fairness**	**Actuarial fairness**	**Causality**	**Solidarity**	**Solidarity**

prepricing local metrics. Reveal potential unfairness in the dataset (Section 5.1).

risk spread. The spread of best estimates across values of D:

$$\Delta_{\text{risk}}(\mathbf{x}) = \left|\mu^B(\mathbf{x},1) - \mu^B(\mathbf{x},0)\right|.$$

proxy vulnerability. The difference between ignoring D and controlling for it:

$$\Delta_{\text{proxy}}(\mathbf{x}) = \mu^U(\mathbf{x}) - \mu^A(\mathbf{x}).$$

fairness range. The range of estimates across the spectrum of fair premiums:

$$\Delta_{\text{fair}}(\mathbf{x},d) = \max\left\{\mu^B(\mathbf{x},d), \mu^U(\mathbf{x}), \mu^A(\mathbf{x}), \mu^H(\mathbf{x}), \mu^C(\mathbf{x},d)\right\} - \min\left\{\mu^B(\mathbf{x},d), \mu^U(\mathbf{x}), \mu^A(\mathbf{x}), \mu^H(\mathbf{x}), \mu^C(\mathbf{x},d)\right\}.$$

parity cost. The (monetary) cost of shifting from actuarial fairness to solidarity:

$$\Delta_{\text{parity}}(\mathbf{x},d) = \mu^C(\mathbf{x},d) - \mu^B(\mathbf{x},d).$$

postpricing local metrics. Relate a given price $\pi(\mathbf{X}, D)$ to the spectrum (Section 5.2):

commercial loading. The difference between π and a chosen reference premium μ from the spectrum:

$$\Delta_{\text{load}}(\mathbf{x}, d; \pi, \mu) = \pi(\mathbf{x}, d) - \mu(\mathbf{x}, d).$$

commercial burden. The commercial loading as a percentage of μ:

$$\rho_{\text{burden}}(\mathbf{x}, d; \pi, \mu) = \frac{\pi(\mathbf{x}, d)}{\mu(\mathbf{x}, d)} - 1 = \frac{\Delta_{\text{load}}(\mathbf{x}, d; \pi, \mu)}{\mu(\mathbf{x}, d)}.$$

implied propensity (nondirectly discriminatory π). The implicit weight on $D = 1$:

$$\tilde{P}_D(\mathbf{x}; \pi) = \frac{\pi(\mathbf{x}) - \mu^B(\mathbf{x}, 0)}{\mu^B(\mathbf{x}, 1) - \mu^B(\mathbf{x}, 0)}.$$

excess lift (directly discriminatory π). The excess differentiation on D relative to μ^B:

$$\Delta_{\text{excess}}(\mathbf{x}; \pi) = \left|\pi(\mathbf{x}, 1) - \pi(\mathbf{x}, 0)\right| - \left|\mu^B(\mathbf{x}, 1) - \mu^B(\mathbf{x}, 0)\right|.$$