Exposing indirect discrimination

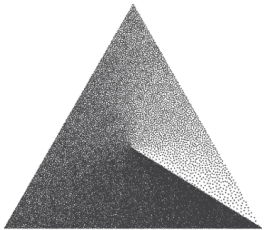
A scalable toolbox for fairness assessment of insurance rates

Presented by

Olivier Côté

Joint work with

Prof. Marie-Pier Côté (U. Laval) **and**
Prof. Arthur Charpentier (UQÀM)



Séminaire étudiant UQAM

Montréal

23 Juillet 2025

UNIVERSITÉ LAVAL

Fairness and equity

The strategic goal of the Council of Europe in the field of anti-discrimination, diversity and inclusion is to ensure genuine equality and **full access to rights and opportunities for all members** of society.

2021 Report by the Secretary General of the Council of Europe entitled “State of democracy, human rights and the rule of law : A democratic renewal for Europe”

Fairness and equity : no consensus

The strategic goal of the Council of Europe in the field of anti-discrimination, diversity and inclusion is to ensure genuine equality and **full access to rights and opportunities for all members** of society.

2021 Report by the Secretary General of the Council of Europe entitled “State of democracy, human rights and the rule of law : A democratic renewal for Europe”

The public release of these [Federal Equity Action] plans demonstrated immense public waste and **shameful discrimination**. That ends today. Americans deserve a government committed to **serving every person with equal dignity and respect** [...]

Executive order of The White House issued on January 20, 2025 entitled “Ending Radical And Wasteful Government DEI Programs And Preferencing”

Differentiation at the core of insurance pricing

Insurance pricing relies on grouping policyholders by risk to set adequate premiums.

Modern predictive models excel at detecting statistical associations to differentiate between risks, but they can learn spurious correlations.

This raises concerns when sensitive variables may (intentionally or inadvertently) affect the **fairness of insurance pricing**.

Notation

Age	Vehicle	Occupation	Gender	Religion	Credit	Claim
					800	
					700	
					650	
					435	
			Europe	California	Ontario	

Allowed variables
X

Prohibited variables
D
(Collected)

Response
Y

Some fair premiums

Given a prohibited variable D,

- Frees and Huang (2023) project the allowed variables X in a space orthogonal to the prohibited D in a **pre-processing** step;

Some fair premiums

Given a prohibited variable D,

- Frees and Huang (2023) project the allowed variables X in a space orthogonal to the prohibited D in a **pre-processing** step;
- Lindholm et al. (2022) propose a **post-processing** of the premium in order to prevent D from indirectly influencing the calculation.

Some fair premiums

Given a prohibited variable D,

- Frees and Huang (2023) project the allowed variables X in a space orthogonal to the prohibited D in a **pre-processing** step;
- Lindholm et al. (2022) propose a **post-processing** of the premium in order to prevent D from indirectly influencing the calculation.
-

These different premiums all aim to be *fair*. Again, no consensus.

Fair premiums

Given a prohibited variable D,

- Frees and Huang (2023) project allowed variables X orthogonally to prohibited D in **pre-processing**;

Fair premiums

Given a prohibited variable D,

- Frees and Huang (2023) project allowed variables X orthogonally to prohibited D in **pre-processing**;
- Lindholm et al. (2022) **post-process** premiums to limit D's indirect influence.

Fair premiums

Given a prohibited variable D,

- Frees and Huang (2023) project allowed variables X orthogonally to prohibited D in **pre-processing**;
- Lindholm et al. (2022) **post-process** premiums to limit D's indirect influence.
- Boucher and Pigeon (2024) find that **including telematics data** in X dillutes the significance of D.

Methods differ fundamentally but **all aim for fairness**.

Fair premiums

Given a prohibited variable D,

- Frees and Huang (2023) project allowed variables X orthogonally to prohibited D in **pre-processing**;
- Lindholm et al. (2022) **post-process** premiums to limit D's indirect influence.
- Boucher and Pigeon (2024) find that **including telematics data** in X dillutes the significance of D.

Methods differ fundamentally but **all aim for fairness**.

- Frees and Huang (2023) argue that **using variables related to** D constitutes indirect discrimination.
- Lindholm et al. (2024) equate indirect and **proxy discrimination**, where formulas implictly infer D.

Unclear fairness terms impede progress.

Goals

1. Understand the behaviour of fair pricing methodologies.
   - A Fair price to pay : exploiting causal graphs for fairness in insurance, *Journal of Risk and Insurance*

Goals

1. Understand the behaviour of fair pricing methodologies.
   - A Fair price to pay : exploiting causal graphs for fairness in insurance, *Journal of Risk and Insurance*
2. Quantify the **confounding bias** on real data.
   - A scalable toolbox for exposing indirect discrimination in insurance rates, *Forthcoming*

Goals

1. Understand the behaviour of fair pricing methodologies.
   - A Fair price to pay : exploiting causal graphs for fairness in insurance, *Journal of Risk and Insurance*
2. Quantify the **confounding bias** on real data.
   - A scalable toolbox for exposing indirect discrimination in insurance rates, *Forthcoming*

Goals

1. Understand the behaviour of fair pricing methodologies.
   - A Fair price to pay : exploiting causal graphs for fairness in insurance, *Journal of Risk and Insurance*
2. Quantify the **confounding bias** on real data.
   - A scalable toolbox for exposing indirect discrimination in insurance rates, *Forthcoming*

This joint work with Marie-Pier Côté and Arthur Charpentier is supported by a Canadian insurance company.

Exploiting causal graphs for fairness

Causal graphs for insurance

Actuarial predictive models **are based on proxies** rather than risk factors (Embrechts and Wüthrich, 2022).

Causal graphs for insurance

Actuarial predictive models **are based on proxies** rather than risk factors (Embrechts and Wüthrich, 2022).

But, **causality is key** for fairness (Makhlouf et al., 2024) :

Causal graphs for insurance

Actuarial predictive models **are based on proxies** rather than risk factors (Embrechts and Wüthrich, 2022).

But, **causality is key** for fairness (Makhlouf et al., 2024) :

- Araiza Iturria et al. (2024) relate discrimination-free premiums to causal inference techniques.
- Fernandes Machado et al. (2024) explore fairness through counterfactuals.

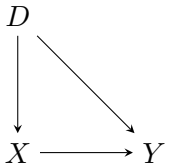
Causal graphs for insurance

Actuarial predictive models **are based on proxies** rather than risk factors (Embrechts and Wüthrich, 2022).

But, **causality is key** for fairness (Makhlouf et al., 2024) :

- Araiza Iturria et al. (2024) relate discrimination-free premiums to causal inference techniques.
- Fernandes Machado et al. (2024) explore fairness through counterfactuals.

Both are **supported by causal graphs** (DAG).



Motivation

- The causal graph that relates X, D and Y
  - has properties that are **not fully exploited**.
  - is too simple,
  - does not relate the premium (score) with the other variables,

Motivation

- The causal graph that relates X, D and Y
  - has properties that are **not fully exploited**.
  - is too simple,
  - does not relate the premium (score) with the other variables,

- There is **confusion** around fairness terms.
  - Indirect discrimination lacks a consensual definition.
  - The interplay between fairness criteria and indirect discrimination is ambiguous.

Motivation

- The causal graph that relates X, D and Y
  - has properties that are **not fully exploited**.
  - is too simple,
  - does not relate the premium (score) with the other variables,

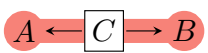
- There is **confusion** around fairness terms.
  - Indirect discrimination lacks a consensual definition.
  - The interplay between fairness criteria and indirect discrimination is ambiguous.

- Many fairness methodologies exist and do not match. The *Pre-In-Post* categorization **does not clarify the expected fairness properties**.

Dependence implications of a causal graph

		Causal type of C		
		Confounder	Mediator	Collider
Status of C	Unconditioned	$A \leftarrow C \rightarrow B$		
	Conditioned	$A \leftarrow \boxed{C} \rightarrow B$		

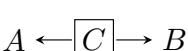
Dependence implications of a causal graph

Status of C	Causal type of C: Confounder	Mediator	Collider
Unconditioned			
Conditioned			

Dependence implications of a causal graph

Status of C	Causal type of C: Confounder	Mediator	Collider
Unconditioned	$A \leftarrow C \rightarrow B$	$A \rightarrow C \rightarrow B$	
Conditioned	$A \leftarrow \boxed{C} \rightarrow B$	$A \rightarrow \boxed{C} \rightarrow B$	

Dependence implications of a causal graph

Status of C	Causal type of C: Confounder	Mediator	Collider
Unconditioned			
Conditioned			

Dependence implications of a causal graph

Status of C	Causal type of C: Confounder	Mediator	Collider
Unconditioned	$A \leftarrow C \rightarrow B$	$A \rightarrow C \rightarrow B$	$A \rightarrow C \leftarrow B$
Conditioned	$A \leftarrow \boxed{C} \rightarrow B$	$A \rightarrow \boxed{C} \rightarrow B$	$A \rightarrow \boxed{C} \leftarrow B$

Dependence implications of a causal graph

Status of C		Causal type of C: Confounder	Mediator	Collider
	Unconditioned	$A \leftarrow C \rightarrow B$	$A \rightarrow C \rightarrow B$	$A \rightarrow C \leftarrow B$
	Conditioned	$A \leftarrow \boxed{C} \rightarrow B$	$A \rightarrow \boxed{C} \rightarrow B$	$A \rightarrow \boxed{C} \leftarrow B$

Dependence implications of a causal graph

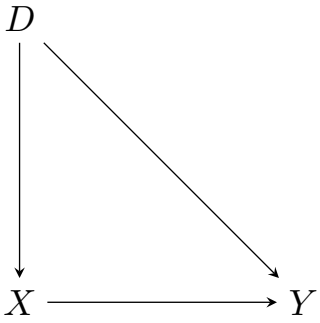
Status of C	Causal type of C: Confounder	Mediator	Collider
Unconditioned	$A \leftarrow C \rightarrow B$	$A \rightarrow C \rightarrow B$	$A \rightarrow C \leftarrow B$
Conditioned	$A \leftarrow \boxed{C} \rightarrow B$	$A \rightarrow \boxed{C} \rightarrow B$	$A \rightarrow \boxed{C} \leftarrow B$

Dependence implications of a causal graph

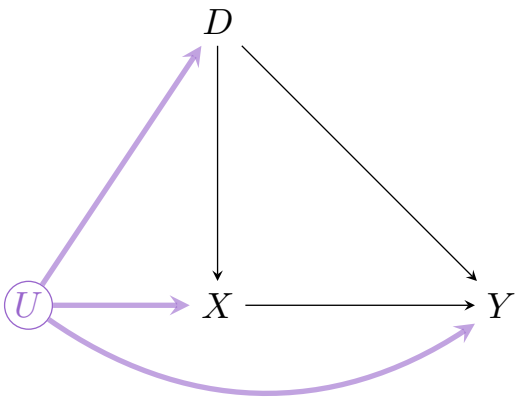
Status of C		Causal type of C: Confounder	Mediator	Collider
	Unconditioned	$A \not\perp\!\!\!\perp B$ $A \leftarrow C \rightarrow B$	$A \not\perp\!\!\!\perp B$ $A \rightarrow C \rightarrow B$	$A \perp\!\!\!\perp B$ $A \rightarrow C \leftarrow B$
	Conditioned	$A \perp\!\!\!\perp B \vert C$ $A \leftarrow \boxed{C} \rightarrow B$	$A \perp\!\!\!\perp B \vert C$ $A \rightarrow \boxed{C} \rightarrow B$	$A \not\perp\!\!\!\perp B \vert C$ $A \rightarrow \boxed{C} \leftarrow B$

Causal relationships have **dependence implications**.

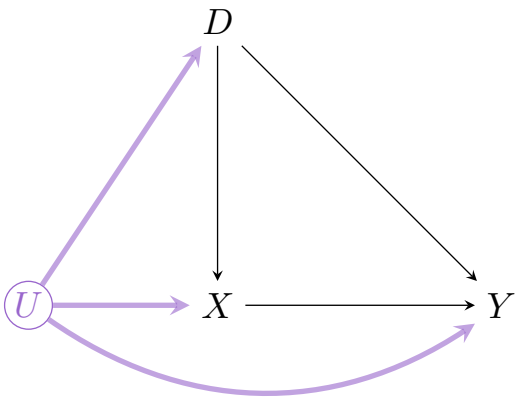
Potential bias : unobserved confounders



Potential bias : unobserved confounders

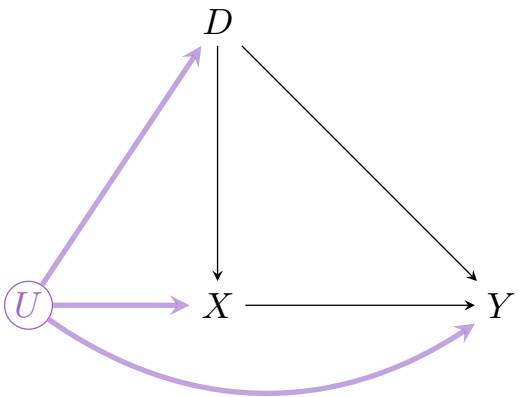


Potential bias : unobserved confounders



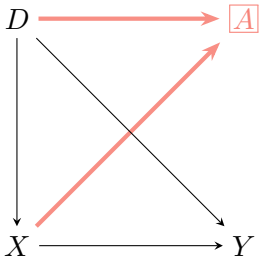
- **Omitted variable bias** arises because $D \leftarrow U \rightarrow X$ is indistinguishable from $D \rightarrow X$.

Potential bias : unobserved confounders



- **Omitted variable bias** arises because $D \leftarrow U \rightarrow X$ is indistinguishable from $D \rightarrow X$.
- Detecting unobserved confounders relies on expert judgment, making these confounders **unverifiable**.

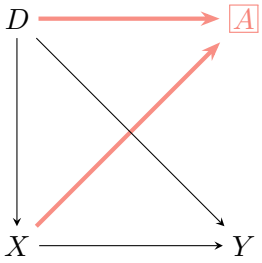
Portfolio selection bias



Selection bias in insurance : why portfolio-specific fairness fails to extend market-wide, *available on SSRN*

- The insurer's portfolio is conditional on $A = 1$, where A **is the insured's inclusion**; 1 if included, 0 otherwise.

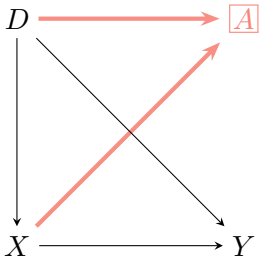
Portfolio selection bias



Selection bias in insurance : why portfolio-specific fairness fails to extend market-wide, *available on SSRN*

- The insurer's portfolio is conditional on $A = 1$, where A **is the insured's inclusion**; 1 if included, 0 otherwise.
- The path $D \to A \leftarrow X$ represents portfolio-specific **selection bias**.

Portfolio selection bias



Selection bias in insurance : why portfolio-specific fairness fails to extend market-wide, *available on SSRN*

- The insurer's portfolio is conditional on $A = 1$, where A **is the insured's inclusion**; 1 if included, 0 otherwise.
- The path $D \to A \leftarrow X$ represents portfolio-specific **selection bias**.
- Should insurers focus on **portfolio or market-wide** fairness ?

Separation into components according to D

Fairness studies are always centered on D.

Separation into components according to D

Fairness studies are always centered on D.

We decompose X and Y into : the part generated by D and the residual independent of D

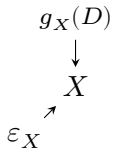
Separation into components according to D

Fairness studies are always centered on D.

We decompose X and Y into : the part generated by D and the residual independent of D, i.e.,

$$X = g_X(D) + \varepsilon_X$$

where $\varepsilon_X \perp\!\!\!\perp (D, g_X(D))$,



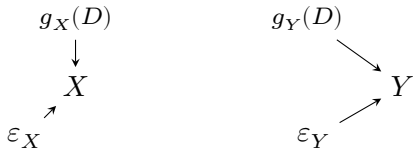
Separation into components according to D

Fairness studies are always centered on D.

We decompose X and Y into : the part generated by D and the residual independent of D, i.e.,

$$X = g_X(D) + \varepsilon_X \quad \text{and} \quad Y = g_Y(D) + \varepsilon_Y,$$

where $\varepsilon_X \perp\!\!\!\perp (D, g_X(D))$, $\varepsilon_Y \perp\!\!\!\perp (D, g_Y(D))$



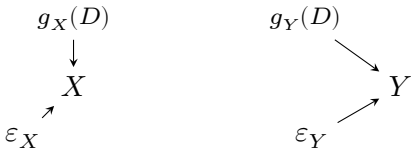
Separation into components according to D

Fairness studies are always centered on D.

We decompose X and Y into : the part generated by D and the residual independent of D, i.e.,

$$X = g_X(D) + \varepsilon_X \quad \text{and} \quad Y = g_Y(D) + \varepsilon_Y,$$

where $\varepsilon_X \perp\!\!\!\perp (D, g_X(D))$, $\varepsilon_Y \perp\!\!\!\perp (D, g_Y(D))$, and $\varepsilon_X, \varepsilon_Y$ should be “large”.



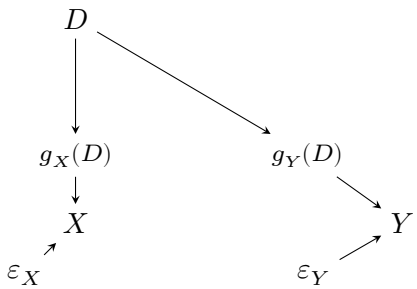
Separation into components according to D

Fairness studies are always centered on D.

We decompose X and Y into : the part generated by D and the residual independent of D, i.e.,

$$X = g_X(D) + \varepsilon_X \quad \text{and} \quad Y = g_Y(D) + \varepsilon_Y,$$

where $\varepsilon_X \perp\!\!\!\perp (D, g_X(D))$, $\varepsilon_Y \perp\!\!\!\perp (D, g_Y(D))$, and $\varepsilon_X, \varepsilon_Y$ should be “large”.



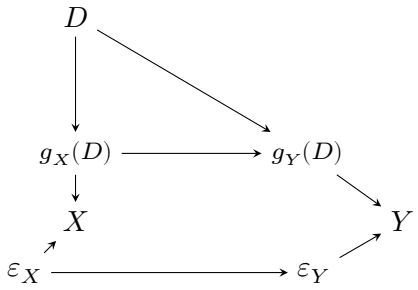
Separation into components according to D

Fairness studies are always centered on D.

We decompose X and Y into : the part generated by D and the residual independent of D, i.e.,

$$X = g_X(D) + \varepsilon_X \quad \text{and} \quad Y = g_Y(D) + \varepsilon_Y,$$

where $\varepsilon_X \perp\!\!\!\perp (D, g_X(D))$, $\varepsilon_Y \perp\!\!\!\perp (D, g_Y(D))$, and $\varepsilon_X, \varepsilon_Y$ should be “large”.



What is a premium?

Discrimination is defined relative to a score, **premium**, or formula.

What is a premium?

Discrimination is defined relative to a score, **premium**, or formula.

Definition (premium)

Let a premium μ be a deterministic function $(x, d) \mapsto \mu(x, d)$

Examples of premiums include estimates of $\mathbb{E}[Y|X, D]$ and final market prices.

What is a premium?

Discrimination is defined relative to a score, **premium**, or formula.

Definition (premium)

Let a premium μ be a deterministic function $(x, d) \mapsto \mu(x, d)$

Examples of premiums include estimates of $\mathbb{E}[Y|X, D]$ and final market prices.

A premium can be placed in a causal graph (Rosenbaum and Rubin, 1983) if it encapsulates a causal relationships.

(a) Causal relationship

What is a premium?

Discrimination is defined relative to a score, **premium**, or formula.

Definition (premium)

Let a premium μ be a deterministic function $(x, d) \mapsto \mu(x, d)$

Examples of premiums include estimates of $\mathbb{E}[Y|X, D]$ and final market prices.

A premium can be placed in a causal graph (Rosenbaum and Rubin, 1983) if it encapsulates a causal relationships.

(a) Causal relationship **(b)** **Sufficient** premium $\mu(A)$

Indirect discrimination from the legal perspective

"A situation in which an **apparently neutral provision**, criterion or practice would put persons of a racial or ethnic origin at a **particular disadvantage compared with other persons**, unless that provision, criterion or practice is **objectively justified** by a legitimate aim and the means of achieving that aim are appropriate and necessary."

—European Union (2000)

Indirect discrimination from the legal perspective

"A situation in which an **apparently neutral provision**, criterion or practice would put persons of a racial or ethnic origin at a **particular disadvantage compared with other persons**, unless that provision, criterion or practice is **objectively justified** by a legitimate aim and the means of achieving that aim are appropriate and necessary."

—European Union (2000)

According to Fredman (2022), recurring elements from legal definitions are :

Indirect discrimination from the legal perspective

"A situation in which an **apparently neutral provision**, criterion or practice would put persons of a racial or ethnic origin at a **particular disadvantage compared with other persons**, unless that provision, criterion or practice is **objectively justified** by a legitimate aim and the means of achieving that aim are appropriate and necessary."

—European Union (2000)

According to Fredman (2022), recurring elements from legal definitions are :

1. " equal treatment", i.e., no direct discrimination;

Indirect discrimination from the legal perspective

"A situation in which an **apparently neutral provision**, criterion or practice would put persons of a racial or ethnic origin at a **particular disadvantage compared with other persons**, unless that provision, criterion or practice is **objectively justified** by a legitimate aim and the means of achieving that aim are appropriate and necessary."

—European Union (2000)

According to Fredman (2022), recurring elements from legal definitions are :

1. " equal treatment", i.e., no direct discrimination;
2. "treatment must have disparate results";

Indirect discrimination from the legal perspective

"A situation in which an **apparently neutral provision**, criterion or practice would put persons of a racial or ethnic origin at a **particular disadvantage compared with other persons**, unless that provision, criterion or practice is **objectively justified** by a legitimate aim and the means of achieving that aim are appropriate and necessary."

—European Union (2000)

According to Fredman (2022), recurring elements from legal definitions are :

1. " equal treatment", i.e., no direct discrimination;
2. "treatment must have disparate results";
3. "the disparate impact can be justified if [...]".

Indirect discrimination from the legal perspective

"A situation in which an **apparently neutral provision**, criterion or practice would put persons of a racial or ethnic origin at a **particular disadvantage compared with other persons**, unless that provision, criterion or practice is **objectively justified** by a legitimate aim and the means of achieving that aim are appropriate and necessary."

—European Union (2000)

According to Fredman (2022), recurring elements from legal definitions are :

1. " equal treatment", i.e., no direct discrimination;
2. "treatment must have disparate results";
3. "the disparate impact can be justified if [...]".

We formalize for elements 1 and 2.

Defining indirect discrimination

Definition (indirect discrimination)

A premium $\mu(X, D)$ **discriminates indirectly** on D if :

1. it does not discriminate directly on D;
2. it exhibits at least one **source of disparate impact**.

Defining indirect discrimination

Definition (indirect discrimination)

A premium $\mu(X, D)$ **discriminates indirectly** on D if :

1. it does not discriminate directly on D;
2. it exhibits at least one **source of disparate impact**.

Defining indirect discrimination

Definition (indirect discrimination)

A premium $\mu(X, D)$ **discriminates indirectly** on D if :

1. it does not discriminate directly on D;
2. it exhibits at least one **source of disparate impact**.

- A **disparate impact** implies a decision $\mu(X, D)$ affecting protected groups D.

Defining indirect discrimination

Definition (indirect discrimination)

A premium $\mu(X, D)$ **discriminates indirectly** on D if :

1. it does not discriminate directly on D;
2. it exhibits at least one **source of disparate impact**.

- A **disparate impact** implies a decision $\mu(X, D)$ affecting protected groups D.
  - We formalise these **sources** as the paths between $\mu(X, D)$ and D.

Defining indirect discrimination

Definition (indirect discrimination)

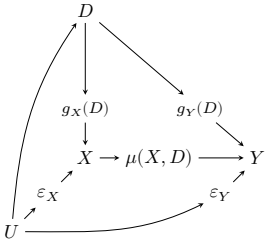
A premium $\mu(X, D)$ **discriminates indirectly** on D if :

1. it does not discriminate directly on D;
2. it exhibits at least one **source of disparate impact**.

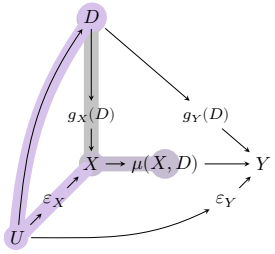
- A **disparate impact** implies a decision $\mu(X, D)$ affecting protected groups D.
  - We formalise these **sources** as the paths between $\mu(X, D)$ and D.

- We assume there is no direct discrimination.

Sources of disparate impact

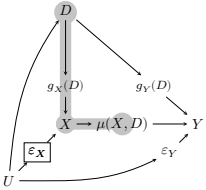


Sources of disparate impact

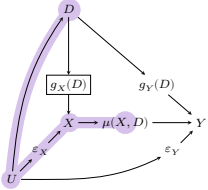


Disp. Impact through X

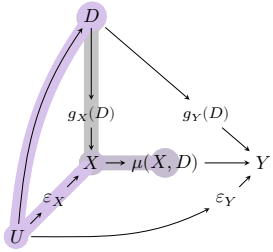
Sources of disparate impact



Purely through X

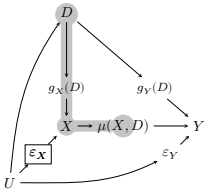


Through an conf. of $D \to X$

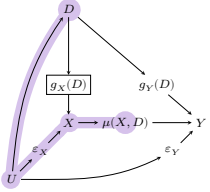


Disp. Impact through X

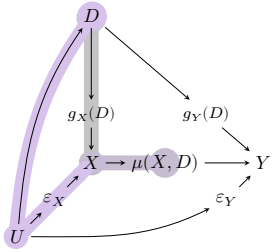
Sources of disparate impact



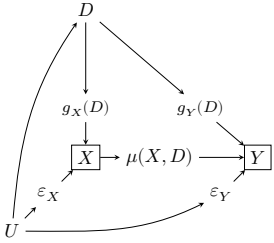
Purely through X



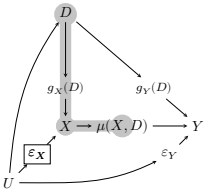
Through an conf. of $D \to X$



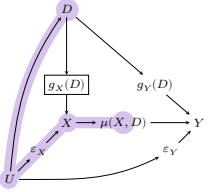
Disp. Impact through X



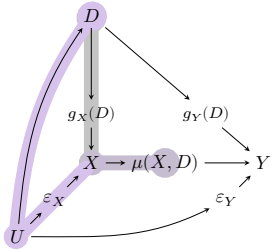
Sources of disparate impact



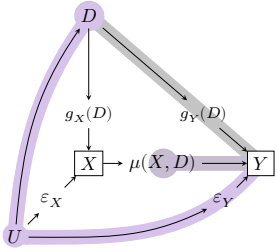
Purely through X



Through an conf. of $D \to X$

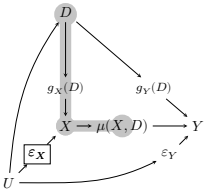


Disp. Impact through X

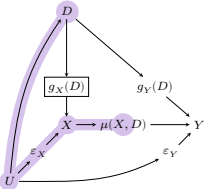


Disp. Impact through $\boxed{Y}$

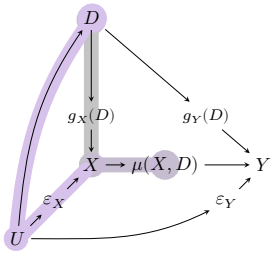
Sources of disparate impact



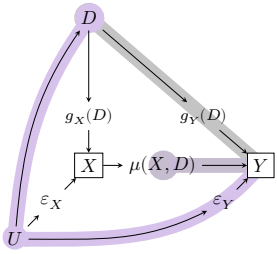
Purely through X



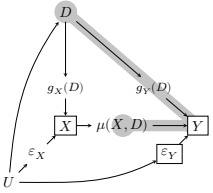
Through an conf. of $D \to X$



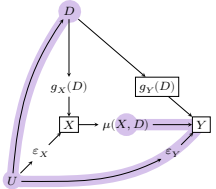
Disp. Impact through X



Disp. Impact through $\boxed{Y}$

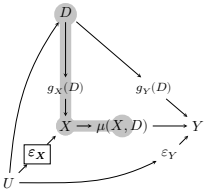


Purely through $\boxed{Y}$

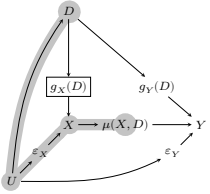


Through a conf. of $D \to Y$

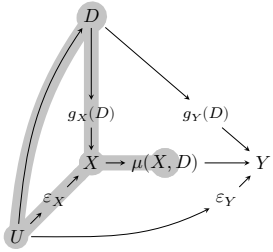
Sources of disparate impact



Purely through X



Through an conf. of $D \to X$

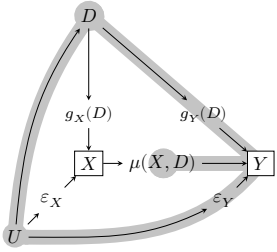


Disp. Impact through X

Mitigation leads to

Solidarity

(Levelled premiums)

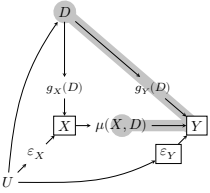


Disp. Impact through $\boxed{Y}$

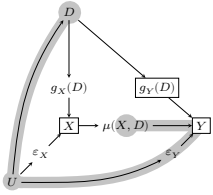
Mitigation leads to

Actuarial Fairness

(Levelled profits)



Purely through $\boxed{Y}$



Through a conf. of $D \to Y$

Discrimination is inevitable

Proposition. Unless $D \perp\!\!\!\perp Y$ and $D \perp\!\!\!\perp Y|X$, all premiums disparately impact D.

Discrimination is inevitable

Proposition. Unless $D \perp\!\!\!\perp Y$ and $D \perp\!\!\!\perp Y|X$, all premiums disparately impact D.

Corollary.

Unless $D \perp\!\!\!\perp Y$ and $D \perp\!\!\!\perp Y|X$, all premiums **must discriminate** either directly or indirectly on D.

Discrimination is inevitable

Proposition. Unless $D \perp\!\!\!\perp Y$ and $D \perp\!\!\!\perp Y|X$, all premiums disparately impact D.

Corollary.

Unless $D \perp\!\!\!\perp Y$ and $D \perp\!\!\!\perp Y|X$, all premiums **must discriminate** either directly or indirectly on D.

The **impossibility** theorem by Kleinberg et al. (2016) **extends to discrimination**.

Examples from the literature

Premium of Lindholm et al. (2022)

Premium of Frees and Huang (2023)

Examples from the literature

Premium of Lindholm et al. (2022)

The premium is

$$\pi^L(X) = \mathrm{E}_D\{\mathrm{E}(Y|X,D)\}$$

Premium of Frees and Huang (2023)

For $X^* = (\mathbf{I} - H_D)X$, the premium is

$$\pi^F(X,D) = X^\top(\mathbf{I} - H_D)(X^{*\top}X^*)^{-1}X^{*\top}Y.$$

Examples from the literature

Premium of Lindholm et al. (2022)

The premium is

$$\pi^L(X) = \mathrm{E}_D\{\mathrm{E}(Y|X,D)\}$$

It is a **causal inference** technique.

Premium of Frees and Huang (2023)

For $X^* = (\mathbf{I} - H_D)X$, the premium is

$$\pi^F(X,D) = X^\top(\mathbf{I} - H_D)(X^{*\top}X^*)^{-1}X^{*\top}Y.$$

X^* is an **estimator of** ε_X.

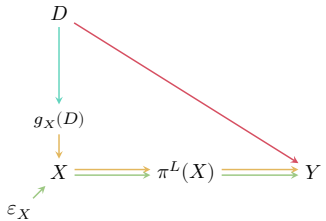
Examples from the literature

Premium of Lindholm et al. (2022)

The premium is

$$\pi^L(X) = \mathrm{E}_D\{\mathrm{E}(Y|X,D)\}$$

It is a **causal inference** technique.

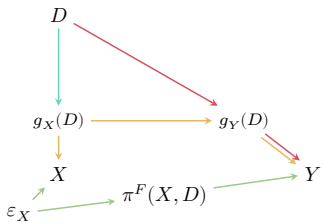


Premium of Frees and Huang (2023)

For $X^* = (\mathbf{I} - H_D)X$, the premium is

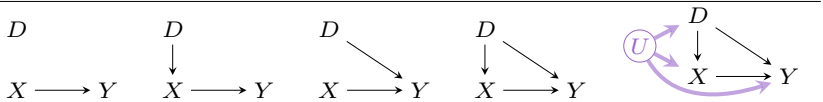
$$\pi^F(X,D) = X^\top(\mathbf{I} - H_D)(X^{*\top}X^*)^{-1}X^{*\top}Y.$$

X^* is an **estimator of** ε_X.



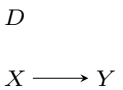
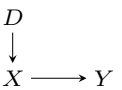
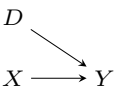
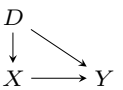
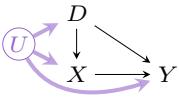
Discrimination for the premium $\pi^L(X)$

Data generating mechanism

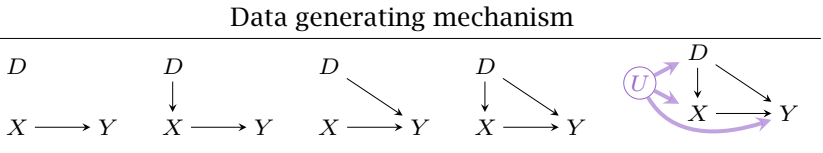


Discr.						
	Direct	○	○	○	○	○
	Indirect	○	●	●	●	●

Discrimination for the premium $\pi^L(X)$

			Data generating mechanism				
							
Discr.		Direct	○	○	○	○	○
		Indirect	○	●	●	●	●
Disp. Impact	X	Purely	○	●	○	●	●
		Conf.	○	○	○	○	●
	$\boxed{Y}$	Purely	○	○	●	●	●
		Conf.	○	○	○	○	●

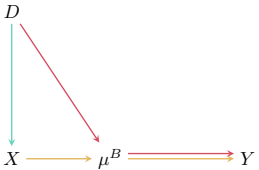
Discrimination for the premium $\pi^F(X, D)$



Discr.		Direct	○	●	○	●	●
		Indirect	○	○	●	○	○
Disp. Impact	X	Purely	○	○	○	○	○
		Conf.	○	○	○	○	●
	$\boxed{Y}$	Purely	○	●	●	●	●
		Conf.	○	○	○	○	●

Best-estimate premiums

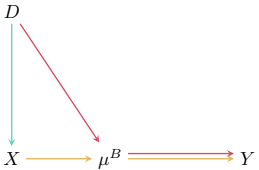
The **best-estimate** premium $(X, D) \mapsto \mu^B(X, D)$ uses (X, D) to predict Y.



$$\mu^B(x, d) = E(Y|X = x, D = d)$$

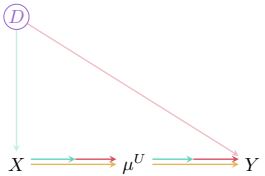
Best-estimate, unaware premiums

The **best-estimate** premium $(X, D) \mapsto \mu^B(X, D)$ uses (X, D) to predict Y.



$$\mu^B(x, d) = E(Y|X = x, D = d)$$

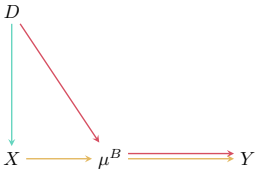
The **unaware** premium $X \mapsto \mu^I(X)$ uses X, ignoring D, to predict Y.



$$\mu^U(x) = E(Y|X = x)$$

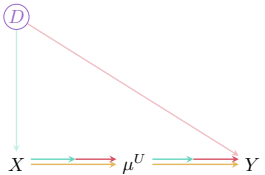
Best-estimate, unaware, and aware premiums

The **best-estimate** premium $(X, D) \mapsto \mu^B(X, D)$ uses (X, D) to predict Y.



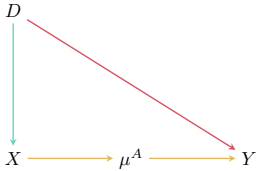
$$\mu^B(x, d) = E(Y|X = x, D = d)$$

The **unaware** premium $X \mapsto \mu^I(X)$ uses X, ignoring D, to predict Y.



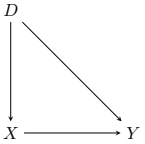
$$\mu^U(x) = E(Y|X = x)$$

The **aware** premium $X \mapsto \mu^A(X)$ uses X to predict Y while controling for D.

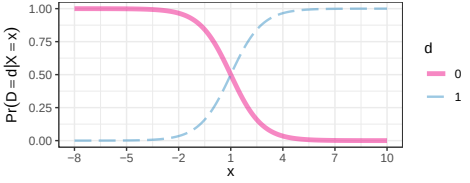


$$\mu^A(x) = E_D\{E(Y|X = x, D)\}$$

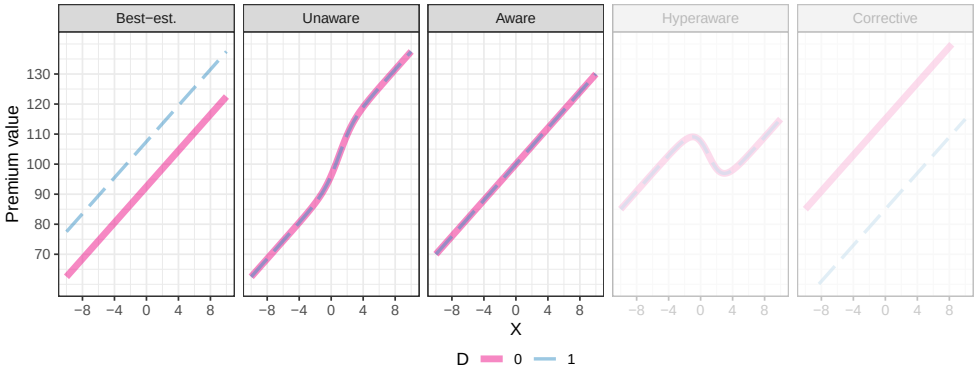
Configuration of the example



- Let $D \in \{0, 1\}$ be Bernoulli with $\Pr(D = 1) = 0.5$.
- The variables X and Y are Gaussian and the DAG is satisfied.

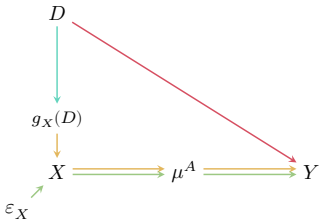


Premiums in terms of x and d



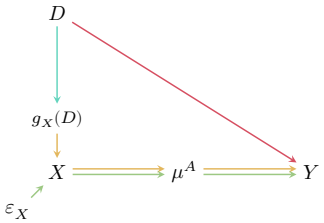
Hyperaware and corrective premiums

The **aware** premium $X \mapsto \mu^A(X)$ uses all of X to predict Y while correcting for the omitted variable bias associated with D.

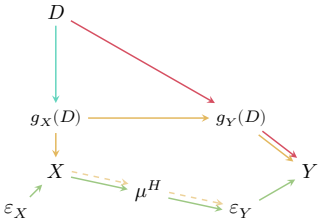


Hyperaware and corrective premiums

The **aware** premium $X \mapsto \mu^A(X)$ uses all of X to predict Y while correcting for the omitted variable bias associated with D.

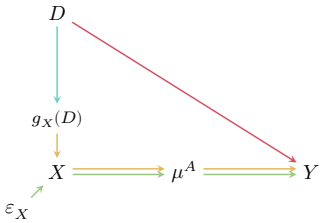


A **hyperaware** premium $X \mapsto \mu^H(X)$ use all of X to predict ε_Y.

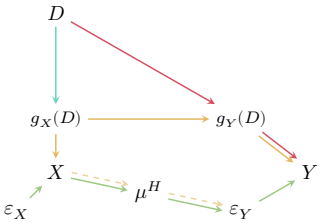


Hyperaware and corrective premiums

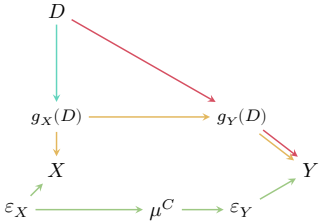
The **aware** premium $X \mapsto \mu^A(X)$ uses all of X to predict Y while correcting for the omitted variable bias associated with D.



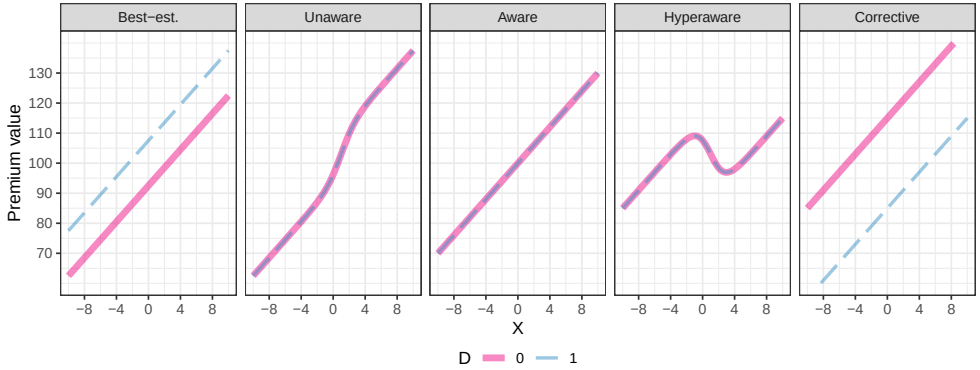
A **hyperaware** premium $X \mapsto \mu^H(X)$ use all of X to predict ε_Y.



The **corrective** premium $\varepsilon_X \mapsto \mu^C(\varepsilon_X)$ models the residual information independent from D.



Premiums in terms of x and d



Our classification : a few examples

Expected premium	Type of intervention: *Pre-processing*	*In-processing*	*Post-processing*
Aware	Plečko et al. (2021) Charpentier et al. (2023a)	Lindholm et al. (2024) Gabric et al. (2024)	Pope and Sydnor (2011) Aseervatham et al. (2016) **Lindholm et al. (2022)**
Hyperaware	Zemel et al. (2013)	Grari et al. (2022)	
Corrective	Calmon et al. (2017) Komiyama and Shimao (2017) **Frees and Huang (2023)**	Fermanian and Guegan (2021)	Charpentier et al. (2023b) Hu et al. (2023) Hu et al. (2024)

Our classification : a few examples

Expected premium (*what*)	Type of intervention (*when*)		
	Pre-processing	***In-processing***	***Post-processing***
Aware	Plečko et al. (2021) Charpentier et al. (2023a)	Lindholm et al. (2024) Gabric et al. (2024)	Pope and Sydnor (2011) Aseervatham et al. (2016) **Lindholm et al. (2022)**
Hyperaware	Zemel et al. (2013)	Grari et al. (2022)	
Corrective	Calmon et al. (2017) Komiyama and Shimao (2017) **Frees and Huang (2023)**	Fermanian and Guegan (2021)	Charpentier et al. (2023b) Hu et al. (2023) Hu et al. (2024)

So far

- We detailed the causal graph used for fairness in insurance to adequately place the different fair premiums in the graph.

So far

- We detailed the causal graph used for fairness in insurance to adequately place the different fair premiums in the graph.
- We defined direct and indirect discrimination for a premium.

So far

- We detailed the causal graph used for fairness in insurance to adequately place the different fair premiums in the graph.
- We defined direct and indirect discrimination for a premium.
- We define five families of premiums according to their fairness properties, therefore giving a **new categorization** of fairness methodologies.

So far

- We detailed the causal graph used for fairness in insurance to adequately place the different fair premiums in the graph.
- We defined direct and indirect discrimination for a premium.
- We define five families of premiums according to their fairness properties, therefore giving a **new categorization** of fairness methodologies.

How does this framework make a difference **in practice**?

A scalable toolbox for exposing indirect discrimination

1 Exploiting causal graphs for fairness

2 A scalable toolbox for exposing indirect discrimination
- What does it mean to be fair ?
- Can one benchmark for fairness ?
- Is unfairness truly material ?
- Where indirect discrimination concentrates ?

Context around this project

Context around this project

Figure 2 – Journée de collaboration au local **Desjardins** du MILA, 20 février 2025, Montréal

Context around this project

Figure 2 – Journée de collaboration au local **Desjardins** du MILA, 20 février 2025, Montréal

Outlook

Congratulations on Achieving Your ACAS! CRM:0100683

De CAS Portal - NO REPLY <cobaltams@casact.org>
Date Ven 2025-06-06 12:26
À Olivier Côté <olivierc_actuariat@outlook.com>

Olivier Côté, ACAS
870 av Marguerite-Bourgeoys
Lévis QC G1S3X2
Canada

Expertise. Insight. Solutions.®

Dear Olivier,

On behalf of the Casualty Actuarial Society, it is our pleasure to notify you that you are now an Associate of the CAS. You may begin using the designation "Associate, Casualty Actuarial Society" or "ACAS" immediately. CAS credentials are unmatched for their rigor, integrity, and relevance, making it the gold standard among property & casualty actuarial credentials. Congratulations on attaining this milestone in your actuarial career!

Sincerely,

David Cummings, FCAS
President

Victor Carter-Bey, DM
Chief Executive Officer

Motivation

- There is growing interest in ensuring fairness in insurance pricing.
  - E.g. Financial Services Regulatory Authority of Ontario (2024)

Motivation

- There is growing interest in ensuring fairness in insurance pricing.
  - E.g. Financial Services Regulatory Authority of Ontario (2024)

- Fairness in insurance is difficult to operationalize due to its ambiguity.
  - « there is no general agreement on what constitutes an “equitable” classification system or “fair” discrimination. » — ASOP No. 12, ASB (2005)

Motivation

- There is growing interest in ensuring fairness in insurance pricing.
  - E.g. Financial Services Regulatory Authority of Ontario (2024)
- Fairness in insurance is difficult to operationalize due to its ambiguity.
  - « there is no general agreement on what constitutes an “equitable” classification system or “fair” discrimination. » — ASOP No. 12, ASB (2005)
- Current fairness metrics are not suited for actuarial needs.
  - They don’t identify problematic segments or variables.
  - Their units are abstract.

Motivation

- There is growing interest in ensuring fairness in insurance pricing.
  - E.g. Financial Services Regulatory Authority of Ontario (2024)
- Fairness in insurance is difficult to operationalize due to its ambiguity.
  - « there is no general agreement on what constitutes an “equitable” classification system or “fair” discrimination. » — ASOP No. 12, ASB (2005)
- Current fairness metrics are not suited for actuarial needs.
  - They don’t identify problematic segments or variables.
  - Their units are abstract.
- Fairness is debated in theory, but **unfairness unfolds in practice**.
  - Few Canadian insurance case studies exist.

Vulnerability to proxies in Québec auto insurance

Objective : Evaluate a given **pseudo price** for fairness regarding financial vulnerability in material damage premiums for at-fault accidents (Chapter B2) in Quebec.

Vulnerability to proxies in Québec auto insurance

Objective : Evaluate a given **pseudo price** for fairness regarding financial vulnerability in material damage premiums for at-fault accidents (Chapter B2) in Quebec.

Data : $\approx$ 768,000 insured vehicles in Quebec, from 2016–2017. Data obtained via insurer partnership.

Note : Personal data anonymized ; strict confidentiality measures applied.

Data overview

Notation	Concept	Domain	Notes
Y	Claim amount ($)	$\mathbb{R}^+$	$\overline{Y} \approx 200$, with 97% at 0

Data overview

Notation	Concept	Domain	Notes
Y	Claim amount ($)	$\mathbb{R}^+$	$\overline{Y} \approx 200$, with 97% at 0
D	Low credit indicator	$\{0,1\}$	1 indicates low credit, with $\overline{D} \approx 0.40$

Data overview

Notation	Concept	Domain	Notes
Y	Claim amount ($)	$\mathbb{R}^+$	$\overline{Y} \approx 200$, with 97% at 0
D	Low credit indicator	$\{0,1\}$	1 indicates low credit, with $\overline{D} \approx 0.40$
$\mathbf{X}$	Policyholder info	Dim. 16	E.g., gender, driving experience, mileage, education, occupation
	Geographic info	Dim. 4	E.g., FSA and territorial risk score
	Vehicle info	Dim. 4	E.g., vehicle age, new purchase, vehicle risk score
	Policy info	Dim. 3	E.g., home insurance, endorsements

Data overview

Notation	Concept	Domain	Notes
Y	Claim amount ($)	$\mathbb{R}^+$	$\overline{Y} \approx 200$, with 97% at 0
D	Low credit indicator	$\{0,1\}$	1 indicates low credit, with $\overline{D} \approx 0.40$
$\mathbf{X}$	Policyholder info	Dim. 16	E.g., gender, driving experience, mileage, education, occupation
	Geographic info	Dim. 4	E.g., FSA and territorial risk score
	Vehicle info	Dim. 4	E.g., vehicle age, new purchase, vehicle risk score
	Policy info	Dim. 3	E.g., home insurance, endorsements

We aim to analyze **PseudoPrice(x,d)**.

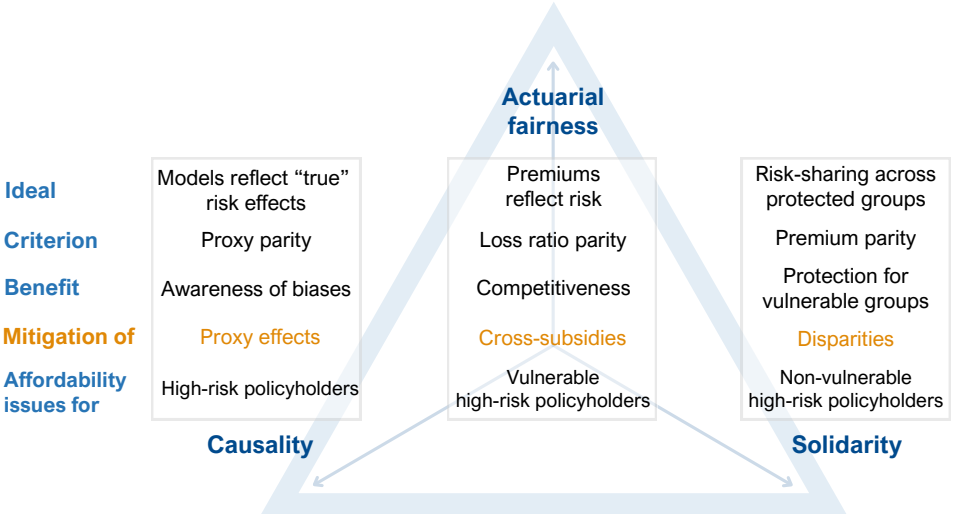
Data overview

Notation	Concept	Domain	Notes
Y	Claim amount ($)	$\mathbb{R}^+$	$\overline{Y} \approx 200$, with 97% at 0
D	Low credit indicator	$\{0,1\}$	1 indicates low credit, with $\overline{D} \approx 0.40$
$\mathbf{X}$	Policyholder info	Dim. 16	E.g., gender, driving experience, mileage, education, occupation
	Geographic info	Dim. 4	E.g., FSA and territorial risk score
	Vehicle info	Dim. 4	E.g., vehicle age, new purchase, vehicle risk score
	Policy info	Dim. 3	E.g., home insurance, endorsements

We aim to analyze **PseudoPrice(x,d)**.

The **non-equivalence of the pseudoprice to the actual pricing function** precludes any conclusion regarding the fairness of the partner insurer's pricing.

The dimensions of fairness



Actuarial Fairness

A premium is actuarially fair if “it represents an unbiased estimate of the expected value of all future costs associated with the risk transfer” (Casualty Actuarial Society, 1988).

Actuarial Fairness

A premium is actuarially fair if “it represents an unbiased estimate of the expected value of all future costs associated with the risk transfer” (Casualty Actuarial Society, 1988).

- Premium matches expected cost.

Actuarial Fairness

A premium is actuarially fair if "it represents an unbiased estimate of the expected value of all future costs associated with the risk transfer" (Casualty Actuarial Society, 1988).

- Premium matches expected cost.
- Self-sustaining loss ratios (no cross-subsidies).

Actuarial Fairness

A premium is actuarially fair if "it represents an unbiased estimate of the expected value of all future costs associated with the risk transfer" (Casualty Actuarial Society, 1988).

- Premium matches expected cost.
- Self-sustaining loss ratios (no cross-subsidies).
- Avoids **non-risk-based adjustments.**

Solidarity

Solidarity is the foundation of insurance : pooling and sharing risks among **all** policyholders.

Solidarity

Solidarity is the foundation of insurance : pooling and sharing risks among **all** policyholders.

In the presence of admissible variables $\mathbf{X}$ and sensitive ones D,

- we can leverage the effect of $\mathbf{X}$ on the response variable Y
- while aiming for solidarity on D : equal premiums (in expectation or distribution) across all protected groups.

This is referred to as **demographic parity of premiums**.

Causality and proxy effects

Avoiding proxy effects requires two actions :

- Exclude factors that do not determine risk,
- Limit effect of risk factors to their “**true**” risk relevance.

Causality and proxy effects

Avoiding proxy effects requires two actions :

- Exclude factors that do not determine risk,
- Limit effect of risk factors to their “**true**” risk relevance.

Even valid risk factors can suffer from proxy effects. (e.g., territory and ethnicity)

A variable’s use – not the variable itself – determines its role as a proxy.

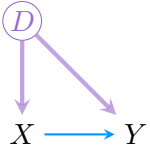
Causality and proxy effects

Avoiding proxy effects requires two actions :

- Exclude factors that do not determine risk,
- Limit effect of risk factors to their "**true**" risk relevance.

> Even valid risk factors can suffer from proxy effects. (e.g., territory and ethnicity)
>
> A variable's use – not the variable itself – determines its role as a proxy.

In fairness analysis with respect to D, causality seeks to identify the **effect of X on Y** without **proxy effects from D**.



Five fair benchmarks

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	$\mathbb{E}(Y\|\mathbf{X}=\mathbf{x}, D=d)$	$\mathbb{E}(Y\|\mathbf{X}=\mathbf{x})$	$\mathbb{E}_D\{\mu^B(\mathbf{x}, D)\}$	$\mathbb{E}\{\mu^C(\mathbf{x}, D)\|\mathbf{X}=\mathbf{x}\}$	$\mathcal{T}^{d\to\star}\{\mu^B(\mathbf{x}, d)\}$
Direct discrimination	✔	✘	✘	✘	✔
Proxy discrimination	–	✔	✘	✔	–
Demographic disparities	✔	✔	✔	✘	✘
Pillar	**AF**	**AF**	**C**	**S**	**S**

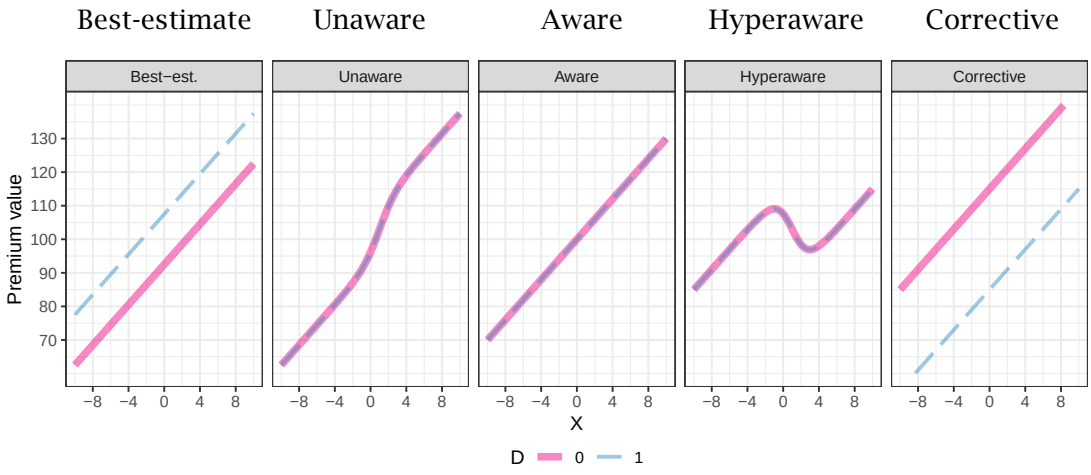
Five fair benchmarks

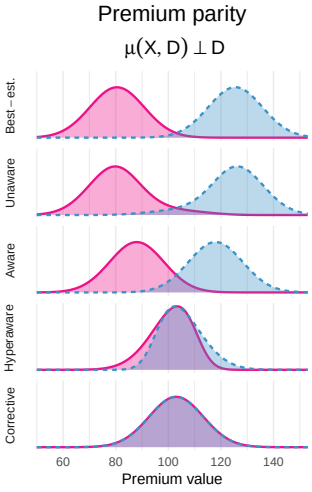
Premium	Best-estimate	Unaware	Aware	Hyperaware	Corrective
Notation	$\mu^B(\mathbf{x}, d)$	$\mu^U(\mathbf{x})$	$\mu^A(\mathbf{x})$	$\mu^H(\mathbf{x})$	$\mu^C(\mathbf{x}, d)$
Formula	$\mathbb{E}(Y\|\mathbf{X}=\mathbf{x}, D=d)$	$\mathbb{E}(Y\|\mathbf{X}=\mathbf{x})$	$\mathbb{E}_D\{\mu^B(\mathbf{x}, D)\}$	$\mathbb{E}\{\mu^C(\mathbf{x}, D)\|\mathbf{X}=\mathbf{x}\}$	$\mathcal{T}^{d\to\star}\{\mu^B(\mathbf{x}, d)\}$
Direct discrimination	✔	✘	✘	✘	✔
Proxy discrimination	–	✔	✘	✔	–
Demographic disparities	✔	✔	✔	✘	✘
Pillar	**AF**	**AF**	**C**	**S**	**S**

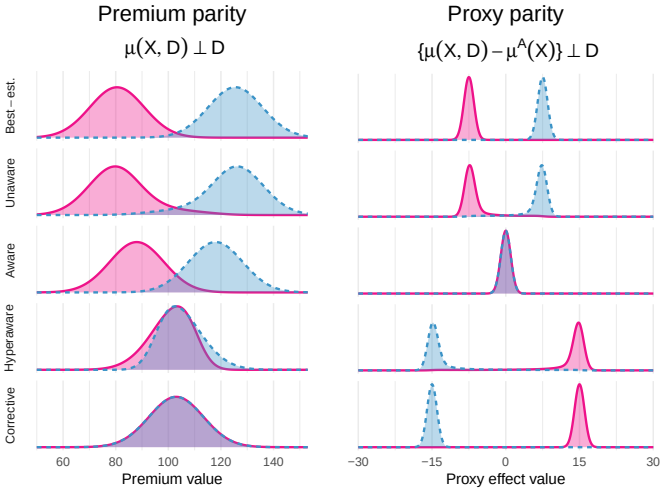
We estimate the premium spectrum using :

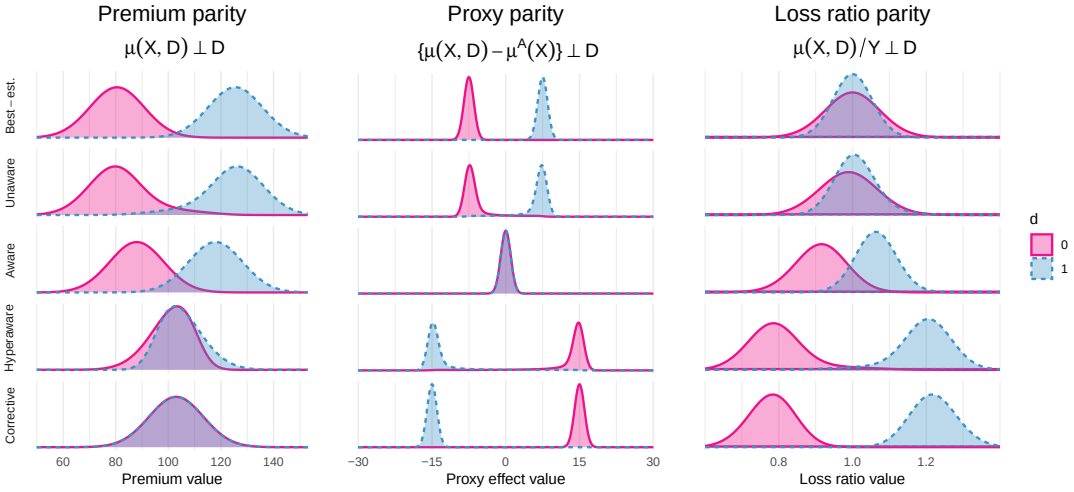
- `lightGBM` (Ke et al., 2017) to learn conditional expectations,
- empirical marginals of D for population-level integration, and
- optimal transport mappings via `Equipy` (Fernandes Machado et al., 2025).

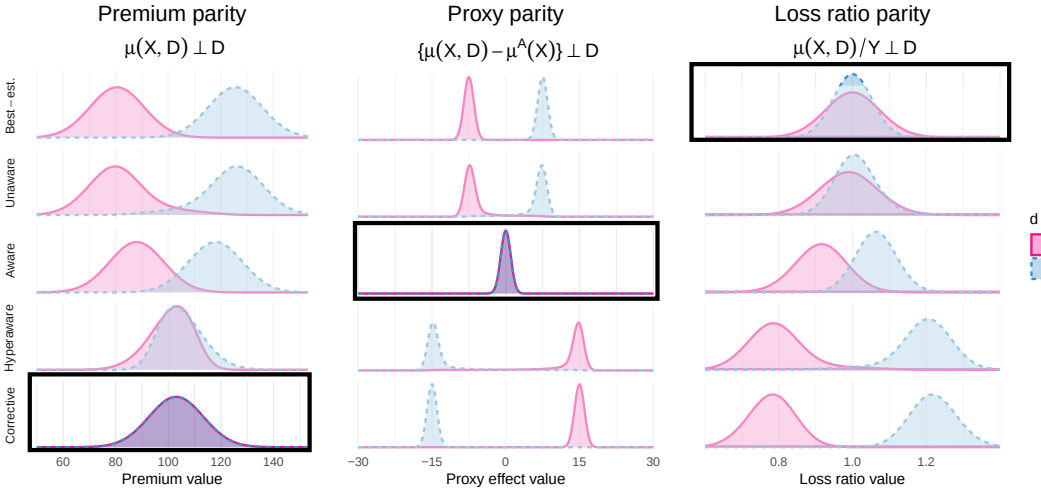
Example : Premiums in terms of x and d

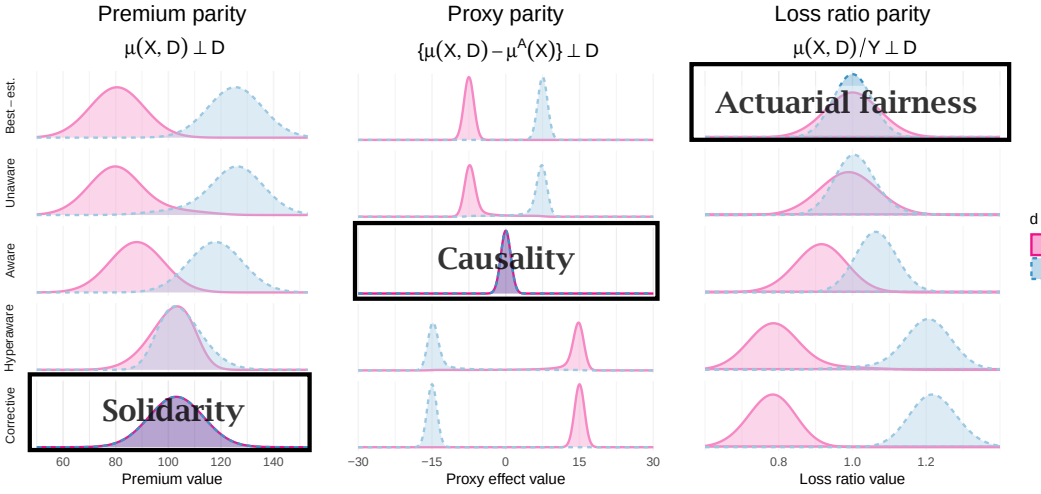


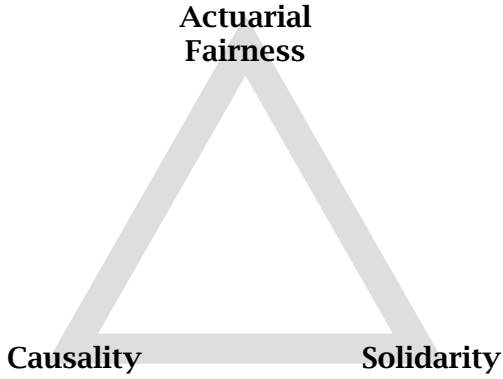

Premium parity
μ(X, D) ⊥ D
Best − est.
Unaware
Aware
Hyperaware
Corrective
60
80
100
120
140
Premium value
d
0
1

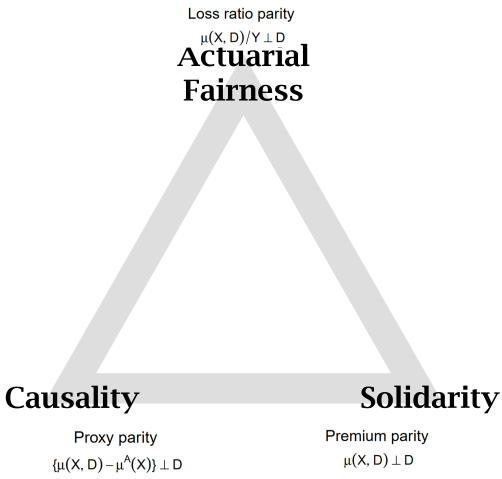

Premium parity
$\mu(X, D) \perp D$
Proxy parity
$\{\mu(X, D) - \mu^A(X)\} \perp D$
Best – est.
Unaware
Aware
Hyperaware
Corrective
60
80
100
120
140
Premium value
−30
−15
0
15
30
Proxy effect value
d
0
1

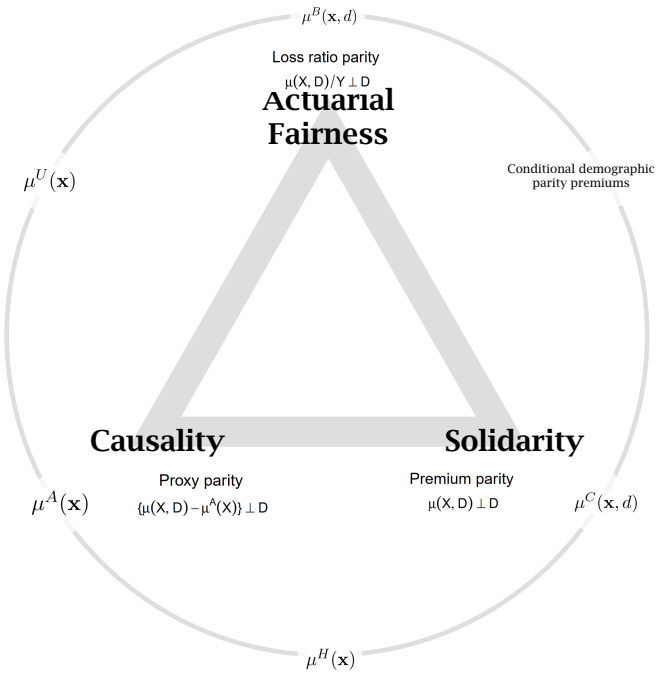

Premium parity
$\mu(X, D) \perp D$
Proxy parity
$\{\mu(X, D) - \mu^A(X)\} \perp D$
Loss ratio parity
$\mu(X, D)/Y \perp D$
Best – est.
Unaware
Aware
Hyperaware
Corrective
60
80
100
120
140
Premium value
−30
−15
0
15
30
Proxy effect value
0.8
1.0
1.2
Loss ratio value
d
0
1


Premium parity
$\mu(X, D) \perp D$
Proxy parity
$\{\mu(X, D) - \mu^A(X)\} \perp D$
Loss ratio parity
$\mu(X, D)/Y \perp D$
Best – est.
Unaware
Aware
Hyperaware
Corrective
60
80
100
120
140
Premium value
−30
−15
0
15
30
Proxy effect value
0.8
1.0
1.2
Loss ratio value
d
0
1

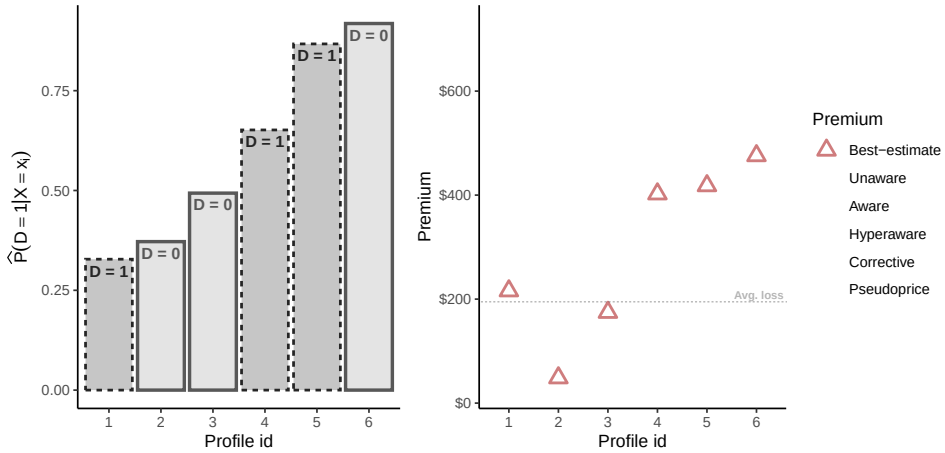

Premium parity
$\mu(X, D) \perp D$
Proxy parity
$\{\mu(X, D) - \mu^A(X)\} \perp D$
Loss ratio parity
$\mu(X, D)/Y \perp D$
Actuarial fairness
Causality
Solidarity
Best – est.
Unaware
Aware
Hyperaware
Corrective
60
80
100
120
140
Premium value
−30
−15
0
15
30
Proxy effect value
0.8
1.0
1.2
Loss ratio value
d
0
1


Actuarial
Fairness
Causality
Solidarity

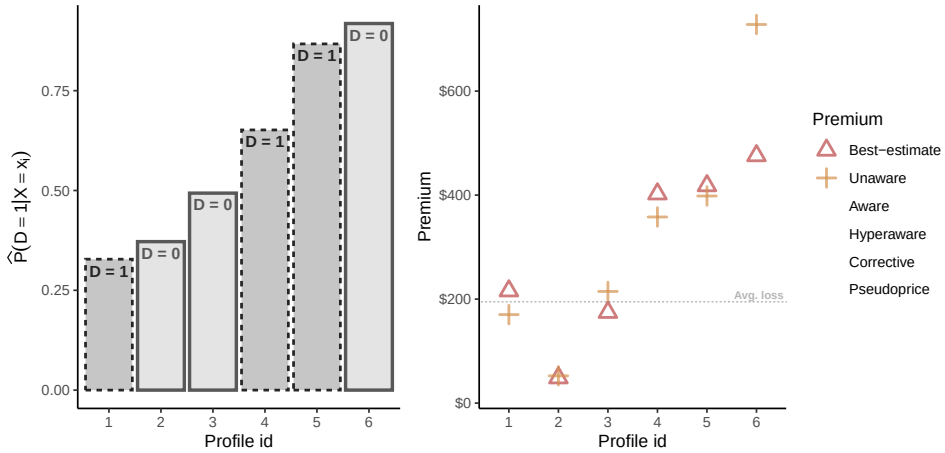

Loss ratio parity
$\mu(X, D)/Y \perp D$
Actuarial Fairness
Causality
Solidarity
Proxy parity
$\{\mu(X, D) - \mu^{A}(X)\} \perp D$
Premium parity
$\mu(X, D) \perp D$


$\mu^B(\mathbf{x}, d)$
Loss ratio parity
$\mu(X, D)/Y \perp D$
Actuarial Fairness
$\mu^U(\mathbf{x})$
Conditional demographic parity premiums
Causality
Solidarity
Proxy parity
$\{\mu(X, D) - \mu^A(X)\} \perp D$
Premium parity
$\mu(X, D) \perp D$
$\mu^A(\mathbf{x})$
$\mu^C(\mathbf{x}, d)$
$\mu^H(\mathbf{x})$

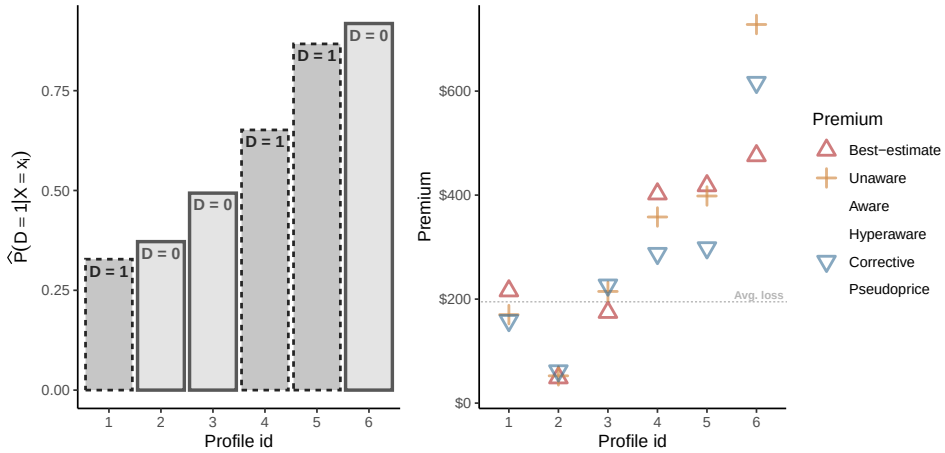
Fairness range for six profiles



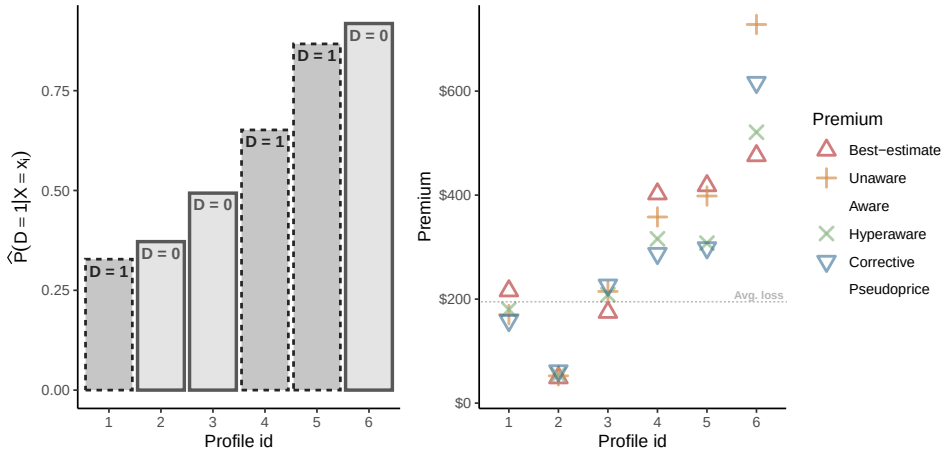
Fairness range for six profiles



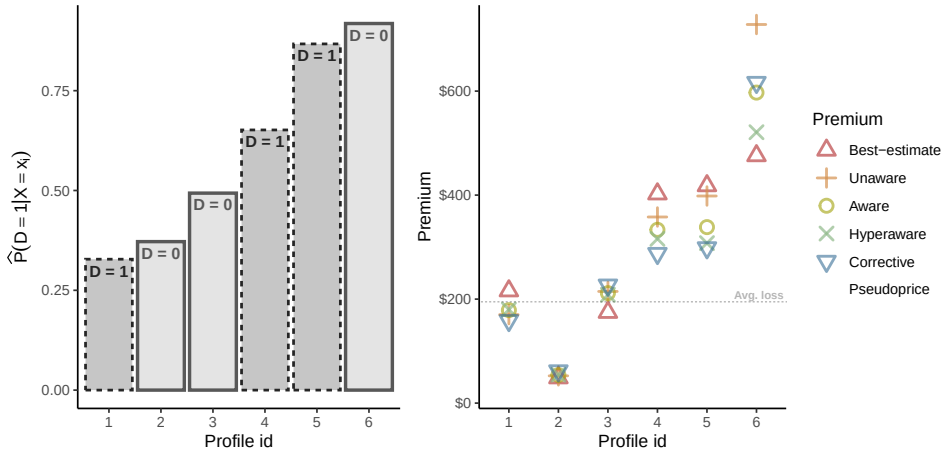
Fairness range for six profiles



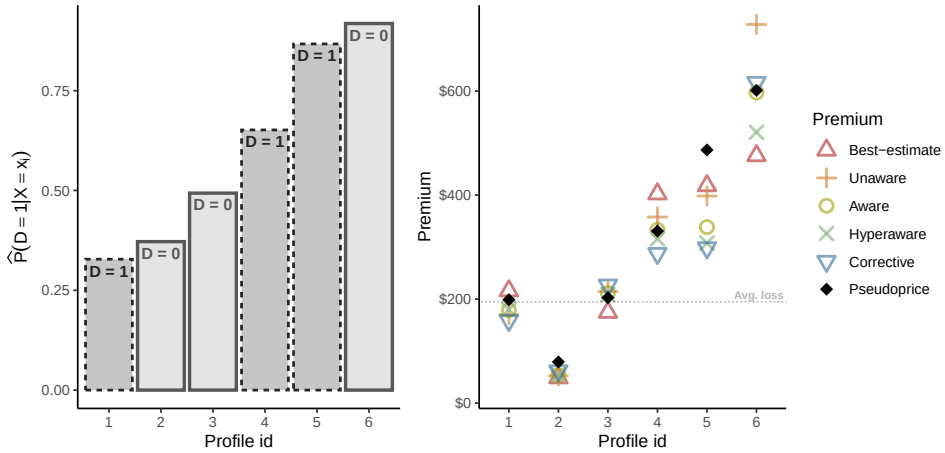
Fairness range for six profiles



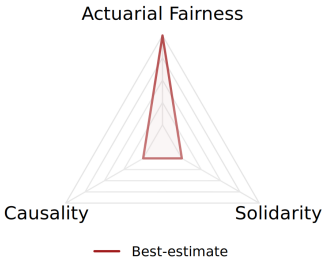
Fairness range for six profiles

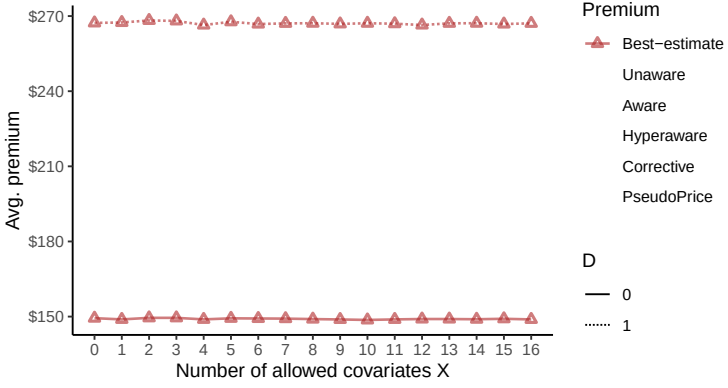


Fairness range for six profiles

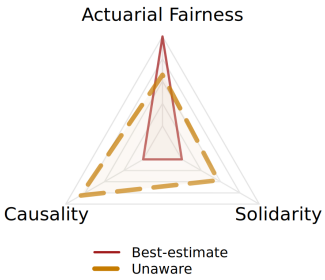


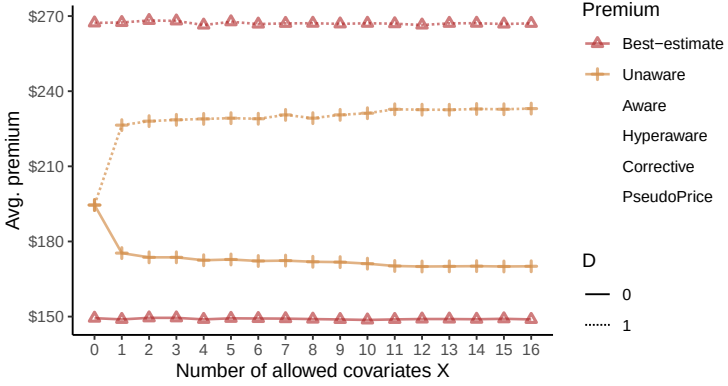
Grasping behavior of families



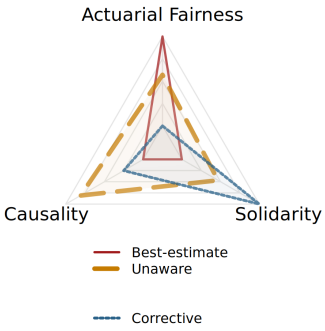


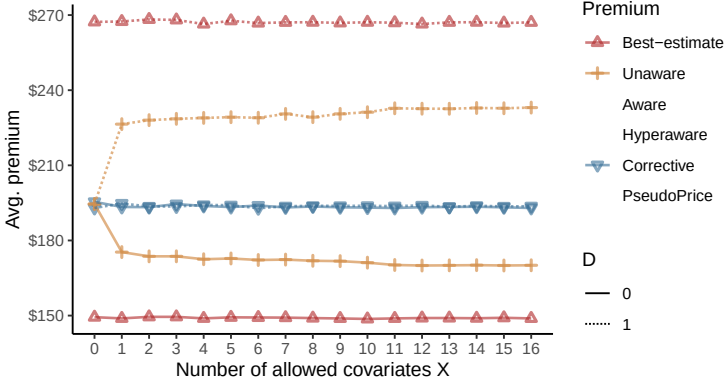
Grasping behavior of families



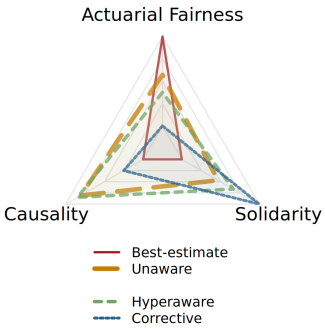


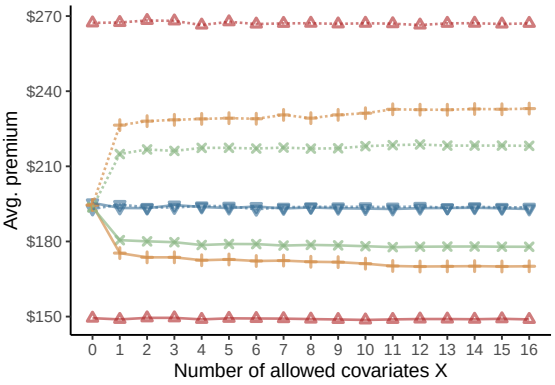
Grasping behavior of families



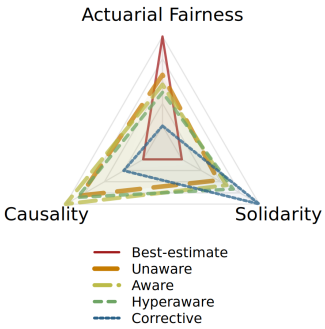


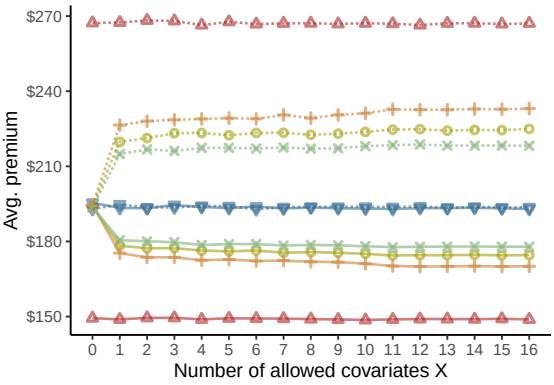
Grasping behavior of families



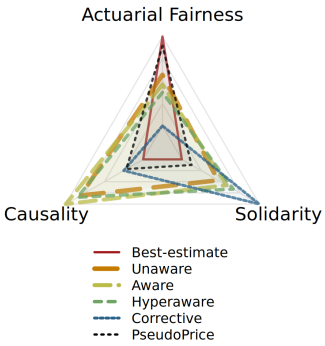


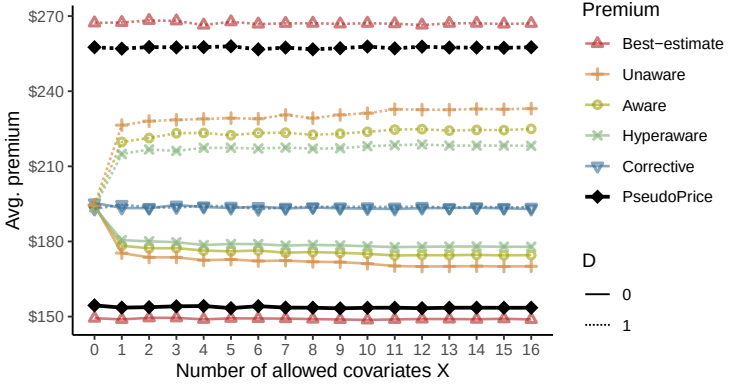
Grasping behavior of families



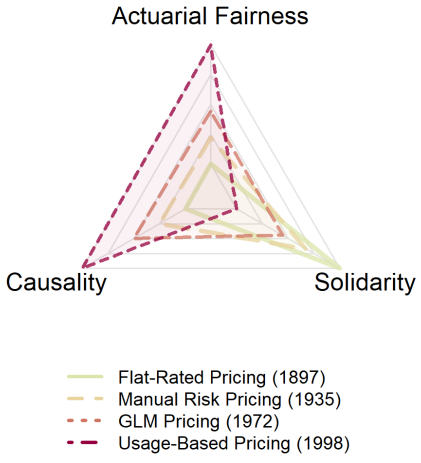


Grasping behavior of families



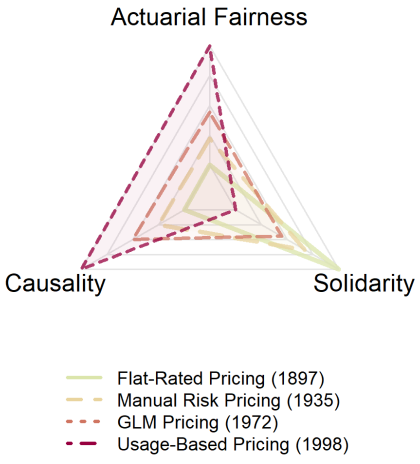


The history of risk classification revisited



The history of risk classification revisited

As data granularity increases, so does the potential for actuarial justification in perpetuating disparities.



Prior-pricing local fairness metrics

The potential for disparate treatment on D is the **risk spread** :

$$\Delta_{\text{risk}}(\mathbf{x}) = \sup_{d\in\mathcal{D}} \mu^B(\mathbf{x}, d) - \inf_{d\in\mathcal{D}} \mu^B(\mathbf{x}, d).$$

Prior-pricing local fairness metrics

The potential for disparate treatment on D is the **risk spread** :

$$\Delta_{\text{risk}}(\mathbf{x}) = \sup_{d\in\mathcal{D}} \mu^B(\mathbf{x}, d) - \inf_{d\in\mathcal{D}} \mu^B(\mathbf{x}, d).$$

The vulnerability of a segment $\mathbf{x}$ to proxy effect is the **proxy vulnerability** :

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

Prior-pricing local fairness metrics

The potential for disparate treatment on D is the **risk spread** :

$$\Delta_{\text{risk}}(\mathbf{x}) = \sup_{d\in\mathcal{D}} \mu^B(\mathbf{x}, d) - \inf_{d\in\mathcal{D}} \mu^B(\mathbf{x}, d).$$

The vulnerability of a segment $\mathbf{x}$ to proxy effect is the **proxy vulnerability** :

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

The **fairness range** shows how much prices vary across fairness methods

$$\Delta_{\text{fair}}(\mathbf{x}, d) = \sup_{a\in\{B,U,A,H,C\}} \mu^a(\mathbf{x}, d) - \inf_{b\in\{B,U,A,H,C\}} \mu^b(\mathbf{x}, d)$$

The **parity cost** is the (monetary) cost of enforcing premium parity :

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

Prior-pricing local fairness metrics

The potential for disparate treatment on D is the **risk spread** :

$$\Delta_{\text{risk}}(\mathbf{x}) = \sup_{d\in\mathcal{D}} \mu^B(\mathbf{x}, d) - \inf_{d\in\mathcal{D}} \mu^B(\mathbf{x}, d).$$

The vulnerability of a segment $\mathbf{x}$ to proxy effect is the **proxy vulnerability** :

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

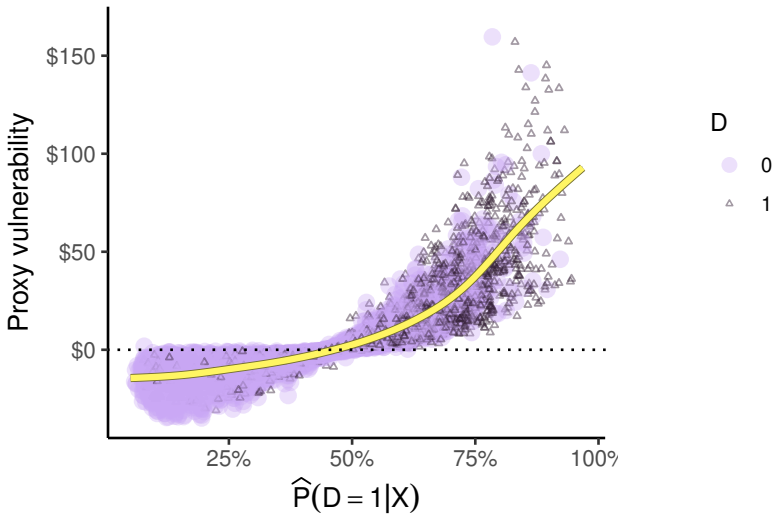
The **fairness range** shows how much prices vary across fairness methods

$$\Delta_{\text{fair}}(\mathbf{x}, d) = \sup_{a\in\{B,U,A,H,C\}} \mu^a(\mathbf{x}, d) - \inf_{b\in\{B,U,A,H,C\}} \mu^b(\mathbf{x}, d)$$

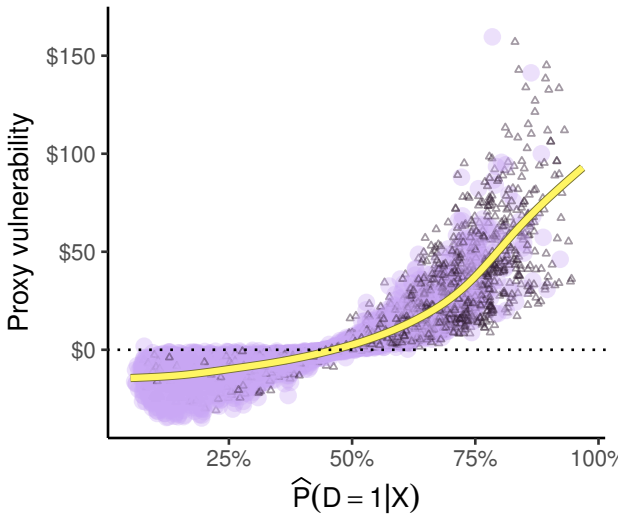
The **parity cost** is the (monetary) cost of enforcing premium parity :

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

Visualizing the proxy vulnerability

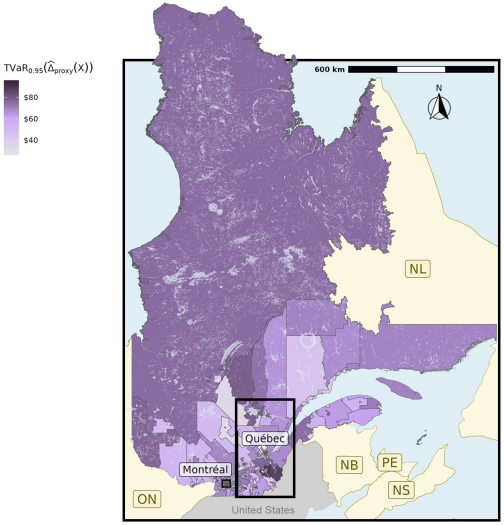


Visualizing the proxy vulnerability

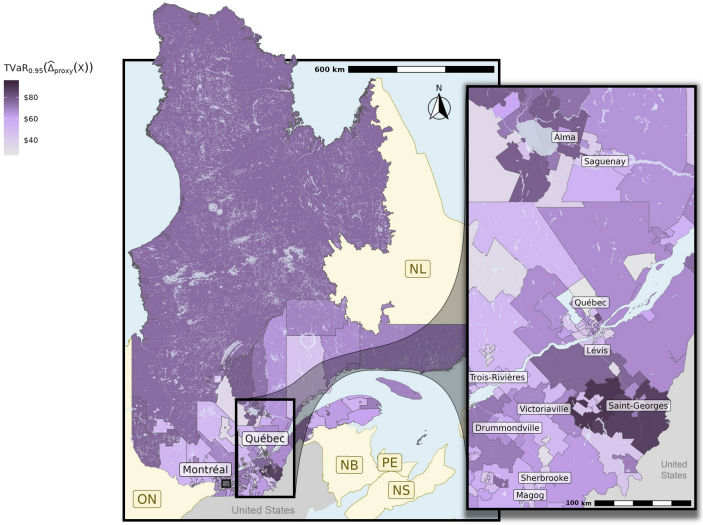


Proxy vulnerability is both material and skewed

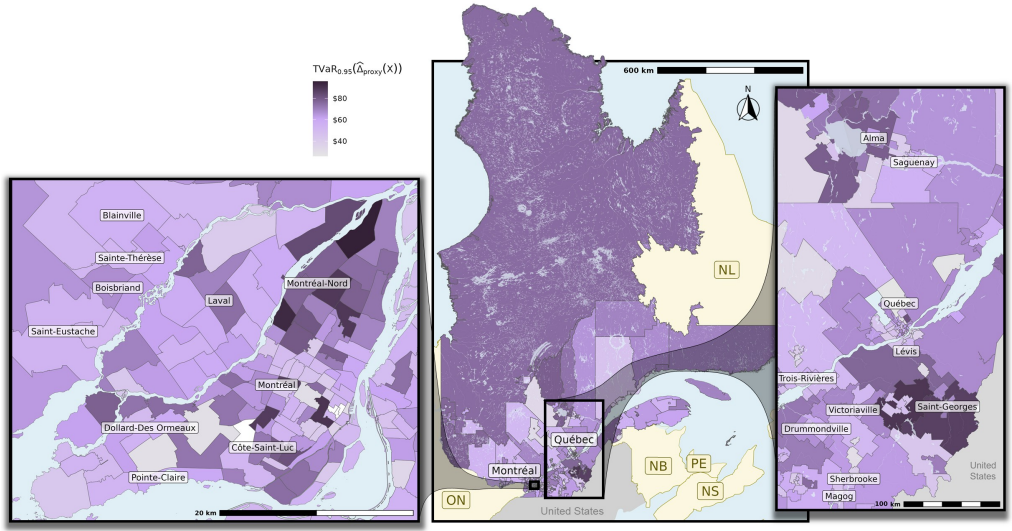
Geographic distribution of the 95% TVaR of proxy vulnerability



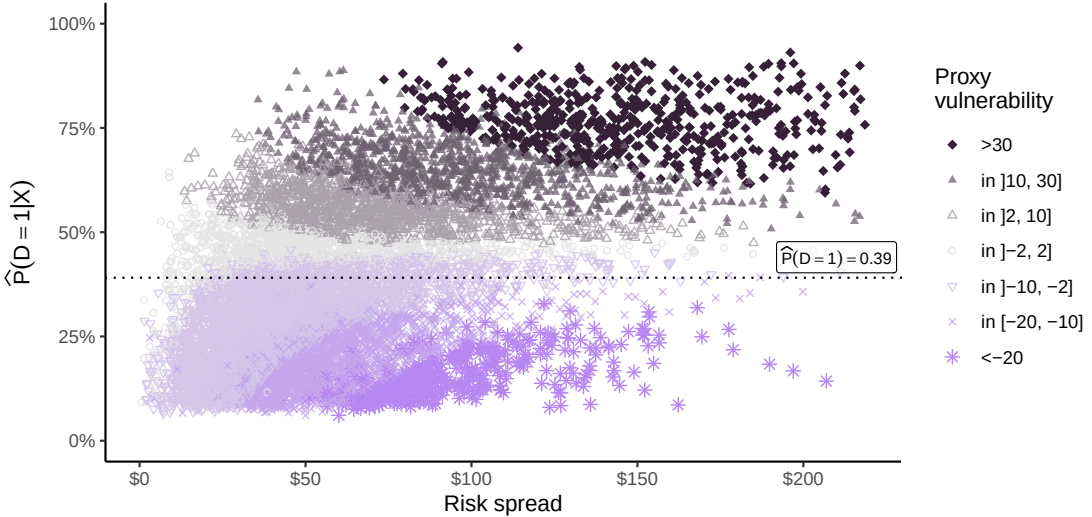
Geographic distribution of the 95% TVaR of proxy vulnerability



Geographic distribution of the 95% TVaR of proxy vulnerability



Decomposing proxy vulnerability



Local metrics between commercial premium and fair benchmarks

The **commercial loading** relative to the reference premium μ^A is :

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

The **implied propensity** is the implicit* weight of $D = 1$ in $\pi(\mathbf{x})$:

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

The **excess lift** measures excessive price differentiation on D :

$$\Delta_{\text{excess}}(\mathbf{x}; \pi) = \left\{ \sup_d \pi(\mathbf{x}, d) - \inf_d \pi(\mathbf{x}, d) \right\} - \Delta_{\text{risk}}(\mathbf{x}).$$

Revealing disparities via partitioning

Problem : Fairness metrics over large groups can hide disparities within vulnerable subgroups.

Revealing disparities via partitioning

Problem : Fairness metrics over large groups can hide disparities within vulnerable subgroups.

Solution : Partition policyholders based on fairness-relevant metrics.

Revealing disparities via partitioning

Problem : Fairness metrics over large groups can hide disparities within vulnerable subgroups.

Solution : Partition policyholders based on fairness-relevant metrics.

Method :

- Use (optimal) decision trees for simplicity and interpretability.
- Partition based on **proxy vulnerability** $\Delta_{\text{proxy}}(\mathbf{X})$.
- Partition based on **commercial loading** $\Delta_{\text{load}}(\mathbf{X})$.

Revealing disparities via partitioning

Problem : Fairness metrics over large groups can hide disparities within vulnerable subgroups.

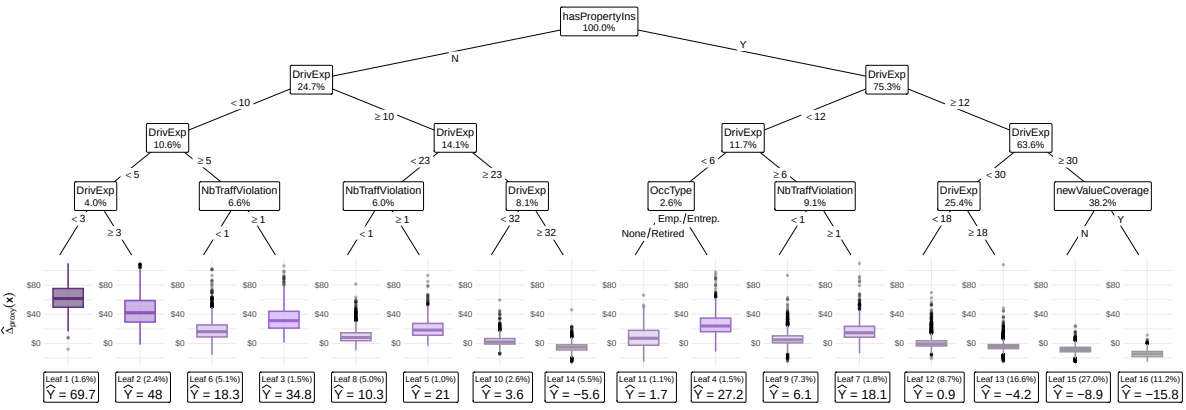
Solution : Partition policyholders based on fairness-relevant metrics.

Method :

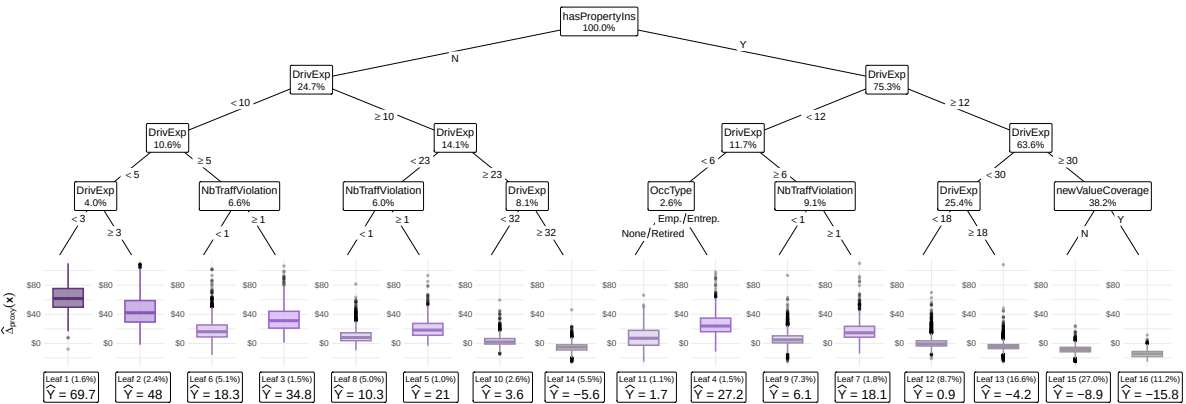
- Use (optimal) decision trees for simplicity and interpretability.
- Partition based on **proxy vulnerability** $\Delta_{\text{proxy}}(\mathbf{X})$.
- Partition based on **commercial loading** $\Delta_{\text{load}}(\mathbf{X})$.

Goal : Identify groups that are systematically **vulnerable** and **overloaded**.

Partitioning policyholders following proxy vulnerability

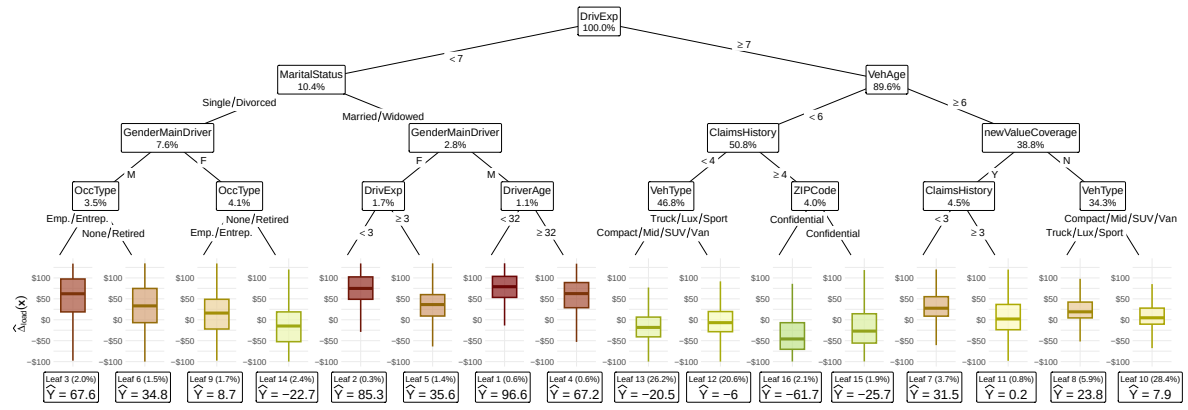


Partitioning policyholders following proxy vulnerability

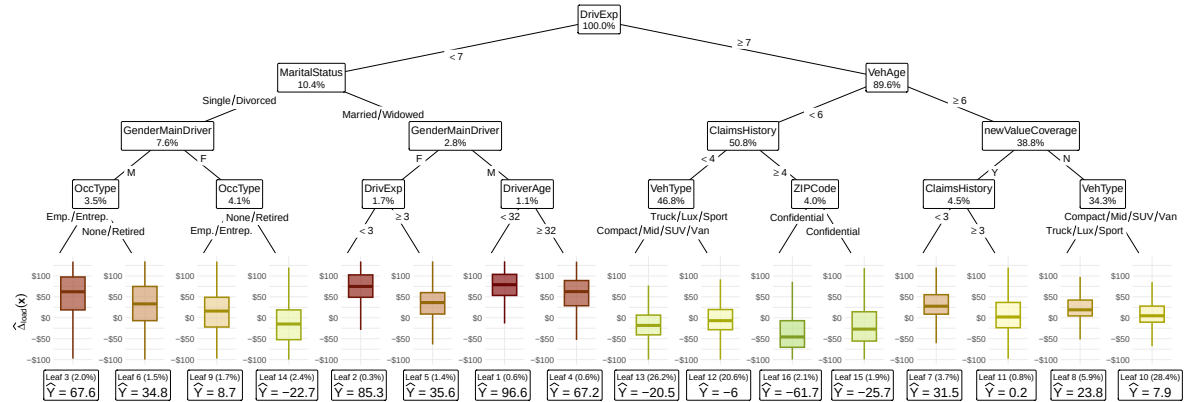


Leaves with high predictions signal **segments where proxy effects are most likely** to be exploited by a model.

Partitioning policyholders following commercial loading



Partitioning policyholders following commercial loading



Monitoring commercial loadings means **monitoring the cumulative impact of pricing decisions** – likely beyond the insurer's intent or awareness.

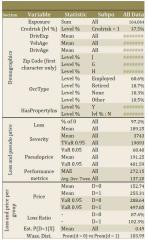
An integrated fairness assessment framework

Is the pseudo-price fair ?

Section	Variable	Statistic	Subpo	All Data
Demographics	Exposure	Sum	All	164.064
	Credrisk (lvl %)	Level %	Credrisk = 1	37.5%
	DrivExp	Mean	All	#######
	VehAge	Mean	All	#######
	DrivAge	Mean	All	#######
	Zip Code (first character only)	Level %	J	#######
		Level %	G	#######
		Level %	H	#######
	OccType	Level %	Employed	60.6%
		Level %	Retired	18.7%
		Level %	None	10.3%
		Level %	Other	10.5%
	HasPropertyIns	Level %	Y	#######
		Level %	lvl % : N	#######
Loss and pseudo price	Loss	% of 0	All	97.2%
		Mean	All	189.25
	Severity	Mean	All	3743
		TVaR 0.95	All	19093
	Pseudoprice	VaR 0.05	All	68.40
		Mean	All	191.25
		VaR 0.95	All	401.59
	Performance metrics	MAE	All	272.15
		Avg. Dev. Twee.	All	137.20
Loss and price per group	Price	Mean	D=0	152.74
		Mean	D=1	255.31
		VaR 0.95	D=0	288.64
		VaR 0.95	D=1	497.85
	Loss Ratio	-	D=0	87.4%
		-	D=1	102.3%
	Est. P(D=1|X)	Mean	All	0.49
	Wass. Dist.	Prem(d = 0) vs Prem(d = 1)		103.99

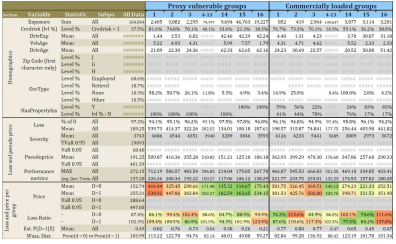
An integrated fairness assessment framework

Is the pseudo-price fair ?



Section	Variable	Statistic	Subpa	All Data
Demographics	Exposure	Sum	All	664,644
	CrossVeh (M %)	Level %	CrossVeh = 1	37.5%
	DrivExp	Mean	All	########
	VehAge	Mean	All	########
	DrivAge	Mean	All	########
	Zip Code (first character only)	Level %	J	########
		Level %	G	########
		Level %	H	########
	OccType	Level %	Employed	68.6%
		Level %	Retired	18.7%
		Level %	None	18.3%
		Level %	Other	10.5%
	HasProperty(m	Level %	Y	########
		Level %	Int % > N	########
Loss and pseudo_price	Loss	% of O	All	97.2%
	Severity	Mean	All	189.25
		Mean	All	3763
		TVaR 0.95	All	19093
		VaR 0.95	All	60.44
	Pseudoprice	Mean	All	191.25
		VaR 0.95	All	461.56
	Performance metrics	MAE	All	272.15
		Avg. Dev. Tweed	All	137.18
Loss and price per group	Price	Mean	D=0	152.74
		Mean	D=1	253.31
		VaR 0.95	D=0	289.64
		VaR 0.95	D=1	497.68
	Loss Ratio	-	D=0	97.4%
		-	D=1	102.9%
	Est. P(D=1|X)	Mean	All	0.49
	Wass. Dist.	Price(X = 0) vs Price(X = 1)		183.99

An integrated fairness assessment framework

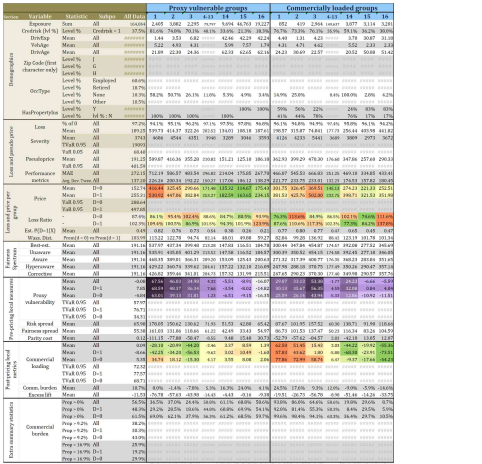


					Proxy vulnerable groups						Commercially loaded groups					
Version	Variable	Statistic	Subpop	All Data	1	2	3	4-13	14	15	16	1	2	3	4-13	14	15	16
Demographics	Exposure	Sum	All	164,644	3,605	3,082	2,295	########	9,094	44,763	19,227	852	419	2,944	########	3,077	53,114	5,281
	Credtrisk (M-%)	Level %	Credtrisk = 1	37.5%	81.6%	74.0%	70.1%	48.1%	33.0%	21.3%	18.3%	76.7%	73.3%	70.1%	35.9%	53.1%	26.2%	30.0%
	DrivExp	Mean	All	########	1.44	3.53	6.82	########	42.66	42.29	42.24	4.40	1.31	4.21	########	3.79	30.87	31.18
	VehAge	Mean	All	########	5.22	4.93	4.21	########	5.98	7.57	1.79	4.31	4.71	4.62	########	5.52	2.33	2.33
	DrivAge	Mean	All	########	21.89	22.20	24.56	########	62.59	62.65	62.16	24.23	30.49	22.57	########	20.52	50.88	51.42
	Zip Code (first character only)	Level %	J	########	########	########	########	########	########	########	########	########	########	########	########	########	########	########
		Level %	G	########	########	########	########	########	########	########	########	########	########	########	########	########	########	########
		Level %	H	########	########	########	########	########	########	########	########	########	########	########	########	########	########	########
	OccType	Level %	Employed	68.6%	########	########	########	########	########	########	########	########	########	########	########	########	########	########
		Level %	Retired	18.7%	########	########	########	########	########	########	########	########	########	########	########	########	########	########
		Level %	None	10.3%	58.2%	53.7%	20.1%	11.0%	5.3%	4.9%	3.4%	18.9%	25.6%	########	########	8.4%	108.0%	2.0%	4.2%
		Level %	Other	10.5%	########	########	########	########	########	########	########	########	########	########	########	########	########	########
	HasProperty(ies)	Level %	Y	########	########	########	########	########	########	100%	100%	75%	56%	22%	########	24%	83%	83%
		Level %	Int % / N	########	100%	100%	100%	########	100%	########	########	41%	44%	78%	########	76%	17%	17%
Loss and pseudo-price	Loss	% of	All		97.2%	94.1%	95.1%	94.2%	93.1%	97.9%	97.0%	94.0%	94.1%	94.8%	94.9%	97.6%	95.0%	94.1%	94.2%
		Mean	All		189.25	339.73	414.37	322.26	283.52	134.03	180.18	187.61	190.87	315.87	74.81	177.73	256.44	485.98	441.82
	Severity	Mean	All		375.1	606.1	654.4	619.1	574.0	320.9	329.9	304.4	359.9	612.6	323.3	544.1	369.9	509.9	497.1	367.2
		TVaR 0.95	All		1909.9	########	########	########	########	########	########	########	########	########	########	########	########	########	########
		TVaR 0.05	All		60.44	########	########	########	########	########	########	########	########	########	########	########	########	########	########
	Pseudoprice	Mean	All		191.25	599.87	616.36	355.20	210.81	151.21	125.18	186.18	162.93	399.29	478.30	176.68	347.86	257.68	290.33
		TVaR 0.95	All		481.56	########	########	########	########	########	########	########	########	########	########	########	########	########	########
	Performance metrics	MAE	All		272.15	712.19	586.57	460.54	296.81	214.04	175.65	247.70	466.87	585.53	664.83	351.29	469.18	334.85	410.41
		Avg. Sev/Tree	All		137.28	226.16	200.34	192.22	150.57	117.06	146.12	150.29	121.27	210.75	219.81	137.31	179.58	157.82	180.45
Loss and price per group	Price	Mean	D=0		152.74	516.48	525.45	290.66	171.49	125.12	114.67	175.44	201.75	326.45	366.51	146.11	274.23	221.33	252.51
		Mean	D=1		253.13	810.92	847.86	502.84	353.27	182.59	161.64	236.15	161.33	625.76	702.80	235.75	390.71	321.53	351.94
		VaR 0.95	D=0		288.64	########	########	########	########	########	########	########	########	########	########	########	########	########	########
		VaR 0.95	D=1		497.69	########	########	########	########	########	########	########	########	########	########	########	########	########	########
	Loss Ratio	-	D=0		97.6%	86.1%	95.6%	102.4%	98.6%	84.7%	88.9%	93.9%	76.3%	115.6%	88.9%	86.5%	102.1%	74.6%	111.4%
		-	D=1		102.9%	109.8%	110.5%	86.9%	101.5%	94.3%	101.9%	123.9%	87.6%	110.6%	117.3%	102.9%	77.7%	84.2%	117.8%
	Est. P(D=1|S)	Mean	All		0.69	0.82	0.76	0.75	0.64	0.38	0.26	0.21	0.57	0.80	0.77	0.41	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	59.23	82.84	99.20	136.92	86.61	125.19	101.78	101.34

Is the pseudo-price fair ?

- Partitions define relevant subgroups (columns).

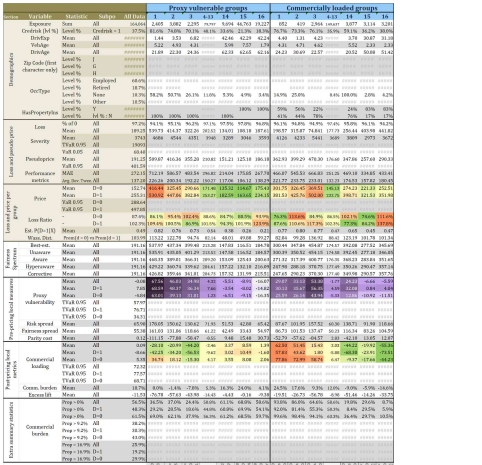
An integrated fairness assessment framework



Is the pseudo-price fair ?

- Partitions define relevant subgroups (columns).
- Local metrics quantify dissect unfairness (rows).

An integrated fairness assessment framework



Is the pseudo-price fair ?

- Partitions define relevant subgroups (columns).
- Local metrics quantify dissect unfairness (rows).

The toolbox guide **fairness assessment**.

Question

Is the given pricing function $\pi(\mathbf{x}, d)$ fair?

Inputs

A pricing function: $\pi(\mathbf{x}, d)$

Data: $\{\mathbf{X}_i, D_i, Y_i\}_{i=1}^n$

Question

Is the given pricing function $\pi(\mathbf{x}, d)$ fair?

Inputs

A pricing function: $\pi(\mathbf{x}, d)$

Data: $\{\mathbf{X}_i, D_i, Y_i\}_{i=1}^n$

→ What does it mean for pricing to be fair?

→ Can one benchmark fairness?

→ Is unfairness truly material?

→ Where does unfairness concentrate?

Question

Is the given pricing function $\pi(\mathbf{x}, d)$ fair?

What does it mean for pricing to be fair?

Can one benchmark fairness?

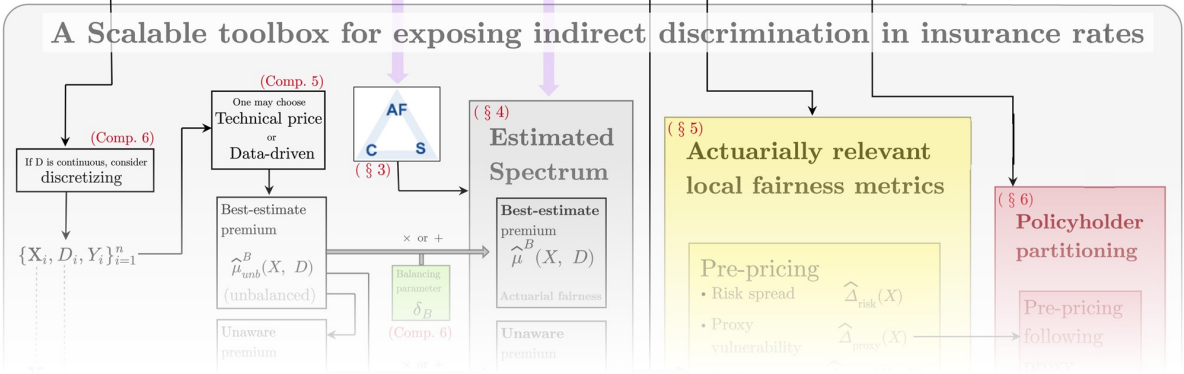
Is unfairness truly material?

Where does unfairness concentrate?

Inputs

A pricing function: $\pi(\mathbf{x}, d)$

Data: $\{\mathbf{X}_i, D_i, Y_i\}_{i=1}^n$



Question

Is the given pricing function $\pi(\mathbf{x}, d)$ fair?

Inputs

A pricing function: $\pi(\mathbf{x}, d)$

Data: $\{\mathbf{X}_i, D_i, Y_i\}_{i=1}^n$

What does it mean for pricing to be fair?

Can one benchmark fairness?

Is unfairness truly material?

Where does unfairness concentrate?

The toolbox guides assessment.

(Fig. 14)

A Scalable toolbox for exposing indirect discrimination in insurance rates

(Comp. 6)

If D is continuous, consider discretizing

$\{\mathbf{X}_i, D_i, Y_i\}_{i=1}^n$

(Comp. 5)

One may choose Technical price or Data-driven

Best-estimate premium

$\widehat{\mu}^B_{unb}(X, D)$

(unbalanced)

Unaware premium

AF

C S

(§ 3)

× or +

Balancing parameter δ_B

(Comp. 6)

(§ 4)

Estimated Spectrum

Best-estimate premium

$\widehat{\mu}^B(X, D)$

Actuarial fairness

Unaware premium

(§ 5)

Actuarially relevant local fairness metrics

Pre-pricing

- Risk spread $\widehat{\Delta}_{\text{risk}}(X)$
- Proxy vulnerability $\widehat{\Delta}_{\text{proxy}}(X)$

(§ 6)

Policyholder partitioning

Pre-pricing following

Moving forward

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

1. Fairness often assumes **access to protected attributes**, which may unavailable.

Moving forward

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

1. Fairness often assumes **access to protected attributes**, which may unavailable.
2. Applying anti-discrimination **regulations** meets resistance where actuarial justification holds authority.

Moving forward

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

1. Fairness often assumes **access to protected attributes**, which may unavailable.
2. Applying anti-discrimination **regulations** meets resistance where actuarial justification holds authority.
3. Fairness is typically studied as a one-year objective, but its **long-term welfare** effects remain unclear (Shimao et al., 2022).

Moving forward

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

1. Fairness often assumes **access to protected attributes**, which may unavailable.
2. Applying anti-discrimination **regulations** meets resistance where actuarial justification holds authority.
3. Fairness is typically studied as a one-year objective, but its **long-term welfare** effects remain unclear (Shimao et al., 2022).
4. Portfolio fairness may conflict with market fairness (Côté et al., 2024).

Thank you


CAS


Desjardins
Assurances


NSERC
CRSNG


SOCIETY OF
ACTUARIES


mitacs


Québec
Fonds de recherche – Nature et technologies

References i

Araiza Iturria, C. A., Hardy, M., and Marriott, P. (2024). A discrimination-free premium under a causal framework. *North American Actuarial Journal*, pages 1–21.

ASB (2005). Actuarial Standard of Practice No. 12 : Risk Classification (for All Practice Areas). actuarialstandardsboard.org/asops/risk-classification-practice-areas/. Accessed : March 31, 2025.

Aseervatham, V., Lex, C., and Spindler, M. (2016). How do unisex rating regulations affect gender differences in insurance premiums ? *The Geneva Papers on Risk and Insurance-Issues and Practice*, 41(1) :128–160.

Boucher, J.-P. and Pigeon, M. (2024). Balancing risk assessment and social fairness : an auto telematics case study. *Casualty Actuarial Society Research Paper Series on Race and Insurance Pricing*. Retrieved from casact.org/publications-research/research/research-paper-series-race-and-insurance-pricing. Accessed : March 31, 2025.

References ii

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30.

Casualty Actuarial Society (1988). Statement of Principles Regarding Property and Casualty Insurance Ratemaking. casact.org/statement-principles-regarding-property-and-casualty-insurance-ratemaking. Originally published in 1988, rescinded in 2020, and reinstated in 2021 for reference. Accessed : March 31, 2025.

Charpentier, A., Flachaire, E., and Gallic, E. (2023a). Optimal transport for counterfactual estimation : A method for causal inference. In *Optimal Transport Statistics for Economics and Related Topics*, pages 45–89. Springer.

Charpentier, A., Hu, F., and Ratz, P. (2023b). Mitigating discrimination in insurance with wasserstein barycenters. *BIAS, 3rd Workshop on Bias and Fairness in AI, International Workshop of ECML PKDD*.

References iii

Côté, M.-P., Côté, O., and Charpentier, A. (2024). Selection bias in insurance : why portfolio-specific fairness fails to extend market-wide. *Available at SSRN : 10.2139/ssrn.5018749*.

Côté, O., Côté, M.-P., and Charpentier, A. (2025). A fair price to pay : Exploiting causal graphs for fairness in insurance. *Journal of Risk and Insurance*, 92 :33–75.

Embrechts, P. and Wüthrich, M. V. (2022). Recent challenges in actuarial science. *Annual Review of Statistics and Its Application*, 9(1) :119–140.

European Union (2000). Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. `https://eur-lex.europa.eu/eli/dir/2000/43/oj`. Accessed : December 4, 2024.

Fermanian, J.-D. and Guegan, D. (2021). Fair learning with bagging. *Documents de travail du Centre d'Économie de la Sorbonne*.

Fernandes Machado, A., Charpentier, A., and Gallic, E. (2024). Sequential conditional transport on probabilistic graphs for interpretable counterfactual fairness.

References iv

Fernandes Machado, A., Grondin, S., Ratz, P., Charpentier, A., and Hu, F. (2025). Equipy : Sequential fairness using optimal transport in python. *arXiv preprint*, arXiv :2503.09866.

Financial Services Regulatory Authority of Ontario (2024). Proposed Guidance : Automobile Insurance Rating and Underwriting Supervision. fsrao.ca/industry/auto-insurance/regulatory-framework/guidance-auto-insurance/proposed-guidance-automobile-insurance-rating-and-underwriting-supervision-guidance. Accessed : March 31, 2025.

Fredman, S. (2022). *Discrimination Law*. Oxford University Press, Oxford, 3rd edition.

Frees, E. W. and Huang, F. (2023). The discriminating (pricing) actuary. *North American Actuarial Journal*, 27(1) :2–24.

Gabric, L. J., Zhou, S., and Zhou, K. Q. (2024). A Bayesian approach to discrimination-free insurance pricing. *Available at SSRN 4785927*.

References v

Grari, V., Charpentier, A., Lamprier, S., and Detyniecki, M. (2022). A fair pricing model via adversarial learning. *arXiv preprint arXiv :2202.12008*.

Hu, F., Ratz, P., and Charpentier, A. (2023). Fairness in multi-task learning via Wasserstein barycenters. In Koutra, D., Plant, C., Gomez Rodriguez, M., Baralis, E., and Bonchi, F., editors, *Machine Learning and Knowledge Discovery in Databases : Research Track*, pages 295–312, Cham. Springer Nature Switzerland.

Hu, F., Ratz, P., and Charpentier, A. (2024). A sequentially fair mechanism for multiple sensitive attributes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11) :12502–12510.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM : A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30 :3146–3154.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint*, arXiv :1609.05807.

References vi

Komiyama, J. and Shimao, H. (2017). Two-stage algorithm for fairness-aware machine learning. *arXiv preprint arXiv :1710.04924*.

Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M. V. (2022). Discrimination-free insurance pricing. *ASTIN Bulletin*, 52(1) :55–89.

Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M. V. (2024). What is fair ? Proxy discrimination vs. demographic disparities in insurance pricing. *Scandinavian Actuarial Journal*, 2024(9) :935–970.

Makhlouf, K., Zhioua, S., and Palamidessi, C. (2024). When causality meets fairness : A survey. *Journal of Logical and Algebraic Methods in Programming*, 141 :101000.

Plečko, D., Bennett, N., and Meinshausen, N. (2021). fairadapt : Causal reasoning for fair data pre-processing. *arXiv preprint arXiv :2110.10200*.

Pope, D. G. and Sydnor, J. R. (2011). Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal : Economic Policy*, 3(3) :206–31.

References vii

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1) :41–55.

Shimao, H., Huang, F., and Khern-am nuai, W. (2022). Welfare implications of fairness regulations in insurance cost modeling : A multi-method study. *Available at SSRN : 10.2139/ssrn.5112616*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 325–333.