Sensitivity-based measures of discrimination in insurance pricing

Mathias Lindholm* Ronald Richman† Andreas Tsanakas‡ Mario V. Wüthrich§

December 23, 2024

Abstract

Different notions of fairness and discrimination have been extensively discussed in the machine learning, operations research, and insurance pricing literatures. As not all fairness criteria can be concurrently satisfied, metrics are needed that allow assessing the materiality of discriminatory effects and the trade-offs between various criteria. Methods from sensitivity analysis have been deployed for the measurement of demographic unfairness, that is, the statistical dependence of risk predictions on protected attributes. We produce a sensitivity-based measure for the distinct phenomenon of proxy discrimination, referring to the implicit inference of protected attributes from other covariates. For this, we first define a set of admissible prices that avoid proxy discrimination. Then, the measure is defined as the normalised L^2-distance of a price from the closest element in that set. We use arguments from variance-based sensitivity analysis, to attribute the proxy discrimination measure to individual (or subsets of) covariates and investigate how properties of the data generating process are reflected in those metrics. Furthermore, we build on the global (i.e., portfolio-wide) measures of demographic unfairness and proxy discrimination to propose local (i.e., instance- or policyholder-specific) measures, which allow a fine-grained understanding of discriminatory effects. Finally, we apply the methods developed in the paper to a real-world insurance dataset, where ethnicity is a protected variable. We observe substantial proxy-discriminatory effects for one ethnic group and identify the key variables driving this.

Keywords: Sensitivity analysis, proxy discrimination, demographic parity, insurance pricing, algorithmic fairness.

*Corresponding author. Department of Mathematics, Stockholm University, SE-106 91, Sweden. lindholm@math.su.se

†insureAI, Floor 2, 30 Melrose Boulevard, Melrose Arch, Gauteng, South Africa, 2196. ron@insureai.co.

‡Bayes Business School, City St George's, University of London, 106 Bunhill Row, London, EC1Y 8TZ, United Kingdom. A.Tsanakas.1@city.ac.uk

§RiskLab, Department of Mathematics, ETH Zurich, Switzerland. mario.wuethrich@math.ethz.ch

1 Introduction

Questions of fairness and discrimination have become central to the machine learning literature (Barocas & Selbst, 2016; Mehrabi et al., 2021) and its applications. Increasingly, the wider operational research literature is building fairness considerations into problems as diverse as organ allocation (Bertsimas et al., 2013), efficiency measurement (Radovanović et al., 2022), hiring (Komiyama & Noda, 2024), and stress testing (Glasserman & Li, 2024). In particular, in applications such as credit scoring (Kozodoi et al., 2022; Hurlin et al., 2024) or insurance pricing (Lindholm et al., 2022; Frees & Huang, 2023), the risk of financial exclusion for particular demographic groups has a high societal salience. Furthermore, questions of fairness are intrinsically linked to questions of model interpretability, since the latter can be a pre-requisite for effectively identifying and addressing the former (De Bock et al., 2024; Kraus et al., 2024).

A variety of criteria have been formulated, which model predictions should satisfy in order to be considered fair (indicatively, Barocas & Selbst, 2016; Mehrabi et al., 2021; Charpentier, 2024). On the one hand, *group fairness* criteria interrogate the statistical relationship between responses, predictions, and protected attributes such as gender or ethnicity; these criteria are typically formulated via (conditional) independence statements. On the other hand, *individual fairness* criteria focus on whether individuals with similar risk profiles are treated similarly, e.g., are quoted comparable insurance prices. Here different fairness criteria arise from different notions of similarity and information restrictions. Hence, while group fairness criteria consider the outcome of a prediction algorithm, individual fairness notions revolve more around the way that these predictions are generated. Furthermore, as part of the rich literature on fairness and discrimination, an understanding has developed that such criteria are not necessarily consistent with each other and can even be mutually exclusive (Kleinberg et al., 2016; Lindholm et al., 2024). Consequently, the need emerges to construct metrics for different forms of discrimination and unfairness: one does not just need to know whether such phenomena take place as part of a prediction/decision process, but also whether the effects are material enough to justify concern and eventual (e.g., regulatory) action. Furthermore, the potential incompatibility of different fairness notions means that one cannot require for all of them to hold at the same time. This creates the need to monitor many aspects of possible unfairness and measure the respective trade-offs, including those arising from any adjustments to model predictions.

Here we develop measures for two distinct phenomena, *demographic unfairness* and *proxy discrimination*, with a clear emphasis on the latter. While our application context is insurance pricing, the measures are more broadly applicable to settings such as, e.g., credit scoring. Demographic unfairness relates to violations of demographic parity, that is, the requirement that prices are statistically independent of policyholders' protected attributes. While the deployment of this particular group fairness notion in insurance has been criticised (Lindholm et al., 2024), we consider it for two reasons: first, because it is easy to explain and politically salient, therefore a potential source of reputational risk for insurers (e.g., Cook et al., 2022); and second because it helps with introduc-

ing the construction of the measures we are using. The notion of proxy discrimination builds on the understanding that some policyholder attributes, like gender or ethnicity, should not be used to calculate the price for individual policyholders, as this would constitute *direct discrimination*. Based on that premise, it is additionally desirable to avoid the effective proxying of protected attributes by other variables (e.g., car engine size, postal code) that are correlated to them. Several conceptualisations of proxy discrimination exist (Prince & Schwarcz, 2019; Tschantz, 2022), with the causal structure of covariates often taking centre stage (Araiza Iturria et al., 2024; Côté et al., 2024). Here we take a view of proxy discrimination as a form of omitted variable bias, which is not contingent on assumptions of causality, but focuses on the indirect inference of protected attributes from other covariates, in the sense of Lindholm et al. (2022, 2023, 2024).

The measures of discrimination we develop are based on methods from Global Sensitivity Analysis, originating in the work of Sobol' (2001) and having been deeply studied by a variety of authors (e.g., Saltelli et al., 2008, 2010; Borgonovo & Plischke, 2016; Fissler & Pesenti, 2023). Sensitivity analysis is deployed to provide insight into complex computational models, evaluate the relative importance of model inputs and identify model vulnerabilities; for applications specifically to insurance risk portfolios and insurance regulation, see respectively Rabitti & Borgonovo (2020); Vallarino et al. (2024) and Borgonovo et al. (2024). The variable importance metrics used in sensitivity analysis can thus be suitable tools for evaluating the direct and indirect impact of protected attributes on insurance prices. This is already recognised in the work of Bénesse et al. (2024) who provide sensitivity-based measures for a variety of fairness criteria, though, to our knowledge, applications of sensitivity analysis to the problem of measuring proxy discrimination are currently lacking in the literature (a brief relevant discussion is found in Hiabu et al. (2023)).

In Section 2 we formally introduce the ideas of demographic unfairness and proxy discrimination, using the paradigm of insurance pricing in the exposition. Specifically, in Sections 2.3 and 2.4 we define, respectively, measures of demographic unfairness and proxy discrimination and discuss their properties. The former is already found in Bénesse et al. (2024) and we only deal with it briefly. The measure of proxy discrimination is new and addresses the challenging task of quantifying a phenomenon related to the way prices are calculated from observable quantities. The measure is based on the distance between any given price and the closest element in a set of prices that avoid proxy discrimination. This set of admissible prices arises as a convex combination of the discrimination-free prices in Lindholm et al. (2022) and constant prices that do not depend on any policyholder characteristics. This idea is operationalised through a constrained regression of the price on best-estimate prices, calculated using different scenarios regarding the value of a protected attribute. The measure of proxy discrimination takes values between 0 and 1 for an insurance portfolio and can be evaluated for any system of prices, without reference to how these were calculated. As a result it lends itself to empirical evaluation and price auditing.

While a global measure of proxy discrimination for a portfolio is useful, it is also necessary to understand which covariates are the sources for such discrimination. In Section 2.5, we discuss how to attribute the measure of proxy discrimination to covariates. Given that the measure is a result

of an L^2-projection, this can be achieved by a simple adaptation of the variance-based Sobol' and total sensitivity indices (Saltelli et al., 2008).

In Section 3.1, we discuss properties of the underlying data generating process and explain how the sensitivity measures respond to such properties. Subsequently, in Sections 3.2 and 3.3 we introduce, respectively, local measures of demographic unfairness and proxy discrimination. These measures are evaluated for individual policies and allow a more fine-grained understanding and visualisation of the way in which discriminatory effects may arise in a portfolio. The local measures are given by the price of a policy minus a benchmark price that is free of the particular type of unfairness considered. For the case of demographic unfairness, the benchmark price comes from an Output Optimal Transport transformation of the portfolio's prices (Lindholm et al., 2024; Charpentier, 2024). For the case of proxy discrimination, the benchmark is given by the closest element in the set of prices that avoid proxy discrimination.

In Section 4, we apply the measures developed to a real-world motor insurance dataset, where ethnicity is the protected characteristic. Attribution of the proxy discrimination metric to covariates helps with identifying the key drivers of discriminatory effects. The use of local measures and the consideration of global metrics on sub-portfolios give a more granular view, with specific groups (e.g. younger drivers of one particular ethnicity) seen to be more heavily affected by proxy discrimination. The application shows how the metrics introduced in this paper can be applied in practice and provide insights into the potentially discriminatory effects of algorithmic decisions.

Brief conclusions are stated in Section 5. Some additional technical details are given as appendices in the Supplementary materials.

2 Measures of proxy discrimination and demographic unfairness

2.1 Setup and notation

We work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathbb{P}$ being the real-world probability measure. On that space we consider the random vector $(Y, \boldsymbol{X}, D)$. The random variable Y represents a quantity to be predicted based on covariates $(\boldsymbol{X}, D)$. Of these, $\boldsymbol{X}$ is a vector of *non-protected covariates* (non-discriminatory characteristics), while D reflects a *protected attribute* (discriminatory or sensitive characteristic). From now on, we will refer specifically to an insurance pricing context, where Y is a loss or loss frequency, while $(\boldsymbol{X}, D)$ capture policyholder characteristics. The variability of $(\boldsymbol{X}, D)$ under $\mathbb{P}$ represents heterogeneity in the population of policyholders, while the variability of Y conditional on $\{\boldsymbol{X} = \boldsymbol{x}, D = d\}$ reflects loss uncertainty for a policyholder with known features $(\boldsymbol{x}, d)$. We will assume throughout that D is discrete, taking values in the finite set $\mathfrak{D}$. Elements of $\mathfrak{D}$ can be thought of as representing different demographic groups, e.g., by ethnicity and/or gender.

Throughout, for a generic random variable Z, we represent its distribution function by $\mathbb{P}(z)$; the conditional distribution of Z given $W = w$ is denoted accordingly by $\mathbb{P}(z|w)$. In the case of absolutely continuous random variables Z, probability density functions are given by $\mathrm{d}\mathbb{P}(z)/\mathrm{d}z$. In

the case of discrete Z we have probability weights $\mathbb{P}(z) = \mathbb{P}(Z = z) > 0$.

2.2 Pricing functions and discriminatory effects

We define the *best-estimate price* as

$$\mu(\boldsymbol{x}, d) := \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, D = d], \tag{1}$$

such that $\mu(\boldsymbol{x}, d)$ is the optimal (in L^2-norm) prediction of the loss Y for a policyholder with features $(\boldsymbol{x}, d)$. Best-estimate prices have discriminatory effects because of their direct dependence on protected characteristics D. The most straightforward way of correcting for such *direct discrimination*, is to calculate insurance prices without including the information D as a covariate for prediction of Y. The resulting conditional expectation based only on non-protected characteristics $\boldsymbol{X}$ is termed the *unawareness price* and is defined by

$$\mu(\boldsymbol{x}) := \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]. \tag{2}$$

Nonetheless, the unawareness price may still have discriminatory effects, arising from the potential dependence between the random vectors $\boldsymbol{X}$ and D. Such dependence need not have a causal source, but just be a feature of a particular portfolio structure of insurance policies, e.g., when female and male policyholders have different age distributions.

Out of many different notions of unfairness or discrimination we focus on two complementary perspectives on the impact of the dependence of protected attributes and non-protected covariates. First, given such dependence it will generally hold that $\mu(\boldsymbol{X})$ is also dependent on D. This means that insurance prices may vary across demographic groups, such that for some $d \neq d'$, $d, d' \in \mathfrak{D}$, we have that $\mathbb{E}[\mu(\boldsymbol{X}) \mid D = d] \neq \mathbb{E}[\mu(\boldsymbol{X}) \mid D = d']$. For short, we call such insurance prices *demographically unfair*. Second, removing D from the set of covariates does not imply that these are not used *indirectly* in pricing. A concern is that, if D can be partially predicted from $\boldsymbol{X}$, it is possible that insurance prices, derived with the aim of maximising predictive accuracy, implicitly use $\boldsymbol{X}$ to infer D. In fact, by observing that we can write unawareness prices as

$$\mu(\boldsymbol{x}) = \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{x}, d)\, \mathbb{P}(d \mid \boldsymbol{x}), \tag{3}$$

it becomes clear that unawareness prices do indeed rely on implicit inference of D from $\boldsymbol{X}$, via the conditional probability $\mathbb{P}(d \mid \boldsymbol{x})$ used in averaging over best-estimates (if there is dependence between D and $\boldsymbol{X}$). We say that prices utilising such inference of protected characteristics are subject to *proxy discrimination*.

2.3 Measuring demographic unfairness

We now put the previous ideas on a more formal footing, which applies to a general pricing functional $\boldsymbol{X} \mapsto \pi(\boldsymbol{X})$. We start with the well-known idea of demographic parity.

Definition 1. *The pricing functional* $\boldsymbol{X} \mapsto \pi(\boldsymbol{X})$ *satisfies* demographic parity *with respect to* $\mathbb{P}(\boldsymbol{X}, D)$*, if the random variable* $\pi(\boldsymbol{X})$ *is independent of* D *under* $\mathbb{P}$*. If* π *violates demographic parity, we say that it is* demographically unfair*.*

Note that independence of $\boldsymbol{X}$ and D is a sufficient but not a necessary condition for demographic parity. Furthermore, establishing that a pricing functional π is demographically unfair does not necessarily mean that the resulting impact is substantial. To understand the materiality of unfairness, one can visualise the changes in (empirical estimates of) the conditional density of $\pi(\boldsymbol{X})|D = d$ across demographic groups $d \in \mathfrak{D}$. Sometimes though, a global numerical metric is useful. We now present such a metric, following Bénesse et al. (2024) who apply ideas from sensitivity analysis to evaluate various concepts of algorithmic (un)fairness.

Definition 2. *The demographic unfairness metric* UF *is defined as*

$$\mathrm{UF}(\pi) = \frac{\mathrm{Var}(\mathbb{E}[\pi(\boldsymbol{X}) \mid D])}{\mathrm{Var}(\pi(\boldsymbol{X}))}, \tag{4}$$

with the convention that if $\mathrm{Var}(\pi(\boldsymbol{X})) = 0$*, then* $\mathrm{UF}(\pi) = 0$*.*

The rationale for the construction (4) is well established in the different context of Global Sensitivity Analysis and the metric is known as a *Sobol' Index* (e.g. Sobol', 2001; Saltelli et al., 2008). Noting that $\mathrm{Var}(\pi(\boldsymbol{X})) = \mathrm{Var}(\mathbb{E}[\pi(\boldsymbol{X}) \mid D]) + \mathbb{E}[\mathrm{Var}(\pi(\boldsymbol{X}) \mid D)]$, the numerator in (4) represents the amount of variation in $\pi(\boldsymbol{X})$ attributable to the protected variable D.

The following easily derived properties are stated without proof.

Proposition 1. *The unfairness metric* UF *satisfies the following properties.*

i) $0 \leq \mathrm{UF}(\pi) \leq 1$*. Furthermore, for all* $a, b \in \mathbb{R}$ *it holds that* $\mathrm{UF}(a + b\pi) = \mathrm{UF}(\pi)$*.*

ii) If π *satisfies demographic parity with respect to* $\mathbb{P}(\boldsymbol{X}, D)$*, then* $\mathrm{UF}(\pi) = 0$*.*

iii) If $\sigma(\pi(\boldsymbol{X})) \subseteq \sigma(D)$*, i.e.,* $\pi(\boldsymbol{X})$ *is* D*-measurable, then* $\mathrm{UF}(\pi) = 1$*.*

Example 1. Let $D \in \{0, 1\}$, $X \sim \mathrm{U}(0, 1)$ and $\mathbb{P}(D = 1|X) = \mathbb{E}[D \mid X] = X$. This implies $\mathbb{P}(D = 1) = \frac{1}{2}$. Assume that the best-estimate price is

$$\mu(X, D) = \frac{1}{2} + X + D,$$

which includes a fixed cost. In this model there is the potential for demographic unfairness, since (X, D) are dependent. The unawareness price equals

$$\mu(X) = \mathbb{E}\left[\frac{1}{2} + X + D \;\middle|\; X\right] = \frac{1}{2} + 2X.$$

Straightforward calculations lead to

$$\mathbb{E}[\mu(X) \mid D = 0] = \frac{1}{2} + 2\mathbb{E}[X \mid D = 0] = \frac{1}{2} + \frac{2}{3},$$
$$\mathbb{E}[\mu(X) \mid D = 1] = \frac{1}{2} + 2\mathbb{E}[X \mid D = 1] = \frac{1}{2} + \frac{4}{3}.$$

It then follows that $\mathrm{Var}(\mathbb{E}[X \mid D]) = 1/36$. Hence, if the unawareness price is used, the average premium for $D = 0$ is different compared to $D = 1$. Consequently demographic unfairness arises. We can quantify this effect via the UF metric:

$$\mathrm{UF} = \frac{\mathrm{Var}(\mathbb{E}[1/2 + 2X \mid D])}{\mathrm{Var}(1/2 + 2X)} = \frac{\mathrm{Var}(\mathbb{E}[X \mid D])}{\mathrm{Var}(X)} = \frac{1}{3} > 0.$$

■

2.4 Defining and quantifying proxy discrimination

We now turn our attention to the measurement of proxy discrimination. Here the situation is different, compared to unfairness. Recall that the best-estimate prices $\mu(\boldsymbol{X}, D)$ are not used, as they would give rise to direct discrimination and some other pricing functional $\pi(\boldsymbol{X})$ must be used instead. Effectively this corresponds to merging rating classes with the same non-protected profile $\{\boldsymbol{X} = \boldsymbol{x}\}$ but different protected characteristics $\{D = d\}$. Hence, one can calculate the price in each of those new classes as a weighted average over d of the corresponding costs $\mu(\boldsymbol{x}, d)$. Thus, pricing relates to reallocating the claims costs $\mu(\boldsymbol{x}, d)$, following the removal of the explanatory effect of D. To avoid proxy discrimination – unlike the situation of unawareness prices (3) – the selected weights used should not depend on $\boldsymbol{x}$.

To address this issue, Lindholm et al. (2022) suggest the pricing formula

$$h^*(\boldsymbol{X}) = \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{X}, d)\mathbb{P}^*(d), \tag{5}$$

for some distribution $\mathbb{P}^*(d)$ on $\mathfrak{D}$. Here, we build on Lindholm et al. (2022), to construct an expanded set of admissible pricing functionals that we consider to be free from proxy discrimination. We accept that a constant price $\pi(\boldsymbol{X}) \equiv \pi$ that does not depend on the covariates $\boldsymbol{X}$ cannot proxy-discriminate, as it does not discriminate between policyholders in any sense. Consequently, we would not like to exclude convex combinations of the form $(1 - \alpha)\pi + \alpha h^*(\boldsymbol{X})$, $\alpha \in [0, 1]$, from the admissible set of prices. Hence, we both allow for an additive constant offset and also that the weights $v_d := \alpha\mathbb{P}^*(d)$ that are applied on the best-estimate prices $\mu(\boldsymbol{X}, d)$ may sum to less than 1.

Proxy discrimination is now defined with the above arguments in mind. Define the set $\mathcal{V} := \{\boldsymbol{v} \in [0, 1]^{|\mathfrak{D}|} \ : \ \sum_{d \in \mathfrak{D}} v_d \leq 1\}$.

Definition 3. *The pricing functional* $\boldsymbol{X} \mapsto \pi(\boldsymbol{X})$ avoids proxy discrimination *with respect to* $\mu(\boldsymbol{X}, D)$*, if for* $\mathbb{P}$*-almost every* $\boldsymbol{X}$ *we can write*

$$\pi(\boldsymbol{X}) = c + \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{X}, d)v_d, \tag{6}$$

for some $c \in \mathbb{R}$ *and* $\boldsymbol{v} \in \mathcal{V}$ *that do not depend on* $\boldsymbol{X}$*. If* π *does not have that structure, we say that it is* proxy-discriminatory*.*

Following Definition 3, we can now define a measure of proxy discrimination.

Definition 4. *The proxy discrimination metric* PD *is defined as*

$$\mathrm{PD}(\pi) = \frac{\min_{c\in\mathbb{R},\ \boldsymbol{v}\in\mathcal{V}} \mathbb{E}\left[\left(\pi(\boldsymbol{X}) - c - \sum_{d\in\mathfrak{D}} \mu(\boldsymbol{X},d)v_d\right)^2\right]}{\mathrm{Var}(\pi(\boldsymbol{X}))}, \tag{7}$$

with the convention that if $\mathrm{Var}(\pi(X)) = 0$*, then* $\mathrm{PD}(\pi) = 0$*.*

The metric PD quantifies the extent to which a pricing functional cannot be expressed as a weighted average of best-estimate cost terms $\mu(\boldsymbol{x}, d)$, allowing also for a fixed cost term. Note that the presence of the intercept c, even with the constraints on $\boldsymbol{v}$, ensures that any solution $(c^*, \boldsymbol{v}^*)$ of the regression problem (7), satisfies $\mathbb{E}\left[\pi(\boldsymbol{X}) - c^* - \sum_{d\in\mathfrak{D}} \mu(\boldsymbol{X},d)v_d^*\right] = 0$. We note that even though $c^*, \boldsymbol{v}^*$ need not be unique, the quantity $c^* + \sum_{d\in\mathfrak{D}} \mu(\boldsymbol{X},d)v_d^*$ is. Hence, we can explicitly solve for c and express (7) as

$$\mathrm{PD}(\pi) = \frac{\mathrm{Var}\left(\pi(\boldsymbol{X}) - \sum_{d\in\mathfrak{D}} \mu(\boldsymbol{X},d)v_d^*\right)}{\mathrm{Var}(\pi(\boldsymbol{X}))}, \tag{8}$$

with $\mathrm{PD}(\pi)$ reflecting the residual variance for the constrained regression of $\pi(\boldsymbol{X})$ on $\mu(\boldsymbol{X}, d),\ d \in \mathfrak{D}$.

The quantity $\pi^*(\boldsymbol{X}) := c^* + \sum_{d\in\mathfrak{D}} \mu(\boldsymbol{X},d)v_d^*$ is the closest element to $\pi(\boldsymbol{X})$ in the set of prices that are free from proxy discrimination. We do *not* specifically suggest $\pi^*(\boldsymbol{X})$ as a suitable price correction for $\pi(\boldsymbol{X})$, given that the optimal $(c^*, \boldsymbol{v}^*)$ generally depends on the joint distribution of $(\boldsymbol{X}, D)$. This would violate the requirements on avoiding proxy discrimination, as formulated by Lindholm et al. (2024).

Again, simple properties of the metric can be stated.

Proposition 2. *The proxy discrimination metric* PD *satisfies the following properties.*

i) $0 \le \mathrm{PD}(\pi) \le 1$*. Furthermore, for all* $a \in \mathbb{R}, b \in \mathbb{R}_+$ *it holds that* $\mathrm{PD}(a + b\pi) = \mathrm{PD}(\pi)$*.*

ii) If π *avoids proxy discrimination with respect to* $\mu(\boldsymbol{X}, D)$*, then* $\mathrm{PD}(\pi) = 0$*.*

iii) If $\pi(\boldsymbol{X})$ *is uncorrelated with* $\mu(\boldsymbol{X}, d)$ *for all* $d \in \mathfrak{D}$*, then* $\mathrm{PD}(\mu) = 1$*.*

Proof. Parts i) and ii) are immediate.

For iii), uncorrelatedness implies that in the regression (7) we have $v_d^* = 0$ for all $d \in \mathfrak{D}$ and $c^* = \mathbb{E}[\pi(\boldsymbol{X})]$. Consequently $\min_{c\in\mathbb{R},\ \boldsymbol{v}\in\mathcal{V}} \mathbb{E}\left[\left(\pi(\boldsymbol{X}) - c - \sum_{d\in\mathfrak{D}} \mu(\boldsymbol{X},d)v_d\right)^2\right] = \mathrm{Var}(\pi(\boldsymbol{X}))$. □

Parts i)-ii) of Proposition 2 show that PD is an interpretable metric for proxy discrimination, while part iii) describes a situation where proxy discrimination is maximal: the insurance prices are not at all explained by claim costs and, thus, any discrimination they achieve between policyholders must be undesirable.

In the definition of the proxy discrimination metric (7) we constrained the weights $\boldsymbol{v}$ to be less than one. The reason for this is that higher weights on terms $\mu(\boldsymbol{X}, d)$ may also produce proxying effects, as the next example shows.

Example 2. We continue with the simple model of Example 1. In that model we expect the unawareness price to be subject to proxy discrimination, as (X, D) are dependent and the best-estimate prices are sensitive in D. Let us evaluate the numerator of (7), for $\pi(X) = \mu(X) = \frac{1}{2} + 2X$. We have that

$$\mu(X) - c - \sum_{d\in\{0,1\}} \mu(X,d)v_d = (2 - v_0 - v_1)X - c - \frac{1}{2}(v_0 + 3v_1 - 1).$$

Since c can be chosen to remove the bias for any choices of v_0, v_1, we have that

$$\mathbb{E}\left[\left(\mu(X) - c^* - \sum_{d\in\{0,1\}} \mu(X,d)v_d^*\right)^2\right] = (2 - v_0^* - v_1^*)^2 \operatorname{Var}(X).$$

From this it is clear that the minimum is achieved by $v_0^* + v_1^* = 1$. Hence, the proxy discrimination metric (7) becomes

$$\mathrm{PD}(\mu) = \frac{\operatorname{Var}(X)}{\operatorname{Var}(1/2 + 2X)} = \frac{1}{4}.$$

We now consider the alternative price $\pi(X) = 3X$, noting that it agrees on average with the unawareness price, i.e., $\mathbb{E}[\mu(X)] = \mathbb{E}[\pi(X)] = 3/2$. This price penalises further policyholders with X close to 1, which we know are more likely to satisfy $D = 1$. Hence, we expect the price to proxy-discriminate even more than $\mu(X)$; moreover this will be in a gratuitous way, as the increased level of proxy discrimination does not benefit prediction accuracy. Let us now calculate $\mathrm{PD}(\pi)$. Using similar arguments as above, it follows that $\mathbb{E}\left[\left(\pi(X) - c^* - \sum_{d\in\{0,1\}} \mu(X,d)v_d^*\right)^2\right] = \operatorname{Var}(2X)$. Then,

$$\mathrm{PD}(\pi) = \frac{\operatorname{Var}(2X)}{\operatorname{Var}(3X)} = \frac{4}{9} > \frac{1}{4} = \mathrm{PD}(\mu),$$

such that the increase in the degree of proxy discrimination is reflected in our metric.

Finally, any price that is free of proxy discrimination according to Definition 3 will, for some $c \in \mathbb{R}$, $\boldsymbol{v} \in \mathcal{V}$, take the form

$$\mu^*(X) = c + v_0\mu(X,0) + v_1\mu(X,1) = c + \frac{1}{2}(v_0 + 3v_1) + (v_0 + v_1)X.$$

This allows for different choices of prices that avoid proxy discrimination. For example

$$\mu_1^*(X) := 1 + X \text{ (for } v_0 + v_1 = 1) \quad \text{or} \quad \mu_2^*(X) := \frac{5}{4} + \frac{1}{2}X \text{ (for } v_0 + v_1 = 1/2),$$

where $\mathbb{E}[\mu_1^*(X)] = \mathbb{E}[\mu_2^*(X)] = \mathbb{E}[\mu(X)]$. For the price $\mu_2^*(X)$ by choosing $v_0 + v_1 < 1$ we have a decreased sensitivity to claim costs and we compensate by a higher flat premium part c.

■

2.5 Attribution of proxy discrimination to individual covariates

Given the measurement of proxy discrimination by (7), a next question of interest is which (subsets of) covariates – elements of $\boldsymbol{X}$ – are mostly responsible. Here we draw again from literature on

Global Sensitivity Analysis (e.g. Saltelli et al., 2008). Let the dimension of $\boldsymbol{X}$ be q and $\mathcal{S} \subseteq \{1, \dots, q\} =: \mathcal{Q}$ a set of indices, such that $\boldsymbol{X}_{\mathcal{S}}$ is the corresponding sub-vector of $\boldsymbol{X}$. Analogously, denote $\mathcal{S}^c = \mathcal{Q} \setminus \mathcal{S}$, and $\boldsymbol{X}_{\mathcal{S}^c}$, such that $\boldsymbol{X} = (\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}^c})$.

Noting the form (8), we can attribute the variance in the numerator to the subset $\mathcal{S}$ of covariates, by conditioning on sub-vectors. Following Definition 4, denote the regression residual by

$$\Lambda(\pi, \boldsymbol{X}) := \pi(\boldsymbol{X}) - c^* - \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{X}, d) v_d^*. \tag{9}$$

We now define two metrics that reflect the contribution of (a subset of) covariates to proxy discrimination.

Definition 5. *For the proxy discrimination metric* PD *of* (7) *and* $\Lambda(\pi, \boldsymbol{X})$ *as in* (9)*, we define the contribution of the sub-vector* $\boldsymbol{X}_{\mathcal{S}}$, $\mathcal{S} \subseteq \mathcal{Q}$ *to proxy discrimination by the two metrics,*

$$\mathrm{PD}_{\mathcal{S}}(\pi) = \frac{\mathrm{Var}(\mathbb{E}[\Lambda(\pi, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}])}{\mathrm{Var}(\pi(\boldsymbol{X}))}, \tag{10}$$

$$\widetilde{\mathrm{PD}}_{\mathcal{S}}(\pi) = \frac{\mathrm{Var}(\Lambda(\pi, \boldsymbol{X})) - \mathrm{Var}(\mathbb{E}[\Lambda(\pi, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}^c}])}{\mathrm{Var}(\pi(\boldsymbol{X}))}. \tag{11}$$

When $\mathcal{S} = \{i\}$*, we write* $\mathrm{PD}_i(\pi)$, $\widetilde{\mathrm{PD}}_i(\pi)$.

The metric $\mathrm{PD}_{\mathcal{S}}$ is thus understood as the sensitivity of the residual $\Lambda(\pi, \boldsymbol{X})$ to the subset of covariates $\boldsymbol{X}_{\mathcal{S}}$, reflecting the amount of variability in $\Lambda(\pi, \boldsymbol{X})$ driven by $\boldsymbol{X}_{\mathcal{S}}$. The metric $\widetilde{\mathrm{PD}}_{\mathcal{S}}$ reflects the expected reduction in the variance of $\Lambda(\pi, \boldsymbol{X})$ achieved by averaging out $\boldsymbol{X}_{\mathcal{S}}$. When $\mathcal{S} = \{i\}$, then PD_i is identified by a (rescaled) Sobol' Index (or first-order sensitivity), while $\widetilde{\mathrm{PD}}_i$ is known as a *total sensitivity* (Saltelli et al., 2008). A difference to standard sensitivity measures is that here we are normalising with $\mathrm{Var}(\pi(\boldsymbol{X}))$ – rather than $\mathrm{Var}(\Lambda(\pi, \boldsymbol{X}))$ – to maintain the direct connection with the global PD metric (7) (note that this suppresses the scale of $\mathrm{PD}_{\mathcal{S}}$, $\widetilde{\mathrm{PD}}_{\mathcal{S}}$).

The metrics introduced in Definition 5 have the following properties, which we state without proof.

Proposition 3. *The metrics* $\mathrm{PD}_{\mathcal{S}}$*, and* $\widetilde{\mathrm{PD}}_{\mathcal{S}}$ *satisfy the following properties.*

i) $0 \leq \mathrm{PD}_{\mathcal{S}}(\pi), \widetilde{\mathrm{PD}}_{\mathcal{S}}(\pi) \leq \mathrm{PD}(\pi)$.

ii) a) If $\boldsymbol{X}_{\mathcal{S}} \perp\!\!\!\perp \Lambda(\pi, \boldsymbol{X})$*, then* $\mathrm{PD}_{\mathcal{S}}(\pi) = 0$.

b) If $\Lambda(\pi, \boldsymbol{X})$ *is* $\boldsymbol{X}_{\mathcal{S}}$*-measurable, then* $\mathrm{PD}_{\mathcal{T}}(\pi) = \mathrm{PD}(\pi)$ *for all* $\mathcal{S} \subseteq \mathcal{T} \subseteq \mathcal{Q}$.

iii) a) If $\Lambda(\pi, \boldsymbol{X})$ *is* $\boldsymbol{X}_{\mathcal{S}^c}$*-measurable, then* $\widetilde{\mathrm{PD}}_{\mathcal{T}}(\pi) = 0$*, for all* $\mathcal{T} \subseteq \mathcal{S}$.

b) If $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp \Lambda(\pi, \boldsymbol{X})$*, then* $\widetilde{\mathrm{PD}}_{\mathcal{S}}(\pi) = \mathrm{PD}(\pi)$.

iv) If there are functions g, h *such that* $\Lambda(\pi, \boldsymbol{X}) = g(\boldsymbol{X}_{\mathcal{S}}) + h(\boldsymbol{X}_{\mathcal{S}^c})$ *and* $\boldsymbol{X}_{\mathcal{S}} \perp\!\!\!\perp \boldsymbol{X}_{\mathcal{S}^c}$*, then* $\mathrm{PD}_{\mathcal{S}}(\pi) + \mathrm{PD}_{\mathcal{S}^c}(\pi) = \mathrm{PD}(\pi)$ *and* $\widetilde{\mathrm{PD}}_{\mathcal{S}}(\pi) + \widetilde{\mathrm{PD}}_{\mathcal{S}^c}(\pi) = \mathrm{PD}(\pi)$.

Part i) of Proposition 3 is a natural condition for stating that the metrics $\mathrm{PD}_{\mathcal{S}}(\pi)$, $\widetilde{\mathrm{PD}}_{\mathcal{S}}(\pi)$ reflect contributions to the overall proxy discrimination $\mathrm{PD}(\pi)$. Parts ii) and iii) give conditions for the metrics taking their extremal values. The conditions are complementary, using different independence or measurability assumptions to reflect irrelevance or full relevance of $\boldsymbol{X}_{\mathcal{S}}$. Finally, part iv) gives strong conditions for being able to additively decompose the proxy discrimination metric $\mathrm{PD}(\pi)$. If an additive decomposition is required, methods such as the Shapley value can be used; for a detailed discussion see the Supplementary materials. We resume the discussion of the properties of $\mathrm{PD}_{\mathcal{S}}$, $\widetilde{\mathrm{PD}}_{\mathcal{S}}(\pi)$ in Section 3.1.

3 Structural properties, price adjustments, and local measures

3.1 Structural properties

The ideas of demographic unfairness and proxy discrimination discussed in Section 2 related to the statistical properties and construction of pricing functionals π. Here we associate such properties with structural properties of the data generating process, that is, with features of the joint distribution $\mathbb{P}(Y, \boldsymbol{X}, D)$. First, we define a number of such properties.

Definition 6 (Structural properties)**.**

P1	$\boldsymbol{X} \perp\!\!\!\perp D$	*(independence)*
P2	$Y \perp\!\!\!\perp D \mid \boldsymbol{X}$	*($\boldsymbol{X}$-sufficiency)*
P3	$\mu(\boldsymbol{X}, D) = \mu(\boldsymbol{X})$	*(weak $\boldsymbol{X}$-sufficiency)*
P4	$\sigma(\boldsymbol{X}) \subseteq \sigma(D)$	*($\boldsymbol{X}$-irrelevance)*
P5	$Y \perp\!\!\!\perp \boldsymbol{X} \mid D$	*(D-sufficiency)*
P6	$\mu(\boldsymbol{X}, D) = \mu(D)$	*(weak D-sufficiency)*

The formulation of properties P1-P6 is not reliant on any assumed causal relation between Y, $\boldsymbol{X}$ and D. Nonetheless, one can formulate Directed Acyclical Graphs (DAG) representing causal relations such that, e.g. properties P2 or P5 are satisfied – for causal perspectives on proxy discrimination see Tschantz (2022); Araiza Iturria et al. (2024); Côté et al. (2024).

In the light of Definitions 1, 2, 3, 4, the relationships summarised in Proposition 4, below, hold. These show the implications of structural properties P1-P6 for the metrics UF and PD, for either a general price $\pi(\boldsymbol{X})$ or, more specifically, the unawareness price $\mu(\boldsymbol{X})$. As the properties P1-P6 are formulated with respect to the data generating process rather than arbitrary pricing functionals, they are strong and only give rise to sufficient conditions.

Proposition 4.

i) If P1 holds, any pricing functional $\pi(\boldsymbol{X})$ *satisfies demographic parity and* $\mathrm{UF}(\pi) = 0$.

ii) If any of P1, P2 or P3 hold, the unawareness price $\mu(\boldsymbol{X})$ *avoids proxy discrimination and* $\mathrm{PD}(\mu) = 0$.

iii) If P4 holds, any pricing functional $\pi(\boldsymbol{X})$ *is demographically unfair and* $\mathrm{UF}(\pi) = 1$, *except if* $\pi(\boldsymbol{X}) \equiv \pi$, *a constant.*

iv) If any of P4, P5 or P6 hold, any pricing functional $\pi(\boldsymbol{X})$ *is proxy discriminatory and* $\mathrm{PD}(\pi) = 1$, *except if* $\pi(\boldsymbol{X}) \equiv \pi$, *a constant.*

Proof. Part i) follows from P1 $\implies$ $\pi(\boldsymbol{X}) \perp\!\!\!\perp D$. For part ii) note from (3), that P1 $\implies$ $\mathbb{P}(D = d|\boldsymbol{x}) = \mathbb{P}(D = d)$, from which $\mathrm{PD}(\mu) = 0$ follows. Finally, P2 $\implies$ P3. Then the regression in Definition 4 reduces to $\min_{c,\boldsymbol{v}} \mathbb{E}\left[\left(\mu(\boldsymbol{X}) - c - \mu(\boldsymbol{X})\sum_d v_d\right)^2\right] = 0$.

Part iii) is immediate. For part iv) note that either of P4 or P5 implies P6. The implication from P6 is a special case of $\min_{c,\boldsymbol{v}} \mathbb{E}\left[\left(\pi(\boldsymbol{X}) - c - \sum_d \mu(d)v_d\right)^2\right] = \mathrm{Var}(\pi(\boldsymbol{X}))$. □

Proposition 4, parts i)-ii), gives conditions for avoiding demographic unfairness or proxy discrimination; in ii) limiting to the case of unawareness prices. Demographic fairness relates to the joint law of the pricing functional π and the response Y, hence the strong requirement P1 arises as a natural sufficient condition. Property P2 means that Y depends on D only via $\boldsymbol{X}$. Hence measurement of non-protected characteristics $\boldsymbol{X}$ eliminates any benefit to predictions from collecting protected characteristics D. This in turn implies P3, which means that the best-estimate prices are insensitive in D. Hence, if we restrict to unawareness prices, these conditions guarantee the absence of proxy discrimination. Conversely, parts iii)-iv) of Proposition 4 give conditions for maximal levels of demographic unfairness and proxy discrimination. Here, properties P4-P6, in different ways, mean that knowing $\boldsymbol{X}$ in addition to D adds no new information useful for predicting the claims Y.

We now turn our attention to the way that structural properties of the data generating process impact on the sensitivity of the PD measure to covariate sub-vectors $\boldsymbol{X}_{\mathcal{S}}$. In Proposition 5 below, we deal with the common case of unawareness prices, providing dependence scenarios, under which the contributions of sub-vectors of $\boldsymbol{X}$ are either zero or equal to the total level of proxy discrimination $\mathrm{PD}(\mu)$. These scenarios hence represent cases of full relevance of $\boldsymbol{X}_{\mathcal{S}}$ and irrelevance of $\boldsymbol{X}_{\mathcal{S}^c}$ with respect to the alternative metrics $\mathrm{PD}_{\mathcal{S}}$ and $\widetilde{\mathrm{PD}}_{\mathcal{S}}$.

Proposition 5. *i) Assume that* $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp (Y, D) \mid \boldsymbol{X}_{\mathcal{S}}$. *Then, for the unawareness price* $\pi(\boldsymbol{X}) = \mu(\boldsymbol{X})$ *the following hold.*

a) $\mathrm{PD}_{\mathcal{S}}(\mu) = \mathrm{PD}(\mu)$ *and* $\widetilde{\mathrm{PD}}_{\mathcal{S}^c}(\mu) = 0$.

b) If additionally $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp \boldsymbol{X}_{\mathcal{S}}$, *then it it also holds that* $\widetilde{\mathrm{PD}}_{\mathcal{S}}(\mu) = \mathrm{PD}(\mu)$ *and* $\mathrm{PD}_{\mathcal{S}^c}(\mu) = 0$.

ii) Let the best-estimate price take the form $\mu(\boldsymbol{X}, D) = g(\boldsymbol{X}) + h(D)$ *and assume that* $\mathrm{Cov}(g(\boldsymbol{X}), h(D)) \geq 0$ *and* $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp D \mid \boldsymbol{X}_{\mathcal{S}}$*. Then, for the unawareness price* $\pi(\boldsymbol{X}) = \mu(\boldsymbol{X})$ *the following hold.*

a) $\mathrm{PD}_{\mathcal{S}}(\mu) = \mathrm{PD}(\mu)$ *and* $\widetilde{\mathrm{PD}}_{\mathcal{S}^c}(\mu) = 0$*.*

b) If additionally $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp \boldsymbol{X}_{\mathcal{S}}$*, then it also holds that* $\widetilde{\mathrm{PD}}_{\mathcal{S}}(\mu) = \mathrm{PD}(\mu)$ *and* $\mathrm{PD}_{\mathcal{S}^c}(\mu) = 0$*.*

Proof. i) a) We have that $\mathbb{P}(D = d \mid \boldsymbol{X}) = \mathbb{P}(D = d \mid \boldsymbol{X}_{\mathcal{S}})$ by the conditional independence assumption and also $\mu(\boldsymbol{X}, D) = \mathbb{E}[Y \mid \boldsymbol{X}_{\mathcal{S}}, D] =: \mu(\boldsymbol{X}_{\mathcal{S}}, D)$. Consequently, noting (3), we have that

$$\mu(\boldsymbol{X}) = \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{X}_{\mathcal{S}}, d)\mathbb{P}(D = d \mid \boldsymbol{X}_{\mathcal{S}}) \implies \Lambda(\mu, \boldsymbol{X}) = \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{X}_{\mathcal{S}}, d)\left(\mathbb{P}(D = d \mid \boldsymbol{X}_{\mathcal{S}}) - v_d^*\right) - c^*.$$

As the last expression is $\boldsymbol{X}_{\mathcal{S}}$-measurable, it holds that $\mathbb{E}\left[\Lambda(\mu, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}\right] = \Lambda(\mu, \boldsymbol{X})$. Hence,

$$\mathrm{PD}_{\mathcal{S}}(\mu) = \frac{\mathrm{Var}(\Lambda(\mu, \boldsymbol{X}))}{\mathrm{Var}(\pi(\boldsymbol{X}))} = \mathrm{PD}(\mu), \quad \widetilde{\mathrm{PD}}_{\mathcal{S}^c}(\pi) = \frac{\mathrm{Var}(\Lambda(\pi, \boldsymbol{X})) - \mathrm{Var}(\mathbb{E}[\Lambda(\pi, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}])}{\mathrm{Var}(\pi(\boldsymbol{X}))} = 0.$$

b) Here it is sufficient to show that $\mathbb{E}[\Lambda(\mu, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}^c}]$ is a constant. It has already been shown that $\Lambda(\mu, \boldsymbol{X})$ is $\boldsymbol{X}_{\mathcal{S}}$-measurable; hence the result follows from $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp \boldsymbol{X}_{\mathcal{S}}$.

ii) a) The unawareness price will take the form $\mu(\boldsymbol{X}) = g(\boldsymbol{X}) + \mathbb{E}[h(D) \mid \boldsymbol{X}]$. We consider the quantity to be minimised in the numerator of the PD measure. From the specific form of the best-estimate and unawareness prices we have:

$$\begin{aligned} &\mu(\boldsymbol{X}) - \mathbb{E}[\mu(\boldsymbol{X})] - \sum_{d \in \mathfrak{D}} v_d \left(\mu(\boldsymbol{X}, d) - \mathbb{E}[\mu(\boldsymbol{X}, d)]\right) \\ &= \left(1 - \sum_{d \in \mathfrak{D}} v_d\right)\left(g(\boldsymbol{X}) - \mathbb{E}[g(\boldsymbol{X})]\right) + \mathbb{E}[h(D) \mid \boldsymbol{X}] - \mathbb{E}[h(D)]. \end{aligned}$$

Consequently, we can just let $v := \sum_{d \in \mathfrak{D}} v_d \in [0, 1]$ and optimise over that. By checking the Karush-Kuhn-Tucker conditions (calculations not documented here), we find that the condition $\mathrm{Cov}(g(\boldsymbol{X}), h(D)) \geq 0$ is necessary and sufficient for $v = 1$. Thus we obtain $\Lambda(\mu, \boldsymbol{X}) = \mathbb{E}[h(D) \mid \boldsymbol{X}] - \mathbb{E}[h(D)] = \mathbb{E}[h(D) \mid \boldsymbol{X}_{\mathcal{S}}] - \mathbb{E}[h(D)]$, where we used the additional assumption $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp D \mid \boldsymbol{X}_{\mathcal{S}}$ in the second equation. Furthermore, $\mathbb{E}[\Lambda(\mu, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}] = \Lambda(\mu, \boldsymbol{X})$, from which the stated result follows.

b) Again, it is sufficient to show that $\mathbb{E}[\Lambda(\mu, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}^c}]$ is a constant, which follows from the additional independence assumption.

□

3.2 Local measurement of demographic unfairness

We introduce local (i.e., policyholder-specific) measures, based on the difference of the price-in-use and the a benchmark fair price. For demographically fair prices, a standard construction has been

via optimal transport (OT) methods (Gordaliza et al., 2019; Chiappa et al., 2020). In an insurance context, such methods aim at achieving independence between prices $\pi(\boldsymbol{X})$ and the protected characteristics D by suitable variable transformations; for insurance-specific investigations see Lindholm et al. (2024); Charpentier (2024). The broader mathematical problem of approximating a random variable with another, subject to an independence constraint, is treated by Delbaen & Majumdar (2024).

Here, to create a demographically fair benchmark price, we follow an *Output OT* approach. This refers to a transformation (post-processing) of prices $\pi(\boldsymbol{X})$ in order to achieve independence from D. Denote the conditional distributions of the prices by $G_d(m) = \mathbb{P}(\pi(\boldsymbol{X}) \leq m \mid D = d),\ d \in \mathfrak{D}$, and assume for simplicity that they are continuous. Then, for any continuous distribution G, we may construct the prices

$$\tilde{\pi}(\boldsymbol{X}, D) = \sum_{d \in \mathfrak{D}} \mathbb{1}_{\{D=d\}} G^{-1} \circ G_d\big(\pi(\boldsymbol{X})\big). \tag{12}$$

The price $\tilde{\pi}(\boldsymbol{X}, D)$ satisfies $\mathbb{P}(\tilde{\pi}(\boldsymbol{X}, D) \leq m) = G(m)$ and $\mathbb{P}(\tilde{\pi}(\boldsymbol{X}, D) \leq m \mid D = d) = G(m),\ d \in \mathfrak{D}$, such that $D \perp\!\!\!\perp \tilde{\pi}(\boldsymbol{X}, D) \sim G$; the construction (12) works by making the conditional distribution of prices the same on each demographic subgroup $D = d$. Note that the transformed price $\tilde{\pi}(\boldsymbol{X}, D)$ explicitly depends on D – even as $\pi(\boldsymbol{X})$ does not. This is a form of direct discrimination arising in the process of engineering demographic parity; see Lindholm et al. (2024) for more discussion of this point. Finally, to construct a demographically fair benchmark, we need to select the target distribution G. A standard choice is given by Chzhen et al. (2020)

$$G^{-1}(u) = \sum_{d' \in \mathfrak{D}} \mathbb{P}\left(D = d'\right) G_{d'}^{-1}(u), \tag{13}$$

From now on we will consistently refer to the *Output OT price* as the construction $\tilde{\pi}(\boldsymbol{X}, D)$ from (12) and (13).

We can now proceed with the definition of a local measure of demographic unfairness.

Definition 7. *Consider a pricing functional π and the Output OT price $\tilde{\pi}(\boldsymbol{X}, D)$. Then, for the policyholder with profile $\boldsymbol{X} = \boldsymbol{x}$, $D = d$, the local measure of demographic unfairness is defined as:*

$$\delta_{\mathrm{UF}}(\boldsymbol{x}, d; \pi) = \pi(\boldsymbol{x}) - \tilde{\pi}(\boldsymbol{x}, d). \tag{14}$$

If $\pi(\boldsymbol{X}) = \mu(\boldsymbol{X})$ is the unawareness price, we just write $\delta_{\mathrm{UF}}(\boldsymbol{x}, d)$.

A value of $\delta_{\mathrm{UF}}(\boldsymbol{x}, d; \pi) > 0$ implies that policyholders with attributes $\boldsymbol{X} = \boldsymbol{x}$, $D = d$ suffer from demographic unfairness in the sense that the price they are charged is higher than the corresponding benchmark demographically fair price. Clearly, if $\boldsymbol{X}$ and D are already independent, the prices $\pi(\boldsymbol{X})$ and $\tilde{\pi}(\boldsymbol{X}, D)$ coincide such that the measure becomes zero. This is stated formally below.

Proposition 6. *If P1 in Definition 6 holds, then $\delta_{\mathrm{UF}}(\boldsymbol{x}, d) = 0$.*

Furthermore, it is of interest to establish conditions for the sign of the metric $\delta_{\mathrm{UF}}(\boldsymbol{x}, d; \pi)$. In the simple but common case of a binary D, this is straightforward. Denote by $\preceq_{\mathrm{st}}$ precedence in

the usual stochastic order, such that for two distributions F, G, we have that $F \preceq_{\text{st}} G \iff F(x) \geq G(x)$ for all x; this is a strong condition not allowing the crossing of distributions.

Proposition 7. *Let* $\mathfrak{D} = \{0, 1\}$ *and* $G_0 \preceq_{\text{st}} G_1$.

i) $\delta_{\text{UF}}(\boldsymbol{x}, 0; \pi) \leq 0$ *and* $\delta_{\text{UF}}(\boldsymbol{x}, 1; \pi) \geq 0$.

ii) For $\boldsymbol{x}$ *such that* $\mathbb{P}(D = 0 \mid \boldsymbol{X} = \boldsymbol{x}) > 0$, *it holds that:*

$$\mathbb{E}[\delta_{\text{UF}}(\boldsymbol{X}, D; \pi) \mid \boldsymbol{X} = \boldsymbol{x}] > 0 \iff$$
$$\frac{\mathbb{P}(D = 1 \mid \boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}(D = 0 \mid \boldsymbol{X} = \boldsymbol{x})} > \frac{-\delta_{\text{UF}}(\boldsymbol{x}, 0; \pi)}{\delta_{\text{UF}}(\boldsymbol{x}, 1; \pi)} = \frac{G^{-1} \circ G_0(\pi(\boldsymbol{x})) - \pi(\boldsymbol{x})}{\pi(\boldsymbol{x}) - G^{-1} \circ G_1(\pi(\boldsymbol{x}))} \geq 0,$$

where G *is given by* (13).

Proof.

i) The statement follows by noting that construction (13) implies $G_0 \preceq_{\text{st}} G \preceq_{\text{st}} G_1$ and consequently

$$\delta_{\text{UF}}(\boldsymbol{x}, 0; \pi) = \pi(\boldsymbol{x}) - G^{-1} \circ G_0(\pi(\boldsymbol{x})) \leq 0,$$
$$\delta_{\text{UF}}(\boldsymbol{x}, 1; \pi) = \pi(\boldsymbol{x}) - G^{-1} \circ G_1(\pi(\boldsymbol{x})) \geq 0.$$

ii) We have that

$$\mathbb{E}[\delta_{\text{UF}}(\boldsymbol{X}, D; \pi) \mid \boldsymbol{X} = \boldsymbol{x}] =$$
$$= \mathbb{P}(D = 0 \mid \boldsymbol{X} = \boldsymbol{x})\delta_{\text{UF}}(\boldsymbol{x}, 0; \pi) + \mathbb{P}(D = 1 \mid \boldsymbol{X} = \boldsymbol{x})\delta_{\text{UF}}(\boldsymbol{x}, 1; \pi),$$

with the stated result following directly from the inequalities of part i).

□

To interpret part i) of Proposition 7, first note that $G_0 \preceq_{\text{st}} G_1$ means that the policyholders with protected attribute $D = 1$ tend to be considered as higher risk, according to best-estimate prices. So these are the policyholders for whom the use of the Output OT price (12) should confer a discount, compared with the unawareness price. For part ii), we consider a situation where for a policyholder with $\boldsymbol{X} = \boldsymbol{x}$ the value of D may not be known. The left-hand side of the stated condition implies that, on average, the local measure of unfairness will be positive when $\mathbb{P}(D = 1 \mid \boldsymbol{X} = \boldsymbol{x})/\mathbb{P}(D = 0 \mid \boldsymbol{X} = \boldsymbol{x})$ is high, such that there is a high chance that, given $\boldsymbol{X} = \boldsymbol{x}$, the policyholder belongs to the demographically disadvantaged group $D = 1$. Furthermore, the condition is more likely to be satisfied when the ratio $-\delta_{\text{UF}}(\boldsymbol{x}, 0; \pi)/\delta_{\text{UF}}(\boldsymbol{x}, 1; \pi)$ is low. That fraction becomes small if the comparative disadvantage for group $D = 1$ (denominator) becomes much higher than the comparative advantage of group $D = 0$ (numerator), given the information $\boldsymbol{X} = \boldsymbol{x}$.

These ideas are illustrated in the following example.

Example 3. We continue from Example 2, where $\mu(X, D) = \frac{1}{2} + X + D$ and $X \sim \mathrm{U}(0, 1)$. Note that $\mu(x, 1) > \mu(x, 0)$ for all x. However, we now allow a variety of positive and negative dependence relations between (X, D) by assuming that

$$\mathbb{P}(D = 1 \mid X = x) = \frac{1 - a}{2} + ax, \quad a \in (-1, 1].$$

By setting $a = 1$, we recover the exact setting of Example 2; $0 < a < 1$ gives a weaker positive dependence, while $-1 < a < 0$ gives a negative dependence. With this modification the unawareness price changes to $\mu(X) = \frac{2-a}{2} + (a + 1)X$.

We now evaluate the local measure of unfairness (14) for the unawareness price $\mu(X)$ and various levels of a. The calculations are simple but tedious and are not reported here. In Figure 1 we plot the functions $\delta_{\mathrm{UF}}(x, d)$ (blue for $d = 0$, red for $d = 1$), as well as their conditional mean $\mathbb{E}[\delta_{\mathrm{UF}}(X, D) \mid X = x]$ (black) for $a = 0.75$ (positive dependence) and $a = -0.75$ (negative dependence). While the shapes appear similar, there are two observations. First, for $a = 0.75$, the red line is above zero and the blue line below, showing that policyholders with $D = 1$ are adversely affected by demographic unfairness, while policyholders with $D = 0$ are benefiting. This pattern is reversed when the dependence of (X, D) becomes negative ($a = -0.75$). Second, the two plots are at very different scales, with the absolute value of local demographic unfairness being an order of magnitude higher in the case of positive dependence. The reason is that, for $a = 0.75$, the positive dependence of (X, D) works in the same direction as the impact of each of those two variables on claims costs. However, when dependence is negative, then the unawareness price becomes less sensitive to X (in the extreme $a \to -1$ leads to a constant $\mu(X)$). As a result, for negative dependence much smaller disparities between demographic groups emerge. This point is reinforced by Figure 2, where we plot $\mathbb{E}[\delta_{\mathrm{UF}}(X, D) \mid X = x]$ against different values of the dependence parameter a.

■

3.3 Local measurement of proxy discrimination

Analogously to the last section, we propose a local measure of proxy discrimination. Again, we need for that purpose a benchmark price that avoids proxy discrimination. Any such price takes the form (6); nonetheless, to produce a benchmark one needs to choose the values of $c, v_d,\ d \in \mathfrak{D}$. Here we choose values that minimise the numerator in (7), such that the benchmark is the price closest to the original price $\pi(\boldsymbol{X})$, which is free from proxy discrimination. The difference between a price and its closest proxy-discrimination-free approximation has already been defined in Section 2.5, as the regression residual $\Lambda(\boldsymbol{X}, \pi)$. Hence we re-purpose this quantity as a local measure of proxy discrimination.

Definition 8. *Consider a pricing functional π and $\Lambda(\boldsymbol{X}, \pi)$ in equation (9). Then, for the policyholder with profile $\boldsymbol{X} = \boldsymbol{x}$, the local measure of proxy discrimination is defined as:*

$$\delta_{\mathrm{PD}}(\boldsymbol{x}; \pi) = \Lambda(\boldsymbol{x}, \pi). \tag{15}$$

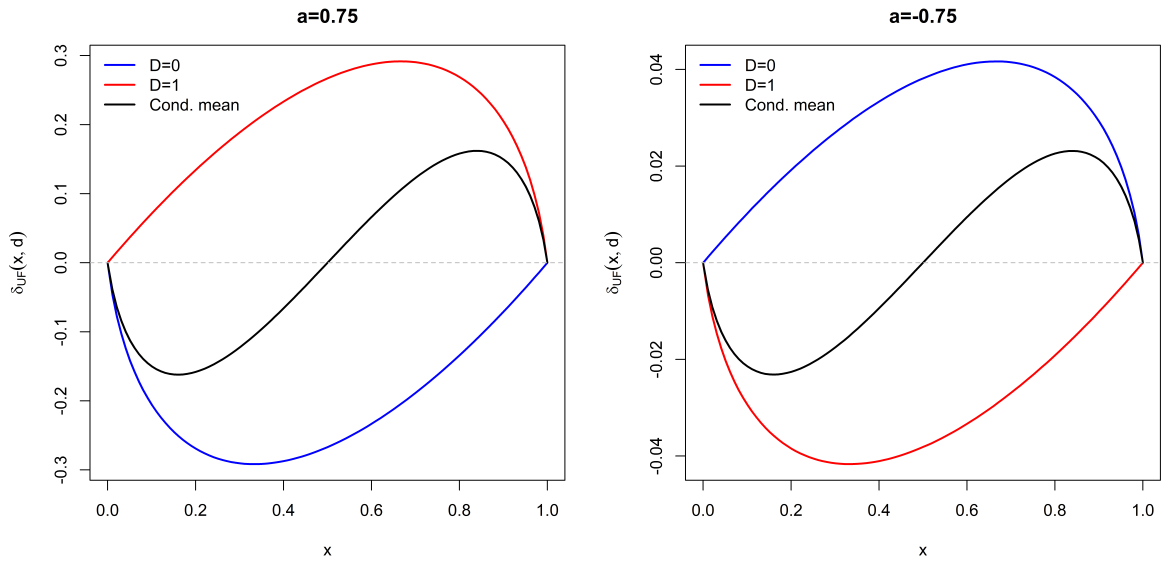

Figure 1: Local unfairness metric $\delta_{\mathrm{UF}}(x, d)$ (blue for $d = 0$, red for $d = 1$) and conditional mean $\mathbb{E}[\delta_{\mathrm{UF}}(X, D) \mid X = x]$ (black) for $a = 0.75$ (left) and $a = -0.75$ (right).

If $\pi(\boldsymbol{X}) = \mu(\boldsymbol{X})$ is the unawareness price, we just write $\delta_{\mathrm{PD}}(\boldsymbol{x})$.

We state Proposition 8 below without proof – for each property it is easy to show that the unawareness price is already free from proxy discrimination, such that $\Lambda(\boldsymbol{X}, \pi)$ is identically zero.

Proposition 8. *If any one of P1, P2 or P3 in Definition 6 holds, then $\delta_{\mathrm{PD}}(\boldsymbol{x}) = 0$.*

We conclude this section with a continuation of our running example.

Example 4. We continue from Example 3, considering the evaluation of our local measure of proxy discrimination. We make a qualitative argument, omitting a formal proof. Recall the forms of the best-estimate and unawareness prices, $\mu(X, D) = 1/2 + X + D$ and $\mu(X) = 1 - a/2 + (1 + a)X$ respectively. If $-1 < a \leq 0$, the slope of the unawareness price in X is less or equal to that of the best-estimate price. As a result $\Lambda(X, \mu) = 0$ and there is no proxy discrimination. However, when $0 < \alpha \leq 1$, proxy discrimination arises. The closest approximation to $\mu(X)$ that avoids proxy discrimination is the price $\mu^*(X) = 1 + X$, which reduces the slope in X to 1. Consequently we have

$$\delta_{\mathrm{PD}}(x) = \begin{cases} 0, & -1 \leq a \leq 0 \\ -\frac{a}{2} + ax, & 0 < a \leq 1, \end{cases}$$

showing how the metric increases linearly with x, at a slope determined by the dependence parameter a. Because of the positive dependence, policyholders with $x > 0.5$ are implicitly inferred to have protected attribute $D = 1$ and hence are disadvantaged in the sense of proxy discrimination; the reverse happens for $x < 0.5$.

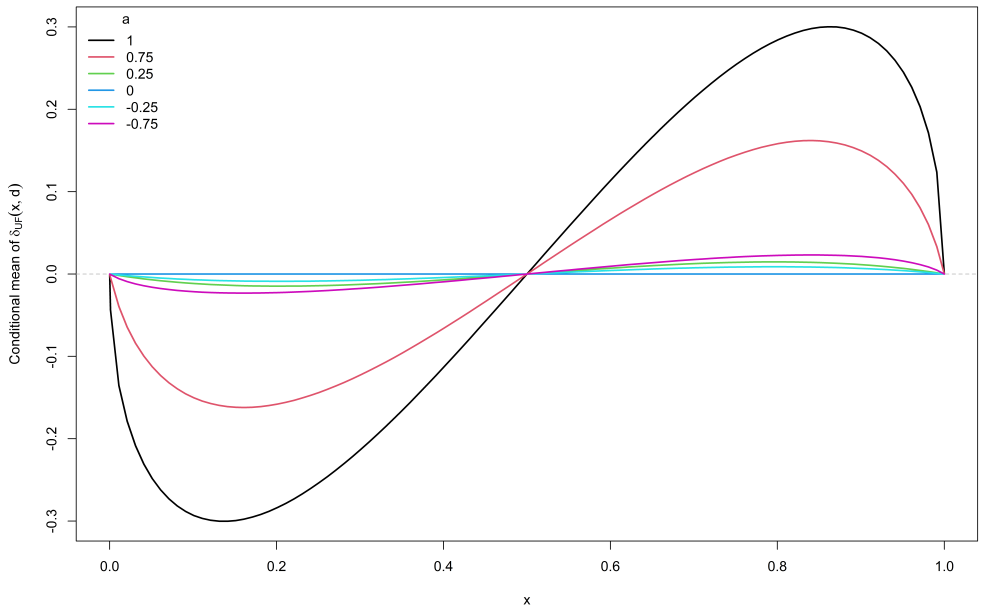

Figure 2: Conditional mean of the local unfairness metric $\mathbb{E}[\delta_{\mathrm{UF}}(X, D) \mid X = x]$ for the unawareness price and a range of dependence parameters a.

4 Case study: insurance prices and ethnicity in a real-world motor portfolio

We now apply the metrics developed in this paper to a real-world non-life motor insurance dataset, previously discussed in Lindholm et al. (2023). The dataset consists of policyholder and vehicle covariates, as well as exposure and claims data for a large portfolio, observed over a single year of exposure. Moreover, the data include protected information $D \in \mathfrak{D} = \{1, 2, 3, 4, 5\}$, relating to the ethnicity code of the policyholder, as defined in the jurisdiction in which the policies were written. The anonymous insurer contributing the data – a large multi-national insurance company – records this information at the time of underwriting the policy. The ethnicity data are not used by the insurer for pricing, but are collected to enable the insurer to monitor and report insurance penetration. To preserve confidentially of the contributing company and to ensure that no commercially sensitive information is provided, the ethnicity categories relating to the levels of D have not been used in our analysis; instead we use an integer code randomly assigned to each category.

The data relate to a period close to the turn of the century and contain 165,511 years of exposure of comprehensive motor insurance policies, with 41,608 claims arising in the same period of exposure. The claims can be assumed to be fully run-off, i.e., no incurred but not reported adjustments need to be made, and they relate to property (motor-hull), liability (third-party property and/or bodily injury) and other associated coverages; we do not disclose the exact coverages to avoid

disclosing potentially commercially sensitive information. Moreover, information on the excesses and deductibles (which influence the frequency) in this portfolio are also not disclosed, thus, the information shown is not commercially useful.

Table 1 shows a summary of the claims, exposures and frequencies for each ethnicity code $D \in \mathfrak{D}$ in the data.

Ethnicity code	Number of claims	Exposure	Frequency
1	5,223	14,317	36.48%
2	965	3,925	24.59%
3	3,354	14,363	23.35%
4	5,249	20,240	25.93%
5	26,817	112,667	23.80%

Table 1: Summary of real-world non-life insurance claims, exposures and frequencies of claims according to the ethnicity codes $D \in \mathfrak{D}$.

The portfolio exhibits substantial heterogeneity in the size of demographic groups. The largest group corresponds to $D = 5$, while other groups range from $D = 2$ with the smallest population to $D = 4$ with approximately one fifth of the size of the largest group. Analysis of claims experience reveals broadly similar frequencies across groups, with the notable exception of policyholders with $D = 1$ who display markedly higher claim frequencies.

The vector of non-protected characteristics $\boldsymbol{X} = (X_1, \ldots, X_q)^\top$ comprises standard rating factors used in non-life insurance pricing, including policyholder and driver demographics (age and gender), vehicle characteristics, and geographical information (note that, within the context of this application, we do not treat gender as a protected characteristic). For each policy i, where $1 \leq i \leq n$, we observe the tuple $(Y_i, \boldsymbol{X}_i, D_i, v_i)$, where v_i represents the exposure period (the fraction of the year for which the policy was active) and serves as a weight in the calculation of claim frequencies.

Throughout this application we compare three types of price: (a) best-estimate prices (1); (b) unawareness prices (2); and (c) discrimination-free prices (6) under the canonical choices $c = 0$, $v_d = \mathbb{P}(D = d)$ as in Lindholm et al. (2022) – note that these choices are exogenous and not the product of any optimisation. These data were modelled by Lindholm et al. (2023), using a plain-vanilla Feed-forward Neural Network (FNN) with embedding layers for the categorical covariates and standard training procedures to regularize the network; we refer the interested reader to that paper and its appendices for further details. Here, using the predictions from the fitted models, we estimate the unfairness metric UF and the proxy discrimination metric PD, using empirical versions of equations (4) and (7).

Table 2 shows the estimated values of the two measures, for the best-estimate, unawareness and discrimination-free prices, with bootstrap standard deviations in parentheses. We observe that the best-estimate price exhibits the highest level of the unfairness metric UF, with UF = 0.0892, indicating that approximately 9% of price variability can be attributed to systematic differences

Price	UF	PD
Best-estimate	0.0892 (0.00095)	0.00924 (0.0000443)
Unawareness	0.0639 (0.00079)	0.00277 (0.0000203)
Discrimination-free	0.0533 (0.00070)	0

Table 2: Demographic unfairness metric UF and proxy discrimination metric PD, for the motor dataset and for the three pricing rules considered. Bootstrap standard errors are in parentheses.

between ethnic groups. The unawareness price yields an intermediate value of UF = 0.0639, while the lowest level of the unfairness measure, UF = 0.0533, comes from the discrimination-free prices. This ordering is not surprising. First, the unawareness price leads to a reduction in demographic unfairness, since it eliminates the direct impact of D on prices. Nonetheless, demographic unfairness persists, because of the statistical dependence of $\mu(\boldsymbol{X})$ and D. Second, beyond removing the direct impact of D on predictions, discrimination-free prices aim to also remove indirect proxying impacts. Nonetheless, it is not obvious that the latter effect holds more generally, given that demographic unfairness and proxy discrimination are deeply different concepts, and there are no guarantees that using a discrimination-free price will also lead to a reduction in (let alone an elimination of) demographic unfairness – see Lindholm et al. (2024) for an extensive discussion.

In Figure 3 we take a more granular view, by evaluating UF for the unawareness and discrimination free prices, on different sub-portfolios arising from segmentation by age. We select age given its importance in terms of potential discriminatory effects, as identified in Lindholm et al. (2023). The demographic unfairness metric exhibits its highest values for young drivers around age 20, with UF reaching approximately 0.3 for both unawareness and discrimination-free prices. This effect diminishes substantially with age, stabilizing at much lower levels (around 0.05) for drivers over 40. While both pricing approaches show similar patterns, the unawareness price consistently demonstrates slightly higher levels of unfairness compared to the discrimination-free price, with the difference being most pronounced for drivers of ages 20-40.

Table 2 also presents the proxy discrimination metrics PD for the three pricing approaches. The standard best-estimate price shows the highest level of proxy discrimination with PD = 0.00924. Note that in this case, the PD measure also implicitly reflects some of the direct discrimination arising from using D as a rating factor. For unawareness prices, the value of the metric is lower, at PD = 0.00277, now reflecting only proxy discrimination. Finally, by construction, the discrimination-free price achieves PD = 0.

The relatively low, though still statistically significant, value of the PD metric for the unawareness price, suggests that proxy discrimination represents a relatively small portion of the overall price variation in this portfolio. This is consistent with the nature of modern motor insurance pricing, where key rating factors like vehicle characteristics and driver age have strong direct effects on expected claims, beyond any relation they may have with protected characteristics. Nonetheless, even such low levels of proxy discrimination may be considered problematic from a regulatory

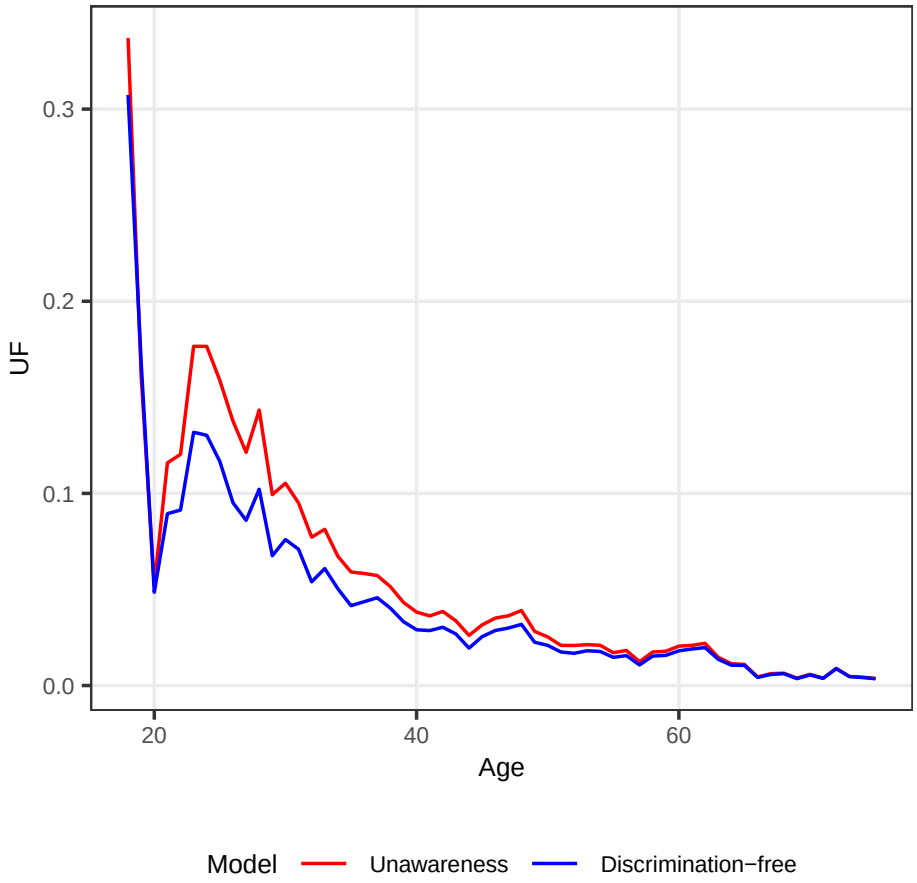

Figure 3: Unfairness metric UF evaluated for the unawareness prices, evaluated on sub-portfolios split by age.

perspective, particularly given the systematic nature of the bias they represent.

In particular, heterogeneity in the portfolio means that substantial discriminatory effects may be suffered by a specific sub-population of policyholders. To investigate this further, we evaluate PD on subsets of the portfolio. We split the portfolio into 5 parts, by ethnicity and calculate PD on each of these. This is of interest, as the results reflects the comparative extent to which different ethnicities are impacted by proxy discrimination, with reference to variation within their own group (segmentation will tend to increase the value of the PD metric by reducing the value of the denominator in (7)). Furthermore, for each ethnicity and for the portfolio as a whole, we also segment the data by age and calculate PD for each age group.

The results are summarised Figure 4; panels correspond, first, to the portfolio as a whole and, then, to individual ethnicity codes. Within each panel we state the value of the PD metric and show graphically PD by age. It is seen that the proxy discrimination metric shows pronounced heterogeneity across both age and ethnicity groups. Most notably, ethnicity code 1 exhibits the highest levels of proxy discrimination, with a sharp peak for young drivers around age 20. While

other ethnicity groups also show some proxy discrimination effects, these are generally of lower magnitude. The age-related pattern of declining discrimination is consistent across all ethnic groups, though the rate and extent of this decline varies. These findings align with and provide additional insight to the observations in Lindholm et al. (2023), who noted the vulnerability of young drivers with ethnicity code 1 in this dataset, to discriminatory pricing effects.

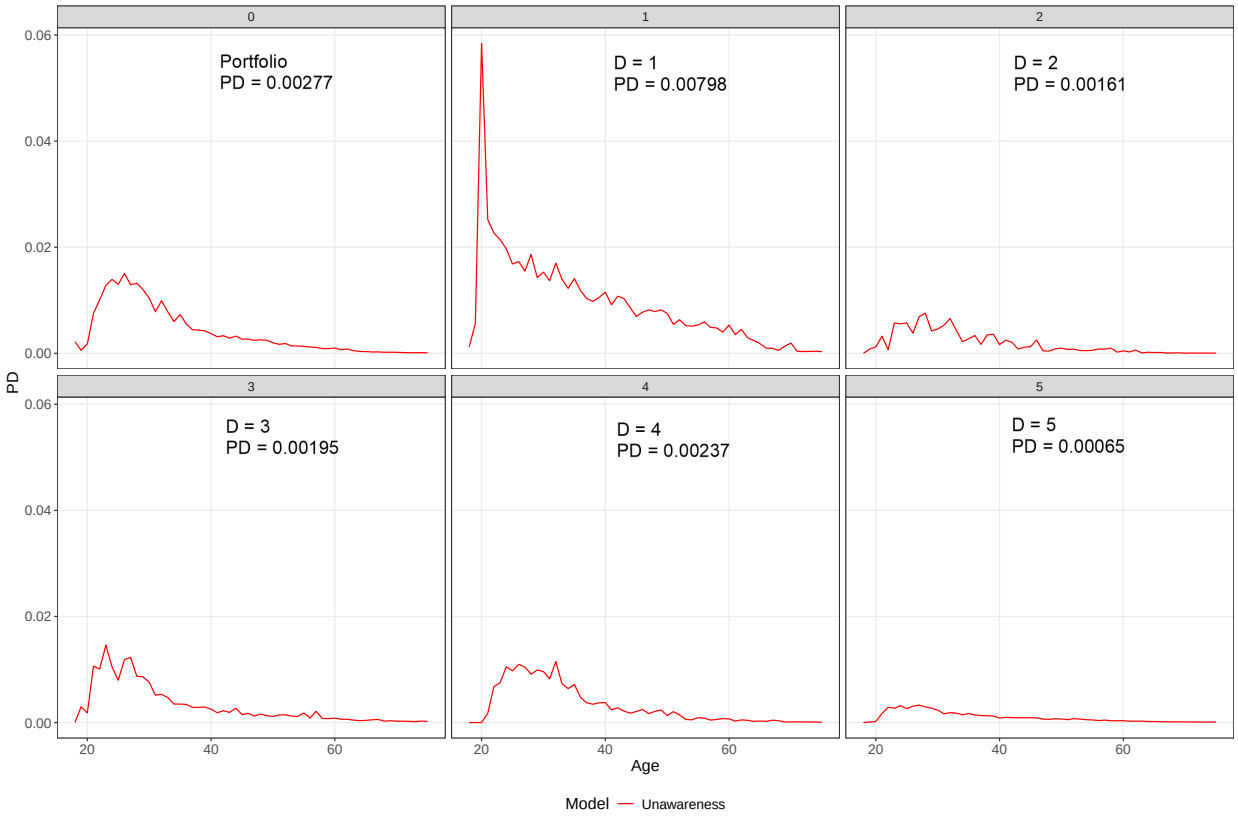

Figure 4: Proxy discrimination metric PD evaluated for the age variable, for each ethnicity code separately and for the portfolio as a whole.

While Figure 4 shows how the heterogeneity of proxy-discriminatory effects across the portfolio, it does not provide an interpretation of the absolute scale of the PD metric. For example, the portfolio shows PD $= 0.00277$ – but is that a lot? To gain insight into this question, we distort the model under which the unawareness prices are calculated. Specifically, we add a fixed cost equal to α to best-estimate prices, for those policyholders with $D = 1$, such that the distorted unawareness price becomes $\sum_{d\in\mathfrak{D}} \left(\mu(\boldsymbol{X}, D) + \alpha \mathbb{1}_{\{D=1\}}\right) \mathbb{P}(D = d \mid \boldsymbol{X})$. Details on this technique are given in the Supplementary materials. There, we also provide an alternative model distortion that controls the extent to which an ethnicity code is inferred more accurately from the data $\boldsymbol{X}$, compared to the baseline model.

In Figure 5, we show how the PD metric for the whole portfolio changes, with the level α of fixed costs added to the costs of ethnicity code 1. As one would expect, the PD measure increases for

positive α, since this means additional costs that are proxied by the covariates $\boldsymbol{X}$ in the calculation of unawareness prices. The PD measure approximately doubles from its baseline of 0.00277 to 0.00531 for $\alpha = 0.02$, while it reaches its approximate minimum of 0.00102 for a negative value $\alpha = -0.04$. From a practical perspective, since α represents the increase in claims frequency for members of demographic group $D = 1$, we can use this analysis to benchmark what levels of the PD metric mean in real-world terms. From the first row of Table 1, it can be seen that the value $\alpha = 0.02$ represents a 5.5% increase in the claims costs for ethnicity 1, while $\alpha = -0.04$ represents a 11% reduction. These are substantial changes for a real-world portfolio. Thus, by evaluation of the sensitivity of PD to model distortions, we can establish that the values of PD we observe are practically significant. More broadly this type of concrete interpretation can allow pricing actuaries and other stakeholders to assess whether observed values of the PD metric constitute problematic levels of proxy discrimination.

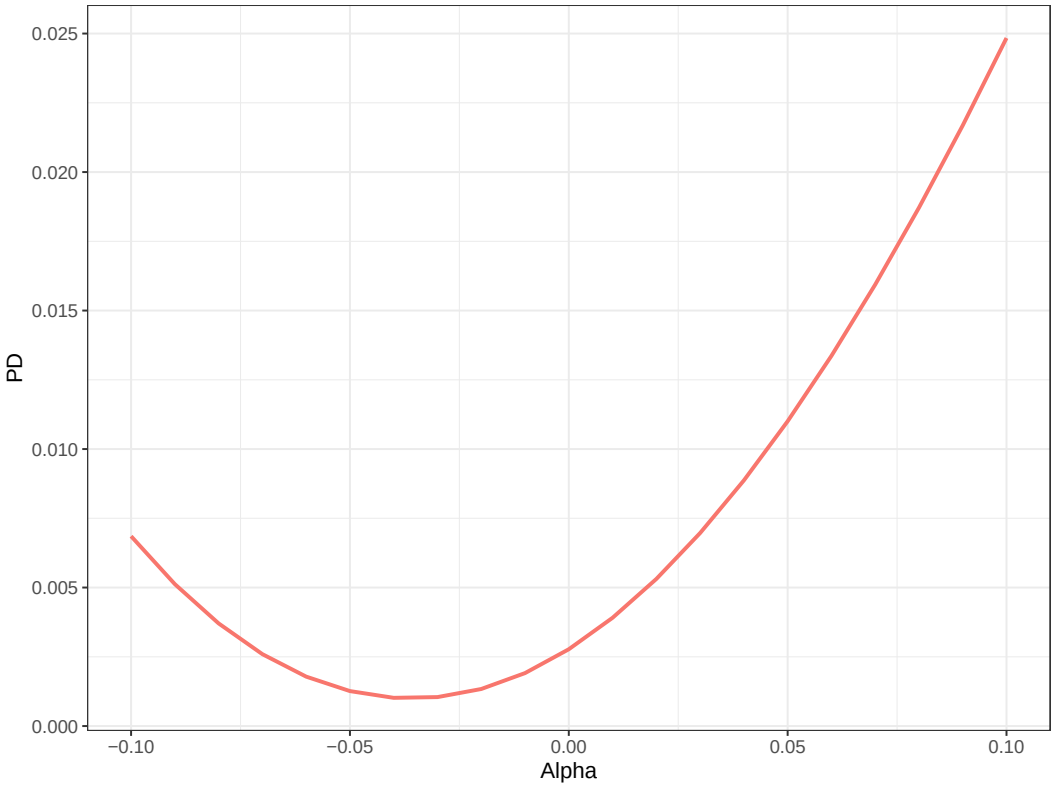

Figure 5: Sensitivity of the PD metric, subject to distortions of the underlying model. The value of α represents additional claims cost applied to ethnicity $D = 1$.

We now demonstrate the practical application of the PD attribution methodology of Section 2.5 using real insurance data. Specifically, we use attributions to identify the most important (single) covariates in our dataset, which contribute most substantially to proxy discrimination when the unawareness price is used. We calculate both the first-order (PD_i) and total sensitivity ($\widetilde{\text{PD}}_i$) metrics defined in Section 2.5 for each covariate. To approximate the conditional expectations in

equations (10), we use a random forest model (Breiman, 2001), fitted to a 50,000-record subset of the dataset. Figure 6 shows these results for each covariate, where, for commercial reasons, we have disguised the covariate name (except for the age). We observe that the first-order and total sensitivity metrics produce different orderings of the covariates, though there is a broad agreement between them regarding the most significant ones. Furthermore, the attribution analysis reveals substantial variation in the extent to which different covariates contribute to proxy discrimination. Several covariates demonstrate notable proxying effects, with covariate 17 showing the strongest first-order contribution to proxy discrimination.

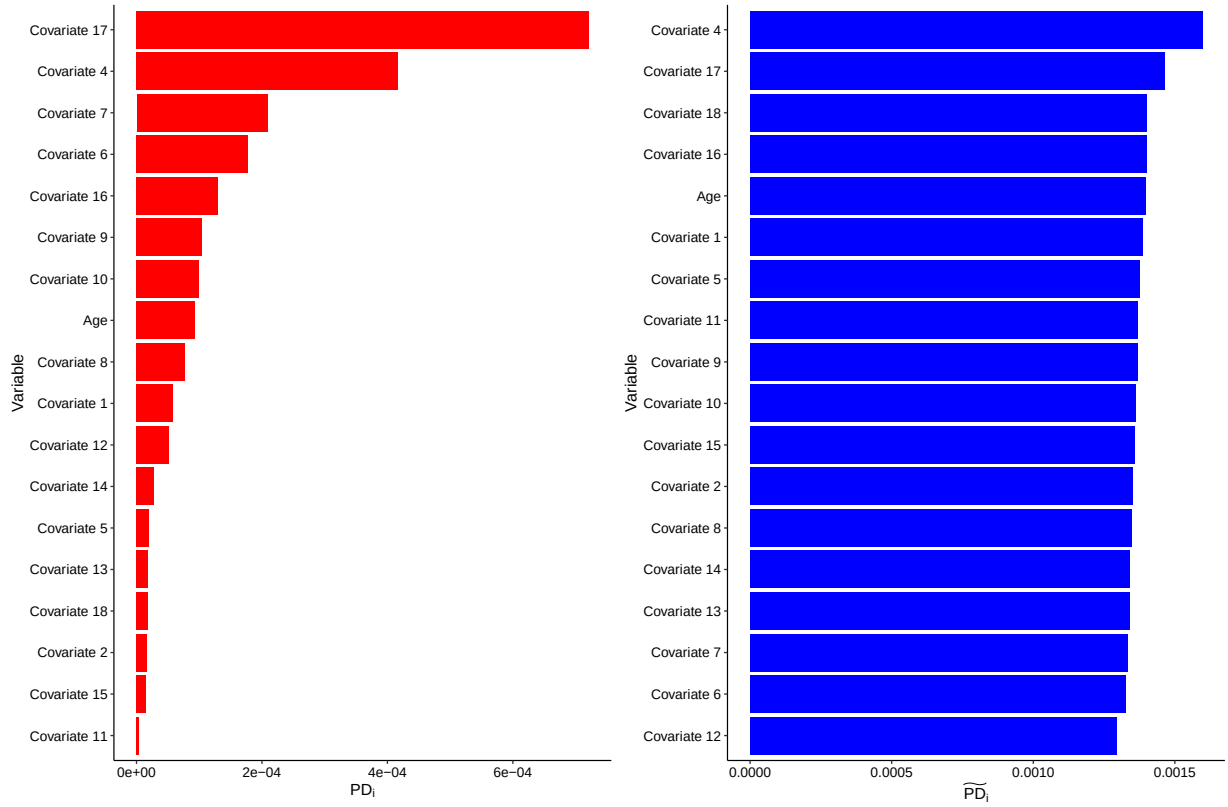

Figure 6: First-order (PD_i) and total sensitivity ($\widetilde{\mathrm{PD}}_i$) of unawareness prices, for each covariate in the real world dataset.

The total sensitivity measures generally exceed their first-order counterparts and, furthermore, total sensitivities differentiate less between covariates. This indicates substantial interaction effects (including statistical dependence) between covariates. Age demonstrates moderate proxy effects by both measures, suggesting that while demographic variables naturally correlate with ethnicity, their contribution to proxy discrimination is not necessarily dominant. Some covariates, e.g., 4 and 12, show notably different rankings between first-order and total sensitivity measures, highlighting the complex interplay between rating factors in insurance portfolios.

In Table 3, we consider covariate 17, to which the value of the PD metric is most strongly attributed using first-order sensitivities, and report its conditional distribution given each ethnicity

code. Marked differences in the distribution of this variable across ethnic groups can be observed – in particular the conditional distribution of the covariate is different for ethnicity code 1, compared to other ethnicities. For instance, Level 4 represents 60.7% of ethnicity code 1 but ranges between 82.0% and 85.8% for other ethnicity codes. Correspondingly, for ethnicity code 1, there is a higher frequency of observing Level 1 of covariate 17. The different distribution for ethnicity code 1 is consistent with the higher proxy discrimination observed for this ethnicity in the earlier analysis (Figure 4) together with the higher observed claim rates for this ethnicity shown in Table 1.

	Levels of covariate 17				
Ethnicity Code	Level 1	Level 2	Level 3	Level 4	Level 5
1	0.201	0.004	0.171	0.607	0.017
2	0.153	0.003	0.016	0.820	0.009
3	0.156	0.001	0.009	0.823	0.011
4	0.137	0.002	0.018	0.832	0.011
5	0.124	0.002	0.006	0.858	0.011

Table 3: Conditional distribution of covariate 17, given ethnicity code (the table rows sum to 1).

Finally, we consider local measurements of demographic unfairness and proxy discrimination using the tools developed in Section 3. First, we illustrate the local measure of demographic unfairness by estimating empirical versions of the distributions underlying (12), calculating $\delta_{\text{UF}}(\boldsymbol{x}, d)$ w.r.t. the unawareness prices predicted for the real-world portfolio and then plotting this quantity against the age covariate, as well as against covariate 17 identified above. This analysis is shown in Figure 7.

For the age covariate (left panel), we observe clear patterns of demographic unfairness that vary both by age and ethnicity. Policyholders with ethnicity 1 (red points) experience consistently positive values of $\delta_{\text{UF}}(\boldsymbol{x}, d)$, particularly in the 20-50 age range where values reach up to 0.12, indicating these policyholders are charged higher prices compared to the demographically fair benchmark. In contrast, policyholders with ethnicities 3-5 (green, blue and purple points) show predominantly negative or near-zero values of $\delta_{\text{UF}}(\boldsymbol{x}, d)$, suggesting they benefit from demographic unfairness in pricing. The magnitude of unfairness appears to decrease with age across all ethnic groups, with the spread of $\delta_{\text{UF}}(\boldsymbol{x}, d)$ values notably smaller above age 60.

The right panel shows the relationship between $\delta_{\text{UF}}(\boldsymbol{x}, d)$ and covariate 17, which takes discrete levels 1-5, as analyzed in Table 3. Here we observe that the magnitude of unfairness varies substantially across different levels of this covariate. The effect is particularly pronounced for ethnicity 1, where level 4 shows the highest concentration of positive $\delta_{\text{UF}}(\boldsymbol{x}, d)$ values. This suggests that the interaction between ethnicity and covariate 17 is an important driver of demographic unfairness in the portfolio. These patterns suggest that demographic unfairness is not uniform across the real world portfolio but rather concentrated in particular combinations of age, ethnicity and other rating factors.

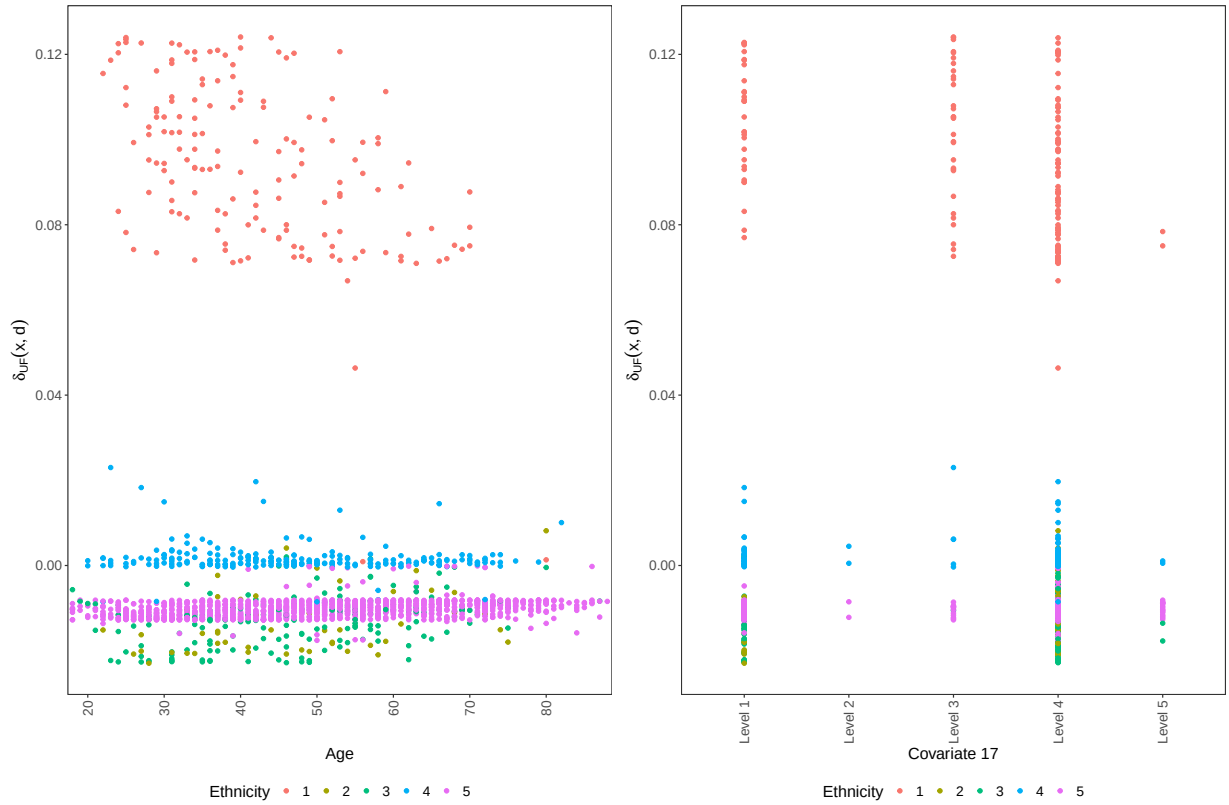

Figure 7: Local unfairness metric $\delta_{\text{UF}}(\boldsymbol{x}, d)$ plotted against the values of age (left panel) and covariate 17 (right panel).

A different picture emerges when we consider the local measure of proxy discrimination, $\delta_{\text{PD}}(x)$, shown in Figure 8. The left panel shows $\delta_{\text{PD}}(x)$ plotted against age. There is a notable concentration of positive values for ethnicity 1 (red points) in the younger age ranges, particularly between ages 20-40, suggesting these segments are most affected by proxy discrimination, which is a finding in line with Lindholm et al. (2023). The effect appears to diminish with age, with values converging closer to zero beyond age 60. Unlike the demographic unfairness measure, we observe both positive and negative values of $\delta_{\text{PD}}(x)$ across all ethnic groups, indicating a more complex pattern of discriminatory effects.

The right panel displays $\delta_{\text{PD}}(x)$ against the levels of covariate 17, revealing distinct clustering patterns. Level 3 shows the highest concentration of proxy discrimination effects, with values reaching up to 0.06 particularly for ethnicity 1. This suggests that this level of covariate 17 may be particularly effective at proxying protected characteristics, as discussed above. These visualizations demonstrate how the local measure $\delta_{\text{PD}}(x)$ can identify specific segments where proxy discrimination manifests most strongly, while also highlighting areas where pricing appears to avoid such effects. The patterns suggest that proxy discrimination, while present, operates in a more nuanced way than demographic unfairness. This is because the latter quantifies systematic differences in outcomes

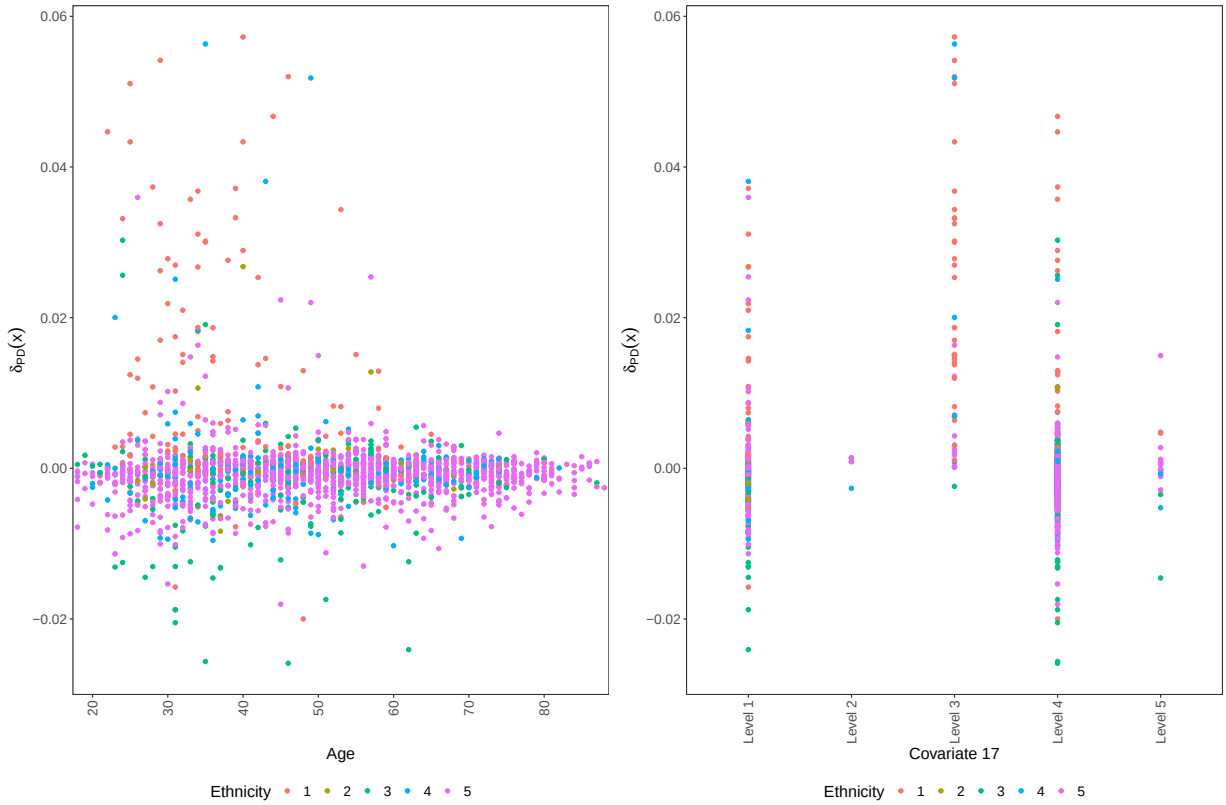

Figure 8: Local proxy discrimination metric $\delta_{\mathrm{PD}}(x)$ plotted against the values of age (left panel) and covariate 17 (right panel).

across groups without considering the level of risk within each subset of the portfolio, while the former involves the implicit inference of protected characteristics from other variables within the risk rating process.

5 Conclusions

We propose measures of demographic unfairness and proxy discrimination that can be widely applied to decision-making contexts were algorithmic predictions are used and these forms of unfairness are of concern. The measure of demographic unfairness is already present in the literature (Bénesse et al., 2024), while the measure of proxy discrimination is new. For that measure, we also propose methods for attributing any proxying effects to the different covariates. These measures are global, in the sense that they quantify unfairness or discrimination across a portfolio. In addition to studying the properties of these measures, we develop related local measures, which allow a quantification of demographic unfairness and proxy discrimination at the granular policy level.

The proposed methods were successfully applied to a real-world insurance claims data set, where ethnicity is the protected characteristic. Importance analysis of the proxy discrimination metric revealed the key covariates generating discriminatory effects, through implicit inference of ethnicity.

We note that the importance of those variables was not a priori known to us – in other words, the analysis led to a deeper understanding of the portfolio. Furthermore, given the heterogeneity of the portfolio, application of the PD measure to sub-populations of policyholders provides a more granular understanding of the groups mostly affected by proxy discrimination.

Author contributions

M. Lindholm, R. Richman, A. Tsanakas, and M.V. Wüthrich have all contributed to: "Conceptualization", "Methodology", "Software", and "Writing – original draft".

Data statement

The real insurance data used in the numerical illustrations is not allowed to be shared due the confidentiality reasons.

Funding

Parts of this research was carried out while M.V. Wüthrich was a KAW guest professor at Stockholm University.

References

Araiza Iturria, C. A., Hardy, M., & Marriott, P. (2024). A discrimination-free premium under a causal framework. *North American Actuarial Journal*, 24(4), 801–821.

Barocas, S. & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, (pp. 671–732).

Bénesse, C., Gamboa, F., Loubes, J.-M., & Boissin, T. (2024). Fairness seen as global sensitivity analysis. *Machine Learning*, 113(5), 3205–3232.

Bertsimas, D., Farias, V. F., & Trichakis, N. (2013). Fairness, efficiency, and flexibility in organ allocation for kidney transplantation. *Operations Research*, 61(1), 73–87.

Borgonovo, E., Clemente, G. P., & Rabitti, G. (2024). Why insurance regulators need to require sensitivity settings of internal models for their approval. *Finance Research Letters*, 60, 104859.

Borgonovo, E. & Plischke, E. (2016). Sensitivity analysis: A review of recent advances. *European Journal of Operational Research*, 248(3), 869–887.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

Charpentier, A. (2024). *Insurance, biases, discrimination and fairness*. Springer.

Chiappa, S., Jiang, R., Stepleton, T., Pacchiano, A., Jiang, H., & Aslanides, J. (2020). A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34 (pp. 3633–3640).

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., & Pontil, M. (2020). Fair regression with Wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33, 7321–7331.

Cook, T., Greenall, A., & Sheehy, E. (2022). *Discriminatory pricing: Exploring the 'ethnicity penalty' in the insurance market*. Technical report, Citizens Advice.

Côté, O., Côté, M.-P., & Charpentier, A. (2024). A fair price to pay: exploiting causal graphs for fairness in insurance. *Available at SSRN 4709243*.

De Bock, K. W., Coussement, K., De Caigny, A., Słowiński, R., Baesens, B., Boute, R. N., Choi, T.-M., Delen, D., Kraus, M., Lessmann, S., et al. (2024). Explainable AI for operational research: A defining framework, methods, applications, and a research agenda. *European Journal of Operational Research*, 317(2), 249–272.

Delbaen, F. & Majumdar, C. (2024). Approximation with independent variables. In *Peter Carr Gedenkschrift: Research Advances in Mathematical Finance* (pp. 311–327). World Scientific.

Fissler, T. & Pesenti, S. M. (2023). Sensitivity measures based on scoring functions. *European Journal of Operational Research*, 307(3), 1408–1423.

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

Glasserman, P. & Li, M. (2024). Should bank stress tests be fair? *Management Science*.

Gordaliza, P., Del Barrio, E., Fabrice, G., & Loubes, J.-M. (2019). Obtaining fairness using optimal transport theory. In *International conference on machine learning* (pp. 2357–2365). PMLR.

Hiabu, M., Meyer, J. T., & Wright, M. N. (2023). Unifying local and global model explanations by functional decomposition of low dimensional structures. In *International Conference on Artificial Intelligence and Statistics* (pp. 7040–7060). PMLR.

Hurlin, C., Pérignon, C., & Saurin, S. (2024). The fairness of credit scoring models. *Management Science*.

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

Komiyama, J. & Noda, S. (2024). On statistical discrimination as a failure of social learning: A multiarmed bandit approach. *Management Science*.

Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3), 1083–1094.

Kraus, M., Tschernutter, D., Weinzierl, S., & Zschech, P. (2024). Interpretable generalized additive neural networks. *European Journal of Operational Research*, 317(2), 303–316.

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

Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2023). A multi-task network approach for calculating discrimination-free insurance prices. *European Actuarial Journal*, 14, 329–369.

Lindholm, M., Richman, R., Tsanakas, A., & Wuthrich, M. V. (2024). What is fair? proxy discrimination vs. demographic disparities in insurance pricing. *Scandinavian Actuarial Journal*,

2024(9), 935–970.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6), 1–35.

Prince, A. E. & Schwarcz, D. (2019). Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev.*, 105, 1257.

Rabitti, G. & Borgonovo, E. (2020). Is mortality or interest rate the most important risk in annuity models? a comparison of sensitivity analysis methods. *Insurance: Mathematics and Economics*, 95, 48–58.

Radovanović, S., Savić, G., Delibašić, B., & Suknović, M. (2022). FairDEA – removing disparate impact from efficiency scores. *European Journal of Operational Research*, 301(3), 1088–1098.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer physics communications*, 181(2), 259–270.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). *Global sensitivity analysis: the primer*. John Wiley & Sons.

Sobol', I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3), 271–280.

Tschantz, M. C. (2022). What is proxy discrimination? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1993–2003).

Vallarino, A., Rabitti, G., & Chokami, A. K. (2024). Construction of rating systems using global sensitivity analysis: A numerical investigation. *ASTIN Bulletin: The Journal of the IAA*, 54(1), 25–45.

Supplementary material: Appendices to 'Sensitivity-based measures of discrimination in insurance pricing'

Mathias Lindholm* Ronald Richman† Andreas Tsanakas‡ Mario V. Wüthrich§

December 23, 2024

Numbered references in these Appendices point to equations, theorems, etc in the main paper.

A Additive decomposition of the PD and $\widetilde{\text{PD}}$ metrics

The question of additivity is important for interpreting the contributions of covariates to proxy discrimination. Specifically, for the importance measures developed in Section 2.5 it will generally be $\sum_{i=1}^q \text{PD}_i(\pi) \neq \text{PD}(\pi)$ and $\sum_{i=1}^q \widetilde{\text{PD}}_i(\pi) \neq \text{PD}(\pi)$. Nonetheless, an additive decomposition of $\text{PD}(\pi)$ is achievable by employing the game-theoretical concept of the *Shapley value* (Shapley et al., 1953) for the value functional $\mathcal{S} \mapsto \text{Var}(\mathbb{E}[\Lambda(\pi, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}])$. While recent literature on model interpretability has focused on the use of Shapley values to derive local model explanations for given instances $\boldsymbol{X} = \boldsymbol{x}$ (Lundberg & Lee, 2017; Aas et al., 2021), we use the Shapley value for a decomposition of a global sensitivity measure, following Owen (2014); Owen & Prieur (2017); Song et al. (2016). This leads to following definition.

Definition A. *For the proxy discrimination metric* PD *of* (7) *and* $\Lambda(\pi, \boldsymbol{X})$ *as in* (9)*, denote*

$$w(\mathcal{S}) = \text{Var}(\mathbb{E}[\Lambda(\pi, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}]), \ \mathcal{S} \subseteq \mathcal{Q}.$$

Then, we define the Shapley attribution of the covariate X_i *to proxy discrimination as the metric,*

$$\text{PD}_i^{\text{sh}}(\pi) = \frac{1}{\text{Var}(\pi(\boldsymbol{X}))\, q} \sum_{\mathcal{S} \subseteq \mathcal{Q} \setminus \{i\}} \binom{q-1}{|\mathcal{S}|}^{-1} \big(w(\mathcal{S} \cup \{i\}) - w(\mathcal{S})\big). \tag{i}$$

*Corresponding author. Department of Mathematics, Stockholm University, SE-106 91, Sweden. lindholm@math.su.se

†insureAI, Floor 2, 30 Melrose Boulevard, Melrose Arch, Gauteng, South Africa, 2196. ron@insureai.co.

‡Bayes Business School, City St George's, University of London, 106 Bunhill Row, London, EC1Y 8TZ, United Kingdom. A.Tsanakas.1@city.ac.uk

§RiskLab, Department of Mathematics, ETH Zurich, Switzerland. mario.wuethrich@math.ethz.ch

As the Shapley value is a well-known concept across literatures, we do not review its properties here. The key practical feature is that by the use of Shapley values we achieve an additive attribution.

$$\sum_{i=1}^{q} \mathrm{PD}_i^{\mathrm{sh}}(\pi) = \mathrm{PD}(\pi).$$

Furthermore, we note that while Definition A calculates the Shapley value with respect to $w(\mathcal{S}) = \mathrm{Var}(\mathbb{E}[\Lambda(\pi, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}])$ and is thus based on $\mathrm{PD}_{\mathcal{S}}$, a result identical to (i) is obtained if instead we additivise the alternative metric $\widetilde{\mathrm{PD}}_{\mathcal{S}}$ (Song et al., 2016). Hence the distinction between the two metrics of Definition A collapses when Shapley values are employed.

Furthermore, Shapley-based decompositions can also be applied to the local measures of Section 3. In Example 3, we had a single non-protected covariate $\boldsymbol{X} = X$, but, in general, $\delta_{\mathrm{UF}}(\boldsymbol{x}, d)$ will be a multivariate function. Then the question arises as to how individual covariates contribute to this metric. This can be done by using standard local model explainability methods, e.g., by, analogously to (i), calculating Shapley values with respect to the alternative value functional $\mathcal{S} \mapsto \mathbb{E}[\delta_{\mathrm{UF}}(\boldsymbol{X}, D) \mid \boldsymbol{X}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}]$, where $\boldsymbol{x}_{\mathcal{S}}$ is a sub-vector of $\boldsymbol{x}$ for the specific instance of interest (Lundberg & Lee, 2017; Aas et al., 2021). The same comments apply to the local metric for proxy discrimination introduced in the next section.

B Response of PD metric to model distortions

To gain more insights into the scale of the PD metric as applied to the unawareness price, we design distortions of the underlying model, which enable re-evaluations of the metric without needing to re-fit the model. We assume that we have available the best-estimate prices $\mu(\boldsymbol{X}, d)$, and the conditional distribution $\mathbb{P}(D = d \mid \boldsymbol{X})$ for all $d \in \mathfrak{D}$. The unawareness price can be calculated by averaging

$$\mu(\boldsymbol{X}) = \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{X}, d)\mathbb{P}(D = d \mid \boldsymbol{X}),$$

rather than a direct regression of claims on $\boldsymbol{X}$ (Lindholm et al., 2022).

There are two ways in which we will intervene in the model. First, we make the best-estimate prices more sensitive to a particular demographic group, e.g., $D = 1$. Denote by $\nu(\boldsymbol{X}, D)$ and $\nu(\boldsymbol{X})$ the best-estimate and unawareness prices of the distorted model. We set

$$\nu_\alpha(\boldsymbol{X}, D) := \mu(\boldsymbol{X}, D) + \alpha \mathbb{1}_{\{D=1\}}.$$

Hence α represents an additional fixed predicted claim cost corresponding to the demographic group $D = 1$.

Second, we increase the model's ability to proxy the demographic group by being able to infer more accurately whether or not $D = 1$, given $\boldsymbol{X}$. To do this consider an alternative probability $\mathbb{Q}_\beta$ under which the event $D = 1$ can be better inferred from $\boldsymbol{X}$. The model distortion operates by increasing the variability of the random variable $\mathbb{P}(D = 1 \mid \boldsymbol{X})$ (since constancy implies independence). For $\beta > 0$, let $G_\beta(\cdot)$ by the cumulative distribution of a $\mathrm{Beta}(1 + \beta, 1 + \beta)$ variable.

This has a sigmoid shape, resembling a step function as β gets large. Hence, we can apply the transformation:

$$\mathbb{Q}_\beta(D = 1 \mid \boldsymbol{X}) = G_\beta(\mathbb{P}(D = 1 \mid \boldsymbol{X}))$$
$$\mathbb{Q}_\beta(D = d \mid \boldsymbol{X}) = \mathbb{P}(D = d \mid \boldsymbol{X}) \left(1 + \frac{\mathbb{P}(D = 1 \mid \boldsymbol{X}) - \mathbb{Q}_\beta(D = 1 \mid \boldsymbol{X})}{\sum_{d' \neq 1} \mathbb{P}(D = d' \mid \boldsymbol{X})}\right), \quad d \neq 1$$

It can be checked that $\mathbb{Q}_\beta(D = d \mid \boldsymbol{X})$ is a proper probability mass function in d. This transformation makes the probability of $\mathbb{Q}_\beta(D = 1 \mid \boldsymbol{X})$ closer to 1 or 0 for specific ranges of $\boldsymbol{X}$ meaning that there is a higher potential for proxy discrimination. (Note though that the marginal probabilities are also affected i.e., generally $\mathbb{Q}_\beta(D = d) \neq \mathbb{P}(D = d)$.)

With this in place we can now work out the unawareness price of the distorted model:

$$\nu_{\alpha,\beta}(\boldsymbol{X}) = \sum_{d \in \mathfrak{D}} \nu_\alpha(\boldsymbol{X}, d) \mathbb{Q}_\beta(D = d \mid \boldsymbol{X}).$$

By comparing the PD metric under $\nu_{\alpha,\beta}(\boldsymbol{X})$ with that under $\mu(\boldsymbol{X})$, one can understand now the scale of the PD metric changes with the specified model distortions. By plotting the function $(\alpha, \beta) \mapsto \text{PD}(\nu_{\alpha,\beta}(\boldsymbol{X}))$, we can get a sense of how distortions in the model that increase proxying can impact the PD measure.

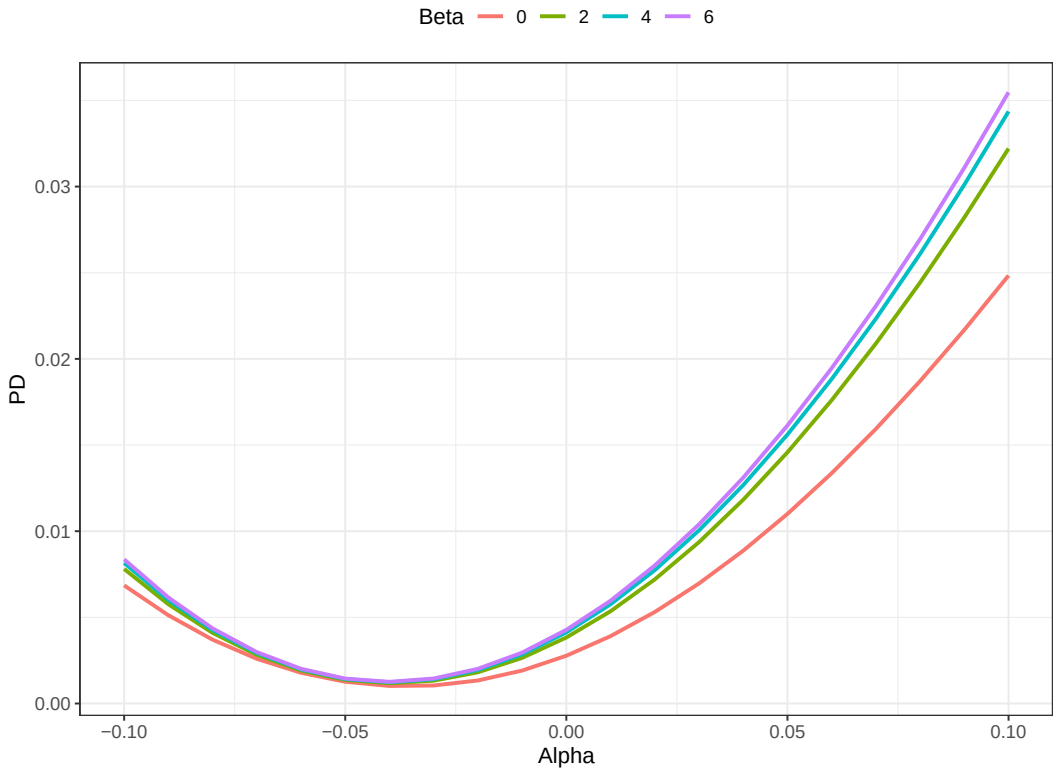

Figure 1: Sensitivity analysis of the PD metric, subject to distortions of the underlying model.

In Figure 1, we produce this analysis for our real-world dataset of 4 in the main paper, for α varying between -0.1 and 0.1, and β varying from 0 to 6. The role of α is already explained in the

paper. The impact of increasing β, which controls the ability to infer demographic characteristics, is positive but shows diminishing returns at higher values. The highest PD value of 0.0926 is achieved when both direct sensitivity to protected attributes and the ability to infer them are highest. This finding validates the PD metric's capacity to detect not only explicit demographic sensitivity in pricing but also the more subtle effects of proxy discrimination arising from enhanced demographic inference. Notably, the metric exhibits higher sensitivity to changes in α than to changes in β, suggesting that direct demographic sensitivity in best-estimate prices has a more substantial impact on proxy discrimination than improvements in demographic inference capability.

References

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298, 103502.

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

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Owen, A. B. (2014). Sobol'indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1), 245–251.

Owen, A. B. & Prieur, C. (2017). On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1), 986–1002.

Shapley, L. S. et al. (1953). A value for n-person games.

Song, E., Nelson, B. L., & Staum, J. (2016). Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1), 1060–1083.