What is Fair? Proxy Discrimination vs. Demographic Disparities in Insurance Pricing

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

May 24, 2023¶

Abstract

Indirect discrimination and fairness are major concerns in algorithmic models. This is particularly true in insurance, where protected policyholder attributes are not allowed to be used for insurance pricing. Simply disregarding protected policyholder attributes is not an appropriate solution, as this still allows for the possibility of inferring protected attributes from non-protected covariates. Such inference leads to so-called proxy or indirect discrimination. Though proxy discrimination is qualitatively different from the group fairness concepts in the machine learning literature, group fairness criteria have been proposed to control the impact of protected attributes on the calculation of insurance prices. The purpose of this paper is to discuss the differences between, on the one hand, direct and indirect discrimination in insurance and, on the other, the most popular group fairness axioms. In particular, we show that one does not imply the other, as these concepts are materially different. Furthermore, we discuss input data pre-processing and model post-processing methods that achieve both discrimination-free insurance prices and group fairness by demographic parity. The main tool in these methods is the theory of optimal transport.

Keywords. Discrimination, indirect discrimination, proxy discrimination, fairness, protected attributes, discrimination-free, unawareness, group fairness, demographic parity, statistical parity, independence axiom, equalized odds, separation axiom, predictive parity, sufficiency axiom, input pre-processing, output post-processing, optimal transport, Wasserstein distance.

1 Introduction

For legal and societal reasons, there are several policyholder attributes that are not allowed to be used in insurance pricing [3, 9, 14, 15, 27], e.g., European law does not allow for the use of gender information in insurance pricing, or ethnicity is a critical attribute that may be declared as a protected characteristics; we also refer to a recent report of the European Insurance and

*Department of Mathematics, Stockholm University
†Old Mutual Insure and University of the Witwatersrand
‡Bayes Business School (formerly Cass), City, University of London
§RiskLab, Department of Mathematics, ETH Zurich

¶An earlier version of this manuscript by the same authors, with the title "A discussion of discrimination and fairness in insurance pricing", is available on SSRN, manuscript ID 4207310.

Occupational Pension Authority (EIOPA) [13], which discusses governance principles towards an ethical and trustworthy use of artificial intelligence in the insurance sector.
Frees–Huang [17] and Xin–Huang [37] give extensive overviews on protected information in insurance and implications for pricing, while Avraham et al. [3], Prince–Schwarcz [27] and Maliszewska-Nienartowicz [23] provide legal viewpoints on this topic. The critical issue is that just ignoring (being unaware of) protected information does not solve the problem, as protected information can be inferred from non-protected characteristics if the corresponding variables are associated. This is especially true in high-dimensional algorithmic models. Such inference implies proxy or indirect discrimination, and is often implicitly performed during the fitting procedure of complex models.
There are several attempts to prevent this inference: from the perspective of causal statistics a counterfactual approach is suggested, see Kusner et al. [20], Charpentier [7], and Araiza Iturria et al. [2]; a probabilistic approach called discrimination-free insurance pricing is put forward by Lindholm et al. [21]; group fairness concepts are built into the training of machine learning models, see, e.g., Grari et al. [18].
The purpose of this paper is to discuss the three most popular group fairness axioms in the light of insurance pricing, through concrete examples that demonstrate potential trade-offs and incompatibilities. Specifically, we present a statistical model producing insurance prices that are free of discrimination, but do not satisfy any of the most popular group fairness axioms of the machine learning literature. Conversely, we provide an example where insurance prices satisfy a group fairness axiom, but directly discriminate. This shows that discrimination considerations, in the context of insurance pricing, and fairness criteria motivated by the machine learning literature are materially different concepts, and the latter do not provide a quick fix for the former. We do not aim to draw sharp distinctions between the two fields and acknowledge that fairness criteria are potentially relevant for insurance pricing, while the idea of proxy discrimination can be important in machine learning applications. Rather, our goal is to examine how we can transfer fairness concepts to insurance, in light of the particular characteristics of insurance markets and current regulatory frameworks.
The theory of optimal transport has recently been promoted as input pre-processing and model post-processing methods to make statistical models fair, see Barrio et al. [5] and Chiappa et al. [8], and an early application of these ideas in an insurance context w.r.t. creating gender neutral policies in life insurance using mean-field approximations can be found in Example 5.1 of Djehiche–Löfdahl [12]. In the second part of this paper we study these pre- and post-processing methods, and conclude that input pre-processing may be a very helpful tool in achieving fairness objectives in insurance pricing. The extent to which the resulting prices can be considered discrimination-free is however a matter of interpretation. Model post-processing, which is more frequently used in machine learning, is simpler to apply and allows for optimal modeling choices from the perspective of predictive accuracy. However, model post-processing can lead to results that are not explainable to insurance customers and policymakers. Therefore, there are substantial challenges to its practical adoption in insurance pricing.

Organization. In Section 2, we discuss discrimination-free insurance pricing and group fairness notions, and show that these do not imply each other. In Section 3, we present optimal transport and its application to data pre-processing and model post-processing. Moreover, we show the usefulness of these methods to achieve fairness and to obtain discrimination-free insurance prices.

In Section 4, we conclude and discuss further aspects. The mathematical results are proved in Appendix A.

2 Discrimination and fairness in insurance pricing

2.1 Discrimination-free insurance pricing

To set the ground, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{P}$ describing the real world probability measure. On this probability space we consider the random triplet $(Y, \boldsymbol{X}, \boldsymbol{D})$. The response variable Y describes the insurance claim that we try to predict (and price). The vector $\boldsymbol{X}$ describes the *non-protected covariates* (non-discriminatory characteristics), and $\boldsymbol{D}$ describes the *protected attributes* (discriminatory characteristics). We assume that the partition into non-protected covariates and protected attributes is given exogenously, e.g., by law or by societal norms and preferences. We use the distribution $\mathbb{P}(\boldsymbol{X}, \boldsymbol{D})$ to describe the insurance portfolio, i.e., the random selection of a policyholder from the insurance portfolio. Different insurance companies may have different insurance portfolio distributions $\mathbb{P}(\boldsymbol{X}, \boldsymbol{D})$, and this insurance portfolio distribution typically differs from the overall population distribution in a given society because the insurance penetration is not uniform across the entire population. For simplicity, in this paper, we assume that the protected attributes $\boldsymbol{D}$ are discrete and finite, only taking values in the finite set $\mathfrak{D}$.

Best-estimate price. For insurance pricing, one aims at designing a regression model that describes the conditional distribution of Y, given the explanatory variables $(\boldsymbol{X}, \boldsymbol{D})$.

Definition 2.1 *The best-estimate price of Y, given full information $(\boldsymbol{X}, \boldsymbol{D})$, is given by*

$$\mu(\boldsymbol{X}, \boldsymbol{D}) := \mathbb{E}\left[Y \middle| \boldsymbol{X}, \boldsymbol{D}\right]. \tag{2.1}$$

This price is called 'best-estimate' because it has minimal prediction variance, i.e., it is the most accurate predictor for Y, given $(\boldsymbol{X}, \boldsymbol{D})$, in the $L^2(\mathbb{P})$-sense; for simplicity, we assume that all considered random variables are square-integrable w.r.t. $\mathbb{P}$.

In general, the best-estimate price *directly discriminates* because it uses the protected attributes $\boldsymbol{D}$ as an input, see (2.1).

Fairness through unawareness. The most simple fairness concept in machine learning is the fairness through unawareness concept that drops the protected attributes $\boldsymbol{D}$ from the pricing functional.

Definition 2.2 *The unawareness price of Y, given $\boldsymbol{X}$, is defined by*

$$\mu(\boldsymbol{X}) := \mathbb{E}\left[Y \middle| \boldsymbol{X}\right]. \tag{2.2}$$

The unawareness price does not directly discriminate because it does not use protected attributes $\boldsymbol{D}$ as an input, i.e., it is blind w.r.t. the protected attributes $\boldsymbol{D}$. However, it may *indirectly discriminate* because the knowledge of $\boldsymbol{X}$ allows inference of $\boldsymbol{D}$ through the tower property

$$\mu(\boldsymbol{X}) = \sum_{\boldsymbol{d} \in \mathfrak{D}} \mu(\boldsymbol{X}, \boldsymbol{d})\, \mathbb{P}(\boldsymbol{D} = \boldsymbol{d} | \boldsymbol{X}). \tag{2.3}$$

This formula shows that if there is statistical dependence (association) between $\boldsymbol{X}$ and $\boldsymbol{D}$ w.r.t. $\mathbb{P}$, we may implicitly use this dependence for inference of $\boldsymbol{D}$ from $\boldsymbol{X}$; in Example 2.6, below, we illustrate this inference on an explicit example which is based solely on statistical dependence between $\boldsymbol{D}$ and $\boldsymbol{X}$, and not on any causal relationship.

Remark 2.3 Formula (2.3) highlights that there are *two necessary conditions* to obtain indirect discrimination in $\mu(\boldsymbol{X})$, for a given $\boldsymbol{X}$. First, we need to have a conditional probability

$$\mathbb{P}(\boldsymbol{D}=\boldsymbol{d}|\boldsymbol{X}) \neq \mathbb{P}(\boldsymbol{D}=\boldsymbol{d}) \qquad \text{for some } \boldsymbol{d} \in \mathfrak{D}, \tag{2.4}$$

i.e., we need to have dependence between $\boldsymbol{X}$ and $\boldsymbol{D}$ that allows us to (partly) infer the protected attributes $\boldsymbol{D}$ from the non-protected covariates $\boldsymbol{X}$. Property (2.4) is the reason for referring indirect discrimination also to *proxy discrimination*, because $\boldsymbol{X}$ is used to proxy $\boldsymbol{D}$. Second, the functional $\boldsymbol{d} \mapsto \mu(\boldsymbol{X}, \boldsymbol{d})$ needs to have a sensitivity in $\boldsymbol{d}$, otherwise, if

$$\mu(\boldsymbol{X}, \boldsymbol{d}) \equiv \mu(\boldsymbol{X}) \qquad \text{for all } \boldsymbol{d} \in \mathfrak{D}, \tag{2.5}$$

the inference potential from $\boldsymbol{X}$ to $\boldsymbol{D}$ is not useful in (2.3), and we do not have indirect discrimination. In fact, under property (2.5) we may choose any portfolio distribution $\mathbb{P}(\boldsymbol{X}, \boldsymbol{D})$ and we receive equal unawareness and best-estimate prices. In that case, there cannot be any (indirect) discrimination because $\boldsymbol{X}$ is *sufficient* to compute the best-estimate price (2.1). As an example, we imagine that (non-protected) telematics data $\boldsymbol{X}$ makes gender information $\boldsymbol{D}$ superfluous to predict claims Y. This would imply a (causal) graph $\boldsymbol{D} \to \boldsymbol{X} \to Y$, which means that $\boldsymbol{D}$ does not carry any additional information to predict claims Y, given $\boldsymbol{X}$. Therefore, (2.5) holds in this telematics data example.

We conclude, in general, fairness through unawareness indirectly discriminates, and it does not solve the problem of non-discriminatory insurance pricing.

Discrimination-free insurance price. Lindholm et al. [21] proposed to break the inference potential in (2.3) and (2.4), respectively, to arrive at a discrimination-free insurance price.

Definition 2.4 *A discrimination-free insurance price of Y, given $\boldsymbol{X}$, is defined by*

$$\mu^*(\boldsymbol{X}) := \sum_{\boldsymbol{d} \in \mathfrak{D}} \mu(\boldsymbol{X}, \boldsymbol{d})\, \mathbb{P}^*(\boldsymbol{D}=\boldsymbol{d}), \tag{2.6}$$

where the pricing measure $\mathbb{P}^*(\boldsymbol{D})$ is dominated by the marginal distribution of the protected attributes $\boldsymbol{D}$.[1]

Replacing the conditional distribution $\mathbb{P}(\boldsymbol{D}=\boldsymbol{d}|\boldsymbol{X})$ in (2.3) by a (marginal) pricing distribution $\mathbb{P}^*(\boldsymbol{D}=\boldsymbol{d})$ breaks the link for proxy discrimination and, therefore, we call the resulting price $\mu^*(\boldsymbol{X})$ 'discrimination-free'.

Remarks 2.5

[1] To make the discrimination-free insurance price (2.6) well-defined we need to assume that $\mu(\boldsymbol{X}, \boldsymbol{D})$ exists for all $(\boldsymbol{X}, \boldsymbol{D})$, a.s.

- Under (2.5), i.e., if $\boldsymbol{d} \mapsto \mu(\boldsymbol{X}, \boldsymbol{d})$ does not have any sensitivity in $\boldsymbol{d}$, the best-estimate price $\mu(\boldsymbol{X}, \boldsymbol{D})$, the unawareness price $\mu(\boldsymbol{X})$ and the discrimination-free insurance price $\mu^*(\boldsymbol{X})$ all coincide, and such a model is generally free of proxy discrimination under best-estimate pricing, because $\boldsymbol{X}$ is sufficient to compute the best-estimate price and the specific dependence structure between $\boldsymbol{X}$ and $\boldsymbol{D}$ becomes irrelevant.
- Under additional assumptions on causal graphs, the discrimination-free insurance price (2.6) coincides with the causal impact of $\boldsymbol{X}$ on Y, see Lindholm et al. [21] and Araiza Iturria et al. [2]. However, causal considerations are often too restrictive in insurance pricing as, generally, they require that there are no unmeasured confounders or that these unmeasured confounders satisfy additional restrictive causal assumptions, otherwise one cannot adjust for the protected attributes $\boldsymbol{D}$; see Pearl [25]. In an insurance pricing context there are always policyholder attributes that cannot be observed and act as unmeasured confounders for which it is difficult/impossible to verify the necessary causal assumptions; e.g., in car driving the current health and mental states may matter to explain propensity to claims. Henceforth, though tempting, causal arguments are not a feasible way in practice of solving non-discriminatory insurance pricing.

To illustrate the ideas of this section, and set the scene for concepts discussed in later sections, we introduce two examples. First, we consider a situation where we have a response variable Y whose conditional expectation is fully described by the non-protected covariates $\boldsymbol{X}$, and the protected attributes $\boldsymbol{D}$ do not carry any additional information about the mean of the response Y. Therefore, in this example, the best-estimate price provides the discrimination-free insurance price, see first item of Remarks 2.5. Moreover, this model is simple enough to be able to calculate all quantities of interest, and, even if it is unrealistic in practice, it allows us to gain intuition about the relationship between indirect discrimination in insurance pricing and the group fairness concepts found in the machine learning literature, which will be introduced in the sequel.

Example 2.6 (No discrimination despite dependence of $(\boldsymbol{X}, \boldsymbol{D})$.)

Assume we have two-dimensional covariates $(\boldsymbol{X}, \boldsymbol{D}) = (X, D)$ having a mixture Gaussian portfolio density

$$(X, D) \;\sim\; f(x, d) = \frac{1}{2} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{ -\frac{1}{2\tau^2} \left(x - x_d\right)^2 \right\}, \tag{2.7}$$

with $d \in \mathfrak{D} = \{0, 1\}$, $x \in \mathbb{R}$, $\tau^2 > 0$, $x_0 > 0$, $\rho > 0$, and where we set

$$x_d = x_0 + \rho d.$$

Thus, D is a Bernoulli random variable taking the values 0 and 1 with probability $1/2$, and X is conditional Gaussian, given $D = d$, with mean x_d and variance $\tau^2 > 0$. Below, we make explicit choices for x_0 and x_1 which are kept throughout all examples.
For the response Y we assume conditionally, given $(\boldsymbol{X}, \boldsymbol{D})$,

$$Y|_{(\boldsymbol{X}, \boldsymbol{D})} \;\sim\; \mathcal{N}(X, 1 + D). \tag{2.8}$$

That is, the mean of the response does *not* depend on the protected attributes $\boldsymbol{D}$, but only on the non-protected covariates $\boldsymbol{X}$. This means that $\boldsymbol{X}$ is sufficient to describe the mean of Y.

The best-estimate and the unawareness prices coincide in this example, see (2.5), and they are given by

$$\mu(\boldsymbol{X}, \boldsymbol{D}) = \mu(\boldsymbol{X}) = X. \tag{2.9}$$

Therefore, in this example, we do not have (indirect) discrimination and the best-estimate price is discrimination-free, see first item of Remarks 2.5.

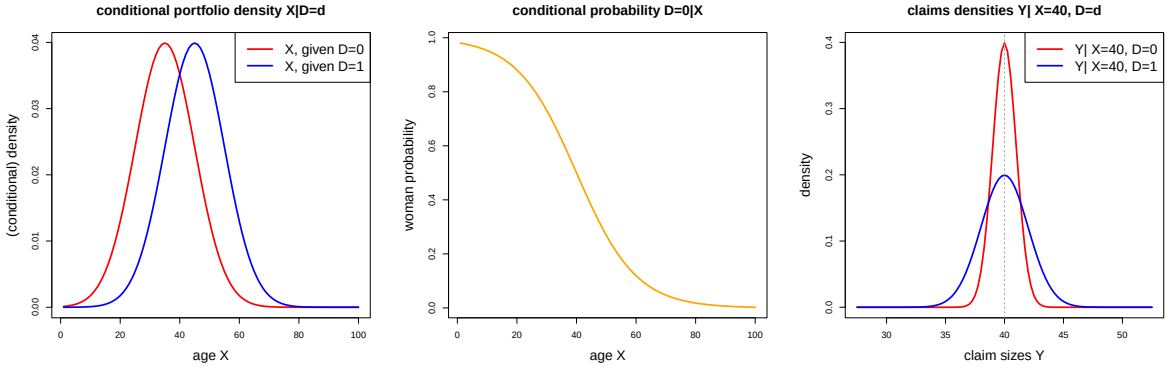

Figure 1: (lhs) Conditional Gaussian densities $f(x|d)$ for $d \in \mathfrak{D} = \{0, 1\}$; (middle) conditional probability $\mathbb{P}(D = 0|X = x)$ as a function of $x \in \mathbb{R}$; (rhs) densities of response Y for age $X = 40$ and genders $D = 0, 1$.

In Figure 1 (lhs) we give an explicit example for model (2.7). This plot shows the conditional Gaussian densities of X, given $D = d \in \{0, 1\}$; we select $x_0 = 35$, $\rho = 10$ (providing $x_1 = 45$), and $\tau = 10$. Here, the non-protected covariate X can be interpreted as the age of the policyholder, and D as the gender of the policyholder with $D = 0$ for women and $D = 1$ for men. We can easily calculate the conditional probability of $D = 0$ (being woman), given X,

$$\mathbb{P}\left(D = 0 \middle| X\right) = \frac{\exp\left\{-\frac{1}{2\tau^2}\left(X - x_0\right)^2\right\}}{\sum_{d \in \mathfrak{D}} \exp\left\{-\frac{1}{2\tau^2}\left(X - x_d\right)^2\right\}} \in (0, 1). \tag{2.10}$$

Figure 1 (middle) shows these conditional probabilities as a function of the age variable $X = x$. For small X we have likely a woman, $D = 0$, and for large X a man, $D = 1$. Figure 1 (rhs) shows the Gaussian densities of the claims Y at the given age $X = 40$ and for both genders $D = 0, 1$. The vertical dotted line shows the resulting means (2.9). These means coincide for both genders $D = 0, 1$, and the protected attribute D only influences the width of the Gaussian densities, see (2.8). ■

We give some general remarks on Example 2.6.

Remarks 2.7

- In Example 2.6, there is no causality involved between $\boldsymbol{X}$ and $\boldsymbol{D}$, but we interpret the dependence between $\boldsymbol{X}$ and $\boldsymbol{D}$ as a purely associational one stemming from the particular choice of the insurance portfolio distribution $\mathbb{P}(\boldsymbol{X}, \boldsymbol{D})$. While one may alternatively interpret $(\boldsymbol{X}, \boldsymbol{D})$ as being *risk factors* of Y, in the causal sense of Definition 3.1 of Araiza Iturria et al. [2], this is not a necessary assumption in our setting.

- A crucial feature of Example 2.6 is that the non-protected covariates $\boldsymbol{X}$ are sufficient to describe the mean of the response Y, and the protected attributes $\boldsymbol{D}$ only impact higher moments of Y. Therefore, there is no indirect discrimination in this example because (2.9) holds. From a practical point of view we may question such a model, but it has the advantage for the subsequent discussions that we do not need to rely on any type of proxy discrimination debiasing for stating the crucial points about group fairness and discrimination. We could modify (2.8) to include $\boldsymbol{D}$ also in the first moment of Y and derive similar conclusions, but then we would first need to convince the reader that the discrimination-free insurance price $\mu^*(\boldsymbol{X})$ is indeed the right way to correct for proxy discrimination.

- A situation where protected attributes $\boldsymbol{D}$ only impact higher moments may arise in the case of a lack of historical data of a demographic group. This may lead to higher uncertainty, reflected in higher moments, but not the means. From an insurance pricing point of view, this manifests in higher risk loadings, which may then be subject to discrimination; however risk loadings are not discussed further in this paper. The situation where predictions for different demographic groups are subject to higher uncertainty finds parallels in the machine learning literature, where there is concern about poor performance of predictive models for populations that are under-represented in training samples, e.g., in the context of facial recognition see Buolamwini–Gebru [6]. The crucial point is whether such increased uncertainty has adverse impacts on these demographic groups, such as a higher likelihood of misidentification leading to systematic penalties, see, e.g., Vallance [33].

We now present a variation of the previous example, where the dependence of $(\boldsymbol{X}, \boldsymbol{D})$ leads to proxy discrimination, which requires correction in the sense of equation (2.6).

Example 2.8 (Proxy discrimination and correction by $\mu^*(\boldsymbol{X})$)
We again assume two-dimensional covariates $(\boldsymbol{X}, \boldsymbol{D}) = (X, D)$ having the same mixture Gaussian distribution as in (2.7). For the response variable Y we now assume that conditionally, given $(\boldsymbol{X}, \boldsymbol{D})$,

$$Y|_{(\boldsymbol{X},\boldsymbol{D})} \;\sim\; \mathcal{N}\left(X + 20(1-D)\mathbb{1}_{X\in[20,40]} - 10D, 100\right). \tag{2.11}$$

The interpretation of this model is that female policyholders ($D = 0$) between ages 20 and 40 generate higher costs due to a potential pregnancy,[2] and male policyholders generally have lower costs.
The resulting best-estimate prices, illustrated in Figure 2 by the red and blue dotted lines, are given by

$$\mu(\boldsymbol{X}, \boldsymbol{D}) = \mathbb{E}\left[Y \middle| \boldsymbol{X}, \boldsymbol{D}\right] = X + 20(1-D)\mathbb{1}_{X\in[20,40]} - 10D.$$

The crucial difference of these best-estimate prices to the ones in Example 2.6 is that we do not have monotonicity in $x \mapsto \mu(X = x, D = 0)$ for women. This will make the current example more interesting.
Next we calculate the unawareness price

$$\mu(\boldsymbol{X}) = X + \frac{20\exp\left\{-\frac{1}{2\tau^2}\left(X - x_0\right)^2\right\}}{\sum_{d\in\mathfrak{D}}\exp\left\{-\frac{1}{2\tau^2}\left(X - x_d\right)^2\right\}}\mathbb{1}_{X\in[20,40]} - \frac{10\exp\left\{-\frac{1}{2\tau^2}\left(X - x_1\right)^2\right\}}{\sum_{d\in\mathfrak{D}}\exp\left\{-\frac{1}{2\tau^2}\left(X - x_d\right)^2\right\}},$$

[2]For simplicity of this exposition, we conflate biological sex and gender such that by "woman"/"female" we identify policyholders who can potentially be pregnant.

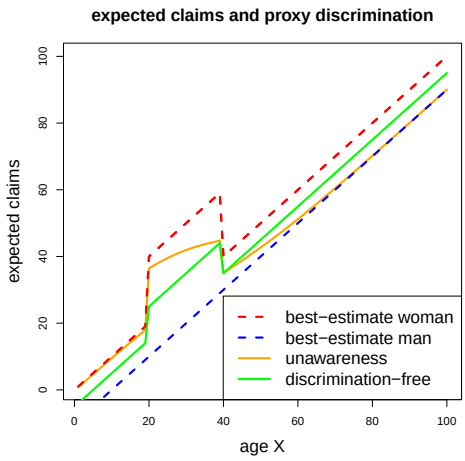

Figure 2: Best-estimate, unawareness and discrimination-free insurance prices in Example 2.8.

where we have used (2.10). This unawareness price is illustrated in orange color in Figure 2. Not surprisingly, it closely follows the best-estimate prices for woman policyholders for small ages and men for large ages, because we can infer the gender D from the age X quite well, see Figure 1 (middle). Thus, except in the age range from 20 to 60, we almost charge the best-estimate price to the corresponding genders, except to a few 'mis-allocated' men at small ages and women at high ages. This is precisely indirect/proxy discrimination as, e.g., described in paragraph 5 of Section 2 of Maliszewska-Nienartowicz [23], and it can be interpreted as generating a disproportionate impact on (woman) policyholders.
Finally, the discrimination-free insurance price for the choice $p^* = \mathbb{P}^*(D = 0) = 1/2$ is shown in green color in Figure 2 and reads as

$$\mu^*(\boldsymbol{X}) = X + 10 \cdot \mathbb{1}_{X \in [20,40]} - 5.$$

This discrimination-free insurance price exactly interpolates between the two best-estimate prices for women and men. As a result we have a cost reallocation between different ages which leads to a loss of predictive power and to cross-financing of claim costs within the portfolio. Here and in the sequel we measure predictive performance of a predictor $\widehat{\mu}$ for Y by the mean squared error (MSE)

$$\mathcal{L}(\widehat{\mu}, Y) := \mathbb{E}\left[(Y - \widehat{\mu})^2\right],$$

and study a potential bias by providing the average prediction $\mathbb{E}[\widehat{\mu}]$ of the predictor $\widehat{\mu}$ for Y, averaged over the portfolio distribution $\mathbb{P}(\boldsymbol{X}, \boldsymbol{D})$.
We calculate the resulting mean squared errors using Monte Carlo simulation with a pseudo-random sample of size 1 million. The results in Table 1 show the negative impact of deviating from the optimal predictors, based on $(\boldsymbol{X}, \boldsymbol{D})$ and $\boldsymbol{X}$, respectively. This is the price we pay for being discrimination-free w.r.t. the protected attributes $\boldsymbol{D}$. Our pricing measure choice $p^* = \mathbb{P}^*(D = 0) = \mathbb{P}(D = 0) = 1/2$ provides a bias which can still be removed by choosing a different pricing measure $\mathbb{P}^*(D)$. By setting $p^* = \mathbb{P}^*(D = 0) = 0.58$, the discrimination-free insurance price is unbiased on a portfolio level, and we receive a smaller mean squared error, see last line of Table 1. ■

	MSE $\mathcal{L}(\widehat{\mu}, Y)$	bias $\mathbb{E}[\widehat{\mu}]$
best-estimate price $\mu(\boldsymbol{X}, \boldsymbol{D})$	100.00	41.25
unawareness price $\mu(\boldsymbol{X})$	197.20	41.25
discrimination-free insurance price $\mu^*(\boldsymbol{X})$ with $p^* = 0.50$	217.66	39.63
discrimination-free insurance price $\mu^*(\boldsymbol{X})$ with $p^* = 0.58$	210.15	41.25

Table 1: Mean squared errors and average prediction of the different prices in Example 2.8.

2.2 Group fairness axioms

In this section, we introduce the three most popular group fairness axioms from machine learning. These are *demographic parity*, *equalized odds* and *predictive parity*; we refer to Barocas et al. [4], Xin–Huang [37] and Grari et al. [18]. In the next section, we show that the discrimination-free insurance price of Example 2.6, given in (2.9), violates all three of these group fairness axioms. Of course, this casts doubt on whether these group fairness axioms are suitable concepts for dealing with (indirect) discrimination in insurance pricing as, e.g., stated in European regulation [14, 15].

We denote by $\widehat{\mu}$ any predictor of Y, which can be the unawareness price (2.2) or any other pricing functional. In insurance pricing, these predictors $\widehat{\mu} = \widehat{\mu}(\boldsymbol{X}, \boldsymbol{D})$ will typically depend on $\boldsymbol{X}$ and/or on $\boldsymbol{D}$, but this is not crucial in the following group fairness definitions.

(i) Independence axiom / demographic parity / statistical parity. Following Definition 1 of Agarwal et al. [1], we have *demographic parity/statistical parity* if

$$\widehat{\mu} \text{ and } \boldsymbol{D} \text{ are independent under } \mathbb{P}.$$

This independence implies for the distribution of the insurance prices $\widehat{\mu}$, a.s.,

$$\mathbb{P}\left(\widehat{\mu} \leq m \middle| \boldsymbol{D}\right) \;=\; \mathbb{P}\left(\widehat{\mu} \leq m\right) \qquad \text{for all } m \in \mathbb{R}. \tag{2.12}$$

(ii) Separation axiom / equalized odds. Equalized odds is sometimes also called disparate mistreatment. It has been introduced by Hardt et al. [19], and it is defined as follows: We have *equalized odds* if

$$\widehat{\mu} \text{ and } \boldsymbol{D} \text{ are conditionally independent under } \mathbb{P}, \text{ given the response } Y.$$

This conditional independence implies for the distribution of the prices $\widehat{\mu}$, a.s.,

$$\mathbb{P}\left(\widehat{\mu} \leq m \middle| Y, \boldsymbol{D}\right) = \mathbb{P}\left(\widehat{\mu} \leq m \middle| Y\right) \qquad \text{for all } m \in \mathbb{R}. \tag{2.13}$$

In general, independence between $\boldsymbol{X}$ and $\boldsymbol{D}$ is not sufficient to receive equalized odds for a $\sigma(\boldsymbol{X})$-measurable predictor $\widehat{\mu} = \widehat{\mu}(\boldsymbol{X})$.

(iii) Sufficiency axiom / predictive parity. For predictive parity we exchange the role of the response Y and the predictor $\widehat{\mu}$ compared to equalized odds. We have *predictive parity* if

$$Y \text{ and } \boldsymbol{D} \text{ are conditionally independent under } \mathbb{P}, \text{ given the prediction } \widehat{\mu}.$$

This conditional independence implies for the distribution of the response Y, a.s.,

$$\mathbb{P}\left(Y \leq y | \, \widehat{\mu}, \boldsymbol{D}\right) = \mathbb{P}\left(Y \leq y | \, \widehat{\mu}\right) \qquad \text{for all } y \in \mathbb{R}. \tag{2.14}$$

The notion of predictive parity is inspired by the definition of a sufficient statistic in statistical estimation theory. We can interpret $\mathcal{P} = \{\mathbb{P}_{\boldsymbol{d}}(Y \in \cdot\,) := \mathbb{P}(Y \in \cdot\,|\boldsymbol{D} = \boldsymbol{d}); \boldsymbol{d} \in \mathfrak{D}\}$ as a family of distributions of Y being parametrized by $\boldsymbol{d} \in \mathfrak{D}$. In statistics we call a $\sigma(\boldsymbol{X})$-measurable predictor $\widehat{\mu} = \widehat{\mu}(\boldsymbol{X})$ sufficient for $\mathcal{P}$ if (2.14) holds. Basically, this means that $\widehat{\mu}(\boldsymbol{X})$ carries all the necessary information to predict Y, and the explicit knowledge of $\boldsymbol{D} = \boldsymbol{d}$ is not necessary.

Remark 2.9 From an actuarial point of view, demographic parity seems to be the most natural group fairness axiom. A sufficient (but not necessary) condition to have demographic parity fairness of a $\sigma(\boldsymbol{X})$-measurable predictor $\widehat{\mu}(\boldsymbol{X})$ is that $\boldsymbol{X}$ and $\boldsymbol{D}$ are independent. This means that the insurance portfolio is composed such that the conditional distribution of the non-protected covariates $\boldsymbol{X}$, given $\boldsymbol{D}$, is the same for all demographic groups $\boldsymbol{D} = \boldsymbol{d} \in \mathfrak{D}$. If $\boldsymbol{D}$ describes gender, there may be general insurance products where this is feasible (property insurance). However, e.g., in commercial accident insurance this may not be possible, because the genders are represented with different frequencies in different job profiles, which may make it impossible to compose a portfolio such that the selected jobs have the same distribution for both genders.
Equalized odds fairness is more difficult to achieve, especially if the protected attributes $\boldsymbol{D}$ have different risk factors in different subsets of the non-protected covariates $\boldsymbol{X}$. Similar to the pregnancy costs in Example 2.8, this may make it impossible to achieve equalized odds, except for a trivial covariate-independent predictor. Note that the portfolio composition $\mathbb{P}(\boldsymbol{X}, \boldsymbol{D})$ is in the hands of the insurers, whereas risk factor design is not always possible through insurance cover design, we again think of pregnancy costs that cannot simply be excluded in health insurance contracts.
Predictive parity seems not suitable for insurance pricing because there is hardly any example in which claims can fully be described by a (single) mean parameter $\widehat{\mu}$, i.e., we do not think that there is a realistic situation where a $\sigma(\boldsymbol{X})$-measurable parameter $\widehat{\mu}(\boldsymbol{X})$ is sufficient to describe the full distribution of the claim Y.

2.3 Discrimination-free vs. fair prices

The following two propositions show that discrimination-free prices are generally not fair, and vice versa. The following proposition is proved in Appendix A.

Proposition 2.10 *The discrimination-free insurance price $\mu(\boldsymbol{X}) = X$ of Example 2.6 satisfies none of demographic parity, equalized odds or predictive parity.*

The crucial property of Example 2.6 is that the non-protected covariates $\boldsymbol{X}$ are sufficient for describing the conditional expectation of the response Y, but they are not sufficient to describe the full conditional distribution of Y, given $(\boldsymbol{X}, \boldsymbol{D})$.

Proposition 2.11 *Assume that $\widehat{\mu}$ is demographic parity fair for the protected attributes $\boldsymbol{D}$. In general, this does not imply that $\widehat{\mu}$ is discrimination-free.*

Similar statements hold for equalized odds and predictive parity fairness. Thus, in view of Propositions 2.10 and 2.11 we cannot conclude that one concept is generally stronger than the other.

To prove Proposition 2.11 it suffices to give a counterexample. In Example 2.12, we provide a situation where an unawareness price that indirectly discriminates is, at the same time, demographic parity fair. Furthermore, in Example 2.13, we offer a slight modification, under which demographically fair prices produce even direct discrimination.

Example 2.12 (Indirectly discriminatory demographically fair prices)
We choose three-dimensional Gaussian covariates

$$(\boldsymbol{X}, \boldsymbol{D}) = (X_1, X_2, D) \;\sim\; \mathcal{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}2 & 1 & 1\\1 & 2 & 1\\1 & 1 & 2\end{pmatrix}\right). \tag{2.15}$$

For the response variable we assume

$$Y|_{(\boldsymbol{X},\boldsymbol{D})} \;\sim\; \mathcal{N}\left(2X_1 - 3D,\, 1\right).$$

This gives us the best-estimate price

$$\mu(\boldsymbol{X}, \boldsymbol{D}) = 2X_1 - 3D. \tag{2.16}$$

A standard result on multivariate Gaussian random variables tells us, see, e.g., Corollary 4.4 in [34],

$$D|_{\boldsymbol{X}} \;\sim\; \mathcal{N}\left(\frac{X_1 + X_2}{3},\, \frac{4}{3}\right).$$

This allows us to calculate the unawareness price by

$$\mu(\boldsymbol{X}) = \mathbb{E}\left[\mu(\boldsymbol{X}, \boldsymbol{D})\middle|\, \boldsymbol{X}\right] = 2X_1 - \mathbb{E}\left[3D\middle|\, \boldsymbol{X}\right] = X_1 - X_2. \tag{2.17}$$

Since the best-estimate price has a sensitivity in D and because there is dependence between $\boldsymbol{X}$ and D, this unawareness price indirectly discriminates.
The random vector $(X_1 - X_2, D)$ is two-dimensional Gaussian with independent components because

$$\mathrm{Cov}\left(X_1 - X_2, D\right) = \mathrm{Cov}(X_1, D) - \mathrm{Cov}(X_2, D) = 0.$$

This implies that the unawareness price $\mu(\boldsymbol{X}) = X_1 - X_2$ is independent of D, hence, it is demographic parity fair. Thus, we have constructed an example where the unawareness price indirectly discriminates and, at the same time, it is demographic parity fair. This also proves Proposition 2.11. ∎

Example 2.13 (Directly discriminatory demographically fair prices)
Consider now a modification of Example 2.12, with the covariate model still given by (2.15), but the response now modeled by

$$Y|_{(\boldsymbol{X},\boldsymbol{D})} \;\sim\; \mathcal{N}(X_1 - D, 1).$$

This gives a best-estimate price of the form

$$\mu(\boldsymbol{X}, \boldsymbol{D}) = X_1 - D.$$

This best-estimate price is demographic parity fair because it is independent of D under model (2.15). But at the same time, it explicitly depends on D and, thus, directly discriminates. ■

We now give some additional remarks on Propositions 2.10 and 2.11.

Remarks 2.14

- At first sight, it seems surprising that an example that is perfectly fine from a proxy discrimination viewpoint does not satisfy any of the three classical group fairness axioms of machine learning; see Proposition 2.10. Likewise, an insurance price that is demographic parity fair, does not tell us anything about proxy discrimination; see Proposition 2.11. This indicates that non-discriminatory insurance pricing and group fairness are rather different concepts, and, in general, one does not imply the other. For this reason, satisfying simultaneously both (discrimination-free and group fairness) is much more restrictive than just complying with one of them – and sometimes even impossible if one wants to have a non-trivial predictor. Currently, many regulators focus on proxy discrimination, though corresponding legislation leaves room for interpretation. Therefore, constraining pricing models with group fairness criteria does not seem to solve this particular regulatory problem.

- Indirect discrimination is caused by two factors that need to hold simultaneously, namely, (1) there needs to be a dependence between the non-protected covariates $\boldsymbol{X}$ and the protected attributes $\boldsymbol{D}$, and (2) there needs to be a sensitivity of the best-estimate price $\mu(\boldsymbol{X}, \boldsymbol{D})$ in $\boldsymbol{D}$, see first item of Remarks 2.5. This does not tell us anything about the dependence structure between a discrimination-free insurance price $\mu^*(\boldsymbol{X})$ and $\boldsymbol{D}$. In general, $\mu^*(\boldsymbol{X})$ and $\boldsymbol{D}$ are correlated, namely, observe that the dependence structure between $\boldsymbol{X}$ and $\boldsymbol{D}$ is completely irrelevant in the discrimination-free insurance price calculation (2.6). Therefore, we can always find a portfolio distribution $\mathbb{P}(\boldsymbol{X}, \boldsymbol{D})$ under which the discrimination-free insurance price $\mu^*(\boldsymbol{X})$ and the protected attributes $\boldsymbol{D}$ are dependent, unless $\mu^*(\boldsymbol{X})$ does not depend on $\boldsymbol{X}$.

- Focusing on the example of demographic parity fairness, this notion solely relates to the independence of the resulting prices $\widehat{\mu}(\boldsymbol{X})$ and protected attributes $\boldsymbol{D}$. Hence, if the predictor $\widehat{\mu}(\boldsymbol{X})$ is demographic parity fair, then $\boldsymbol{X} \mapsto \widehat{\mu}(\boldsymbol{X})$ can be interpreted as a projection that only extracts the information from $\boldsymbol{X}$ that is orthogonal to/independent of $\boldsymbol{D}$; this is similar to the linear adversarial concept erasure of Ravfogel et al. [28, 29]; see also Example 2.12. That $\widehat{\mu}(\boldsymbol{X})$ becomes independent of $\boldsymbol{D}$ is a specific property of the pricing functional $\boldsymbol{X} \mapsto \widehat{\mu}(\boldsymbol{X})$ in relation to $\boldsymbol{D}$, but this does not account for the full dependence structure in $\mathbb{P}(\boldsymbol{X}, \boldsymbol{D})$ nor for the properties in the best-estimate price $\mu(\boldsymbol{X}, \boldsymbol{D})$. Therefore, in general, demographic parity does not constitute evidence regarding proxy discrimination.

  If we wanted all participants in an insurance market to comply with demographic parity, we would need to choose projections $\boldsymbol{X} \mapsto \widehat{\mu}(\boldsymbol{X})$ that vary from company to company because

they all have different portfolio distributions $\mathbb{P}(\boldsymbol{X}, \boldsymbol{D})$. As a result, every company would consider non-protected covariates in a different way. This would be difficult to explain to customers and may be impossible to regulate. Therefore, stronger assumptions are typically explored, like aiming at full independence between $\boldsymbol{X}$ and $\boldsymbol{D}$, see Section 3.2, below.

- A crucial feature of Example 2.12 is that independence between $\boldsymbol{X}$ and $\boldsymbol{D}$ is a sufficient condition to have demographic parity fairness, but not a necessary one. This is used in an essential way, namely, $\boldsymbol{X}$ and $\boldsymbol{D}$ are dependent, but the projection $\boldsymbol{X} \mapsto \mu(\boldsymbol{X})$ only extracts a part of information from $\boldsymbol{X}$ that is independent of $\boldsymbol{D}$. We can even go one step further by designing a model that has a demographic parity fair price that is directly discriminatory. Example 2.13 goes even further, by demonstrating a situation where a demographic parity fair price directly discriminates.

3 Input pre-processing and model post-processing

Example 2.6 provides an instance where the discrimination-free insurance price does not satisfy any of the three group fairness axioms. Recently, it has been proposed to either perform *input (data) pre-processing* or *model post-processing* (output post-processing) to comply with (some of) the fairness axioms; see Barrio et al. [5] and Chiappa et al. [8]. We discuss these procedures in the light of insurance pricing. Note that generally the group fairness axioms cannot hold simultaneously; see Barocas et al. [4]. Therefore, one needs to make a choice and we typically consider demographic parity fairness.

In Sections 3.2 and 3.3, below, we will use the theory of *optimal transport* (OT) for input pre-processing and model post-processing. In both cases, independence of predictions from discriminatory features is achieved by a $\boldsymbol{D}$-dependent transformation of features $\boldsymbol{X}$. An important difference between input pre-processing and model post-processing is that the former transforms the inputs $\boldsymbol{X} \mapsto \boldsymbol{X}_+$, and retains the dimension of the original non-protected covariates $\boldsymbol{X}$. In fact, up to technical conditions (continuity), the OT input transformation $\boldsymbol{X} \mapsto \boldsymbol{X}_+$ is one-to-one (for given $\boldsymbol{D}$) which allows us to reconstruct the original features $\boldsymbol{X}$ from the pre-processed ones $\boldsymbol{X}_+$. Model post-processing, using an OT map, transforms the (one-dimensional) regression output $\mu(\boldsymbol{X}, \boldsymbol{D}) \mapsto \mu_+$, by making μ_+ independent of the protected attributes $\boldsymbol{D}$. We have already seen in Example 2.13 a situation where the best estimate price $\mu(\boldsymbol{X}, \boldsymbol{D}) = X_1 - D$ is independent of $\boldsymbol{D}$, hence demographic parity fair. In that example the best estimate price can be identified with μ_+ and the OT output map is the identity map.

3.1 Discrimination-free insurance pricing, revisited

The discrimination-free insurance price (2.6) can be understood as a model post-processing method as we take the (discriminatory) best-estimate price $\mu(\boldsymbol{X}, \boldsymbol{D})$ and we transform it to a discrimination-free insurance price, that is,

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

This is a way of model post-processing. Under the specific pricing measure choice $\mathbb{P}^*(\boldsymbol{D} = \boldsymbol{d}) = \mathbb{P}(\boldsymbol{D} = \boldsymbol{d})$, we can also directly obtain a discrimination-free insurance price by solving an

appropriately reweighed optimization problem.

Proposition 3.1 *The discrimination-free insurance price is given by, a.s.,*

$$\mu^*(\boldsymbol{X}) \;=\; \underset{\widehat{\mu}(\boldsymbol{X})}{\arg\min}\; \mathbb{E}\left[\frac{\mathbb{P}^*(\boldsymbol{D})}{\mathbb{P}\left(\boldsymbol{D}\middle|\,\boldsymbol{X}\right)}\left(Y-\widehat{\mu}(\boldsymbol{X})\right)^2\middle|\,\boldsymbol{X}\right],$$

where the minimization runs over all $\sigma(\boldsymbol{X})$-measurable predictors $\widehat{\mu}(\boldsymbol{X})$, and supposed we have square integrability w.r.t. $\mathbb{P}$ in the above minimization.

The beauty of this result is that we can estimate the discrimination-free insurance price directly from an i.i.d. sample $(y_i, \boldsymbol{x}_i, \boldsymbol{d}_i)_{i=1}^n$ of $(Y, \boldsymbol{X}, \boldsymbol{D})$, without going via the best-estimate price. Select pricing measure $\mathbb{P}^*(\boldsymbol{D}=\boldsymbol{d}) = \mathbb{P}(\boldsymbol{D}=\boldsymbol{d})$, and assume we have access to (estimated) population probabilities $\widehat{\mathbb{P}}(\boldsymbol{D})$ and $\widehat{\mathbb{P}}(\boldsymbol{D}|\boldsymbol{X})$. Then, we can directly find an estimate for the discrimination-free insurance price by solving the weighted square loss problem

$$\widehat{\mu}^*(\boldsymbol{X}) \;=\; \underset{\widehat{\mu}(\boldsymbol{X})}{\arg\min}\; \frac{1}{n}\sum_{i=1}^n \frac{\widehat{\mathbb{P}}(\boldsymbol{D}=\boldsymbol{d}_i)}{\widehat{\mathbb{P}}\left(\boldsymbol{D}=\boldsymbol{d}_i\middle|\,\boldsymbol{X}=\boldsymbol{x}_i\right)}\left(y_i-\widehat{\mu}(\boldsymbol{x}_i)\right)^2.$$

Thus, from the unweighted loss minimization problem we get an unawareness price estimate for (2.2), and from its appropriately weighted counterpart a discrimination-free insurance price estimate. In this interpretation, model post-processing takes place during the model fitting.

3.2 Input (data) pre-processing

A sufficient way to make an insurance price demographic parity fair (and in a certain sense discrimination-free) is to pre-process the non-protected covariates $\boldsymbol{X} \mapsto \boldsymbol{X}_+$ such that its transformed version $\boldsymbol{X}_+$ becomes independent of the protected attributes $\boldsymbol{D}$ under $\mathbb{P}$. First, we emphasize that this pre-processing is *only* performed on the input data $\boldsymbol{X}$ (and using $\boldsymbol{D}$), but it does *not* consider the response Y. Second, independence between $\boldsymbol{X}_+$ and $\boldsymbol{D}$ is a sufficient condition for demographic parity fairness w.r.t. $(\boldsymbol{X}_+, \boldsymbol{D})$, but not a necessary one, see Example 2.12.

One method of input pre-processing is to apply an OT map to obtain a covariate distribution that is independent of the protected attributes; for references see Barrio et al. [5] and Chiappa et al. [8]. More specifically, for given $\boldsymbol{d} \in \mathfrak{D}$, we change the conditional distribution $F_{\boldsymbol{d}}$

$$\boldsymbol{X}_{\boldsymbol{d}} := \boldsymbol{X}|_{\{\boldsymbol{D}=\boldsymbol{d}\}} \;\sim\; F_{\boldsymbol{d}}(\boldsymbol{x}) := F\left(\boldsymbol{x}\middle|\,\boldsymbol{D}=\boldsymbol{d}\right), \tag{3.1}$$

to an unconditional distribution F_+ for the non-protected covariates

$$\boldsymbol{X}_+|_{\boldsymbol{D}} \;\sim\; F_+(\boldsymbol{x}), \tag{3.2}$$

meaning that the transformed covariates $\boldsymbol{X}_+ \sim F_+$ are independent of $\boldsymbol{D}$. Intuitively, to minimally change the predictive power by this transformation from (3.1) to (3.2), the unconditional distribution F_+ should be as similar as possible to the conditional ones $F_{\boldsymbol{d}}$, for all $\boldsymbol{d} \in \mathfrak{D}$.[3] In this approach, the covariates $\boldsymbol{X}$ and $\boldsymbol{X}_+$ preserve their meanings because they live on the same

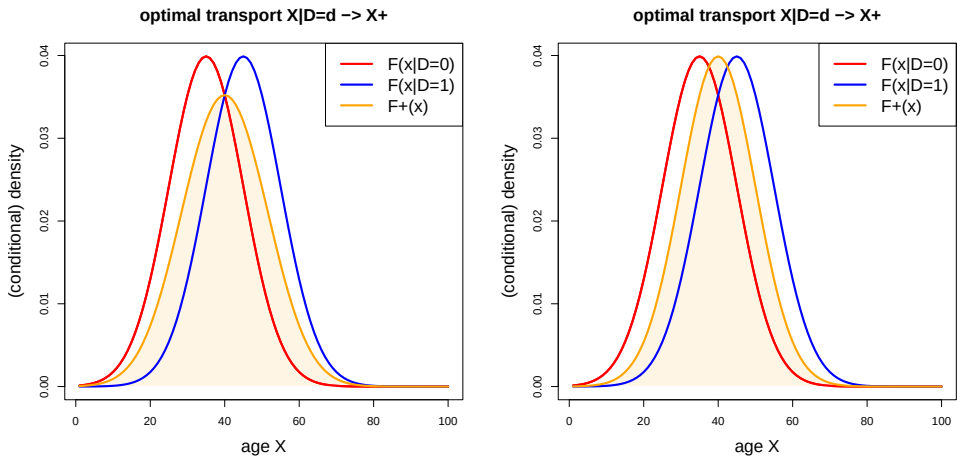

Figure 3: Example 2.8, revisited: conditional densities $f_d(x) = f(x|D = d)$, for $d \in \{0, 1\}$, and two different choices for $f_+(x)$, $x \in \mathbb{R}$; for a formal definition we refer to (3.8)-(3.9).

covariate space, but the OT map locally perturbs the original covariate values $\boldsymbol{X_d} \mapsto \boldsymbol{X}$, based on $\boldsymbol{D} = \boldsymbol{d}$.

We revisit Examples 2.6 and 2.8 illustrated in Figure 1, and give two different proposals for F_+ in Figure 3. The plot on the left hand side shows the average density f_+ of the two Gaussian densities $f_d(x) := f(x|D = d)$, given $D = d \in \{0, 1\}$, i.e., we have a Gaussian mixture for f_+ on the left hand side of Figure 3. The plot on the right hand side shows the Gaussian density for f_+, that averages the means x_0 and x_1; we also refer to (3.8)-(3.9), below. For the moment, it is unclear which of the two choices for F_+ gives a better predictive model for Y.

Assume we have selected an unconditional distribution F_+ to approximate $F_{\boldsymbol{d}}$, $\boldsymbol{d} \in \mathfrak{D}$, and we would like to optimally transform the random variable $\boldsymbol{X_d}$ to its unconditional counterpart $\boldsymbol{X}_+$. This is precisely where OT comes into play. Choose a distance function ϱ on the covariate space. The *(2-)Wasserstein distance* between $F_{\boldsymbol{d}}$ and F_+ is defined by

$$\mathcal{W}_2\left(F_{\boldsymbol{d}}, F_+\right) := \left(\inf_{\pi_{\boldsymbol{d}} \in \Pi_{\boldsymbol{d}}} \int \varrho(\boldsymbol{x}, \boldsymbol{x}_+)^2 \, d\pi_{\boldsymbol{d}}(\boldsymbol{x}, \boldsymbol{x}_+)\right)^{1/2}. \tag{3.3}$$

where $\Pi_{\boldsymbol{d}}$ is the set of all joint probability measures having marginals $F_{\boldsymbol{d}}$ and F_+, respectively. The Wasserstein distance (3.3) measures the difference between the two probability distributions $F_{\boldsymbol{d}}$ and F_+ by optimally coupling them. Colloquially speaking, this optimal coupling means that we try to find the (optimal) transformation $T_{\boldsymbol{d}} : \boldsymbol{X_d} \mapsto \boldsymbol{X}_+$ such that we can perform this change of distribution at a minimal effort;[4] this optimal transformation $T_{\boldsymbol{d}}$ is called an *OT map* or a *push forward*. Under additional technical assumptions, determining the OT map $T_{\boldsymbol{d}} : \boldsymbol{X_d} \mapsto \boldsymbol{X}_+$ is equivalent to finding the optimal coupling $\pi_{\boldsymbol{d}} \in \Pi_{\boldsymbol{d}}$.

[3]This intuition is motivated by the fact that we have maximal information $(\boldsymbol{X}, \boldsymbol{D})$ and we try to retain as much as possible of this information. However, there is no general result that verifies this intuition because the numerical results on sample data will also depend on the chosen class of regression functions. This is in contrast to output post-processing, see Proposition 3.8, below.

[4]The common explanation relates a probability distribution to a pile of soil: a minimal effort can then be understood by transforming this pile of soil of a certain shape into a pile of soil of a given different shape.

Remarks 3.2

- The input OT approach can also be thought of in relation to context-sensitive covariates. For example, the European Commission [14], footnote 1 to Article 2.2(14) – life and health underwriting – mentions the waist-to-hip ratio as a non-protected (useful) context-sensitive covariate for health prediction. Note that the waist-to-hip ratio is gender-, age- and race-dependent. Furthermore the impact of the waist-to-hip ratio on predictions of health outcomes depends specifically on factors like gender, age, and race, that is, the same value should be interpreted differently depending on the demographic group the policyholder belongs to. This means that a $\boldsymbol{D}$-dependent transformation of waist-to-hip ratio is desirable to achieve consistency.
  Applying an OT map will modify the waist-to-hip ratio such that it has the same distribution for both genders, which can then be treated coherently as an input to a predictive model. However, this does not mean that the transformed variable will reflect health impacts in a demographic-group-appropriate way, if the OT map produces a transformation specifically with the aim of removing dependence between $\boldsymbol{X}$ and $\boldsymbol{D}$ and, therefore, depends on the rather arbitrary dependence of those features in a particular portfolio. This also means that care should be taken more generally when considering OT-transformed covariates $\boldsymbol{X}_+$, since their interpretation may not be straightforward. Still, if a transport map is derived from a population distribution of $(\boldsymbol{X}, \boldsymbol{D})$ (e.g., of policyholders across a market), then demographic parity is expected to hold across the market (rather than individual portfolios), and the transformed variables $\boldsymbol{X}_+$ can be interpreted as $\boldsymbol{D}$-agnostic versions of features $\boldsymbol{X}$.

- In many situations the OT map $T_{\boldsymbol{d}} : \boldsymbol{X}_{\boldsymbol{d}} \mapsto \boldsymbol{X}_+$, $\boldsymbol{d} \in \mathfrak{D}$, can explicitly be calculated, e.g., in the discrete covariate case it requires to solve a linear program (LP); see Cuturi–Doucet [11]. The only difficulty in this discrete case is a computational one. Furthermore, the OT map is deterministic for continuous distributions, while in the case of discrete distributions we generally have a random OT map, see also (3.6) below.

- The Wasserstein distance (3.3) can also be defined for categorical covariates. The main difficulty in that case is that one needs to have a suitable distance function ϱ that captures the distance betweeen categorical levels in a meaningful way.

- In general, this OT map should be understood as a local transformation of the covariate space, so that the main structure remains preserved, but the local assignments are perturbed differently for different realizations of $\boldsymbol{D}$. In that, the non-protected covariates $\boldsymbol{X}_{\boldsymbol{d}}$ and $\boldsymbol{X}_+$ keep their original interpretation, e.g., age of policyholder, but through a local perturbation some policyholders receive a slightly smaller or bigger age to make their distributions identical for all $\boldsymbol{D} = \boldsymbol{d}$, $\boldsymbol{d} \in \mathfrak{D}$; note that these perturbations do not use the response Y, i.e., it is a pure input data transformation.

- Assume we have a (one-dimensional) real-valued non-protected covariate $\boldsymbol{x} = x \in \mathbb{R}$ and we choose the Euclidean distance for ϱ. The dual formulation of the Wasserstein distance

(3.3) gives in this special case the simpler formula

$$\begin{aligned}\mathcal{W}_2\left(F_{\boldsymbol{d}}, F_+\right) &= \left(\int_0^1 \left(F_{\boldsymbol{d}}^{-1}(q) - F_+^{-1}(q)\right)^2 dq\right)^{1/2} \\ &= \mathbb{E}\left[\left(F_{\boldsymbol{d}}^{-1}(U) - F_+^{-1}(U)\right)^2\right]^{1/2},\end{aligned} \tag{3.4}$$

where U has a uniform distribution on the unit interval $(0,1)$. The OT map $T_{\boldsymbol{d}}$, $\boldsymbol{d} \in \mathfrak{D}$, is then in the one-dimensional continuous covariate case given by

$$X \;\mapsto\; X_+ = T_{\boldsymbol{d}}(X) = F_+^{-1} \circ F_{\boldsymbol{d}}(X). \tag{3.5}$$

This justifies the statement in the previous bullet point that the OT map is a local transformation, since the topology is preserved by (3.5). In case $F_{\boldsymbol{d}}$ is not continuous, the OT map needs randomization. In the one-dimensional case we replace the last term in (3.5) by

$$V := F_{\boldsymbol{d}}(X_-) + U\left(F_{\boldsymbol{d}}(X_-) - F_{\boldsymbol{d}}(X)\right), \tag{3.6}$$

where U is independent of everything else and uniform on $(0,1)$, and where we set for the left limit $F_{\boldsymbol{d}}(X_-) = \lim_{x \uparrow X} F_{\boldsymbol{d}}(x)$ in X. As a result, V is uniform on $(0,1)$, and we set $X_+ = F_+^{-1}(V)$.

We emphasize that (3.5) and (3.6) reflects the OT map only in the one-dimensional case, and for the multidimensional (empirical) case we have to solve a linear program, as indicated in the second bullet point of these remarks.

Next, we state that the OT input pre-processed version of the non-protected covariates is demographic parity fair and discrimination-free with respect to the transformed inputs $\boldsymbol{X}_+$. Also, interestingly, these notions do not touch the response Y, but it is sufficient to know the best-estimate price $\mu(\boldsymbol{X}, \boldsymbol{D})$. The proof of the next lemma is straightforward.

Lemma 3.3 (OT input pre-processing) *Consider the triplet $(Y, \boldsymbol{X}, \boldsymbol{D})$ and choose the OT maps $T_{\boldsymbol{d}} : \boldsymbol{X}_{\boldsymbol{d}} \mapsto \boldsymbol{X}_+$, $\boldsymbol{d} \in \mathfrak{D}$, with $\boldsymbol{X}_+$ being independent of $\boldsymbol{D}$ (under $\mathbb{P}$). The unawareness price*

$$\begin{aligned}\mu\left(\boldsymbol{X}_+\right) \;=\; \mathbb{E}\left[Y \,|\boldsymbol{X}_+\right] &= \sum_{\boldsymbol{d} \in \mathfrak{D}} \mathbb{E}\left[Y \,|\boldsymbol{X}_+ \,, \boldsymbol{D} = \boldsymbol{d}\right] \mathbb{P}\left(\boldsymbol{D} = \boldsymbol{d}\right) \\ &= \sum_{\boldsymbol{d} \in \mathfrak{D}} \mathbb{E}\left[\mu(\boldsymbol{X}, \boldsymbol{D}) \,|\boldsymbol{X}_+ \,, \boldsymbol{D} = \boldsymbol{d}\right] \mathbb{P}\left(\boldsymbol{D} = \boldsymbol{d}\right)\end{aligned}$$

is discrimination-free w.r.t. $(\boldsymbol{X}_+, \boldsymbol{D})$ and satisfies demographic parity fairness.

We emphasize that Lemma 3.3 makes a statement about the transformed input $(\boldsymbol{X}_+, \boldsymbol{D})$ and not about the original covariates $(\boldsymbol{X}, \boldsymbol{D})$. Hence, whether we can consider the price $\mu(\boldsymbol{X}_+)$ to be truly discrimination-free depends on the interpretation we attach to the transformed inputs $\boldsymbol{X}_+$, see the first bullet in Remarks 3.2. Moreover, Lemma 3.3 applies to any transformation $T_{\boldsymbol{d}} : \boldsymbol{X}_{\boldsymbol{d}} \mapsto \boldsymbol{X}_+$, $\boldsymbol{d} \in \mathfrak{D}$ that makes $\boldsymbol{X}_+$ independent of $\boldsymbol{D}$, and which does not add more information to $(\boldsymbol{X}, \boldsymbol{D})$ w.r.t. the prediction of Y; this is what we use in the last equality statement.

Now, we consider one-dimensional OT in the context of our Example 2.8. The method is similar to the (one-dimensional) proposals in Section 4.3 of Xin–Huang [37], called there 'debiasing variables'. However, the OT approach works in any dimension, and also takes care of the dependence structure within $\boldsymbol{X}$, given $\boldsymbol{D}$. Nevertheless, we consider a one-dimensional example for illustrative purposes.

Example 3.4 (Application of input OT)

We apply the OT input pre-processing to the situation of Example 2.8, which considered age- and gender-dependent costs, including excess costs for women between 20 and 40. Our aim is to obtain an insurance price that is both demographic parity fair and discrimination-free (with respect to the transformed inputs). In this set-up we have a real-valued non-protected covariate $\boldsymbol{X} = X$, and we can directly apply the one-dimensional OT formulations (3.4) and (3.5). For the conditional distributions we have for $d = 0, 1$ and for given x_d and $\tau > 0$, see (2.7),

$$X_d = X|_{\{D=d\}} \;\sim\; F_d(x) = \Phi\left(\frac{x - x_d}{\tau}\right), \tag{3.7}$$

where Φ denotes the standard Gaussian distribution. For the transformed distribution F_+ we select the two examples of Figure 3; the first one is given by

$$F_+(x) = \frac{1}{2}\,\Phi\left(\frac{x - x_0}{\tau}\right) + \frac{1}{2}\,\Phi\left(\frac{x - x_1}{\tau}\right), \tag{3.8}$$

and the second one by

$$F_+(x) = \frac{1}{2}\,\Phi\left(\frac{x - (x_0 + x_1)/2}{\tau}\right). \tag{3.9}$$

Selections (3.8) and (3.9) are two possible choices by the modeler, but any other choice for F_+ which does not depend on D is also possible. The first choice is the average of the two conditional distributions (3.7), the second one is their Wasserstein barycenter; we refer to Proposition 3.8 and Remarks 3.9, below.

We start by calculating the Wasserstein distances (3.4) using Monte Carlo simulation and a discretized approximation to F_+^{-1} in the case of the Gaussian mixture distribution (3.8). The results are presented in Table 2. We observe that the second option (3.9) is closer to the conditional distributions F_d, $d = 0, 1$, in Wasserstein distance; in fact, in this second option we have $|F_d^{-1}(u) - F_+^{-1}(u)| = (x_1 - x_0)/2$ for all $u \in (0, 1)$, and there is no randomness involved in the calculation of the expectation in (3.4).

	$D = 0$	$D = 1$
input OT example (3.8) for F_+	5.14	5.14
input OT example (3.9) for F_+	5.00	5.00

Table 2: Wasserstein distances $\mathcal{W}_2(F_d, F_+)$ for the two examples (3.8)-(3.9) for F_+.

Figure 4 shows the OT maps (3.5) for the two choices of F_+ given by (3.8)-(3.9). We observe that in the second option we generally make women older by $(x_1 - x_0)/2 = 5$ years, and we generally make men younger by $(x_1 - x_0)/2 = 5$ years, so that the distributions F_+ of the OT

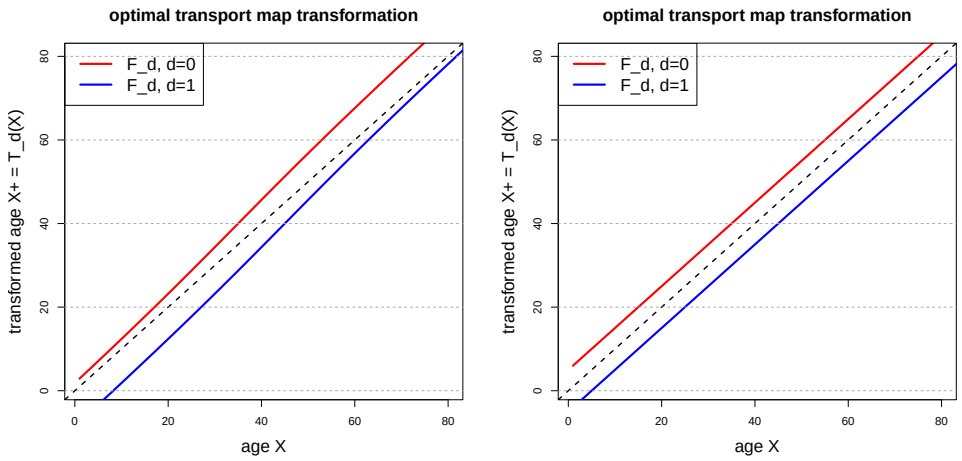

Figure 4: OT maps T_d for examples (3.8)-(3.9) of F_+; the black dotted lines is the 45° diagonal.

transformed ages $X_+ = T_d(X)$ coincide for both genders $d = 0, 1$. The first option (3.8) leads to an age dependent transformation. If we focus on the y-axis in Figure 4, we can identify the ages of women and men that are assigned to the same age cohort. For instance, following the horizontal gray dotted line at level $X_+ = 40$, we find for the second option (3.9) that women of age 35 and men of age 45 will be in the same age cohort (and hence same price cohort). This seems a comparably large age shift which may be difficult to explain to customers. However, in real insurance portfolios we expect more similarity between women and men so that we need smaller age shifts; for mortality related products such an age shift may even be sensible. Additionally, this picture will be superimposed by more non-protected covariates which will require the multidimensional OT map framework.

Based on this OT input transformed data, we construct a regression model $\boldsymbol{X}_+ \mapsto \widehat{\mu}(\boldsymbol{X}_+)$. In this (simple) one-dimensional problem $\boldsymbol{X}_+ = X_+$ we simply fit a cubic spline to the data $(Y, \boldsymbol{X}_+)$ using the `locfit` package in R; see [22]. The resulting model is discrimination-free and demographic parity fair w.r.t. $(\boldsymbol{X}_+, \boldsymbol{D})$, see Lemma 3.3.

	MSE $\mathcal{L}(\widehat{\mu}, Y)$	bias $\mathbb{E}[\widehat{\mu}]$
best-estimate price $\mu(\boldsymbol{X}, \boldsymbol{D})$	100.00	41.25
unawareness price $\mu(\boldsymbol{X})$	197.20	41.25
input OT map of (3.8) for $\widehat{\mu}(\boldsymbol{X}_+)$	162.77	41.25
input OT map of (3.9) for $\widehat{\mu}(\boldsymbol{X}_+)$	162.72	41.25
input OT map of (3.9) for best-estimate $\widehat{\mu}(\boldsymbol{X}_+, \boldsymbol{D})$	100.60	41.24

Table 3: Mean squared errors and average prediction of the different prices in Example 2.8.

Table 3 presents the prediction accuracy of the OT input transformed models. At first sight surprising, the OT input transformed model $\widehat{\mu}(\boldsymbol{X}_+)$ has a better predictive performance than the unawareness price model $\mu(\boldsymbol{X})$. However, by understanding the true model, this is not that surprising. Women have generally higher costs than men under model assumption (2.11). The OT maps (3.8) and (3.9) make women older and men younger, and as a result their risk profiles

w.r.t. the transformed inputs $\boldsymbol{X}_+ = T_{\boldsymbol{d}}(\boldsymbol{X})$ become more similar in this example. This precisely leads, in this case, to a smaller mean squared error of $\widehat{\mu}(\boldsymbol{X}_+)$ over $\mu(\boldsymbol{X})$. This statement can be verified by switching the age profiles by setting $x_0 = 45$ and $x_1 = 35$, and keeping everything else unchanged, as seen by the results in Table 4.

	MSE $\mathcal{L}(\widehat{\mu}, Y)$	bias $\mathbb{E}[\widehat{\mu}]$
best-estimate price $\mu(\boldsymbol{X}, \boldsymbol{D})$	100.00	38.01
unawareness price $\mu(\boldsymbol{X})$	197.12	38.01
input OT map of (3.8) for $\widehat{\mu}(\boldsymbol{X}_+)$	290.68	38.01
input OT map of (3.9) for $\widehat{\mu}(\boldsymbol{X}_+)$	290.64	38.01

Table 4: Changed role of ages of women and men, setting $x_0 = 45$ and $x_1 = 35$.

We emphasize that the OT map $T_{\boldsymbol{d}}$ is selected solely based on the inputs $(\boldsymbol{X}, \boldsymbol{D})$ and not considering the response Y. As a result, we can receive a predictive model that is either better or worse than the unawareness price model. It is important to mention that the selection of the OT map is not allowed to consider the response Y, otherwise it may (and will) imply a sort of indirect model selection discrimination.

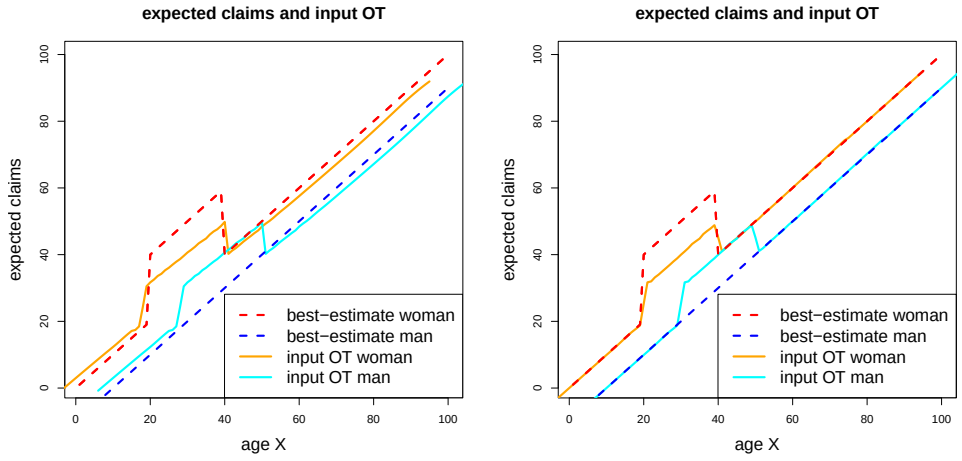

Figure 5: OT input transformed model $\widehat{\mu}(\boldsymbol{X}_+)$ for examples (3.8)-(3.9) of F_+.

Figure 5 illustrates the OT input transformed model prices $\widehat{\mu}(\boldsymbol{X}_+)$ for choices (3.8)-(3.9) for F_+. For Figure 5 we map these prices back to the original features $\boldsymbol{X}$, separated by gender $\boldsymbol{D}$. This back-transformation can be done because the OT maps $T_{\boldsymbol{d}}$ are one-to-one under continuous non-protected covariates $\boldsymbol{X}$, and for given $\boldsymbol{D} = \boldsymbol{d}$, see Remarks 3.2. Figure 5 then evaluates the prices $\widehat{\mu}(\boldsymbol{X}_+)$, where we consider $\boldsymbol{X}_+ = \boldsymbol{X}_+(\boldsymbol{x}; \boldsymbol{d}) = T_{\boldsymbol{d}}(\boldsymbol{x})$ as a function of age $\boldsymbol{x}$ for fixed gender $\boldsymbol{D} = \boldsymbol{d}$. The right hand side shows choice (3.9) for F_+, which leads to parallel shifts for the transformed age assignments $\boldsymbol{X}_+$, see Figure 4 (rhs). As a consequence, the excess pregnancy costs of women with ages in $[20, 40]$ are shared with men having ages in $[30, 50]$ in our example, see orange and cyan lines in Figure 5 (rhs). This should be contrasted to the discrimination-free insurance price $\mu^*(\boldsymbol{X})$ (green line in Figure 2) which shares the excess pregnancy costs within the age class $[20, 40]$ for both genders. The transformation for choice (3.8) for F_+ leads to a

distortion along the age cohorts as we do not have parallel shifts, see Figure 4 (lhs) and Figure 5 (lhs).
The prices depicted in Figure 5 are demographic parity fair and discrimination-free with respect to the covariates $(\boldsymbol{X}_+, \boldsymbol{D})$, see Lemma 3.3. As discussed in Remarks 3.2, whether one considers these prices desirable in relation to direct and indirect discrimination depends on whether the transformed age $\boldsymbol{X}_+$ can be interpreted/justified as a valid covariate in its own right. If it is seen as just an artifice of the dependence structure of $(\boldsymbol{X}, \boldsymbol{D})$, stakeholders may be more interested in discrimination with respect to the original covariates $(\boldsymbol{X}, \boldsymbol{D})$. From such a perspective it is clear that the prices of Figure 5 are subject to even *direct* discrimination, given the different dashed lines for women and for men on the original scale.
An important difference between the discrimination-free insurance price $\mu^*(\boldsymbol{X})$ and the OT map transformed prices $\widehat{\mu}(\boldsymbol{X}_+)$ is that the latter always provides a (statistically) unbiased model. In fact, the latter does not only satisfy the balance property, but even the more restrictive auto-calibration property; see Wüthrich–Ziegel [36].
Finally, we build a best-estimate model $\widehat{\mu}(\boldsymbol{X}_+, \boldsymbol{D})$ on the transformed information $(\boldsymbol{X}_+, \boldsymbol{D})$. We do this by separately fitting two cubic splines to the women data $(Y, X_+, D = 0)$ and the men data $(Y, X_+, D = 1)$, respectively. The results are presented on the last line of Table 3. Up to estimation error, we rediscover the true model, but on the transformed input data, as the mean squared error only contains the noise part (irreducible risk) of the response Y. Thus, as expected, this one-to-one OT map (in the continuous case), for given gender, does not involve a loss of information, and the predictive performance in the parametrizations $(\boldsymbol{X}, \boldsymbol{D})$ and $(\boldsymbol{X}_+, \boldsymbol{D})$ coincides (up to estimation error). ■

3.3 Model post-processing

Model post-processing to achieve fairness works on the outputs, and not on the inputs like data pre-processing. From a purely technical viewpoint, both methods work in a similar manner. A main difference is that input pre-processing usually is multidimensional and (regression) model post-processing is one-dimensional. Assume, in a first step, we have fitted a best-estimate price model $(\boldsymbol{X}, \boldsymbol{D}) \mapsto \mu(\boldsymbol{X}, \boldsymbol{D})$. Model post-processing applies transformations to these best-estimate prices $\mu(\boldsymbol{X}, \boldsymbol{D}) \mapsto \mu_+$ such that the transformed price μ_+ fulfills a fairness axiom. If we focus on demographic parity, the transformed price μ_+ should be independent of $\boldsymbol{D}$. Note that any of the following steps could equivalently be applied to any other pricing functional, such as the unawareness price $\mu(\boldsymbol{X})$.
If we apply an OT output transformation, we modify (3.1) and (3.2) as follows. For $\boldsymbol{d} \in \mathfrak{D}$, we change the conditional distributions $G_{\boldsymbol{d}}$ on $\mathbb{R}$

$$\mu_{\boldsymbol{d}}(\boldsymbol{X}) := \mu(\boldsymbol{X}, \boldsymbol{D})|_{\{\boldsymbol{D}=\boldsymbol{d}\}} \;\sim\; G_{\boldsymbol{d}}(m) := \mathbb{P}\left(\mu(\boldsymbol{X}, \boldsymbol{D}) \le m \middle| \, \boldsymbol{D} = \boldsymbol{d}\right) \qquad \text{for } m \in \mathbb{R}, \tag{3.10}$$

to an unconditional distribution G_+ for the prices

$$\mu_+|_{\boldsymbol{D}} \;\sim\; G_+(m)\,. \tag{3.11}$$

In particular, this means that the real-valued random variable $\mu_+ \sim G_+$ is independent of $\boldsymbol{D}$. Based on these choices we look for OT maps $T_{\boldsymbol{d}} : \mu_{\boldsymbol{d}}(\boldsymbol{X}) \mapsto \mu_+$, given $\boldsymbol{d} \in \mathfrak{D}$, providing the corresponding distribution. Since everything is one-dimensional here, we can directly work with

the versions (3.5) and (3.6), respectively, depending whether our price functionals $\mu_{\boldsymbol{d}}(\boldsymbol{X})$ have continuous marginals $G_{\boldsymbol{d}}$ or not. Thus, in the continuous case we have OT maps

$$\mu_{\boldsymbol{d}}(\boldsymbol{X}) \;\mapsto\; \mu_+ = T_{\boldsymbol{d}}(\mu_{\boldsymbol{d}}(\boldsymbol{X})) = G_+^{-1} \circ G_{\boldsymbol{d}}\left(\mu_{\boldsymbol{d}}(\boldsymbol{X})\right), \tag{3.12}$$

for $\boldsymbol{d} \in \mathfrak{D}$. The resulting Wasserstein distance is given by (3.4) with $(F_{\boldsymbol{d}}, F_+)$ replaced by $(G_{\boldsymbol{d}}, G_+)$. With this procedure, since the distribution G_+ does not depend on $\boldsymbol{D}$, the OT transformed price μ_+ fulfills demographic parity. The remaining question is how to choose G_+.

Remark 3.5 $\mu_{\boldsymbol{d}}(\boldsymbol{X}) \sim G_{\boldsymbol{d}}$ is a real-valued random variable, and one should not get confused by the multidimensional covariate $\boldsymbol{X}$ in this expression; also the OT transformed price $\mu_+ \sim G_+$ is a real-valued random variable, independent of $\boldsymbol{D}$. Often, one wants to relate this price μ_+ to the original covariates $(\boldsymbol{X}, \boldsymbol{D})$. In the continuous case we can do this using the OT maps (3.12), namely, we have a measurable map

$$(\boldsymbol{x}, \boldsymbol{d}) \;\mapsto\; \mu_+ = \mu_+(\boldsymbol{x}; \boldsymbol{d}) = G_+^{-1} \circ G_{\boldsymbol{d}}\left(\mu(\boldsymbol{x}, \boldsymbol{d})\right) \;\in\; \mathbb{R}. \tag{3.13}$$

Formula (3.13) gives the OT transformed price μ_+ of a given insurance policy with covariates $(\boldsymbol{X}, \boldsymbol{D}) = (\boldsymbol{x}, \boldsymbol{d})$, and (3.12) describes the distribution of this price, if we randomly select an insurance policy from our portfolio $\boldsymbol{X}|_{\{\boldsymbol{D}=\boldsymbol{d}\}} \sim F_{\boldsymbol{d}}$, for given protected attributes $\boldsymbol{D} = \boldsymbol{d}$.

Example 3.6 (Application of output OT)

We revisit Examples 2.8, 3.4, but now, instead of input pre-processing, we apply model post-processing to the best-estimate $\mu(\boldsymbol{X}, \boldsymbol{D})$. These best-estimates are illustrated in red and blue color in Figure 2. As density g_+ we simply choose the average of the two conditional densities

$$g_+(m) = \frac{1}{2}\left(g_0(m) + g_1(m)\right) \qquad \text{for } m \in \mathbb{R}. \tag{3.14}$$

Note that the distributions of $\mu(X, D)|_{\{D=d\}}$ are absolutely continuous, therefore their densities g_d exist. Figure 6 illustrates the density g_+ and the resulting distribution G_+, respectively.

	MSE $\mathcal{L}(\widehat{\mu}, Y)$	bias $\mathbb{E}[\widehat{\mu}]$
best-estimate price $\mu(\boldsymbol{X}, \boldsymbol{D})$	100.00	41.25
unawareness price $\mu(\boldsymbol{X})$	197.20	41.25
output OT map of (3.14) for μ_+	152.97	41.25

Table 5: Mean squared errors and average prediction of the different prices in Example 2.8.

Table 5 presents the results of the OT output post-processed best-estimate prices using density (3.14) for g_+. The resulting mean squared error is smaller than the corresponding value of the input OT version, see Table 3. This is generally expected for suitable choices of g_+ because the fairness debiasing only takes place in the last step of the (estimation) procedure, and all previous steps deriving the best-estimate price uses full information $(\boldsymbol{X}, \boldsymbol{D})$. Input OT already performs the debiasing procedure in the first step and, therefore, all subsequent steps are generally non-optimal in terms of full information $(\boldsymbol{X}, \boldsymbol{D})$.

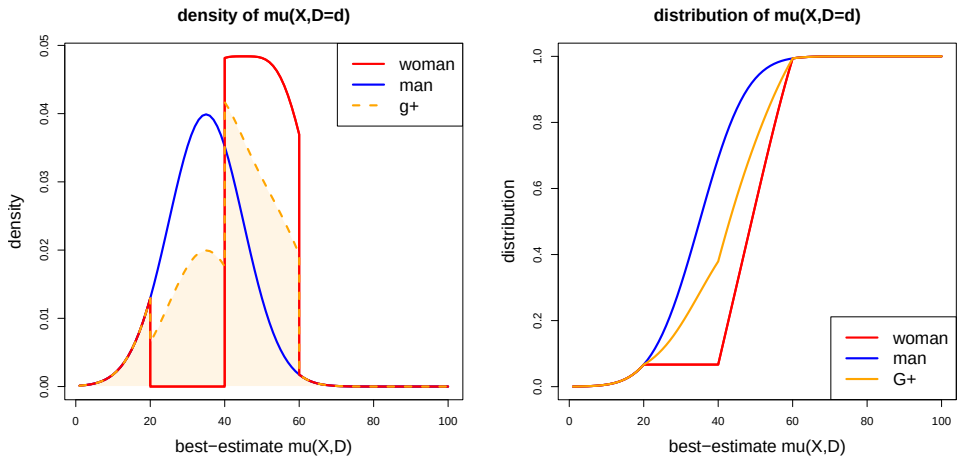

Figure 6: OT output post-processing density g_+ and distribution G_+.

OT output post-processing directly acts on the best-estimate prices $\mu(\boldsymbol{X}, \boldsymbol{D})$. These best-estimate prices can be understood as price cohorts, and for OT output post-processing the specific (multidimensional) value of the non-protected covariates, say $\boldsymbol{X} \in \{\boldsymbol{x}, \boldsymbol{x}'\}$, does not matter as long as $\mu(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{D} = \boldsymbol{d}) = \mu(\boldsymbol{X} = \boldsymbol{x}', \boldsymbol{D} = \boldsymbol{d})$. In case of non-monotone best-estimate prices, this can lead to price distortions that are not explainable to customers and policymakers. In Figure 7 (top) we express the output post-processed prices $\mu_+ = \mu_+(\boldsymbol{x}; \boldsymbol{d})$ as a function of the original age variable $\boldsymbol{X} = \boldsymbol{x}$, separated by gender $\boldsymbol{D} = \boldsymbol{d} \in \{0, 1\}$, we also refer to (3.13). We observe that for women $\boldsymbol{D} = 0$, the best-estimate prices $\mu(\boldsymbol{X} = 30, \boldsymbol{D} = 0) = \mu(\boldsymbol{X} = 50, \boldsymbol{D} = 0) = 50$ coincide (red dots in Figure 7, top), but the underlying risk factors for these high costs are completely different ones. Women at age 30 have high costs because of pregnancy, and women at age 50 have high costs because of aging (women at age 50 are assumed to not be able to get pregnant). Using OT output post-processing, these two age classes (being in the same price cohort) are treated completely equally and obtain the same fairness debiasing discount (orange dot in Figure 7, top). But this discount for women at age 50 cannot be justified if we believe that fairness (or anti-discrimination) should compensate for the excess pregnancy costs which only applies to women but not to men between ages 20 and 40. In fact, this is precisely how the excess pregnancy costs are treated in the discrimination-free insurance price $\mu^*(\boldsymbol{X})$, see green line in Figure 7 (bottom-rhs), and in the OT input pre-processing price $\mu(\boldsymbol{X}_+)$, see Figure 7 (bottom-lhs). (The plots at the bottom of Figure 7 are repeated from Examples 2.8 and 3.4) for ease of comparison). ■

Remark 3.7 From Example 3.6, we conclude that output post-processing should only be used with great care. The price functional $\boldsymbol{x} \mapsto \mu(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{d}) \in \mathbb{R}$ typically leads to a large loss of information (this can be interpreted as a projection), and insurance policies with completely different risk factors may be assigned to the same price cohort by this projection. Therefore, it is questionable if model post-processing should treat different cohorts $\boldsymbol{X} = \boldsymbol{x}$ with equal best-estimate prices equally (which precisely happens in OT output post-processing) or whether we should look for another way of correcting. Of course, one may similarly object to the case of input

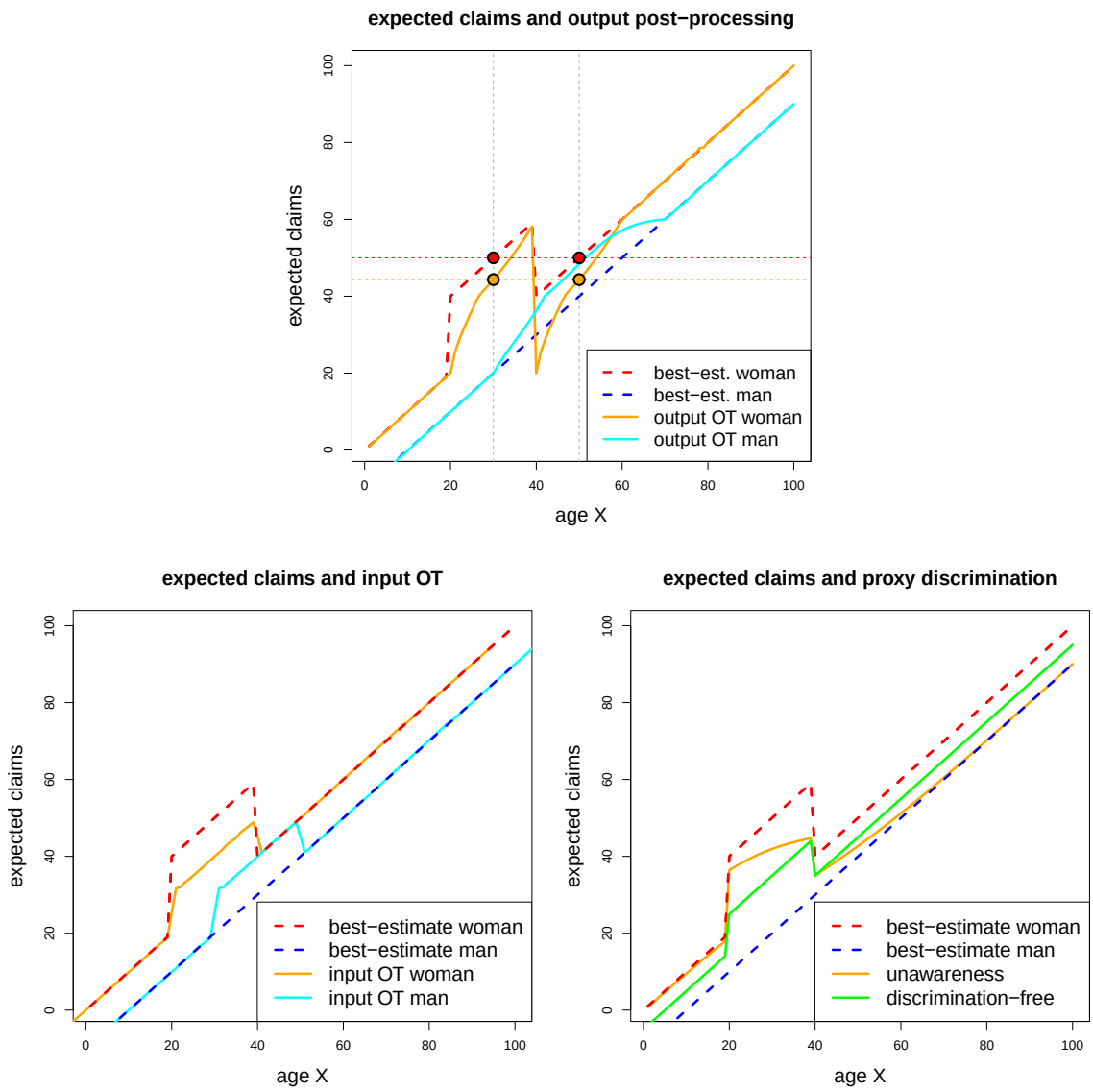

Figure 7: (Top) OT output post-processed prices $\mu_+ = \mu_+(\boldsymbol{x}; \boldsymbol{d})$ expressed in their original features $\boldsymbol{x}$ and separated by gender $\boldsymbol{d}$, see (3.13); (bottom-lhs) OT input pre-processing taken from Figure 5; (bottom-rhs) unawareness and discrimination-free insurance prices taken from Figure 2.

OT, particularly that excess pregnancy costs of women at age 20-40 are shared specifically with men of age 30-50. Nonetheless, note that, at least, the results of input OT, Figure 7 (bottom-lhs), are easier to interpret compared to Figure 7 (top). Note though that when policyholder features $\boldsymbol{X}$ are highly granular, it becomes difficult to assign policies into homogeneous groups. In such circumstances we may find that the new rating classes induced by input OT are also hard to interpret.

If, despite the last criticism, we would like to hold on to OT model post-processing, we may ask the question about the optimal OT transform in (3.12) and (3.11), respectively, to receive

maximal predictive power of $\widehat{\mu}$ for Y. Of course, the same question applies to OT input pre-processing (3.2), but this latter question cannot be generally answered because, in OT input pre-processing, the OT transformed non-protected covariates $\boldsymbol{X}_+$ then run through a general, typically non-linear, regression function $\boldsymbol{X}_+ \mapsto \mu(\boldsymbol{X}_+)$. This makes it impossible to give criteria for optimal pre-processing of the inputs.
For optimal model post-processing with OT we can rely on analytical results in one-dimensional OT. In particular, Theorem 2.3 of Chzhen et al. [10] states the following.

Proposition 3.8 *Assume $\mu_{\boldsymbol{d}}(\boldsymbol{X}) \sim G_{\boldsymbol{d}}$ are absolutely continuous for all $\boldsymbol{d} \in \mathfrak{D}$. Consider*

$$\mu_+(\boldsymbol{x};\boldsymbol{d}) = \left(\sum_{\boldsymbol{d}' \in \mathfrak{D}} \mathbb{P}\left[\boldsymbol{D} = \boldsymbol{d}'\right] G_{\boldsymbol{d}'}^{-1} \right) \circ G_{\boldsymbol{d}}\left(\mu(\boldsymbol{x},\boldsymbol{d})\right). \tag{3.15}$$

Then, $\mu_+ = \mu_+(\boldsymbol{X},\boldsymbol{D})$ is the $\sigma(\boldsymbol{X},\boldsymbol{D})$-measurable and demographic parity fair predictor of Y that has minimal mean squared error.

Remarks 3.9

- The big round brackets in (3.15) give the inverse of the optimal distribution for G_+, see also (3.12). In fact, this specific choice of G_+ corresponds to the *barycenter* of the conditional distributions $(G_{\boldsymbol{d}})_{\boldsymbol{d}\in\mathfrak{D}}$ w.r.t. the Wasserstein distance (3.4). From this we conclude that if we choose this barycenter, we receive the L^2-optimal $\boldsymbol{D}$-independent (demographic parity fair) $\sigma(\boldsymbol{X},\boldsymbol{D})$-measurable predictor for Y. Since choice (3.14) is not the barycenter in that example, predictive performance could still be improved in our OT model post-processing example. On the other hand, we have used the barycenter in (3.9), see also Table 2, but for input pre-processing this is not a crucial choice and other choices may perform better (depending on the specific regression model class being used).

- In (3.15) we have a measurable function of type (3.13). We can relate this back to conditional expectations similar to Lemma 3.3. Consider the random variable

$$\mu^\dagger(\boldsymbol{X};\boldsymbol{d}') := G_{\boldsymbol{d}'}^{-1} \circ G_{\boldsymbol{d}}\left(\mu_{\boldsymbol{d}}(\boldsymbol{X})\right) \;\sim\; G_{\boldsymbol{d}'},$$

  i.e., this random variable $\mu^\dagger(\boldsymbol{X};\boldsymbol{d}')$ has the same conditional distribution as $\mu_{\boldsymbol{d}'}(\boldsymbol{X})$. We can then rewrite (3.15) as follows

$$\mu_+(\boldsymbol{X};\boldsymbol{d}) = \left(\sum_{\boldsymbol{d}' \in \mathfrak{D}} \mathbb{P}\left[\boldsymbol{D} = \boldsymbol{d}'\right] G_{\boldsymbol{d}'}^{-1} \right) \circ G_{\boldsymbol{d}}\left(\mu_{\boldsymbol{d}}(\boldsymbol{X})\right) = \sum_{\boldsymbol{d}' \in \mathfrak{D}} \mu^\dagger(\boldsymbol{X};\boldsymbol{d}')\, \mathbb{P}\left[\boldsymbol{D} = \boldsymbol{d}'\right].$$

  That is, similar to the discrimination-free insurance price and the OT input pre-processed price of Lemma 3.3, we take an unconditional expectation in protected attributes $\boldsymbol{D}$ over $\mu^\dagger(\boldsymbol{X};\boldsymbol{d}')$. Moreover, we can relate the latter to best-estimate prices, i.e., to any realization of $\boldsymbol{X}_{\boldsymbol{d}} = \boldsymbol{x}$ we can assign a covariate value $\boldsymbol{x}^\dagger_{\boldsymbol{d}'}$ such that

$$\mu^\dagger(\boldsymbol{x};\boldsymbol{d}') = \mathbb{E}\left[Y \,\middle|\, \boldsymbol{X} = \boldsymbol{x}^\dagger_{\boldsymbol{d}'}, \boldsymbol{D} = \boldsymbol{d}' \right] = \mu(\boldsymbol{x}^\dagger_{\boldsymbol{d}'}, \boldsymbol{d}').$$

  This implies,

$$\mu_+(\boldsymbol{x};\boldsymbol{d}) = \sum_{\boldsymbol{d}' \in \mathfrak{D}} \mu(\boldsymbol{x}^\dagger_{\boldsymbol{d}'}, \boldsymbol{d}')\, \mathbb{P}\left[\boldsymbol{D} = \boldsymbol{d}'\right].$$

Thus, formally we can write the OT post-processed price as a discrimination-free insurance price. However, this line of argument suffers the same deficiency as Figure 7 (top), namely, the assignment $\boldsymbol{x}^{\dagger}_{\boldsymbol{d}'}$ is non-unique, and we may select different non-protected covariate values for this assignment that have completely different risk factors.

4 Conclusions and discussion

We have shown that discrimination and (group) fairness are materially different concepts. We can have discrimination-free insurance prices that do not satisfy any of the group fairness axioms in machine learning, and, vice versa, we can have, e.g., demographic parity fair prices that are not discrimination-free. In particular, in Example 2.13 we gave an example of a demographic parity fair price that directly discriminates from an insurance regulation view. This clearly questions the direct application of group fairness axioms to insurance pricing, as they do not provide a quick fix for (and may even conflict with) mitigating discrimination.
In a next step, we presented OT input pre-processing and OT output post-processing. These methods can be used to make distributions of non-protected characteristics independent of protected attributes. Input pre-processing locally perturbs the non-protected covariates $\boldsymbol{X}|_{\boldsymbol{D}}$ such that the resulting conditional distributions become independent of the protected attributes $\boldsymbol{D}$. If we only work with these transformed covariates, we receive demographic parity fairness and non-discriminatory insurance prices; however note that there will generally be direct discrimination with respect to the original covariates, as depicted in Figure 7. Output post-processing is different as it acts on the real-valued best-estimates $\mu(\boldsymbol{X}, \boldsymbol{D})$, which should be seen as a summary statistic for pricing that already suffers from a loss of information, i.e., we can no longer fully distinguish the underlying risk factors that lead to these best-estimate prices. This may make output post-processing problematic because we may receive fairness debiasing that cannot be explained to customers and policymakers.

The following table compares the crucial differences between discrimination-free insurance pricing and group fairness through OT input pre-processing.

Addressing indirect discrimination	Addressing fairness
Model post-processing of prices $\mu(\boldsymbol{X}, \boldsymbol{D})$	Input pre-processing of features $\boldsymbol{X}$
Change of probability from $\mathbb{P}$ to $\mathbb{P}^*$	Deformation of $\boldsymbol{X}$ to $\boldsymbol{X}_+$
Independence of $\boldsymbol{X}$ and $\boldsymbol{D}$ under $\mathbb{P}^*$	Independence of $\boldsymbol{X}_+$ and $\boldsymbol{D}$ under $\mathbb{P}$
Dependence of $\boldsymbol{X}$ and $\boldsymbol{D}$ under $\mathbb{P}$ ...	Dependence of $\boldsymbol{X}$ and $\boldsymbol{D}$ under $\mathbb{P}$ ...
... does not matter for price adjustments	... matters for price adjustments

We list further points that require careful considerations in any attempt to regulate insurance prices with reference to non-discrimination and group fairness concepts:

- One difficulty in this field is that there are many different terms that do not have precise (mathematical) definitions or, even worse, their definitions contradict. Therefore, it would be beneficial to have a unified framework and consistent definitions, e.g., for terms such as disparate effect, disparate impact, disproportionate impact, etc.; see, e.g., Chibanda [9]. Some of these terms are already occupied in a legal context.

- Adverse selection and unwanted economic consequences of non-discriminatory pricing should be explored, see e.g. Shimao–Huang [30]. Discrimination-free insurance prices typically fail to fulfill the auto-calibration property which is crucial for having homogeneous risk classes. However, the OT input pre-processed data allows for auto-calibrated regression models, for auto-calibration see Wüthrich–Merz [35].

- All considerations above have been based on the assumption that we know the true model. Clearly, in statistical modeling, there is model uncertainty which may impact different protected classes differently because, e.g., they are represented differently in historical data (statistical and historical biases). There are several examples of this type in the machine learning literature; see, e.g., Barocas et al. [4], Mehrabi et al. [24] and Pessach–Shmueli [26].

- Our considerations so far presented a black-and-white picture of direct/indirect discrimination or group unfairness either taking place or not. Nonetheless, especially in the context of a possible regulatory intervention, it is important to quantify the materiality of those potential problems within a given insurance portfolio. Such an approach requires the use of discrimination and unfairness metrics, pointing more towards formalizing notions like disproportional and disparate impact.

- We have been speaking about (non-)discrimination of insurance prices. These insurance prices are actuarial or statistical prices (technical premium), i.e., they directly result as an output from a statistical procedure. These prices are then modified to commercial prices, e.g., administrative costs are added, etc. An interesting issue is raised in Thomas [31, 32], namely, by converting actuarial prices into commercial prices one often distorts these prices with elasticity considerations, i.e., insurance companies charge higher prices to customers who are (implicitly) willing to pay more. This happens, e.g., with new business and contract renewals that are often priced differently, though the corresponding customers may have exactly the same risk profile – a situation that can also be understood as unfair, see FCA [16]. In the light of discrimination and fairness one should clearly question such practice of elasticity pricing as this leads to discrimination that cannot be explained by risk profiles (no matter whether we consider protected or non-protected information).

  Often, an actuarial pricing system $\boldsymbol{X} \mapsto \pi(\boldsymbol{X})$ is called *actuarially fair* if any price difference $\pi(\boldsymbol{X}_1) \neq \pi(\boldsymbol{X}_2)$ can be explained by differences in the distributions of (propensity to) claims $Y|_{\boldsymbol{X}_i}$, $i = 1, 2$. Price elasticity considerations are not actuarially fair.

- Given all the above arguments, in general we maintain that demographic fairness is not a reasonable requirement for insurance portfolios. Nonetheless a word of caution is needed. Consider the use of individualized data (e.g., wearables, telematics) for accurate quantification of the risk of insurance policies. Using such data may diminish the contribution of protected attributes to predictions, effectively leading to a lack of sensitivity of best-estimate prices in $\boldsymbol{D}$, see (2.5). Quite aside of concerns around surveillance and privacy, such individualized data may capture policyholder attributes (e.g., night-time driving) that are not just associated with, e.g., race, but are a constituent part of racialized experience within a particular society, not least because of historical constraints in employment or housing opportunities. In such situations, the non-protected covariates $\boldsymbol{X}$ become uncom-

fortably entangled with the protected attributes $\boldsymbol{D}$. For that reason, it still makes sense to monitor demographic unfairness within an insurance portfolio and to try to understand its sources. If the extent and source of group unfairness is considered problematic, OT input pre-processing becomes a valuable option for removing demographic disparities while, in a certain sense, still addressing indirect discrimination.

Acknowledgment.
M. Lindholm gratefully acknowledges financial support from the Länsförsäkringar Alliance [projectP9/20 "Machine learning methods in non-life insurance"].

References

[1] Agarwal, A., Dudik, M., Wu, Z.S. (2019). Fair regression: quantitative definitions and reduction-based algorithms. *arXiv*:1905.12843

[2] Araiza Iturria, C.A., Hardy, M., Marriott, P. (2022). A discrimination-free premium under a causal framework. *SSRN Manuscript* ID 4079068.

[3] Avraham, R., Logue, K. D. and Schwarcz, D.B. (2014). Understanding insurance anti-discrimination laws. *Southern California Law Review* **87(2)**, 195-274.

[4] Barocas, S., Hardt, M., Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities.* `https://fairmlbook.org/`

[5] Barrio, del E., Gamboa, F., Grodaliza, P., Loubes, J.-P. (2019). Obtaining fairness using optimal transport theory. In: Proceedings of the 36st International Conference on Machine Learning, Long Beach, California. *Proceedings of Machine Learning Reserach* **97**, 2357-2365.

[6] Buolamwini, J., Gebru, T., (2018). Gender shades: intersectional accuracy disparities in commercial gender classification. In: *Conference on Fairness, Accountability and Transparency*, Proceedings of Machine Learning Research **81**, 77-91.

[7] Charpentier, A. (2022). Insurance: Discrimination, Biases & Fairness. *Institut Louis Bachelier*, Opinions & Débates, No25 – July 2022.

[8] Chiappa, S., Jiang, R., Stepleton, T., Pacchiano, A., Jiang, H., Aslanides, J. (2020). A general approach to fairness with optimal transport. *Proceedings of the 34th AAAI Conference on Artificial Intelligence* **34(04)**, AAAI-20 Technical Tracks 4.

[9] Chibanda, K.F. (2021). Defining discrimination in insurance. *CAS Research Papers: A Special Series on Race and Insurance Pricing.* `https://www.casact.org/publications-research/research/research-paper-series-race-and-insurance-pricing`

[10] 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.

[11] Cuturi, M., Doucet, A. (2014). Fast computation of Wasserstein barycenters. In: Proceedings of the 31st International Conference on Machine Learning, Bejing, China. *Journal of Machine Learning Reserach* **32(2)**, 685-693.

[12] Djehiche, B., Löfdahl, B. (2016). Nonlinear reserving in life insurance: aggregation and mean-field approximation. *Insurance: Mathematics & Economics* **69**, 1-13.

[13] EIOPA (2021). Artificial intelligence governance principles: towards ethical and trustworthy artificial intelligence in the European insurance sector. A report from EIOPA's Consultative Expert Group on Digital Ethics in Insurance.

[14] European Commission (2012). Guidelines on the application of COUNCIL DIRECTIVE 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats). *Official Journal of the European Union* **C11**, 1-11.

[15] European Council (2004). COUNCIL DIRECTIVE 2004/113/EC - implementing the principle of equal treatment between men and women in the access to and supply of goods and services. *Official Journal of the European Union* **L 373**, 37-43.

[16] Financial Conduct Authority (2021). General insurance pricing practices market study: Feedback to CP20/19 and final rules. *Policy Statement PS21/5*.

[17] Frees, E.W.J., Huang, F. (2022). The discriminating (pricing) actuary. *North American Actuarial Journal*, in press.

[18] Grari, V., Charpentier, A., Lamprier, S., Detyniecki, M. (2022). A fair pricing model via adversarial learning. *arXiv*:2202.12008v2.

[19] Hardt, M., Price, E., Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 3315-3323.

[20] Kusner, M.J., Loftus, J., Russell, C. and Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, 4066-4076.

[21] Lindholm, M., Richman, R., Tsanakas, A., Wüthrich, M.V. (2022). Discrimination-free insurance pricing. *ASTIN Bulletin* **52(2)**, 55-89.

[22] Loader, C., Sun, J., Lucent Technologies, Liaw, A. (2022). locfit: local regression, likelihood and density estimation. `https://cran.r-project.org/web/packages/locfit/index.html`

[23] Maliszewska-Nienartowicz, J. (2014). Direct and indirect discrimination in European Union Law - How to draw a dividing line? *International Journal of Social Sciences* **III(1)**, 41-55.

[24] Mehrabi, N., Morstatter, F., Sexana, N., Lerman, K., Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv*:1908.09635v3.

[25] Pearl, J. (2009). *Causality. Models, Reasoning, and Inference*. 2nd edition. Cambridge University Press.

[26] Pessach, D., Erez Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Survey* **55(3)**, article 51.

[27] Prince, A.E.R., Schwarcz, D. (2020). Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review* **105(3)**, 1257-1318.

[28] Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., Goldberg, Y. (2020). Null it out: guarding protected attributes by iterative nullspace projection. *arXiv*:2004.07667.

[29] Ravfogel, S., Twinton, M., Goldberg, Y., Cotterell, R. (2022). Linear adversarial concept erasure. *arXiv*:2201.12091.

[30] Shimao, H., Huang F. (2022). Welfare cost of fair prediction and pricing in insurance market. *SSRN Manuscript* ID 4225159.

[31] Thomas, R.G. (2012). Non-risk price discrimination in insurance: market outcomes and public policy. *Geneva Papers on Risk and Insurance - Issues and Practice* **37**, 27-46.

[32] Thomas, R.G. (2022). Discussion on "The discriminating (pricing) actuary", by E.W.J. Frees and F. Huang. *North American Actuarial Journal*, in press.

[33] Vallance, C. (2021). Legal action over alleged Uber facial verification bias. *BBC News*. `https://www.bbc.co.uk/news/technology-58831373`; accessed 28/04/2023.

[34] Wüthrich, M.V., Merz, M. (2015). Stochastic claims reserving manual: advances in dynamic modeling. *SSRN Manuscript* ID 264905.

[35] Wüthrich, M.V., Merz, M. (2023). *Statistical Foundations of Actuarial Learning and its Applications.* Springer. `https://link.springer.com/book/10.1007/978-3-031-12409-9`

[36] Wüthrich, M.V., Ziegel, J. (2023). Isotonic recalibration under a low signal-to-noise ratio. *arXiv*:2301.02692.

[37] Xin, X., Huang, F. (2021). Anti-discrimination insurance pricing: regulations, fairness criteria, and models. *SSRN Manuscript* ID 3850420.

A Appendix: mathematical proofs

We prove the mathematical results in this appendix.

Proof of Proposition 2.10. We start with demographic parity (the independence axiom). Since the conditional distribution of $\mu(\boldsymbol{X}) = X$, given $\boldsymbol{D} = D$, explicitly depends on the realization of the protected attribute $D = d$ (we have a mixture Gaussian distribution for X), the independence axiom fails to hold, see also (2.10). Sufficiency (2.14) of $\mu(\boldsymbol{X})$ implies that

$$\mathrm{Var}\left(Y\,|\mu(\boldsymbol{X}), \boldsymbol{D}\right) = \mathrm{Var}\left(Y\,|\mu(\boldsymbol{X})\right). \tag{A.1}$$

We calculate the right hand side of (A.1)

$$\begin{aligned}
\mathrm{Var}\left(Y\,|\mu(\boldsymbol{X})\right) &= \mathrm{Var}\left(Y|\,X\right) \\
&= \mathrm{Var}\left(\mathbb{E}\left[Y|\,X, D\right]|\,X\right) + \mathbb{E}\left[\mathrm{Var}\left(Y|\,X, D\right)|\,X\right] \\
&= \mathrm{Var}\left(X|\,X\right) + \mathbb{E}\left[1 + D|\,X\right] \\
&= 1 + \frac{\exp\left\{-\frac{1}{2\tau^2}\left(X - x_1\right)^2\right\}}{\sum_{d\in\mathfrak{D}} \exp\left\{-\frac{1}{2\tau^2}\left(X - x_d\right)^2\right\}} \;\in\; (1,2), \qquad \text{a.s.},
\end{aligned}$$

where we have used (2.10). Next, we calculate the left hand side of (A.1)

$$\mathrm{Var}\left(Y\,|\mu(\boldsymbol{X}), \boldsymbol{D}\right) = \mathrm{Var}\left(Y|\,X, D\right) = 1 + D \;\in\; \{1,2\}, \qquad \text{a.s.}$$

Thus, these two conditional variances have a disjoint range, a.s., and we cannot have sufficiency of $\mu(\boldsymbol{X})$. Finally, there remains to prove the failure of the separation axiom. We aim at proving

$$\mathbb{E}\left[X|\,Y = x_d, D = d\right] \;\neq\; \mathbb{E}\left[X|\,Y = x_d\right], \tag{A.2}$$

for $\mu(\boldsymbol{X}) = X$. We start by analyzing the left hand side of (A.2). We have

$$X|_{D=d} \;\sim\; \mathcal{N}\left(x_d,\, \tau^2\right).$$

The joint density of $(Y, X)|_{D=d} \sim f^{(d)}_{Y,X}$ is given by

$$f^{(d)}_{Y,X}(y,x) \;=\; \frac{1}{\sqrt{2\pi(1+d)}} \exp\left\{-\frac{1}{2}\frac{(y-x)^2}{1+d}\right\} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{1}{2\tau^2}(x - x_d)^2\right\}.$$

This gives for the conditional density of X, given $(Y, D = d)$,

$$\begin{aligned}
f^{(d)}_{X|Y}(x|Y) &\propto \exp\left\{-\frac{1}{2}\frac{(Y-x)^2}{1+d}\right\} \exp\left\{-\frac{1}{2}\frac{(x-x_d)^2}{\tau^2}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{x^2 - 2xY}{1+d} + \frac{x^2 - 2xx_d}{\tau^2}\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{x^2(\tau^2 + 1 + d) - 2x\left(Y\tau^2 + x_d(1+d)\right)}{(1+d)\tau^2}\right)\right\}.
\end{aligned}$$

This is a Gaussian density, and we have

$$X|_{(Y,D=d)} \;\sim\; \mathcal{N}\left(\frac{Y\tau^2 + x_d(1+d)}{\tau^2 + 1 + d},\, \frac{(1+d)\tau^2}{\tau^2 + 1 + d}\right).$$

This implies for $Y = x_d$, for simplicity we set $d = 0$ but the same arguments hold for $d = 1$,

$$\mathbb{E}\left[X|\,Y = x_0, D = 0\right] = x_0.$$

On the other hand,

$$\begin{aligned}
\mathbb{E}\left[X|\,Y = x_0\right] &= \sum_{d=0,1} \mathbb{E}\left[X\,|Y = x_0, D = d\right] \mathbb{P}\left(D = d|\,Y = x_0\right) \\
&= x_0\,\mathbb{P}\left(D = 0|\,Y = x_0\right) + \frac{x_0\tau^2 + 2x_1}{\tau^2 + 2}\,\mathbb{P}\left(D = 1|\,Y = x_0\right) \\
&= x_0\left(1 - \mathbb{P}\left(D = 1|\,Y = x_0\right) + \frac{\tau^2 + 2\frac{x_1}{x_0}}{\tau^2 + 2}\,\mathbb{P}\left(D = 1|\,Y = x_0\right)\right) \;>\; x_0.
\end{aligned}$$

The latter inequality holds because by assumption $0 < x_0 < x_1$ and $\mathbb{P}(D = 1|Y = x) \in (0, 1)$ for all $x \in \mathbb{R}$. This proves (A.2) and that the separation axiom does not hold. □

Proof of Proposition 3.1. We can rewrite the discrimination-free insurance price as follows

$$\begin{aligned}\mu^*(\boldsymbol{X}) &= \sum_{\boldsymbol{d}\in\mathfrak{D}} \mu(\boldsymbol{X},\boldsymbol{d})\,\mathbb{P}^*(\boldsymbol{D}=\boldsymbol{d}) \;=\; \sum_{\boldsymbol{d}\in\mathfrak{D}} \int_y y\, d\mathbb{P}\left(y|\,\boldsymbol{X},\boldsymbol{d}\right)\mathbb{P}^*(\boldsymbol{D}=\boldsymbol{d}) \\ &= \int_y y\, d\mathbb{P}^\dagger\left(y|\,\boldsymbol{X}\right) \;=\; \mathbb{E}^\dagger\left[Y|\,\boldsymbol{X}\right],\end{aligned}$$

where we have defined the distribution (this breaks the dependence between $\boldsymbol{X}$ and $\boldsymbol{D}$)

$$\mathbb{P}^\dagger(Y,\boldsymbol{X},\boldsymbol{D}) := \mathbb{P}\left(Y|\,\boldsymbol{X},\boldsymbol{D}\right)\,\mathbb{P}(\boldsymbol{X})\,\mathbb{P}^*(\boldsymbol{D}).$$

Classical square loss minimization then provides us with

$$\begin{aligned}\mu^*(\boldsymbol{X}) &= \underset{\widehat{\mu}(\boldsymbol{X})}{\arg\min}\; \mathbb{E}^\dagger\left[\left(Y-\widehat{\mu}(\boldsymbol{X})\right)^2\big|\,\boldsymbol{X}\right] \\ &= \underset{\widehat{\mu}(\boldsymbol{X})}{\arg\min}\; \mathbb{E}\left[\frac{\mathbb{P}^*(\boldsymbol{D})}{\mathbb{P}\left(\boldsymbol{D}|\,\boldsymbol{X}\right)}\left(Y-\widehat{\mu}(\boldsymbol{X})\right)^2\Big|\,\boldsymbol{X}\right].\end{aligned}$$

This completes the proof. □