Discrimination-free insurance pricing

by M Lindholm, R Richman, A Tsanakas and MV Wüthrich

Presented at the Actuarial Society of South Africa's 2020 Virtual Convention
6–8 October 2020

ABSTRACT

We consider the following question: given information on individual policyholder characteristics, how can we ensure that insurance prices do not discriminate with respect to protected characteristics, such as gender? We address the issues of direct and indirect discrimination, the latter meaning that we can learn protected characteristics from non-protected ones. We provide rigorous mathematical definitions for direct and indirect discrimination, and we introduce a simple formula for discrimination-free pricing, that avoids both direct and indirect discrimination. Our formula works in any statistical model. We demonstrate its application on a health insurance example, using a state-of-the-art generalised linear model and a neural network regression model. An important conclusion is that discrimination-free pricing in general requires collection of policyholders' discriminatory characteristics, posing potential challenges in relation to policyholder's privacy concerns.

KEYWORDS

Discrimination, differentiation, insurance pricing, individual policy characteristics, discriminatory covariates, direct discrimination, indirect discrimination, neural networks, complex algorithmic models, causal inference, confounding

CONTACT DETAILS

Ronald Richman, QED Actuaries & Consultants; Email: Ronald.Richman@qedact.com
Mathias Lindholm, Dept of Mathematics, Stockholm University; Email: lindholm@math.su.se

Andreas Tsanakas, Faculty of Actuarial Science and Insurance, Cass Business School, City University London; Email: A.Tsanakas.1@city.ac.uk
Mario Wüthrich, RiskLab, Department of Mathematics, ETH Zurich
Email: mario.wuethrich@math.ethz.ch

1. INTRODUCTION

1.1 Motivation

We address the following fundamental question: given information on individual policyholder characteristics and past claims experience, how can we calculate insurance prices that do not discriminate with respect to protected characteristics, such as gender? This is a pertinent question in the context of anti-discrimination legislation; for instance, current European Union (EU) law requires gender neutral pricing, see [9]. The question has become even more relevant with the emergence of big data and associated developments in complex algorithmic models, since such models may be able to infer discriminatory characteristics from other policyholder features. For an overview on anti-discrimination laws we refer to Avraham et al. [2].

We aim at developing pricing formulas that are devoid of discrimination, while the insurer is still able to differentiate between policyholders with respect to non-discriminatory features. We assume that an insurer has access to policyholder data that can be split into discriminatory (e.g. gender, ethnicity) and non-discriminatory features (e.g. age, smoking habits). For the purposes of this paper we distinguish between *direct* and *indirect discrimination* by saying that:[1]

— A pricing model *avoids direct discrimination*, if none of the discriminatory features is used as a rating factor.

— A pricing model *avoids indirect discrimination*, if it avoids direct discrimination and, furthermore, the non-discriminatory features are used in a way that does not allow implicit inference of discriminatory features from them.

These concepts are illustrated in the following simple example.

Example 1 Assume we have observed a health insurance product and obtained the following claim counts $(n_{i,j})_{i,j=0,1}$ and claim exposures $(r_{i,j})_{i,j=0,1}$:

$n_{i,j}$	woman	man	row total
smoker	32	4	36
non-smoker	28	48	76
column total	60	52	112

$r_{i,j}$	woman	man	row total
smoker	133	24	157
non-smoker	131	301	432
column total	264	325	589

[1] The regulatory text [9] provides definitions of direct and indirect discrimination, which to an extent motivate our technical arguments. However, our definitions are drawn from an *actuarial* thinking and we do not make any claim about their correspondence to the *legal* definitions of those terms.

where $i=1$ corresponds to "smoker" and $j=1$ corresponds to "woman". Based on the above contingency tables we estimate the claim frequencies $\lambda_{i,j}$ by the empirical frequency $\hat{\lambda}_{i,j}=n_{i,j}/r_{i,j}$. Assume now that gender is considered a discriminatory characteristic. In order to avoid direct discrimination, its explicit influence on the calculated insurance price needs to be removed. The standard way of doing this is to consider the aggregated estimators (row sums) $\hat{\lambda}_{i,\bullet}=n_{i,\bullet}/r_{i,\bullet}=(n_{i,0}+n_{i,1})/(r_{i,0}+r_{i,1})$. This approach produces, e.g. for smokers,

$$\hat{\lambda}_{1,\bullet}=\frac{36}{157}=0.229.$$

The estimate $\hat{\lambda}_{1,\bullet}$ (and a premium for smokers based on it), thus, can be calculated by completely ignoring policyholders' gender information. But one can note that an alternative representation of $\hat{\lambda}_{1,\bullet}$ is

$$\begin{aligned}\hat{\lambda}_{1,\bullet}&=\hat{\lambda}_{1,1}\frac{r_{1,1}}{r_{1,1}+r_{1,0}}+\hat{\lambda}_{1,0}\frac{r_{1,0}}{r_{1,1}+r_{1,0}}\\&=\hat{\lambda}_{1,1}\hat{\mathbb{P}}(\text{woman}\,|\,\text{smoker})+\hat{\lambda}_{1,0}\hat{\mathbb{P}}(\text{man}\,|\,\text{smoker}),\end{aligned}$$

where $\hat{\mathbb{P}}$ refers to the empirical distribution obtained from the data. Hence, the estimate $\hat{\lambda}_{1,\bullet}$ not only contains information about the influence of smoking on producing a claim, but also, via $\hat{\mathbb{P}}(\text{gender}\,|\,\text{smoking habits})$, about the propensity of smokers to be female or male. In our case, because smoking habits substantially differ between genders (a smoker is a woman with probability $133/157=85\%$, whereas a non-smoker is a woman with probability $131/432=30\%$), it is indeed the case that the above approach implicitly infers the more likely gender from smoking habits. Therefore, indirect discrimination is present; we come back to this in Example 8 below.

Example 1 illustrates that avoiding direct discrimination does not necessarily entail avoiding indirect discrimination. Consequently, just ignoring discriminatory features in the calculation of insurance prices does not necessarily yield discrimination-free prices. Hence, unawareness (or willful ignorance) of discriminatory features is not a solution to the problem of calculating discrimination-free insurance prices.

1.2 Our contributions

First, we formally define direct and indirect discrimination. We are not aware of such a definition in the literature. The ideas and principles we develop are relevant to all situations where predictors correspond to conditionally expected values; hence they are applicable in all fields where discrimination is an important issue, e.g. also in customer credit rating. Second, we give a rigorous probabilistic account of discrimination-free prices and their existence. We propose a simple pricing formula that avoids both direct and indirect discrimination. While the formula only uses non-discriminatory features as rating factors, it introduces an adjustment, which requires knowledge of policyholders' discriminatory features. Third, we justify discrimination-free prices using tools of causal inference. Fourth, we identify bias in aggregate portfolio prices as an unintended consequence of discrimination-free prices. While prices that can be written as conditional expectations under the physical probability measure naturally lead to an unbiased pricing system on a portfolio level, discrimination-free prices do not generally have this property. Therefore, we propose methods for bias corrections. Fifth, we illustrate how the discrimination-free prices can be calculated in practice, using either machine learning algorithms or standard statistical methods like generalised linear models (GLMs).

1.3 Literature review

Although an issue of key relevance for insurance pricing, relatively little attention has been paid to the issue of discrimination-free pricing within the actuarial literature. In a discussion of the implications of EU gender legislation, it is suggested in Guillén [13] that covariates highly correlated with gender can be used as proxies by insurance companies, which from our perspective results in indirect discrimination. Focusing on the case of mortality pricing, Chen and Vigna [5] criticise the industry practice of deriving unisex life tables by mixing the life tables for each gender on the grounds that this does not respect the principles of actuarial fairness, which is to say that the total unisex premiums charged for the portfolio are not equal to the total premiums charged using gender specific life tables. They provide alternative approaches without this shortcoming; note that our proposed discrimination-free price will reproduce the pricing formulas of Chen and Vigna [5]. The implications of unisex pricing on insurer capital requirements in the context of Solvency II are examined in Chen et al. [4], and an ALM approach to unisex pricing is taken in Bruszas et al. [3], where the concept of "gender mix risk" also is discussed. Market implications of unisex tariffs are discussed in Sass and Seifried [21], see also De Jong and Ferris [7] for a discussion of adverse selection stemming from restrictions on risk classification.

The issue of indirect discrimination occurring when using the above-mentioned approach of just ignoring discriminatory covariates, has been discussed in Pope and Sydnor [20] and Kusner et al. [16]. The procedure for discrimination-free pricing provided in Pope and Sydnor [20] is the same as our proposal; this pricing rule is applied in the context of auto insurance pricing by Aseervatham et al. [1]. However these authors do not provide a probabilistic justification for the prices used nor do they address the critical issue of a potential bias at portfolio level (and associated corrections).

We are aware of relatively few examples of causal inference applied within an insurance context. For renewals of insurance policies, some insurers seek to estimate policyholder demand elasticity by randomly varying renewal prices for a subset of policyholders (i.e. a form of a randomised controlled trial is conducted) and estimating the impact on the probability of renewal. Once the demand elasticities have been estimated, a profit maximising pricing policy can be established in a practice referred to as price optimisation, see e.g. Krikler et al. [14]. Within that context, Guelman and Guillén [12] apply methods from causal inference to estimate demand elasticity functions from observational data collected by an insurer.

We emphasise that the issues discussed in this paper apply to many other industries;

we refer to e.g. Fuster et al. [11] where a credit rating application is considered. Their study focuses on evaluating the differential impact of prediction technologies on ethnic groups, rather than on a mathematical definition of discrimination.

1.4 Organisation of the paper

In Section 2 we discuss different kinds of insurance prices, comprising the *best-estimate price*, which considers all available information, the *unawareness price*, which avoids direct discrimination, and the *discrimination-free price*, which avoids both direct and indirect discrimination. In particular, Subsection 2.3 gives mathematical definitions of direct and indirect discrimination, which are based on a change of probability measure. Special cases of discrimination-free prices can be interpreted in terms of causal inference, which offers tools to adjust for confounding covariates; this is discussed in Section 3. The bias that discrimination-free prices can induce at portfolio level is discussed in Section 4, along with proposals for bias correction. In Section 5 we describe the calculation of discrimination-free prices based on models estimated from data. This is explored in more detail in Section 6, where a numerical example is given, based on a synthetic health insurance portfolio. Concluding remarks are collected in Section 7.

2. DISCRIMINATION-FREE PRICING

2.1 Definition of discrimination-free prices

We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying probability space with physical probability measure $\mathbb{P}$. For a given portfolio of insurance policies, let $\mathbf{D}$ denote the vector of *discriminatory covariates* (characteristics, features, explanatory variables) of policyholders, and let $\mathbf{X}$ denote the vector of *non-discriminatory covariates*. We assume that $\mathbf{X}$ and $\mathbf{D}$ are random vectors on $(\Omega, \mathcal{F}, \mathbb{P})$; the randomness of these covariate vectors represents variations between policyholders within a given portfolio. A realisation of $(\mathbf{X}, \mathbf{D})$ corresponds to choosing an insurance policy at random from the portfolio; a policyholder profile with specific characteristics is obtained by conditioning on $\mathbf{X} = \mathbf{x}$, $\mathbf{D} = \mathbf{d}$. For simplicity, we denote the marginal and conditional distributions of covariates under $\mathbb{P}$ by $\mathbf{X} \sim \mathbb{P}(\mathbf{x})$, $\mathbf{D} \sim \mathbb{P}(\mathbf{d})$ and $(\mathbf{D} \,|\, \mathbf{X} = \mathbf{x}) \sim \mathbb{P}(\mathbf{d} \,|\, \mathbf{x})$, respectively.

A policyholder claim is denoted by the random variable Y. The claim Y typically depends on (but is not fully determined by) both the discriminatory covariates $\mathbf{D}$ and the non-discriminatory ones $\mathbf{X}$. Our aim is to price such a claim Y, with the resulting price being free from direct as well as indirect discrimination, according to the arguments of Section 1. A technical definition of these concepts will be given in Section 2.3, below.

In the sequel, it will be useful to assume $Y, \mathbf{X}, \mathbf{D} \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. This assumption is not crucial for defining discrimination-free prices, but it will allow us to give more intuitive interpretations in terms of orthogonal projections and minimal distances. Our notion of price will be based on conditional expectations of Y, when conditioning on different subsets of covariates. We first introduce a number of different prices that are important for the subsequent discussions and derivations.

Definition 2 (best-estimate price) *The* best-estimate price *for Y w.r.t.* $(\mathbf{X},\mathbf{D})$ *is defined by*

$$\mu(\mathbf{X},\mathbf{D}) := \mathbb{E}[Y \,|\, \mathbf{X},\mathbf{D}].$$

Remark 3

(a) We call the price $\mu(\mathbf{X},\mathbf{D})$ "best-estimate" because it minimises the $\mathcal{L}^2$-distance of all $(\mathbf{X},\mathbf{D})$-measurable prices to Y, i.e. $\mu(\mathbf{X},\mathbf{D})$ is the orthogonal projection of Y onto the sub-space generated by $(\mathbf{X},\mathbf{D})$.

(b) In general, the best-estimate price is not discrimination-free, unless we are in the special case of $\mu(\mathbf{X},\mathbf{D}) = \mu(\mathbf{X})$.

(c) The best-estimate price is *unbiased* w.r.t. Y, that is,

$$\mu := \mathbb{E}[Y] = \mathbb{E}[\mu(\mathbf{X},\mathbf{D})];$$

by the tower property for conditional expectations. Unbiasedness is important because it indicates that best-estimate prices achieve on average the correct price level for the portfolio.

As discussed in Example 1, the initial attempt at achieving discrimination-free prices arises through simply ignoring discriminatory covariates **D**.

Definition 4 (unawareness price) *The* unawareness price *for Y w.r.t.* **X** *is defined by*

$$\mu(\mathbf{X}) := \mathbb{E}[Y \,|\, \mathbf{X}]. \tag{1}$$

Remark 5

(a) As the price $\mu(\mathbf{X})$ does not depend explicitly on **D**, it avoids direct discrimination. However, in general, the unawareness price will produce indirect discrimination, as was discussed in Example 1; see also Kusner et al. [16]. Specifically, we can write the unawareness price as

$$\mu(\mathbf{X}) = \int_{\mathbf{d}} \mu(\mathbf{X},\mathbf{d})\ d\mathbb{P}(\mathbf{d} \,|\, \mathbf{X}). \tag{2}$$

Indirect discrimination arises because the conditional probability $d\mathbb{P}(\mathbf{d} \,|\, \mathbf{X})$ enables inference of discriminatory covariates **D** from non-discriminatory ones **X**. Such discrimination is avoided in the special case when **D** and **X** are independent, since then it holds that $d\mathbb{P}(\mathbf{d} \,|\, \mathbf{X}) = d\mathbb{P}(\mathbf{d})$.

(b) The price $\mu(\mathbf{X})$ minimises the $\mathcal{L}^2$-distance to Y based solely on X, i.e. it is the best price w.r.t. information X. At the same time, the price $\mu(\mathbf{X})$ also minimises the $\mathcal{L}^2$-distance to $\mu(\mathbf{X},\mathbf{D})$, by a simple application of the Pythagorean theorem. Note that

$$\left\|\mu(\mathbf{X})-\mu(\mathbf{X},\mathbf{D})\right\|_2^2=\mathbb{E}[Var(\mu(\mathbf{X},\mathbf{D})\,|\,\mathbf{X})],$$

which intuitively should decrease with increasing dependence between $\mathbf{X}$ and $\mathbf{D}$. Hence, the quality in the approximation of $\mu(\mathbf{X},\mathbf{D})$ using $\mu(\mathbf{X})$ should be good if $\mathbf{X}$ essentially is a function of $\mathbf{D}$, i.e., if the non-discriminatory covariates $\mathbf{X}$ allow us to almost perfectly infer the discriminatory covariates $\mathbf{D}$.

(c) The unawareness price is unbiased, since

$$\mu=\mathbb{E}[Y]=\mathbb{E}[\mu(\mathbf{X})].$$

We now propose a price that is free of both direct and indirect discrimination.

Definition 6 (discrimination-free price) *A* discrimination-free *price for Y w.r.t.* $\mathbf{X}$ *is defined by*

$$h^*(\mathbf{X}):=\int_{\mathbf{d}}\mu(\mathbf{X},\mathbf{d})\;\mathrm{d}\mathbb{P}^*(\mathbf{d}),\tag{3}$$

where the distribution $\mathbb{P}^*(\mathbf{d})$ *is defined on the same range as the marginal distribution of the discriminatory variables* $\mathbf{D}\sim\mathbb{P}(\mathbf{d})$.

Remark 7

(a) The discrimination-free price (3) is obtained by averaging best-estimate prices over discriminatory covariates, using a (potentially arbitrary) marginal distribution $\mathbb{P}^*(\mathbf{d})$. The crucial step here is the imposed marginalisation w.r.t. $\mathbf{D}$, rather than the specific choice of $\mathbb{P}^*(\mathbf{d})$ (which can be $\mathbb{P}^*(\mathbf{d})=\mathbb{P}(\mathbf{d})$). Given that the price $h^*(\mathbf{X})$ does not explicitly depend on $\mathbf{D}$, it is obviously free from direct discrimination. We argue that the averaging construction proposed in (3) also removes indirect discrimination. While (3) appears similar to (2), there is a key difference: *discrimination-free prices do not in any way depend on the conditional distribution* $\mathbb{P}(\mathbf{d}|\mathbf{X})$ – hence they do not enable to infer the discriminatory covariates from the non-discriminatory ones. This will be further discussed in Section 2.3 and verified in the case study of Section 6. In the special case of $\mathbf{X}$ and $\mathbf{D}$ being independent, it follows that $h^*(\mathbf{X})=\mu(\mathbf{X})$.

(b) Definition 6 is designed to *remove* the possible explanatory power that $\mathbf{X}$ may have for $\mathbf{D}$; it does *not* assume independence between $\mathbf{X}$ and $\mathbf{D}$ in the given portfolio. This point will be made more precise in Section 2.3, and in Section 2.4 where we discuss the existence of discrimination-free prices as well as alternative interpretations of $h^*(\mathbf{X})$.

(c) Definition 6 can also be motivated by arguments from causal inference. Specifically, formulas like (3) are used to quantify the direct causal effect of $\mathbf{X}$ on Y. Our distribution-free price corresponds to the (expected value version) of the so-called "back-door" adjustment formula in causal inference settings, when seeing $\mathbf{D}$ as a potential confounder of $\mathbf{X}$; these ideas will be discussed in more detail in Section 3. Furthermore, formula (3) using the choice $\mathbb{P}^*(\mathbf{d}) = \mathbb{P}(\mathbf{d})$ corresponds precisely to the Partial Dependence Plot (PDP) introduced by Friedman [10], see also Zhao and Hastie [22]. We stress that although causal inference can in many situations serve as an alternative motivation of discrimination-free prices, the reasoning behind our Definition 6 *does not rely on any causal assumptions*. For further discussions of this, see Section 3.

(d) Prices obtained using (3) will in general *not* be unbiased, since

$$\mu = \mathbb{E}[Y] \neq \mathbb{E}\left[h^*(\mathbf{X})\right] = \int_{\mathbf{x},\mathbf{d}} \mu(\mathbf{x},\mathbf{d})\ \mathrm{d}\mathbb{P}^*(\mathbf{d})\mathrm{d}\mathbb{P}(\mathbf{x}), \tag{4}$$

even for the special choice $\mathbb{P}^*(\mathbf{d}) = \mathbb{P}(\mathbf{d})$. For that reason, the need emerges for portfolio level price corrections, which will be discussed in Section 4.

(e) Note that, given the potential arbitrariness of $\mathbb{P}^*$, evaluation of discrimination-free prices only requires knowledge of the mapping $(\mathbf{x},\mathbf{d}) \mapsto \mu(\mathbf{x},\mathbf{d})$, where $\mu(\mathbf{x},\mathbf{d})$ may be an (algorithmically derived implicit) regression function. Nevertheless, as pointed out in the previous remark, if one aims to correct a potential bias of $h^*(\mathbf{X})$, it is necessary to consider (estimate under) the "real-world" probability measure $\mathbb{P}$.

(f) Given the construction (3), $\mathbb{P}^*(\mathbf{d})$ may be inferred from comparing best-estimate prices $\mu(\mathbf{X},\mathbf{D})$ and observed discrimination-free prices $h^*(\mathbf{X})$.

2.2 Choice of weighting distributions for discriminatory covariates

From Definition 6 it follows that the distribution $\mathbb{P}^*(\mathbf{d})$ can be chosen rather freely. A simple choice is $\mathbb{P}^*(\mathbf{d}) = \mathbb{P}(\mathbf{d})$, that is, average in (3) w.r.t. the marginal distribution of the discriminatory characteristics in the portfolio. This choice is supported by causal inference arguments in Section 3. We denote this special case by:

$$h(\mathbf{X}) := \int_{\mathbf{d}} \mu(\mathbf{X},\mathbf{d})\ \mathrm{d}\mathbb{P}(\mathbf{d}). \tag{5}$$

We illustrate how $h(\mathbf{X})$ is evaluated in the context of Example 1.

Example 8 In Example 1 we argued that aggregated estimators (row sums) $\hat{\lambda}_{i,\bullet}$ are discriminatory because gender can be inferred from smoking habits. The price $h(\mathbf{X})$ removes this effect by replacing the conditional probability $\mathbb{P}$ (gender | smoking habits) by $\mathbb{P}$ (gender). This implies that the frequency estimate for smokers $\hat{\lambda}_{1,\bullet}$ is replaced by

$$\begin{aligned}\tilde{\lambda}_{1,\bullet} &= \hat{\lambda}_{1,1}\hat{\mathbb{P}}(\text{woman}) + \hat{\lambda}_{1,0}\hat{\mathbb{P}}(\text{man}) \\ &= \frac{32}{133}\cdot\frac{264}{589} + \frac{4}{24}\cdot\frac{325}{589} \\ &= 0.200 < 0.229 = \hat{\lambda}_{1,\bullet}.\end{aligned} \tag{6}$$

Similarly, for non-smokers

$$\tilde{\lambda}_{0,\bullet} = \hat{\lambda}_{0,1}\hat{\mathbb{P}}(\text{woman}) + \hat{\lambda}_{0,0}\hat{\mathbb{P}}(\text{man}) = 0.184. \tag{7}$$

We demonstrate the potential portfolio bias that discrimination-free prices induce. The total cost of the portfolio, under best-estimate prices, is equal to the observed total claim of 112. For discrimination-free prices, the total cost is given by

$$\tilde{\lambda}_{1,\bullet}(r_{1,1} + r_{1,0}) + \tilde{\lambda}_{0,\bullet}(r_{0,1} + r_{0,0}) = 110.77 < 112.$$

This indicates that the discrimination-free price $h(\mathbf{X})$ leads to an under-pricing of the overall portfolio in the present situation.

Recall that there is some flexibility in the selection of $\mathbb{P}^*(\mathbf{d})$. In this simple example, with $\mathbf{D}$ being a binary classification, we can in fact choose $\mathbb{P}^*(\text{woman})$ and $\mathbb{P}^*(\text{man})$ in a way that eliminates the portfolio bias. Specifically, we can set

$$\tilde{\lambda}^*_{i,\bullet} = \hat{\lambda}_{i,1}\mathbb{P}^*(\text{woman}) + \hat{\lambda}_{i,0}\mathbb{P}^*(\text{man}), \quad \text{for } i = 0,1,$$

and require for the resulting overall portfolio price that it holds

$$\tilde{\lambda}^*_{1,\bullet}(r_{1,1} + r_{1,0}) + \tilde{\lambda}^*_{0,\bullet}(r_{0,1} + r_{0,0}) = 112.$$

The resulting choice is $\mathbb{P}^*(\text{woman}) = 48.3\% > 44.8\% = \mathbb{P}(\text{woman})$.

Finally, we note that in this example, switching to discrimination-free prices leads to a reduction in the share of the portfolio costs covered by women. Women cause $60/112 = 53.6\%$ of the total costs which is exactly the share of the total costs that women have to pay under best-estimate pricing (assuming that the prices coincide with the claims caused). If we use the unawareness price by simply dropping the gender variable, women cover 47.8% of the total costs. If we charge the discrimination-free price (6)–(7), women cover 45.7% of all costs, thus, less than under the unawareness price. This exactly reflects indirect discrimination in the unawareness price: women have on average higher costs than men, and the allocation of these excess costs is bigger to the sub-population where women are more prevalent compared to the population distribution $\mathbb{P}(\mathbf{d})$, i.e. we learn $\mathbf{D}$ from $\mathbf{X}$.

Furthermore, it is useful to consider the extrema of discrimination-free prices. Consider the following distribution-free prices

$$h^{(+)}(\mathbf{X}) := \sup_{\mathbb{P}^*} \int_{\mathbf{d}} \mu(\mathbf{X},\mathbf{d}) \, d\mathbb{P}^*(\mathbf{d}),$$

$$h^{(-)}(\mathbf{X}) := \inf_{\mathbb{P}^*} \int_{\mathbf{d}} \mu(\mathbf{X},\mathbf{d}) \, d\mathbb{P}^*(\mathbf{d}).$$

$h^{(+)}(\mathbf{X})$ and $h^{(-)}(\mathbf{X})$ correspond to the essential supremum and infimum over $\mathbf{d}$ in the range of **D**, respectively. Thus, for non-discriminatory covariates $\mathbf{X}=\mathbf{x}$, this immediately gives us

$$h^{(-)}(\mathbf{x}) \leq h^*(\mathbf{x}), h(\mathbf{x}), \mu(\mathbf{x}) \leq h^{(+)}(\mathbf{x}).$$

Moreover, for the bias property we get the following relationship

$$\int_{\mathbf{x}} h^{(-)}(\mathbf{x})\, d\mathbb{P}(\mathbf{x}) \leq \mathbb{E}\left[h^*(\mathbf{X})\right], \mu \leq \int_{\mathbf{x}} h^{(+)}(\mathbf{x})\, d\mathbb{P}(\mathbf{x}).$$

By definition $h^{(+)}(\mathbf{x})$ corresponds to the "worst" (or most "prudent") price, and has been discussed in the context of unisex pricing in Chen and Vigna [5].

As seen in Example 8, the distribution-free price (3) is generally biased. An alternative possibility for the choice of $\mathbb{P}^*(\mathbf{d})$ is to additionally require unbiasedness in (4). In the simple case of a binary discriminatory covariate like gender in Example 8, this reduced to choosing a suitable $\mathbb{P}^*(\text{woman})$. A more general construction of unbiased prices via choices of $\mathbb{P}^*(\mathbf{d})$ is presented in Section 4.

A special case corresponds to an additive best-estimate price, in the sense that $\mu(\mathbf{X},\mathbf{D}) = \mu_1(\mathbf{X}) + \mu_2(\mathbf{D})$. Then, the simple choice $\mathbb{P}^*(\mathbf{d}) = \mathbb{P}(\mathbf{d})$ is appealing, as it provides an unbiased price. Note that

$$h(\mathbf{X}) = \int_{\mathbf{d}} \mu_1(\mathbf{X})\; d\mathbb{P}(\mathbf{d}) + \int_{\mathbf{d}} \mu_2(\mathbf{d})\, d\mathbb{P}(\mathbf{d}) = \mu_1(\mathbf{X}) + \mathbb{E}\left[\mu_2(\mathbf{D})\right],$$

which implies

$$\mathbb{E}\left[h(\mathbf{X})\right] = \mathbb{E}\left[\mu_1(\mathbf{X})\right] + \mathbb{E}\left[\mu_2(\mathbf{D})\right] = \mathbb{E}\left[\mu(\mathbf{X},\mathbf{D})\right] = \mu.$$

2.3 Revisiting direct and indirect discrimination

In this section, following the development of our ideas so far, we provide more technical definitions of prices that avoid direct and indirect discrimination.

Choose an arbitrary probability measure $\mathbb{P}^*$ on the measurable space $(\Omega, \mathcal{F})$ such that $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}^*)$. Choose a (sub-)vector Z of the covariates $(\mathbf{X},\mathbf{D})$ and define the $(\mathbb{P}^*, \mathbf{Z})$-conditional-expectation price by

$$\mu^*(\mathbf{Z}) := \mathbb{E}^*[Y \,|\, \mathbf{Z}],$$

where $\mathbb{E}^*$ denotes the expectation under $\mathbb{P}^*$.

Definition 9 *A price avoids* direct discrimination*, if it can be written as*

$$\mu^*(\mathbf{Z}) = \mathbb{E}^*[Y \,|\, \mathbf{Z}],$$

where **Z** *is* $\sigma(\mathbf{X})$*-measurable, and where the expectation is taken w.r.t. a probability measure* $\mathbb{P}^*$ *on* $(\Omega, \mathcal{F})$ *such that* $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}^*)$.

Remark 10

(a) Definition 9 says that a price avoids direct discrimination if it can be written as a measurable function of the non-discriminatory covariates $\mathbf{X}$. For $\mathbf{Z}=\mathbf{X}$ we receive maximal use of non-discriminatory information (relative to $\mathbb{P}^*$), therefore, we typically work with $\mathbf{Z}=\mathbf{X}$.

(b) The choice $\mathbb{P}^*=\mathbb{P}$ (and $\mathbf{Z}=\mathbf{X}$) provides the unawareness price $\mu(\mathbf{X})$ of Definition 4 which, thus, avoids direct discrimination.

(c) Importantly, under the choice $\mathbb{P}^*=\mathbb{P}$, the unawareness price $\mu(\mathbf{X})$ can be calculated without explicit knowledge of $\mu(\mathbf{X},\mathbf{D})$ – hence it does *not* require collection of discriminatory policyholder information. This also applies if we need to estimate $\mu(\mathbf{X})$ from data, see (17) below.

Now, indirect discrimination can be defined.

Definition 11 *A price* $\mu^*(\boldsymbol{Z})$ *that avoids direct discrimination is said to avoid* indirect discrimination *if* $\mathbf{Z}$ *and* $\mathbf{D}$ *are independent under* $\mathbb{P}^*$.

Independence under $\mathbb{P}^*$ effects the decoupling of discriminatory covariates from non-discriminatory ones, for specific policyholders. Thus, according to Definition 11, a price that avoids indirect discrimination satisfies

$$\mu^*(\mathbf{Z})=\int_{\mathbf{d}}\mu^*(\mathbf{Z},\mathbf{d})\,\mathrm{d}\mathbb{P}^*(\mathbf{d}\,|\,\mathbf{Z})=\int_{\mathbf{d}}\mu^*(\mathbf{Z},\mathbf{d})\,\mathrm{d}\mathbb{P}^*(\mathbf{d}), \tag{8}$$

where $\mu^*(\mathbf{Z},\mathbf{d})=\mathbb{E}^*[Y\,|\,\mathbf{Z},\mathbf{D}=\mathbf{d}]$.

Remark 12

(a) The independence in Definition 11 is an artifice of the introduced probability measure $\mathbb{P}^*$ under which insurance is priced and does *not* generally reflect the actual observed dependence between $\mathbf{X}$ and $\mathbf{D}$.

(b) For $\mathbf{Z}=\mathbf{X}$, the calculation that avoids indirect discrimination is based on the knowledge of $\mu^*(\mathbf{X},\mathbf{D})$, see (8) – hence it *requires* collection of discriminatory policyholder information. In fact, one of the most critical problems in practice is that discriminatory information is often incomplete, e.g. about ethnicity, which typically results in indirect discrimination.

(c) In statistical applications we usually use the conditional probability measure $\mathbb{P}(y\,|\,\mathbf{X},\mathbf{D})$ to model a claim Y, given the covariates $(\mathbf{X},\mathbf{D})$. The reason for this choice is that Y, given $(\mathbf{X},\mathbf{D})$, is observed under the real world measure $\mathbb{P}$, which allows for direct estimation of the regression function, see Section 5 below,

$$(\mathbf{x},\mathbf{d})\mapsto\mu(\mathbf{x},\mathbf{d}).$$

We could choose the measure $\mathbb{P}^*$ in a way that preserves the (causal) structure of how the covariates impact the response, that is, let $\mathbb{P}^*(y\,|\,\mathbf{X},\mathbf{D})=\mathbb{P}(y\,|\,\mathbf{X},\mathbf{D})$. This then motivates the choice

$$\mathbb{P}^{*}\left(y,\mathbf{x},\mathbf{d}\right)=\mathbb{P}(y\,|\,\mathbf{X}=\mathbf{x},\mathbf{D}=\mathbf{d})\ \mathbb{P}^{*}\left(\mathbf{x}\right)\ \mathbb{P}^{*}\left(\mathbf{d}\right),$$

for $\mathbf{Z}=\mathbf{X}$ in Definition 11. In view of (8), this results in the discrimination-free price

$$\mu^{*}\left(\mathbf{X}\right)=\int_{\mathbf{d}}\mu\left(\mathbf{X},\mathbf{d}\right)\ d\mathbb{P}^{*}(\mathbf{d}\,|\,\mathbf{X})=\int_{\mathbf{d}}\mu\left(\mathbf{X},\mathbf{d}\right)\ d\mathbb{P}^{*}\left(\mathbf{d}\right)=h^{*}\left(\mathbf{X}\right).$$

Thus, the discrimination-free price of Definition 6 does neither allow for direct nor for indirect discrimination.

(d) Linking to Remark 7(e), in practice, we need to know (calibrate under) the real world measure $\mathbb{P}$ in order to study unbiasedness w.r.t. $\mu=\mathbb{E}\left[Y\right]$. Since the actual portfolio that we hold is described by $\mathbf{Z}\sim\mathbb{P}\left(\mathbf{z}\right)$, we need to average discrimination-free prices $\mu^{*}\left(\mathbf{Z}\right)$ w.r.t. the same population $\mathbb{P}\left(\mathbf{z}\right)$ to see whether we receive unbiasedness of discrimination-free prices on the actual portfolio.

2.4 Existence of discrimination-free prices

We discuss the existence of discrimination-free prices according to Definition 6 and the possibility of avoiding indirect discrimination according to Definition 11 in this section.

We emphasise that properties of available data (and the related statistical models) play a crucial role in our considerations:

— Indirect discrimination may be the result of incomplete discriminatory information, see Remark 12(b).

— Indirect discrimination may be the result of nonexistent or insufficient information of certain parts of the population.

In this section we discuss the second item which can enter in different ways. First, not all parts of the population are equally well represented in the development of the statistical model. For instance, there is research in image recognition to discover malignant melanoma (skin cancer). If this research is mainly based on images of people with light skin, the corresponding model will likely fail to discover malignant melanoma for people with dark skin. This form of discrimination results from *insufficient data* of certain parts of the population. In our situation, this may result in poor best-estimate prices $\mu\left(\mathbf{X},\mathbf{D}\right)$ for certain covariate combinations. Note that the quality of the estimation of best-estimate prices directly impacts discrimination-free prices.

In the current section we rather focus on *nonexistent data* of certain parts of the population. The meaning and implications of nonexistent data are going to be discussed in more detail. We start with an example. Assume that the discriminatory covariates $\mathbf{D}$ correspond to gender and the non-discriminatory ones $\mathbf{X}$ to education. Education could be in the ordinal form “secondary school degree”, “high school degree” or “university degree”, but information about education could also be received in the following categorical form “Catholic college degree”, “public college degree” or “girls college degree”. Per definition the last label “girls college degree” contains as only gender “female”. This implies that

$$\mathbb{P}\left(\mathbf{X} = \text{girls college degree}, \mathbf{D} = \text{man}\right) = 0,$$

thus, the event $A = \left\{\mathbf{X} = \text{girls college degree}, \mathbf{D} = \text{man}\right\} \in \mathcal{F}$ is a null set w.r.t. $\mathbb{P}$. In many cases, we do not model responses Y on null sets. Therefore, neither Y nor A may be specified in our model nor the conditional expectation $\mu(\text{girls college degree,man}) = \mathbb{E}[Y \mid A]$ may be determined. But this implies that we cannot evaluate the discrimination-free price

$$h^*\left(\mathbf{X}\right) = \int_{\mathbf{d}} \mu\left(\mathbf{X},\mathbf{d}\right) \, d\mathbb{P}^*\left(\mathbf{d}\right),$$

if $\mathbb{P}^*\left(\mathbf{d}\right)$ has positive probability mass on both genders. In the current situation, the problem may be solved by setting $\mathbb{P}^*\left(\mathbf{D} = \text{woman}\right) = 1$ which gives the discrimination-free price $h^*\left(\mathbf{X}\right) = \mu\left(\mathbf{X}, \text{woman}\right)$.

If the education information **X** has an additional level "boys college degree", the above solution will not work because we have a second $\mathbb{P}$-null set $B = \{\text{X} = \text{boys college degree,D} = \text{woman}\} \in \mathcal{F}$ which makes it impossible to choose a distribution $\mathbb{P}^*\left(\mathbf{d}\right)$ such that the discrimination-free price $h^*\left(\mathbf{X}\right)$ is well-defined.

The simple solution to this problem is to drop the education information, that is, choose a smaller covariate set. This is equivalent to choosing a true subset **Z** of **X** in Definition 11. In practice, we often try to inter- or extrapolate the model assumptions for Y. This is reasonable if unavailable information corresponds to numerical variables (and responses have some continuity in these covariates). In certain cases it may also be justified for categorical variables by, for example, postulating a multiplicative influence structure of covariates, say, women are x% better than men regardless of the college attended. This is similar to a GLM approach where gender may be reflected by a single parameter on the canonical scale. In our situation such an assumption can be made, but it cannot be verified because of a missing control group.

Proposition 13 *Assume there exists a product measure* $\mathbb{P}^*\left(\mathbf{x}\right)\mathbb{P}^*\left(\mathbf{d}\right)$ *on* $\left(\Omega,\mathcal{F}\right)$ *which is absolutely continuous w.r.t. the probability measure* $\mathbb{P}\left(\mathbf{x},\mathbf{d}\right)$ *of the covariates* $\left(\mathbf{X},\mathbf{D}\right)$*. Then, there exists a price* $\mu^*\left(\mathbf{X}\right)$ *that avoids indirect discrimination.*

Proof. Absolute continuity implies that every $\mathbb{P}\left(\mathbf{x},\mathbf{d}\right)$-null set is also a $\mathbb{P}^*\left(\mathbf{x}\right)\mathbb{P}^*\left(\mathbf{d}\right)$-null set. Therefore, $\mu\left(\mathbf{X},\mathbf{D}\right)$ is well-defined on all sets where $\left(\mathbf{X},\mathbf{D}\right)$ has positive $\mathbb{P}^*\left(\mathbf{x}\right)\mathbb{P}^*\left(\mathbf{d}\right)$-probability mass. Since the latter is a product measure we can calculate the discrimination-free price $h^*\left(\mathbf{X}\right)$ by integrating $\mu\left(\mathbf{X},\mathbf{d}\right)$ over $d\mathbb{P}^*(\mathbf{d} \mid \mathbf{X}) = d\mathbb{P}^*\left(\mathbf{d}\right)$, see also (8). This completes the proof.

3. CAUSAL INFERENCE AND DISCRIMINATION

This section offers further motivation for the discrimination-free price of Definition 6 in a causal inference setting. We try to give these arguments in a pedagogical (and somewhat informal) way. For a rigorous treatment we refer to Pearl [18] and Pearl et al. [19, Ch.3.1].

The starting point of causal inference is a hypothesis of a potential covariate relationship which may be described in terms of a directed graph $\mathfrak{G}$. The directed graph $\mathfrak{G}$ consists of a set of *nodes* corresponding to the characteristics, including the response Y, and *directed edges* – "arrows" – indicating directions of potential influence between the characteristics (including the response Y). This informal definition can be made precise, but it is most easily understood by the example given in Figure 1 (lhs).

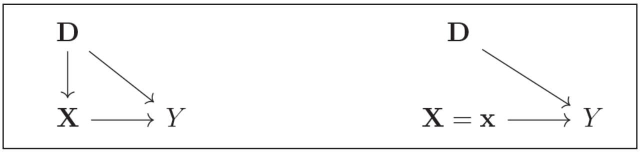

FIGURE 1 (lhs) Causal diagram described by $\mathfrak{G}$
(rhs) causal diagram altered according to the do-operation do($\mathbf{X}=\mathbf{x}$)

The directed graph $\mathfrak{G}$ in Figure 1 (lhs) is an example of a directed acyclic graph (DAG), where acyclic means that when following the direction of the edges, the graph does not contain any loops. For a precise description, see Pearl et al. [19, Ch.1.4]. The graphical representation given in Figure 1 (lhs) corresponds to a situation where the discriminatory characteristics $\mathbf{D}$ may influence both directly on Y, but also indirectly via $\mathbf{X}$. Hence, Figure 1 (lhs) tells us that $\mathbf{D}$ is a *confounder* w.r.t. the effect of $\mathbf{X}$ on Y.

Figure 1 (lhs) already captures a large number of realistic insurance pricing situations. For instance, in view of Example 1, we may identify smoking habits by $\mathbf{X}$ and the gender by the discriminatory factors $\mathbf{D}$. Differences in smoking habits between men and women can be expressed by a directed edge $\mathbf{D} \to \mathbf{X}$. There are intrinsic differences between men and women when it comes to health outcomes (e.g. women cannot get prostate cancer), this is described by $\mathbf{D} \to Y$. Moreover, smoking in itself may cause health problems $\mathbf{X} \to Y$. This situation is exactly covered by the DAG $\mathfrak{G}$ in Figure 1 (lhs).

We remark that the subsequent causality arguments also hold true for more general causal diagrams, for instance, if we have additional (unmeasured) confounding characteristics $\mathbf{U}$ that enter the causal diagram. To keep this discussion short, we do not describe these more general situations but refer to Pearl et al. [19,] and Lauritzen [17, Ch.~3.2.2].

Since the directed edges in the DAG $\mathfrak{G}$ do not act fully deterministically, and to make probabilistic statements, we need to endow the DAG $\mathfrak{G}$ with a probability measure $\mathbb{P}$ that describes the randomness involved. This probability measure $\mathbb{P}$ should be Markovian on $\mathfrak{G}$ which, colloquially speaking, means that all nodes in Figure 1 (lhs) are complemented with independent noisy background variables, see Pearl et al. [18, Ch.3.2.1]; for noisy proxy variables we also refer to Kuroki and Pearl [15].

With this setup in place, one way to approach non-discriminatory pricing is to require the following:

> Given that a policyholder has the set of characteristics $\mathbf{X}=\mathbf{x}$, what is the expected value of Y, after removing the possible confounding effects of discriminatory covariates $\mathbf{D}$?

The idea of the causal (non-confounded) effect[2] of X on Y can be captured precisely by a modification of the DAG $\mathfrak{G}$ given in Figure 1. Specifically, as illustrated on the rhs of Figure 1, remove all directed edges to X and set the value of X to x. The removal of the directed edge $\mathbf{D} \to \mathbf{X}$ allows us to consider only the (direct) causal effect of X = x on Y. Note that this operation is different to conditioning. When conditioning on X = x, the distribution of D is generally affected; but in the modified graph on the rhs of Figure 1, changes in x do not influence D. This is precisely the desired effect of removing the ability of inferring discriminatory covariates from non-discriminatory ones, as was discussed in Remark 12(a) in Section 2. The intervention of removing all directed edges to X and of fixing X = x is denoted by the so-called do-operator do(X = x), see Pearl et al. [19, Ch.3.2.1]. Henceforth, a price that takes into account only the causal (non-confounded) effect of X on Y can be defined by

$$\mathbb{E}[Y \,|\, \mathrm{do}(\mathbf{X} = \mathbf{x})], \tag{9}$$

where the probability $\mathbb{P}(Y \,|\, \mathrm{do}(\mathbf{X} = \mathbf{x}))$ still needs to be defined.

The do-operation do(X = x) going from the lhs to the rhs in the DAG $\mathfrak{G}$ of Figure 1 fulfils the so-called back-door criterion relative to the ordered pair $(\mathbf{X}, Y)$, see Definition 3.3.1 in Pearl et al. [19]. This means that the do-operation do(X = x) blocks every path between X and Y that contains an arrow into X. Under these properties Theorem 3.3.2 (back-door adjustment) of Pearl et al. [18] states that the causal effect of X on Y is identifiable in the Markovian DAG $(\mathfrak{G}, \mathbb{P})$ and given by the formula

$$\mathbb{P}(Y \,|\, \mathrm{do}(\mathbf{X} = \mathbf{x})) = \int_{\mathbf{d}} \mathbb{P}(Y \,|\, \mathbf{X} = \mathbf{x}, \mathbf{D} = \mathbf{d}) \; d\mathbb{P}(\mathbf{d}). \tag{10}$$

The following proposition is an easy consequence of definition (9), the definition of the best-estimate price $\mu(\mathbf{X}, \mathbf{D})$ and formula (10).

Proposition 14 *Assume that the do-operation* do(X = x) *fulfils the back-door criterion relative to the ordered pair* $(\mathbf{X}, Y)$ *in the Markovian DAG* $(\mathfrak{G}, \mathbb{P})$. *It holds that*

$$\mathbb{E}[Y \,|\, \mathrm{do}(\mathbf{X} = \mathbf{x})] = \int_{\mathbf{d}} \mu(\mathbf{x}, \mathbf{d}) \; d\mathbb{P}(\mathbf{d}) = h(\mathbf{x}),$$

where $h(\mathbf{x})$ *is defined by* (5).

Remark 15

(a) Proposition 14 justifies the discrimination-free price $h(\mathbf{X})$ of equation (5) under specific Markovian DAG assumptions, motivating the choice $\mathbb{P}^*(\mathbf{d}) = \mathbb{P}(\mathbf{d})$ in

[2] For our intended application, the interpretation of "causal" will here be restricted to the sub-population which chooses to buy insurance cover. No explicit consideration will be taken w.r.t. which characteristics make individuals belong to this sub-population. However, it is often reasonable to assume that the underlying causal relations should be the same for the general population and the insured sub-population, although the effects are not the same.

Definition 6. In particular, this discrimination-free price is obtained by adjusting for possible confounding of $\mathbf{D}$ w.r.t. $\mathbf{X}$, thus, only measuring the direct causal effect of $\mathbf{X}$ on Y. While violating the assumptions underlying Proposition 14 will remove the causal interpretation of distribution-free prices, these assumptions are *not* needed in order for $h(\mathbf{X})$ to produce proper discrimination-free prices, in the spirit of Section 2.3.

(b) Proposition 14 does not say anything about the quality of the induced price which, of course, will depend on the specific model being used – it merely states that there is no confounding, given that the assumptions of the corollary are fulfilled.

(c) It is possible to extend the covariate relations described by Figure 1 to more general situations, for instance, by including unmeasured characteristics $\mathbf{U}$, see e.g. Pearl [18].

4. ATTRIBUTION OF TOTAL PORTFOLIO PREMIUM TO INDIVIDUAL POLICIES

The difficulty that we still have to deal with is that, in general, a discrimination-free price has a bias, see (4) and Example 8. This bias needs to be corrected because otherwise the premium for the entire portfolio may not be at the appropriate level. There is no canonical way of correcting for this potential bias; moreover, the requirement that the bias correction should be discrimination-free excludes complex cost allocation mechanisms.

The portfolio bias of the $\mathbb{P}^*$-discrimination-free price is defined by

$$B^* := \mu - \mathbb{E}\left[h^*(\mathbf{X})\right] = \mathbb{E}[Y] - \int_{\mathbf{x}\,\mathbf{d}} \mu(\mathbf{x},\mathbf{d})\ d\mathbb{P}^*(\mathbf{d})d\mathbb{P}(\mathbf{x}).$$

Simple bias corrections arise from taking rather different positions. An egalitarian position is taken by distributing the portfolio bias B^* uniformly across the entire portfolio, regardless of any non-discriminatory covariates $\mathbf{X}$. This motivates the *uniformly adjusted* $\mathbb{P}^*$*-discrimination-free price* defined by

$$\pi^{*,u}(\mathbf{X}) := h^*(\mathbf{X}) + B^*. \tag{11}$$

Moreover, if we do not consider any covariates (neither discriminatory nor non-discriminatory ones) we are back in the situation of a homogeneous situation where we charge the same (constant) premium μ to every policyholder. A drawback of the uniformly adjusted price (11) is that it may result in negative prices for certain covariate values $\mathbf{X}$.

A different position is to allocate the bias B^* by differentiating w.r.t. $\mathbf{X}$ in a still discrimination-free fashion (avoiding any inference of $\mathbf{D}$ from $\mathbf{X}$). A natural way is to allocate the total premium proportionally to $h^*(\mathbf{X})$, resulting in the *proportionally adjusted* $\mathbb{P}^*$*-discrimination-free price*

$$\pi^{*,p}(\mathbf{X}) := h^*(\mathbf{X}) \frac{\mu}{\mu - B^*}. \tag{12}$$

In the remainder of this section we discuss a more sophisticated approach which chooses the distribution $\mathbb{P}^*(\mathbf{d})$ specifically such that the discrimination-free price $h^*(\mathbf{X})$ is unbiased, i.e. $B^* = 0$. A simple illustration was given in Example 8. In general, there will be many such distributions that may satisfy this condition, and an additional criterion for choosing $\mathbb{P}^*(\mathbf{d})$ is needed. A standard criterion is to chose the measure $\mathbb{P}^*$ as close as possible to the physical distribution $\mathbb{P}(\mathbf{d})$, subject to the unbiasedness constraint on $h^*(\mathbf{X})$. If the relative entropy (Kullback–Leibler divergence) is chosen as the divergence measure, then standard results can be applied, see Csiszár [6]. Specifically, note that the unbiasedness condition $\mathbb{E}\left[h^*(\mathbf{X})\right] = \mu$ is equivalent to

$$\mathbb{E}^*\left[\zeta(\mathbf{D})\right] = \mu, \tag{13}$$

where we define

$$\zeta(\mathbf{t}) := \mathbb{E}\left[\mu(\mathbf{X},\mathbf{t})\right].$$

Note, we assume the existence of distributions $\mathbb{P}^*(\mathbf{d})$ that fulfil (13) which in view of Section 2.4 needs to be made. The distribution $\mathbb{P}^*(\mathbf{d})$ that minimises the relative entropy with respect to $\mathbb{P}(\mathbf{d})$ – $I(\mathbb{P}^*(\mathbf{d})|| \, \mathbb{P}(\mathbf{d}))$ in the notation of Csiszár [6] – subject to (13), is given by

$$\mathbb{P}^*(\mathbf{d}) = \mathbb{E}\left[1_{\{\mathbf{D} \leq \mathbf{d}\}} \frac{e^{\beta\zeta(\mathbf{D})}}{\mathbb{E}\left[e^{\beta\zeta(\mathbf{D})}\right]}\right],$$

for a suitably chosen parameter β. Hence, the premium for a policyholder with non-discriminatory covariates $\mathbf{X} = \mathbf{x}$ is defined by (subject to existence)

$$\pi^{*,KL}(\mathbf{x}) := h^*(\mathbf{x}) = \mathbb{E}\left[\mu(\mathbf{x},\mathbf{D}) \frac{e^{\beta\zeta(\mathbf{D})}}{\mathbb{E}\left[e^{\beta\zeta(\mathbf{D})}\right]}\right].$$

To ease the interpretation of this formula, let $\mathbf{D} = D$ be one-dimensional and $\mu(\mathbf{x},d) \geq 0$ be increasing in d. Then, for $\beta > 0$, we have

$$\begin{aligned}
\pi^{*,KL}(\mathbf{x}) &= \mathbb{E}\left[\mu(\mathbf{x},D) \frac{e^{\beta\zeta(D)}}{\mathbb{E}\left[e^{\beta\zeta(D)}\right]}\right] \\
&= \mathbb{E}\left[\mu(\mathbf{x},D)\right] + \mathbb{C}\mathrm{ov}\left[\mu(\mathbf{x},D), \frac{e^{\beta\zeta(D)}}{\mathbb{E}\left[e^{\beta\zeta(D)}\right]}\right] \\
&\geq \mathbb{E}\left[\mu(\mathbf{x},D)\right] = h(\mathbf{x}),
\end{aligned}$$

which corresponds to the situation where the choice $\mathbb{P}^* = \mathbb{P}$ would produce a negative bias (under-pricing). The calculation of $\pi^{*,KL}(\mathbf{x})$ assigns a higher premium to policyholders with covariates X = x such that $\mu(\mathbf{x},D)$ is more volatile (this can be made rigorous but is best visible in approximation (14), below). This represents policies for which lack of information on discriminatory covariates matters more, in the sense that there is a higher sensitivity to the uncertainty induced by not using the discriminatory factor D. One can thus view the bias correction in $\pi^{*,KL}(\mathbf{x})$ as an implicit discrimination-free) risk load.

For small β we have approximation

$$h^{*}(\mathbf{x}) \approx \mathbb{E}\left[\mu(\mathbf{x},\mathbf{D})\right] + \beta\mathbb{C}\mathrm{ov}\left[\mu(\mathbf{x},\mathbf{D}),\zeta(\mathbf{D})\right] \tag{14}$$

$$= \mathbb{E}\left[\mu(\mathbf{x},\mathbf{D})\right] + \beta\sqrt{\mathbb{V}\mathrm{ar}\left[\mu(\mathbf{x},\mathbf{D})\right]\mathbb{V}\mathrm{ar}\left[\zeta(\mathbf{D})\right]}\ \mathbb{C}\mathrm{orr}\left[\mu(\mathbf{x},\mathbf{D}),\zeta(\mathbf{D})\right].$$

5. ESTIMATED MODEL

All previous discussion and derivations of discrimination-free prices and indirect discrimination were conducted under the assumption that the "true" probabilistic model underlying the portfolio $(Y,\mathbf{X},\mathbf{D})$ is known, represented by the physical measure $\mathbb{P}$. In practice, an *estimated model* is used because, typically, the data generating mechanism is unknown.

Specifically, one starts from data

$$\mathcal{S} = \left\{(y_1,\mathbf{x}_1,\mathbf{d}_1),\ldots,(y_n,\mathbf{x}_n,\mathbf{d}_n)\right\},$$

assuming that $(y_i,\mathbf{x}_i,\mathbf{d}_i)$ are i.i.d. realisations of $(Y,\mathbf{X},\mathbf{D}) \sim \mathbb{P}$. Subsequently, a regression model (in the broader sense of regression models) is chosen

$$\hat{\mu} : (\mathbf{x},\mathbf{d}) \mapsto \hat{\mu}(\mathbf{x},\mathbf{d}) = \hat{\mu}(\mathbf{x},\mathbf{d};\boldsymbol{\theta}), \tag{15}$$

which typically differs from the (true) best-estimate price functional $(\mathbf{x},\mathbf{d}) \mapsto \mu(\mathbf{x},\mathbf{d})$, given in Definition 2, but which should mimic $\mu(\mathbf{x},\mathbf{d})$ in the best possible way. One may specify a fixed functional form for $\hat{\mu}$ in (15) or, in a wider sense, one can specify an algorithm that generates the mapping (15) from the data $\mathcal{S}$. In either case, $\hat{\mu}$ will still depend on unknown parameters $\boldsymbol{\theta}$ that have to be estimated from the data $\mathcal{S}$ (using a given objective function) yielding estimate $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathcal{S})$.

The resulting $\mathcal{S}$-calibrated regression function

$$(\mathbf{x},\mathbf{d}) \mapsto \hat{\mu}\left(\mathbf{x},\mathbf{d};\hat{\boldsymbol{\theta}}\right) \tag{16}$$

then provides the approximation to the best-estimate price functional $(\mathbf{x},\mathbf{d}) \mapsto \mu(\mathbf{x},\mathbf{d})$. Note that (16) provides an estimate of the best price and, obviously, this estimate is in general discriminatory because it explicitly considers the discriminatory covariate values $\mathbf{d}$. Moreover, since we use the data $\mathcal{S}$ which has been generated under the physical measure $\mathbb{P}$, the regression function (16) also needs to be understood under the physical measure $\mathbb{P}$, we refer to Remark 12(c).

The unawareness price functional $\mathbf{x} \mapsto \mu(\mathbf{x})$ can be approximated in an analogous manner by just dropping $\mathbf{d}$ in (15) and (16), resulting in an estimated regression function

$$\mathbf{x} \mapsto \bar{\mu}\left(\mathbf{x};\hat{\vartheta}\right), \tag{17}$$

where the functional forms $\hat{\mu}$ and $\bar{\mu}$ may differ as well as their parameters $\boldsymbol{\theta}$ and ϑ, respectively. We emphasise that typically $\bar{\mu}\left(\cdot;\hat{\vartheta}\right)$ will indirectly discriminate w.r.t. **d** because in the estimation process of $\hat{\vartheta}$ we implicitly use covariate combinations $\left(\mathbf{x}_i,\mathbf{d}_i\right)$ which (empirically) contain the dependencies $\mathbb{P}(\mathbf{d}\,|\,\mathbf{x})$ that allow inference of **D** from **X**. The estimated unawareness price $\bar{\mu}\left(\mathbf{x};\hat{\vartheta}\right)$ can also be interpreted as an approximation to

$$\mathbb{E}[\hat{\mu}\left(\mathbf{X},\mathbf{D};\hat{\boldsymbol{\theta}}\right)|\,\mathbf{X}=\mathbf{x};\mathcal{S}],$$

using the tower property argument (under the physical measure $\mathbb{P}$).

Typically, also $\mathbb{P}(\mathbf{d})$ is not known. Assuming **D** is discrete, $\mathbb{P}(\mathbf{d})$ can be estimated by the empirical probabilities $n_{\mathbf{d}}/n$ (observed relative frequency of the discriminatory covariate **d** in $\mathcal{S}$). This generates the discrimination-free price

$$\hat{h}(\mathbf{x})=\sum_{\mathbf{d}}\hat{\mu}\left(\mathbf{x},\mathbf{d};\hat{\boldsymbol{\theta}}\right)\frac{n_{\mathbf{d}}}{n}, \tag{18}$$

where we use the estimated best-estimate price functional (16). The price (18) is discrimination-free in the sense of Definition 6, i.e. the discrimination-free property is not affected by the fact that we work with an estimated model. While potential estimation error may result in prices $\hat{h}(\mathbf{x})$ that are not very close to $h(\mathbf{x})$, the property of non-discrimination is preserved within the selected model; we explore this in more detail in Section 6. When choosing the structure of the regression function $\hat{\mu}$ in (15), we should require existence of the discrimination-free price (18) in the sense of Proposition 13.

6. NUMERICAL ILLUSTRATION

6.1 Model and alternative pricing rules

We present a simple health insurance example, demonstrating our proposed approach to discrimination-free pricing. The example we present satisfies the causal relations of Figure 1, such that discrimination-free prices can be understood as reflecting direct (unconfounded) causal effects (in an insured sub-population).

Let $\mathbf{D}=D$ correspond to the single discriminatory characteristic "gender", that is, $D\in\{\text{woman},\text{man}\}$. Furthermore, let $\mathbf{X}=\left(X_1,X_2\right)'$, where $X_1\in\{15,\ldots,80\}$ denotes the age of the policyholder, and $X_2\in\{\text{non}-\text{smoker},\text{smoker}\}$; below we assume that smoking habits are gender related. We consider three different types of health costs: birthing related health costs only affecting women between ages 20 and 40 (type 1), cancer related health costs with a higher frequency for smokers and also for women (type 2), and health costs due to other disabilities (type 3). For simplicity, we only consider claim counts, assuming deterministic claim costs for the three different claim types. Moreover, we model all individuals as independent, having the same exposure (= 1). We assume that the claim counts for the different claim types are described by independent Poisson GLMs with canonical (i.e.

log-) link function. The three different types of claims are governed by the following log-frequencies

$$\log\lambda_1(\mathbf{X},D) := \alpha_0 + \alpha_1 1_{\{X_1\in[20,40]\}} 1_{\{D=\text{woman}\}}, \tag{19}$$

$$\log\lambda_2(\mathbf{X},D) := \beta_0 + \beta_1 X_1 + \beta_2 1_{\{X_2=\text{smoker}\}} + \beta_3 1_{\{D=\text{woman}\}}, \tag{20}$$

$$\log\lambda_3(\mathbf{X},D) := \gamma_0 + \gamma_1 X_1, \tag{21}$$

based on the joint non-discriminatory and discriminatory covariates $(\mathbf{X},D)$. The deterministic claims costs of the different claim types are given by $(c_1,c_2,c_3)=(0.5,0.9,0.1)$ for claims of type 1, type 2, and type 3, respectively.

The best-estimate price (considering all covariates) of Definition 2 is given by

$$\mu(\mathbf{X},D) = c_1\lambda_1(\mathbf{X},D) + c_2\lambda_2(\mathbf{X},D) + c_3\lambda_3(\mathbf{X},D).$$

This best-estimate price is illustrated in Figure 2 for the parameter values $(\alpha_0,\alpha_1)=(-40,38.5)$, $(\beta_0,\beta_1,\beta_2,\beta_3)=(-2,0.004,0.1,0.2)$, and $(\gamma_0,\gamma_1)=(-2,0.01)$. The plots on the lhs of Figure 2 refer to smokers $(X_2=\text{smoker})$, while those on the rhs to non-smokers $(X_2=\text{non-smoker})$. The solid black lines give the best-estimate prices $\mu(\mathbf{X},D)$ for women and the solid red lines for men. Obviously, these best-estimate prices discriminate between genders.

Next, we calculate the discrimination-free price defined in Definition 6 for $\mathbb{P}^*(d)=\mathbb{P}(d)$, see (5), motivated by Proposition 14. It is given by

$$h(\mathbf{X}) = \sum_{d\in\{\text{woman},\text{man}\}} \big(c_1\lambda_1(\mathbf{X},d) + c_2\lambda_2(\mathbf{X},d) + c_3\lambda_3(\mathbf{X},d)\big)\,\mathbb{P}(D=d).$$

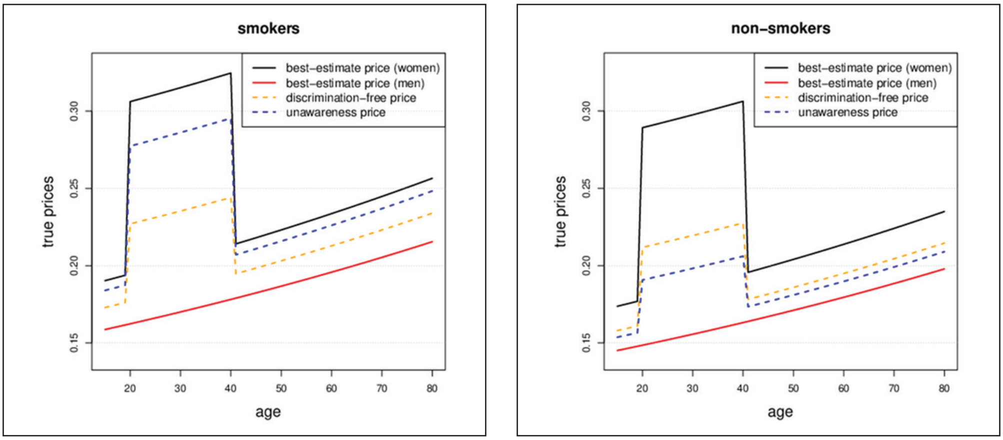

FIGURE 2 True model: (lhs) smokers and (rhs) non-smokers with solid black and red lines giving the best-estimate prices for women and men, respectively. The dotted orange lines show the discrimination-free prices and the dotted blue lines show the unawareness prices.

For the calculation of this discrimination-free price we need the gender proportions within our population. We set $\mathbb{P}(D = \text{woman}) = 0.45$. The orange dotted lines in Figure 2 provide the resulting discrimination-free prices for smokers and non-smokers. Note that these are identical for men and women, i.e. all price differences can be described solely by different ages X_1 and smoking habits X_2, irrespective of gender D. Moreover, the smoking habits do not allow us to infer the gender; note that in the exposition so far, it has not been necessary to describe how smoking habits vary by gender.

We compare this discrimination-free price to the unawareness price obtained by simply dropping the gender covariate D from the calculations (Definition 4). Thus, we calculate

$$\mu(\mathbf{X}) = c_1\mathbb{E}\left[\lambda_1(\mathbf{X},D)\middle|\mathbf{X}\right] + c_2\mathbb{E}\left[\lambda_2(\mathbf{X},D)\middle|\mathbf{X}\right] + c_3\mathbb{E}[\lambda_3(\mathbf{X},D)\,|\,\mathbf{X}]$$

The unawareness price requires *additional information* about the following conditional probabilities

$$\mathbb{P}(D = d\,|\,\mathbf{X} = \mathbf{x}) = \frac{\mathbb{P}(D = d, \mathbf{X} = \mathbf{x})}{\mathbb{P}(\mathbf{X} = \mathbf{x})} = \frac{\mathbb{P}(D = d, X_2 = x_2)}{\mathbb{P}(X_2 = x_2)}, \tag{22}$$

the last equality following from assuming that the age variable X_1 is independent from the random vector (X_2, D). In addition, we set $\mathbb{P}(D = \text{woman} \mid X_2 = \text{smoker}) = 0.8$ and $\mathbb{P}(X_2 = \text{smoker}) = 0.3$. The former assumption tells us that smokers are more likely to be women; this is similar to Example 1. As a consequence, X_2 has explanatory power to predict the gender D, and the unawareness price will therefore be indirectly discriminatory against women. These unawareness prices are illustrated by the blue dotted lines in Figure 2. The blue dotted line lies above the discrimination-free price (orange) for smokers (Figure 2, lhs) and below for non-smokers (rhs). Thus, the unawareness price implicitly allocates a higher price to women because smokers are more likely to be women.

The latter indirect gender discrimination is easily verified by an alternative assumption, namely, that smokers are more likely to be men, say, $\mathbb{P}(D = \text{woman} \mid X_2 = \text{smoker}) = 0.2$. The resulting unawareness prices are plotted by the dotted green lines in Figure 3. We observe that the smokers are below the discrimination-free price (orange dotted line), and for non-smokers we have the opposite sign. That is, in this case women are again indirectly discriminated through their (non-)smoking habits, again serving as a proxy for the explanatory variable of gender. The break-even point is $\mathbb{P}(D = \text{woman} \mid X_2 = \text{smoker}) = 0.45 = \mathbb{P}(D = \text{woman})$ because in this case D and X_2 are independent, which prevents indirect discrimination, and the unawareness price and the discrimination-free price are equal.

6.2 Application on estimated models

The previous discussion has been based on the knowledge of the model generating the data. We now address the more realistic situation where the model needs to be estimated. To this effect, we simulate data from $(\mathbf{X}, D, Y) \sim \mathbb{P}$ consistently with the given model

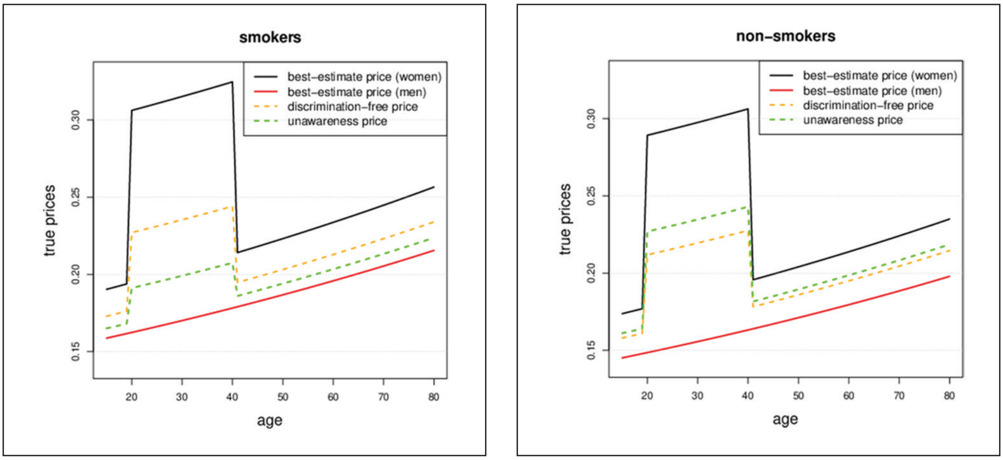

FIGURE 3 True model: (lhs) smokers and (rhs) non-smokers with solid black and red lines giving the best-estimate prices for women and men, respectively. The dotted orange lines show the discrimination-free prices and with dotted green lines show the unawareness prices, for an alternative assumption on $\mathbb{P}(D = \text{woman} \,|\, \text{smoker})$.

assumptions, and subsequently calibrate a neural network regression model to the simulated data.

Specifically, we choose a health insurance portfolio of size $n = 100{,}000$, and simulate claim counts from the Poisson GLMs (19), (20), and (21), with the choice $\mathbb{P}(D = \text{woman} \,|\, X_2 = \text{smoker}) = 0.8$. An age distribution for X_1 is also needed for the simulation – the chosen probability weights are shown in Figure 4. We assume that age X_1 is independent from gender D and smoking habits X_2, as in (22).

Listing 1 gives an excerpt of the simulated data. We have the three covariates X_1 (age), X_2 (smoking habit) and D (gender) on lines 5–7, and lines 2–4 illustrate the number of

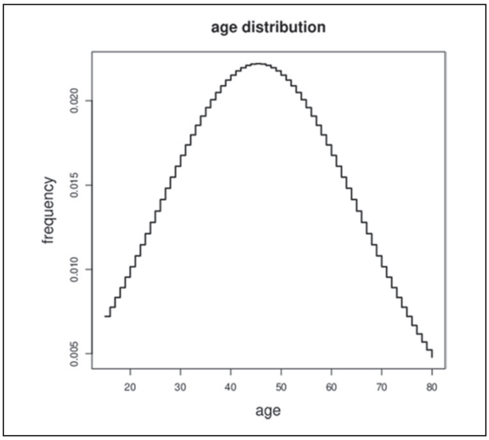

FIGURE 4 The age frequency used for both genders and smoking habits to simulate the data

claims N_1, N_2 and N_3, separated by claim types. The proportion of women in this simulated data is 0.4505 which is close to the true value of $\mathbb{P}(D = \text{woman}) = 0.45$. Our first aim is to fit a regression model to this data, under the assumptions that individual policies are independent, and that the different claim types are independent and Poisson distributed. Beside this, we do not make any structural assumption about the regression functions, but we try to infer them from the data using neural networks. The independence assumption between the claim counts N_1, N_2 and N_3 motivates modelling them separately. Thus, we will fit three different neural networks to model λ_1, λ_2 and λ_3, respectively. As we do not use prior knowledge on the data generating process, we will feed all covariates (X_1, X_2, D) to each of the three networks.

```
’data.frame’:    100000 obs. of  6 variables:
 $ N1: int  0 0 0 0 0 0 0 0 0 0 ...
 $ N2: int  0 0 1 0 0 1 0 0 2 0 ...
 $ N3: int  0 1 0 0 1 0 0 0 0 0 ...
 $ X1: num  36 57 70 49 63 27 41 58 16 34 ...
 $ X2: num  0 0 1 0 0 1 0 0 1 1 ...
 $ D : num  0 1 1 0 0 1 0 0 1 1 ...
```

Listing 1 Simulated health insurance data

```
Design  <- layer_input(shape = c(3), dtype = ’float32’, name = ’Design’)
#
Network = Design %>%
      layer_dense(units=15, activation=’relu’, name=’hidden1’) %>%
      layer_dense(units=15, activation=’relu’, name=’hidden2’) %>%
      layer_dense(units=1, activation=’exponential’, name=’Network’)
#
model <- keras_model(inputs = c(Design), outputs = c(Network))
model %>% compile(loss = ’poisson’, optimizer = ’adam’)
```

Listing 2 Neural network architecture used to infer λ_1, λ_2 and λ_3

Listing 2 illustrates the chosen neural network architecture, using the R library keras, with which the three regression functions (19)–(21) are estimated. We choose neural networks of depth 2 having 15 neurons in both hidden layers, the rectified linear unit (ReLU) activation function, and the canonical link under the Poisson assumption. Moreover, we select the Poisson deviance loss as our objective function. This network involves 316 weights that need to be calibrated. We train these weights of the three networks over 1000 epochs on batches of size 20,000.

Figure 5 illustrates the estimates $\hat{\lambda}_1(\mathbf{X}, D)$, $\hat{\lambda}_2(\mathbf{X}, D)$ and $\hat{\lambda}_3(\mathbf{X}, D)$ of the three regression functions (19), (20) and (21) respectively, obtained by fitting the three neural networks. The lhs of that figure gives claim type 1 which is birthing related. We see a rather accurate shape, with smoking habits correctly ignored, and men not affected by these claims. Figure 5 (middle) gives the cancer related frequencies. Also here we receive the same order w.r.t. gender and smoking habits as in (20). Finally, the rhs illustrates all remaining claims. As,

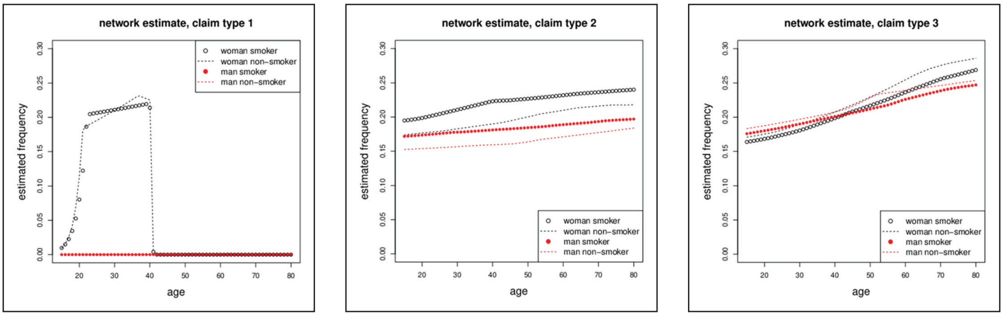

FIGURE 5 Estimated regression functions $\hat{\lambda}_1(\mathbf{X},D)$ (lhs), $\hat{\lambda}_2(\mathbf{X},D)$ (middle) and $\hat{\lambda}_3(\mathbf{X},D)$ (rhs) using the neural network architecture of Listing 2

by (21) claims frequencies should not depend on gender and smoking habits, the variation between lines indicates that the regression model captures a spurious effect.

Using these estimated frequencies, we calculate the estimated best-estimate price, (16),

$$\hat{\mu}\left(\mathbf{X},D;\hat{\boldsymbol{\theta}}\right)=c_1\hat{\lambda}_1(\mathbf{X},D)+c_2\hat{\lambda}_2(\mathbf{X},D)+c_3\hat{\lambda}_3(\mathbf{X},D),$$

and its discrimination-free counterpart (18),

$$\hat{h}(\mathbf{x})=\sum_d \hat{\mu}\left(\mathbf{x},d;\hat{\boldsymbol{\theta}}\right)\frac{n_d}{n},$$

with empirical proportions $n_{\text{woman}}/n=1-n_{\text{man}}/n=0.4505$. These prices are illustrated in Figure 6: black lines give best-estimate prices for women, red lines for men, and with the orange dotted lines showing the discrimination-free counterparts. Comparing Figures 2

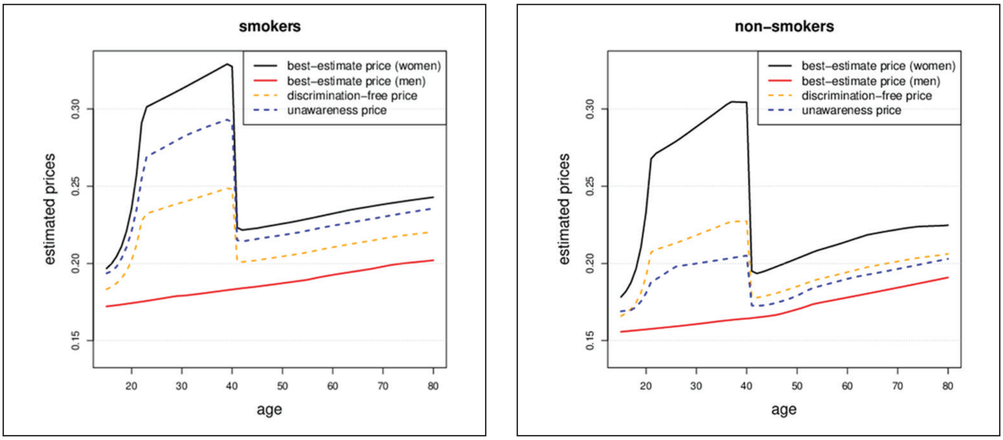

FIGURE 6 Estimated neural network model: (lhs) smokers and (rhs) non-smokers with solid black and red lines giving the best-estimate prices for women and men, respectively. The dotted orange lines show the discrimination-free prices and the dotted blue lines show the unawareness prices.

and 6 we conclude that the resulting true prices and estimated prices are rather similar. Of course, by construction the resulting discrimination-free price is gender neutral within the estimated model, and in our case close to the theoretical one.

We indicate what happens if we drop the gender variable D from the very beginning, i.e. if we train the networks only on the covariates $\mathbf{X}=\left(X_1,X_2\right)$ as considered in (17). We choose exactly the same network architecture as in Listing 2 except that we modify the input dimension on line 1 from 3 for $\left(\mathbf{X},D\right)$ to 2 for $\mathbf{X}$. These networks involve 301 weights that need to be trained. The resulting estimated regression functions $\hat{\lambda}_1\left(\mathbf{X}\right)$, $\hat{\lambda}_2\left(\mathbf{X}\right)$ and $\hat{\lambda}_3\left(\mathbf{X}\right)$, ignoring gender information D, are illustrated in Figure 7. The left-hand side shows that we can no longer distinguish between gender, however, smokers are more heavily punished for birthing related costs, which is an undesired indirect discrimination effect against women because they are more often among the group of smokers (note that the y-scales in Figures 5 and 7 are the same). Finally, merging the different claim types provides the estimated unawareness prices (when first dropping D) as illustrated by the blue dotted lines in Figure 6, which can be compared with the blue dotted lines in Figure 2.

The last step in this example would be to perform a bias correction step, for instance, similarly as in Example 8. We refrain from doing this here explicitly.

In our final analysis we illustrate that the (non-)discrimination property does not depend on the quality of the regression model (15) chosen. We choose a poor model (compared to the neural network above) by just assuming GLMs for $j=1,2,3$

$$\left(\mathbf{x},d\right)\mapsto\ \log\hat{\lambda}_j^{GLM}\left(\mathbf{x},d\right)=\theta_0^{(j)}+\theta_1^{(j)}x_1+\theta_2^{(j)}1_{\{x_2=\text{smoker}\}}+\theta_3^{(j)}1_{\{d=\text{woman}\}}. \tag{23}$$

This model will perform well for $j=2,3$, see (20)–(21), but it will perform poorly for $j=1$, see (19). This is because such a model has difficulties capturing the highly non-linear birthing related effects, as seen in Figure 8 (lhs).

In Figure 9 we present the resulting best-estimate prices, unawareness prices, and discrimination-free prices (in orange dotted lines), as estimated using the GLM. The first observation is that the resulting prices are a poor approximation to the true prices of Figure 2,

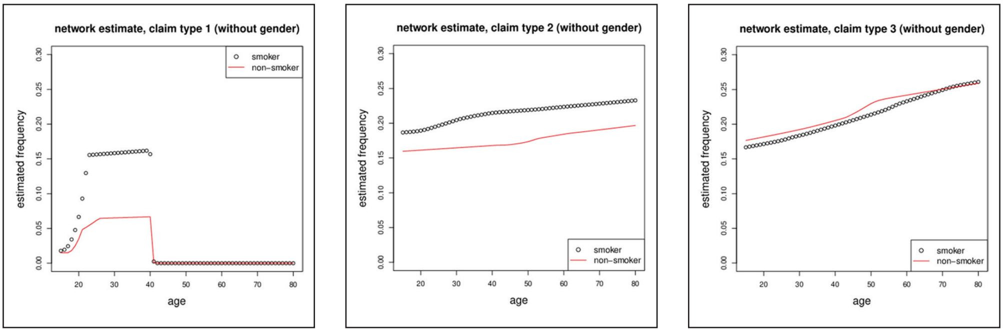

FIGURE 7 Estimated regression functions $\hat{\lambda}_1\left(\mathbf{X}\right)$ (lhs), $\hat{\lambda}_2\left(\mathbf{X}\right)$ (middle) and $\hat{\lambda}_3\left(\mathbf{X}\right)$ (rhs) using neural networks and ignoring the gender information D

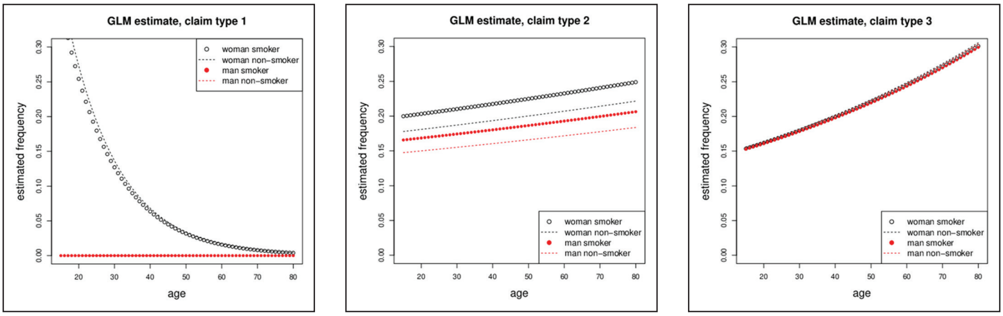

FIGURE 8 GLM estimated regression functions $\hat{\lambda}_1^{GLM}(\mathbf{X},D)$ (lhs), $\hat{\lambda}_2^{GLM}(\mathbf{X},D)$ (middle) and $\hat{\lambda}_3^{GLM}(\mathbf{X},D)$ (rhs)

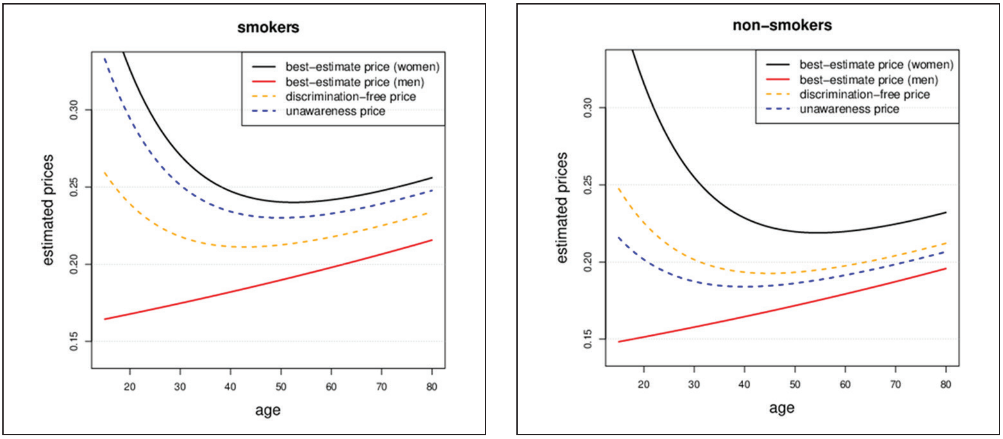

FIGURE 9 Estimated GLM: (lhs) smokers and (rhs) non-smokers with solid black and red lines giving the best-estimate prices for women and men, respectively. The dotted orange lines show the discrimination-free prices and the dotted blue lines show the unawareness prices.

the latter assuming full knowledge of the true model. However, the general discrimination behaviour is the same in both figures, namely, that the unawareness price discriminates indirectly by learning the gender D from smoking habits X_2. This is illustrated by the relative positioning of blue and orange dotted lines, with smokers more heavily charged for birthing related costs due to the fact that smokers are more likely women.

Remark

There is one issue that has not been considered so far, and which has been mentioned in the EU legislation [8], footnote (1) to Article 2.2(14) – life and health underwriting. Namely, we have implicitly assumed that the measurements of the non-discriminatory covariates are independent of the discriminatory characteristics. If we think of gender as a discriminatory covariate, this is not necessarily the case because, for instance, the waist to hip ratios naturally live on different scales for different genders, but they may still have the same impact on health related questions. This implies that non-discriminatory covariates may need pre-processing w.r.t. discriminatory ones, such that the resulting measurements for different discriminatory characteristics are comparable.

7. CONCLUDING REMARKS

The aim of this paper is to provide:

(a) an actuarial perspective of direct and indirect discrimination;
(b) a demonstration that the omission of discriminatory information leads to indirect discrimination in prices;
(c) a proposal for a simple formula that generates discrimination-free prices which works regardless of the choice of the underlying model;
(d) methods that ensure unbiasedness of discrimination-free prices at the portfolio level (the same considerations apply when transforming the actuarial tariff into a commercial one); and
(e) a discussion on the role of available data in obtaining discrimination-free prices.

The starting point to this paper has been an actuarial one. We have intentionally avoided a deeper discussion on "fairness", and, consequently, how fairness may be measured. For more on these topics, we refer to Kusner et al. [16] and the references therein. Moreover, we have also not commented on which factors should be viewed as discriminatory – this is a societal decision that goes far beyond our actuarial discussion, see e.g. Avraham et al. [2]. We (only) provide tools to implement such decisions.

Furthermore, we do not engage in a discussion of the possible systemic implications of the (non-)adoption of discrimination-free prices, be they adverse or beneficial. For example, gender neutral pricing of motor insurance may result in cheaper premiums for more dangerous (male) drivers and vice versa, with the resulting incentives leading to a deterioration of aggregate driving behaviour; telematics can be used to reflect actual driving behaviour, reducing the predictive usefulness of gender, though, this creates new challenges around privacy. Another example relates to the use of post-code information, which often correlates with ethnicity. Here, discrimination-free pricing can prevent the further penalisation of ethnic groups that have suffered historical injustices. The role of insurance in engineering socially beneficial outcomes is yet another discussion we cannot engage with in this paper. Another point worth commenting is whether discrimination-free pricing negatively impacts portfolio mixes (by adverse selection). Such impacts may result in a worse risk landscape for the industry, higher capital demands and, likely, higher premiums for the whole society.

A point worth stressing once again is that, in order to be able to calculate discrimination-free prices, one needs to have access to *all discriminatory characteristics* – otherwise it is not possible to properly adjust for the influence of such characteristics. When it comes to gender, this is generally unproblematic, but if we want to adjust for, e.g., religious beliefs or sexual orientation, such information is in general not readily available. Customers may perceive it as peculiar and intrusive to be approached with questions concerning this type of apparently irrelevant (and possibly sensitive) information. A concrete example is discussed in De Jong and Ferries [7], where sexual preference is discussed as a risk factor relating to AIDS; the authors also highlight the danger of obtaining untruthful answers to questions around sensitive information, undermining the reliability of collected data.

A key position taken in the present paper concerns the role of the overall price prediction at portfolio level. We have argued that the aggregate price for the portfolio may be calculated using all available information, including discriminatory covariates. Given this, it is the allocation of this overall cost that may introduce discrimination, and the discrimination-free pricing may be thought of as generating an allocation that avoids this. From this perspective we know from the start that the allocation is biased w.r.t. the underlying (best-estimate) portfolio risk profile. It is, hence, of interest to analyse how this biased risk profile will affect the performance of the overall portfolio price prediction.

The argument used in the present paper has focused directly on how to obtain a discrimination-free price. This has led us to a procedure which tells us how to adjust the best-estimate price to arrive at a discrimination-free price. In a statistical sense, this could be seen as a "discrimination-free point estimate". A different line of thought instead could be that we try to develop a full statistical model that is discrimination-free, i.e. sacrificing predictive power by appropriately disregarding direct and indirect discrimination, this would result in a full statistical model that provides discrimination-free responses. An example of this approach in a life insurance context are the gender neutral intensities discussed in Chen and Vigna [5]. The main reason for considering prices directly is that we believe that this approach is closer to actuarial thinking, and because maximal predictive accuracy is a desirable feature in risk management.

ACKNOWLEDGEMENTS

The authors would like to thank (in alphabetical order) the following for constructive remarks on an earlier version of the manuscript: Alois Gisler, Andreas Lagerås, Arvid Sjölander, Dirk Tasche.

REFERENCES

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

[2] Ronen Avraham, Kyle D Logue & Daniel B Schwarcz (2014). Understanding insurance anti-discrimination laws. *Southern California Law Review*, **87**(2):195–274.

[3] Sandy Bruszas, Barbara Kaschützke, Raimond Maurer & Ivonne Siegelin (2018). Unisex pricing of German participating life annuities—Boon or bane for customer and insurance company? *Insurance: Mathematics and Economics*, 78:230–245

[4] An Chen, Montserrat Guillén & Elena Vigna (2018). Solvency requirement in a unisex mortality model. *ASTIN Bulletin: The Journal of the IAA*, **48**(3):1219–1243

[5] An Chen & Elena Vigna (2017). A unisex stochastic mortality model to comply with EU Gender Directive. *Insurance: Mathematics and Economics*, **73**:124–136

[6] Imre Csiszár (1975). *I*-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, **3**(1):146–158

[7] Piet De Jong and Shauna Ferris (2006). Adverse selection spirals. *ASTIN Bulletin: The Journal of the IAA*, 36(2):589–628

[8] 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

[9] 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

[10] Jerome H Friedman (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232

[11] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai & Ansgar Walther (2008). Predictably unequal? The effects of machine learning on credit markets. Available at SSRN: https://ssrn.com/abstract=3072038 (Downloaded on 21 February 2020)

[12] Leo Guelman & Montserrat Guillén (2014). A causal inference approach to measure price elasticity in automobile insurance. *Expert Systems with Applications*, **41**(2):387–396

[13] Montserrat Guillén (2012). Sexless and beautiful data: from quantity to quality. *Annals of Actuarial Science*, **6**(2):231–234

[14] Samuel Krikler, Dan Dolberger & Jacob Eckel (2004). Method and tools for insurance price and revenue optimisation. *Journal of Financial Services Marketing*, 9(1):68–79

[15] Manabu Kuroki & Judea Pearl (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, **101**(3):423–437

[16] Matt J Kusner, Joshua Loftus, Chris Russell & Ricardo Silva (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076

[17] Steffen L Lauritzen (1996). *Graphical models*. Oxford Science Publications

[18] Judea Pearl (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146

[19] Judea Pearl, Madelyn Glymour & Nicholas P Jewell (2016). *Causal inference in statistics: A primer*. John Wiley & Sons

[20] Devin G Pope & Justin R Sydnor (2011). Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy*, **3**(3):206–231

[21] Jörn Sass & Frank Thomas Seifried (2014). Insurance markets and unisex tariffs: is the European Court of Justice improving or destroying welfare? *Scandinavian Actuarial Journal*, 2014(3):228–254

[22] Qingyuan Zhao & Trevor Hastie (2019). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*