European Actuarial Journal
https://doi.org/10.1007/s13385-023-00367-z

ORIGINAL RESEARCH PAPER

A multi-task network approach for calculating discrimination-free insurance prices

Mathias Lindholm[1] · **Ronald Richman**[2] · **Andreas Tsanakas**[3] · **Mario V. Wüthrich**[4]

Received: 1 November 2022 / Revised: 9 August 2023 / Accepted: 10 October 2023

Abstract
In applications of predictive modeling, such as insurance pricing, indirect or proxy discrimination is an issue of major concern. Namely, there exists the possibility that protected policyholder characteristics are implicitly inferred from non-protected ones by predictive models and are thus having an undesirable (and possibly illegal) impact on prices. A technical solution to this problem relies on building a best-estimate model using all policyholder characteristics (including protected ones) and then averaging out the protected characteristics for calculating individual prices. However, such an approach requires full knowledge of policyholders' protected characteristics, which may in itself be problematic. Here, we address this issue by using a multi-task neural network architecture for claim predictions, which can be trained using only partial information on protected characteristics and produces prices that are free from proxy discrimination. We demonstrate the proposed method on both synthetic data and a real-world motor claims dataset, in which proxy discrimination can be observed. In both examples we find that the predictive accuracy of the multi-task network is comparable to a conventional feed-forward neural network, when the protected information is available for at least half of the insurance policies. However, the multi-task network has superior performance in the case when the protected information is known for less than half of the insurance policyholders.

Keywords Indirect discrimination · Proxy discrimination · Discrimination-free insurance pricing · Unawareness price · Best-estimate price · Protected information ·

✉ Andreas Tsanakas
a.tsanakas.1@city.ac.uk

1 Department of Mathematics, Stockholm University, Stockholm, Sweden

2 Old Mutual Insure and University of the Witwatersrand, Johannesburg, South Africa

3 Bayes Business School (formerly Cass), City, University of London, London, UK

4 RiskLab, Department of Mathematics, ETH Zurich, Zurich, Switzerland

Published online: 08 November 2023

Discriminatory covariates · Fairness · Incomplete information · Multi-task learning · Multi-output network

1 Introduction

The question of avoiding discrimination in insurance pricing is becoming increasingly important in many markets and jurisdictions. For example, the European Council [9] prohibits using gender information as a rating factor for insurance pricing; for an actuarial overview on discrimination regulation we refer to Frees–Huang [10]. While regulation varies across jurisdictions, it is typically required that both direct and indirect discrimination be avoided. If $\mathbf{D}$ denotes protected information whose use is regarded as discriminatory, *direct discrimination* is avoided by merely not including $\mathbf{D}$ in the regression model used for insurance pricing. However, not including protected information in the regression model is not necessarily sufficient for avoiding discrimination more broadly, because the protected information $\mathbf{D}$ may also be (implicitly) inferred from the non-protected variables, denoted by $\mathbf{X}$. We call the impact of such inference on prices *indirect discrimination*. We note that this is a narrow use of the latter term, equivalent to what is also known as *proxy discrimination*, and does not consider any aspects of fairness and disparate impact; we also refer to Prince–Schwarcz [17], Frees–Huang [10], Lindholm et al. [14, 15], Xin–Huang [25] and Grari et al. [11] for relevant discussions.

In this paper, we address the problem of avoiding indirect discrimination in the calculation of insurance prices. Lindholm et al. [14] give a mathematical definition of direct and indirect discrimination. Their approach for avoiding indirect discrimination amounts to, first, using *all* available information $(\mathbf{X}, \mathbf{D})$ to calculate the so-called *best-estimate price*. In a second step, one removes the potential discriminatory dependence between $\mathbf{X}$ and $\mathbf{D}$ by marginalizing the best-estimate price w.r.t. a pricing distribution which does not allow one to infer (or proxy) the protected information $\mathbf{D}$ from the non-protected variables $\mathbf{X}$. This step removes the statistical dependence between the two sets of information and results in the so-called *discrimination-free insurance price* as defined in Lindholm et al. [14]. This removal of statistical dependence can be motivated (and justified) by concepts of causal statistics, see Lindholm et al. [14] and Araiza Iturria et al. [2]; for an antecedent of this approach in economics, see Pope–Sydnor [16].

An attractive feature of the above discrimination-free insurance pricing approach is that any pricing model can be used to obtain the best-estimate price, which is subsequently adjusted to remove the potential for $\mathbf{D}$ to be proxied by $\mathbf{X}$. Moreover, the suggested procedure ensures that all potential indirect discrimination is removed, where it exists, and this is achieved regardless of the ability of the particular class of regression model used to infer information about $\mathbf{D}$ from $\mathbf{X}$. In particular, there is no need to explicitly quantify the potential impact of indirect discrimination before applying the method.

Thus, the calculation of discrimination-free insurance prices can be carried out using any reasonable pricing model, if one has access to the *full* covariate information $(\mathbf{X}, \mathbf{D})$. In practice, however, one may assume that the protected characteristics

D contain covariates that are considered sensitive, such as, e.g., ethnicity. Then, it will generally not be feasible to collect this information for all insurance contracts in the portfolio. As a consequence, it remains unclear how a model for discrimination-free insurance prices should be fitted, when discriminatory information is incomplete. The goal of this paper is to address precisely this issue. We present a *multi-output neural network* for *multi-task learning*, i.e., the proposed network architecture performs simultaneously different regression tasks. This proposed network architecture can be fitted on incomplete protected information, and still provides accurate results. That is, to fit our network architecture we only need the protected information **D** on a part of the portfolio, but we can still receive a good predictive regression model for discrimination-free insurance pricing. In particular, our proposal allows for a more robust fitting compared to just ignoring insurance policies with missing protected information.

Incomplete protected information **D** is a missing data problem, and there is a broad literature on dealing with missing data. Typical methods to deal with missing data are either imputation methods or surrogate splits in tree based methods; for a survey we refer to Emmanuel et al. [8]. Popular imputation methods are k-nearest neighbor (kNN) imputation, see Batista–Monard [3], Chen–Shao [6] and the references therein, multivariate imputation by chained equations (MICE), see van Buuren–Groothuis-Oudshoorn [22], or generative adversarial imputation nets (GAIN), see Yoon et al. [26].

Classical incomplete data problems present different challenges to the ones arising in our context of discrimination-free insurance pricing. In standard incomplete data settings, one tries to complete the missing information (in our case **D**) on those instances where this information is not available. Typically, this is done as a data pre-processing step using one of the previously mentioned methods. Based on these pre-processed data, one then builds a regression model that optimally predicts responses based on the completed data set (in our case $(\mathbf{X}, \mathbf{D})$). Our problem is fundamentally different in several aspects, which is why it requires a different methodological approach. First, we do not aim at imputing the observations missing from the discriminatory covariates **D**. For the derivation of discrimination-free prices, which are calculated using only knowledge of **X**, it is important to know what the best-estimate prices would be under different values of **D**—however, the value of **D** for a specific instance (policyholder) must be immaterial. Second, and relatedly, imputation methods explicitly rely on proxying **D** from **X**, which is precisely the phenomenon that we need to prevent. In contrast, we aim to work directly on the original, partly incomplete, data and do not engage in any pre-processing step. Third, if imputation of **D** was used to complete the dataset before estimating regression functions, it would remain unclear whether the discrimination-free properties of resulting pricing functionals, discussed in Lindholm et al. [14], could be maintained. This is because the entanglement of estimation with data completion processes creates an additional level of complexity. Our proposed multi-task learning approach is designed specifically to address those issues.

We illustrate our proposed methodology via two detailed case studies, using respectively a synthetic health insurance dataset and a real-world motor claims dataset. In both cases substantial proxy discrimination is present: in the synthetic data this is by construction, while in the real-world data the discriminatory effect of unawareness

pricing is empirically observed. The two cases demonstrate that, first, the multi-output network architecture provides results of comparable accuracy to a conventional feed-forward neural network, when complete information on policyholder characteristics is available. Second, in the more realistic scenario when information on protected policyholder attributes is missing for more than 50% of the insurance policies, the multi-task network clearly outperforms a conventional approach, whereby a regression model is only trained on those instances for which the full information is available. This is observed both in the case where protected data are missing at random and not at random.

Organization of manuscript. In the next section we review the framework of discrimination-free insurance pricing as introduced in Lindholm et al. [14]. Section 3 presents our solution to the problem of having incomplete protected information, by gradually building up towards the multi-task network architecture. Section 4 provides the synthetic data example, which verifies the good performance of our proposal. In Sect. 5 we demonstrate applying the proposed method using a real non-life insurance pricing dataset. Finally, in Sect. 6 we give concluding remarks.

2 Discrimination-free insurance pricing

We first recall the mathematical definitions of the best-estimate, unawareness and discrimination-free insurance prices, as they were introduced in Lindholm et al. [14]. Throughout, we work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that is assumed to be sufficiently rich to carry all the objects that we would like to study, and $\mathbb{P}$ denotes the physical probability measure. Our goal is to employ a regression model that calculates the prices of insurance policies that satisfy the property of being discrimination-free according to Definition 12 of Lindholm et al. [14].

We assume that the vector of covariates $(\mathbf{X}, \mathbf{D})$ can be partitioned into non-discriminatory covariates $\mathbf{X}$ and discriminatory covariates (protected characteristics) $\mathbf{D}$. This split into $\mathbf{X}$ and $\mathbf{D}$ is given exogenously, e.g., by law or by societal norms and preferences. The distribution of the covariates $(\mathbf{X}, \mathbf{D})$ of a randomly selected policyholder is described by $\mathbb{P}$. The (insurance) claim of this policyholder is denoted by Y, and we assume that this claim depends on the covariates $(\mathbf{X}, \mathbf{D})$. That is, we would like to study the conditional distribution function

$$y \in \mathbb{R} \quad \mapsto \quad F_{Y|(\mathbf{x},\mathbf{d})}(y) = \mathbb{P}\left[Y \leq y \mid \mathbf{X} = \mathbf{x}, \mathbf{D} = \mathbf{d}\right],$$

of a selected policyholder having covariates $(\mathbf{X}, \mathbf{D}) = (\mathbf{x}, \mathbf{d})$.

The *best-estimate price* of the policyholder with covariates $(\mathbf{X}, \mathbf{D})$ is defined by the conditional expectation (subject to existence)

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

This price is called best-estimate because, under square integrability, it minimizes the conditional mean squared error of prediction (MSEP), given full information $(\mathbf{X}, \mathbf{D})$. Thus, the best-estimate price (1) is the most accurate price we can calculate under full

covariate information $(\mathbf{X}, \mathbf{D})$. The general statistical problem is to optimally determine (estimate) this regression function

$$(\mathbf{x}, \mathbf{d}) \;\mapsto\; \mu(\mathbf{x}, \mathbf{d}) = \mathbb{E}\left[Y \mid \mathbf{X} = \mathbf{x}, \mathbf{D} = \mathbf{d}\right], \tag{2}$$

from past data (and maybe expert opinion). Below, we are going to use a neural network regression approach for this task.

Obviously this best-estimate price (1) directly discriminates because it uses the discriminatory covariates $\mathbf{D}$ as input. This motivates the definition of the *unawareness price*, ignoring the knowledge about the discriminatory covariates $\mathbf{D}$,

$$\mu(\mathbf{X}) := \mathbb{E}\left[Y \mid \mathbf{X}\right]. \tag{3}$$

The unawareness price (3) avoids *direct discrimination* according to Definition 10 of Lindholm et al. [14] because we no longer need any information about the discriminatory covariates $\mathbf{D}$ to calculate this price. Using the tower property of conditional expectations, we can rewrite the unawareness price as follows

$$\mu(\mathbf{X}) = \int_{\mathbf{d}} \mu\left(\mathbf{X}, \mathbf{d}\right) \, \mathrm{d}\mathbb{P}(\mathbf{d} \mid \mathbf{X}), \tag{4}$$

where $\mathbb{P}(\mathbf{d} \mid \mathbf{X})$ describes the conditional distribution of the discriminatory covariates $\mathbf{D}$, given the non-discriminatory information $\mathbf{X}$. It is exactly this link which is problematic, namely, having broad non-discriminatory information $\mathbf{X}$ may (easily) allow us to infer the discriminatory information $\mathbf{D}$. Such an inference of the protected characteristics $\mathbf{D}$ is therefore implicit in the definition of the unawareness price. This implicit inference has been coined in insurance as *proxy discrimination*, see, e.g., Frees–Huang [10] and Xin–Huang [25], or *indirect discrimination*, see Lindolm et al. [14]. To prevent indirect discrimination, one needs to break the link that allows one to infer $\mathbf{D}$ from $\mathbf{X}$. This can be done purely statistically by just replacing the outer distribution in (4) by an unconditional one. Alternatively, this replacement can be justified by arguments from causal statistics if insurance claims follow a certain causal relationship, see Lindholm et al. [14] and Araiza Iturria et al. [2].

These arguments motivate the definition of the *discrimination-free insurance price*

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

where the pricing distribution $\mathbb{P}^*(\mathbf{d})$ has the same support as the marginal distribution of the discriminatory covariates $\mathbf{D} \sim \mathbb{P}(\mathbf{d})$.

Discrimination-free insurance pricing (5) has two ingredients, namely, the regression function $\mu(\mathbf{x}, \mathbf{d})$, see (2), and the pricing distribution $\mathbb{P}^*(\mathbf{d})$. The most natural choice for this pricing distribution is simply the marginal distribution $\mathbb{P}(\mathbf{d})$, but there may be other (justified) choices, e.g., providing unbiasedness of discrimination-free insurance prices; for a broader discussion on the choice of $\mathbb{P}^*$ we refer to Remark 7 and Section 4 in Lindholm et al. [14].

In this paper we are more concerned about the first issue, namely, about selecting, estimating and applying the (best-estimate) regression function $(\mathbf{x}, \mathbf{d}) \mapsto \mu(\mathbf{x}, \mathbf{d})$. In practice, this requires that we hold both non-discriminatory *and* discriminatory information $(\mathbf{x}, \mathbf{d})$ from the insurance policyholders for regression model fitting, and the discriminatory information is integrated out (adjusted for) only in the subsequent (second) step (5). However, in many cases it is problematic to collect this discriminatory information over the entire insurance portfolio. Therefore, fitting the regression function (5) might not be practical. In the next section, we provide a technical workaround which requires discriminatory information only for part of the portfolio, but it will still equip us with accurate predictive models.

Remark 1 The discrimination-free insurance price (5) is defined within a given model specification, i.e., for a given distributional model for $(Y, \mathbf{X}, \mathbf{D})$, see Definition 12 of Lindholm et al. [14]. This does not consider model error coming from a poorly specified stochastic model for $(Y, \mathbf{X}, \mathbf{D})$, which may result in a different form of discrimination, e.g., arising from a certain sub-population being under-represented in the data. Naturally, on the corresponding part of the covariate space we have greater model uncertainty, because we have less data for an accurate model fit. This may result in forms of demographic discrimination that are outside of our (more narrow) scope which is always attached to a *given* stochastic model. For a discussion of discrimination arising from unrepresentative data in a different context, see Buolamwini–Gebru [4].

3 Multi-output network regression model

We present statistical modeling of the regression function $\mu(\mathbf{x}, \mathbf{d})$ within the framework of feed-forward neural networks (FNNs). We start by introducing a standard (plain-vanilla) FNN architecture, and in a second step we discuss how this FNN architecture can be modified to serve our purpose of deriving discrimination-free insurance prices with partial information of the protected characteristics. The notation and terminology of neural network regression modeling is taken from Chapter 7 of Wüthrich–Merz [24].

3.1 Feed-forward neural network regression model

We start with a short summary of the FNN architecture; a more extended exposition can be found in Appendix A. Assume that the regression function $(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^{q_0} \mapsto \mu(\mathbf{x}, \mathbf{d})$ can be modeled by a FNN architecture taking the following form

$$(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^{q_0} \;\mapsto\; g\left(\mu(\mathbf{x}, \mathbf{d})\right) = \left\langle \boldsymbol{\beta}, \mathbf{z}^{(m:1)}(\mathbf{x}, \mathbf{d}) \right\rangle, \tag{6}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a strictly monotone and smooth link function, $\mathbf{z}^{(m:1)} : \mathbb{R}^{q_0} \to \mathbb{R}^{q_m}$ is a FNN of depth $m \in \mathbb{N}$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{q_m})^\top \in \mathbb{R}^{q_m+1}$ is the readout parameter, and where $\langle \cdot, \cdot \rangle$ denotes the scalar product; for details we refer to (28) below. This FNN architecture $\mathbf{z}^{(m:1)}$ of depth $m \in \mathbb{N}$ is a composition of m hidden FNN layers $\mathbf{z}^{(j)} : \mathbb{R}^{q_{j-1}} \to \mathbb{R}^{q_j}$, $1 \le j \le m$, providing mapping

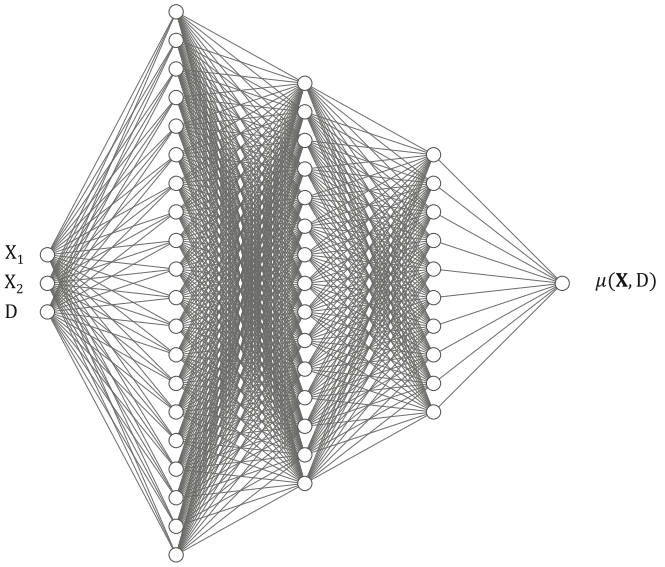

Fig. 1 Illustration of a FNN architecture (6) of depth $m = 3$ with $(q_1, q_2, q_3) = (20, 15, 10)$ hidden neurons in the three hidden layers. This network computes best-estimate prices $\mu(\mathbf{X}, D)$ from the three-dimensional input $(\mathbf{X}, D) = (X_1, X_2, D)$

$$(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^{q_0} \quad \mapsto \quad \mathbf{z}^{(m:1)}(\mathbf{x}, \mathbf{d}) = \left(\mathbf{z}^{(m)} \circ \cdots \circ \mathbf{z}^{(1)}\right)(\mathbf{x}, \mathbf{d}); \tag{7}$$

we also refer to (27) below. This mapping (7) transforms the q_0-dimensional vector-valued input $(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^{q_0}$ to a new q_m-dimensional learned representation $\mathbf{z}^{(m:1)}(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^{q_m}$ of the original non-discriminatory and discriminatory covariates $(\mathbf{x}, \mathbf{d})$. Figure 1 visualizes this standard FNN architecture, in an example with depth $m = 3$ (number of hidden layers) and numbers of neurons $(q_1, q_2, q_3) = (20, 15, 10)$ in these hidden layers.

FNN regression modeling requires specification of the network architecture. This involves the choice of the depth $m \in \mathbb{N}$, the numbers of neurons $q_j \in \mathbb{N}$ in each hidden layer $1 \leq j \leq m$, the activation function in each of these neurons, as well as the link function g. Such a network architecture has network weights $\mathbf{w}$ (these are all parameters in the hidden layers) and readout parameter $\boldsymbol{\beta}$; see Appendix A for more details. These network weights $\mathbf{w}$ and the readout parameter $\boldsymbol{\beta}$ are then fitted (trained) to the available data, so that we obtain a regression model of good predictive power. Successful FNN fitting involves (in most cases) an early stopping strategy to prevent (in-sample) over-fitting of the model to the training data, i.e., targeting for an optimal out-of-sample predictive performance; for a detailed description of network fitting we refer to Section 7.2.3 of Wüthrich–Merz [24].

This fitted FNN (6) provides the best-estimate prices

$$\mu(\mathbf{x}, \mathbf{d}) = \mathbb{E}\left[Y \mid \mathbf{X} = \mathbf{x}, \mathbf{D} = \mathbf{d}\right] = g^{-1}\left\langle \boldsymbol{\beta}, \mathbf{z}^{(m:1)}(\mathbf{x}, \mathbf{d}) \right\rangle, \tag{8}$$

for the insurance claims Y, being described by the covariates $(\mathbf{x}, \mathbf{d})$. From this we can calculate the discrimination-free insurance prices with formula (5) by specifying a suitable pricing distribution $\mathbb{P}^*(\mathbf{d})$. A standard choice is to use the marginal distribution of $\mathbf{D}$ from the part of the portfolio where the protected information of $\mathbf{D}$ is known.

The difficulty in practice with this approach, and similar regression approaches such as generalized linear models (GLMs), is that it requires full knowledge of the discriminatory information $\mathbf{D} = \mathbf{d}$ of the policyholders. Otherwise one cannot fit this FNN (6) on the available (past) data. A naive solution is to just fit this FNN architecture on the sub-portfolio where $\mathbf{D}$ is available; in the missing data literature this approach is called deletion; see, e.g., Emmanuel et al. [8]. We call this the (naive) *plain-vanilla FNN approach* because it is clearly non-optimal to disregard any insurance policy where there is no complete information available about the covariates $(\mathbf{X}, \mathbf{D})$.

Remark 2 In the introduction we have mentioned that discrimination-free insurance pricing can be applied to any pricing (regression) model. Here, we restrict to FNN architectures which seems rather limiting. However, we would like to mention that large FNN architectures provide the universal approximation property. This implies that within FNN architectures we can mimick any other (sufficiently regular) regression model.

3.2 Multi-output neural network regression model

In constructing (and fitting) the plain-vanilla FNN best-estimate prices (8), we directly use the discriminatory information $\mathbf{D} = \mathbf{d}$ of the policyholders as an input variable to the FNN. Our proposal is to change this FNN architecture such that only the non-discriminatory information $\mathbf{X} = \mathbf{x}$ is used as an input variable to the network, but at the same time we generate a whole family of best-estimate prices that reflects the different specifications (levels) of the discriminatory information. In the present section we introduce this new network architecture. We call it a *multi-output FNN* architecture because it has multiple outputs that generate the family of models. In Sect. 3.3, below, we extend this multi-output FNN architecture to a *multi-task FNN* architecture which not only generates a whole family of models, but it also simultaneously fulfills multiple tasks. Namely, it not only calculates best-estimate prices but it also computes unawareness prices by internally accounting for (missing) discriminatory information. It will be exactly this multi-task FNN architecture that we promote for discrimination-free insurance pricing under incomplete discriminatory information, as it can deal with the issue of missing protected information, but still provides good predictive models. In this approach, protected information will only be needed on part of the portfolio for model training.

Assume that the discriminatory information $\mathbf{D}$ only takes finitely many values $\mathbf{d} \in D_K := \{\mathbf{d}_1, \ldots, \mathbf{d}_K\}$. Typically, we think of discriminatory information being of categorical type, e.g., gender or ethnicity. If this is not the case, discriminatory information can be discretized, and one should work with this discretized version. We modify the above plain-vanilla FNN (6) such that it only considers non-discriminatory covariates $\mathbf{x}$ as an input giving us the learned representation

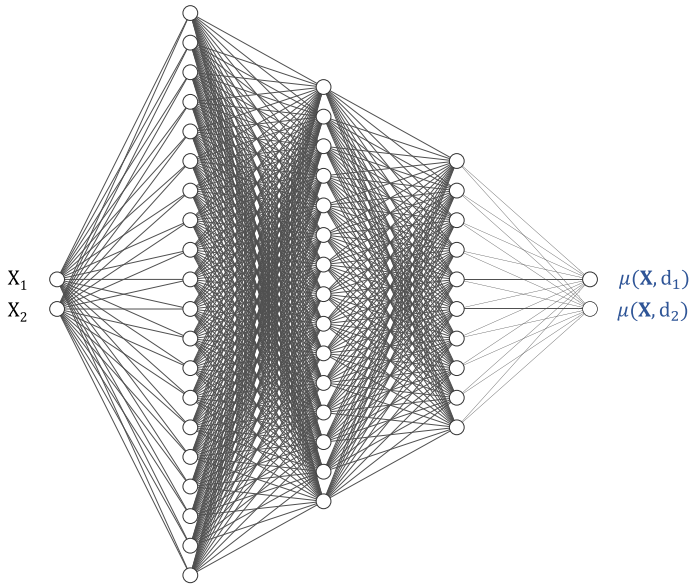

Fig. 2 Illustration of a multi-output FNN architecture (9) of depth $m = 3$ with $(q_1, q_2, q_3) = (20, 15, 10)$ hidden neurons in the three hidden layers. This network computes best-estimate prices $\mu(\mathbf{X}, d)$ for alternative values of $D = d \in \{d_1, d_2\}$, given the bivariate input $\mathbf{X} = (X_1, X_2)$. The key difference to the FNN architecture of Fig. 1 is that the information on the discriminatory covariate D is used only for model fitting but not for carrying out (new) predictions

$$\mathbf{x} \quad \mapsto \quad \mathbf{z}^{(m:1)}(\mathbf{x}) = \left(\mathbf{z}^{(m)} \circ \cdots \circ \mathbf{z}^{(1)}\right)(\mathbf{x}),$$

compare to (7). This learned representation should be sufficiently rich such that it provides a whole family of suitable regression functions, parameterized by $\mathbf{d}_k \in D_K$. This typically requires that the number of hidden neurons, in particular q_m in the last hidden layer, is not too small. This learned representation is now used to calculate the best-estimate prices simultaneously for all discriminatory specifications $\mathbf{d}_k \in D_K$, that is, we set for the *multi-output FNN* architecture

$$\left(\mu(\mathbf{x}, \mathbf{d}_1), \ldots, \mu(\mathbf{x}, \mathbf{d}_K)\right) = \left(g^{-1}\left\langle \boldsymbol{\beta}_1, \mathbf{z}^{(m:1)}(\mathbf{x})\right\rangle, \ldots, g^{-1}\left\langle \boldsymbol{\beta}_K, \mathbf{z}^{(m:1)}(\mathbf{x})\right\rangle\right). \quad (9)$$

The multi-output network architecture is shown in Fig. 2. It is seen that for every non-discriminatory input $\mathbf{x}$ we generate a whole family of outputs that simultaneously provide the best-estimate prices $\mu(\mathbf{x}, \mathbf{d}_k)$ for all levels $\mathbf{d}_k \in D_K, 1 \le k \le K$, of the discriminatory information. The different specifications $\mathbf{d}_k \in D_K$ of the discriminatory covariates $\mathbf{D}$ are encoded in different readout parameters $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K \in \mathbb{R}^{q_m+1}$.

Remark 3 An alternative to the multi-output FNN architecture (9) is to fit K separate FNNs, i.e., a separate model for each level $\mathbf{d}_k \in D_K$, $1 \le k \le K$. This approach is applicable to all types of regression models, and it removes the explicit need of the

(non-trivial) modeling of the interaction effects between $\mathbf{X}$ and $\mathbf{D}$. However, it comes at the price of a very high-dimensional parameter. In contrast, the multi-output FNN architecture (9) shares the network weights $\mathbf{w}$ across all levels $\mathbf{d}_k \in D_K$, $1 \le k \le K$, and only the readout parameters $\boldsymbol{\beta}_k$ may differ. Assume that the non-discriminatory input $\mathbf{x}$ is r-dimensional and that the discriminatory input $\mathbf{d} = d$ is 1-dimensional with K different levels. For the classical FNN architecture (6) we have input dimension $q_0 = r + 1$ which results in network weights $\mathbf{w}$ and readout parameter $\boldsymbol{\beta}$ of total dimension

$$q_1(r + 1 + 1) + \sum_{j=2}^{m} q_j(q_{j-1} + 1) + (q_m + 1), \tag{10}$$

we also refer to Appendix A. The multi-output FNN architecture (9) has a lower-dimensional input which gives fewer network weights $\mathbf{w}$, but uses more readout parameters $\beta_1, \ldots, \beta_K$ resulting in total dimension

$$q_1(r + 1) + \sum_{j=2}^{m} q_j(q_{j-1} + 1) + K(q_m + 1). \tag{11}$$

Finally, fitting a separate FNN (6) for each level $\mathbf{d}_k$ of $\mathbf{D}$ would result in a parameter of total dimension

$$K \left(q_1(r + 1) + \sum_{j=2}^{m} q_j(q_{j-1} + 1) + (q_m + 1) \right).$$

The remaining question is about fitting these parameters in the multi-output FNN case. We start by describing how the plain-vanilla FNN (6) is fit to i.i.d. data $(Y_i, \mathbf{X}_i, \mathbf{D}_i)_{1 \le i \le n}$. For the moment we assume to be in the situation of complete information. We choose a loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ to assess the quality of the fit. This loss function can be the square loss function, a deviance loss function, or any other sensible choice that fits to the estimation problem to be solved. If $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{w})$ collects all parameters to be estimated/fitted, then, typically, an optimal parameter is found by solving (M-estimation)

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} L\Big(Y_i,\ \mu_{\boldsymbol{\theta}}(\mathbf{X}_i, \mathbf{D}_i)\Big), \tag{12}$$

where in the regression function $\mu(\cdot) = \mu_{\boldsymbol{\theta}}(\cdot)$ we highlight its dependence on the parameter $\boldsymbol{\theta}$ to be optimized. This is the process to fit the plain-vanilla FNN given in (8), subject to early stopping to prevent in-sample over-fitting; for a detailed discussion of FNN fitting we refer to Section 7.2.3 in Wüthrich–Merz [24].

For fitting the multi-output FNN (9) we modify this fitting procedure as follows

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\min} \sum_{i=1}^{n} \sum_{k=1}^{K} L\Big(Y_i,\ \mu_{\boldsymbol{\theta}}(\mathbf{X}_i, \mathbf{d}_k)\Big)\, \mathbb{1}_{\{\mathbf{D}_i=\mathbf{d}_k\}} \tag{13}$$
$$= \underset{\boldsymbol{\theta}}{\arg\min} \sum_{i=1}^{n} \sum_{k=1}^{K} L\left(Y_i,\ g^{-1}\left\langle \boldsymbol{\beta}_k, \mathbf{z}^{(m:1)}(\mathbf{X}_i)\right\rangle\right) \mathbb{1}_{\{\mathbf{D}_i=\mathbf{d}_k\}},$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \mathbf{w})$ collects all readout parameters and the network weights, see (9). That is, we add an indicator $\mathbb{1}_{\{\mathbf{D}_i=\mathbf{d}_k\}}$ referring to the discriminatory information $\mathbf{D}_i$ of observation i, which, in turn, trains the corresponding readout parameter $\boldsymbol{\beta}_k \in \mathbb{R}^{q_m+1}$ of the multi-output FNN (9). Note that this is the *only* step where the protected information $\mathbf{D}_i$ is used in the multi-output FNN approach.

Having the fitted multi-output FNN (9) we arrive at the discrimination-free insurance prices

$$\mathbf{x} \ \mapsto\ \mu^*(\mathbf{x}) = \sum_{k=1}^{K} \mu(\mathbf{x}, \mathbf{d}_k)\, \mathbb{P}^*(\mathbf{d}_k) = \sum_{k=1}^{K} g^{-1}\left\langle \boldsymbol{\beta}_k, \mathbf{z}^{(m:1)}(\mathbf{x})\right\rangle \mathbb{P}^*(\mathbf{d}_k), \tag{14}$$

for each possible choice of the pricing distribution $\mathbb{P}^*$ on the finite set $D_K = \{\mathbf{d}_1, \ldots, \mathbf{d}_K\}$.

To conclude, the multi-output FNN (9) generates a whole family of best-estimate prices $(\mu(\cdot, \mathbf{d}_k))_{1\le k\le K}$ of the discriminatory information in D_K, see Fig. 2 for the binary case $K = 2$. This discriminatory information only enters the loss function in the fitting procedure (13), and once this model is fit we no longer need discriminatory information to calculate the discrimination-free insurance price (14) for any (new) insurance policy.

3.3 Multi-task learning and incomplete discriminatory information

We extend the multi-output FNN architecture from the previous section to a multi-task learning model. Firstly, this extension simultaneously solves different prediction tasks, and, secondly, it uses a special loss function for model fitting provided in formula (20), below. This loss function is going to be discussed in the subsequent Remark 4. The proposed multi-task learning model will be suitable for our problem of incomplete discriminatory information, because, besides using maximal available information, it also models the discriminatory information on the policies where this information is not available. This acts as a regularizer for model fitting.

We define the categorical probabilities for $\mathbf{d}_k \in D_K$, $1 \le k \le K$,

$$p_k(\mathbf{x}) := \mathbb{P}[\mathbf{D} = \mathbf{d}_k \mid \mathbf{X} = \mathbf{x}] \ \in\ [0, 1].$$

We choose two separate FNNs for representation learning

$$\mathbf{x} \ \mapsto\ \mathbf{z}_{\mu}^{(m:1)}(\mathbf{x}) \in \mathbb{R}^{q_m} \qquad \text{and} \qquad \mathbf{x} \ \mapsto\ \mathbf{z}_{p}^{(m:1)}(\mathbf{x}) \in \mathbb{R}^{q_m}; \tag{15}$$

these two FNNs are further discussed in the fourth item of Remark 4, below. For simplicity we assume that these two FNNs have exactly the same network architecture, but, typically, their network weights (parameters) $\mathbf{w}_\mu$ and $\mathbf{w}_p$ will differ. The first FNN is used to model the best-estimate prices

$$\mathbf{x} \;\mapsto\; \mu(\mathbf{x}, \mathbf{d}_k) = g^{-1}\left\langle \boldsymbol{\beta}_k, \mathbf{z}_\mu^{(m:1)}(\mathbf{x}) \right\rangle \qquad \text{for } 1 \le k \le K, \tag{16}$$

with readout parameters $\boldsymbol{\beta}_k \in \mathbb{R}^{q_m+1}$. As in the multi-output FNN (9), this architecture only uses the non-discriminatory covariates $\mathbf{x}$ as inputs and it provides a whole family of regression functions (outputs).

The second FNN is used to model the categorical probabilities $(p_k(\mathbf{x}))_{1\le k\le K}$. Using the softmax output function and for readout parameters $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K \in \mathbb{R}^{q_m+1}$, we define the FNN classification model

$$\mathbf{x} \;\mapsto\; p_k(\mathbf{x}) = \frac{\exp\left\langle \boldsymbol{\alpha}_k, \mathbf{z}_p^{(m:1)}(\mathbf{x}) \right\rangle}{\sum_{j=1}^K \exp\left\langle \boldsymbol{\alpha}_j, \mathbf{z}_p^{(m:1)}(\mathbf{x}) \right\rangle} \;\in\; (0,1). \tag{17}$$

At the current stage these two networks are completely unrelated because they run in parallel, and they can be fitted independently from each other. We now make them related by (internally) calculating the unawareness price using the tower property (4), i.e.,

$$\mathbf{x} \;\mapsto\; \mu(\mathbf{x}) = \sum_{k=1}^K \mu(\mathbf{x}, \mathbf{d}_k)\, p_k(\mathbf{x}). \tag{18}$$

Combining (16), (17) and (18) we receive the *multi-task FNN* architecture

$$\mathbf{x} \;\mapsto\; \Big(\mu(\mathbf{x}, \mathbf{d}_1), \ldots, \mu(\mathbf{x}, \mathbf{d}_K);\; p_1(\mathbf{x}), \ldots, p_K(\mathbf{x});\; \mu(\mathbf{x})\Big) \;\in\; \mathbb{R}^{2K+1}, \tag{19}$$

with network parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K, \mathbf{w}_\mu, \mathbf{w}_p)$. As input this multi-task FNN only uses the non-discriminatory information $\mathbf{x}$. Remark that the unawareness price $\mu(\mathbf{x})$ in (19) is calculated internally in the network using (18). The multi-task network architecture is shown in Fig. 3, and we provide further discussion in Remark 4, below.

The crucial part now is how we train this multi-task FNN architecture (19) on incomplete data. We assume that the discriminatory information $\mathbf{D}_i$ is only available on part of the insurance policies $i \in \mathcal{I} \subset \{1, \ldots, n\}$. This requires that we mask $\mathbf{D}_i$ for the policies $i \in \mathcal{I}^c = \{1, \ldots, n\} \backslash \mathcal{I}$ where no discriminatory information is available. As mask we set $\mathbf{D}_i = \texttt{NA}$ for $i \in \mathcal{I}^c$. This will ignore the parts of the following loss function for which the discriminatory information is not available. We set for the new optimization problem

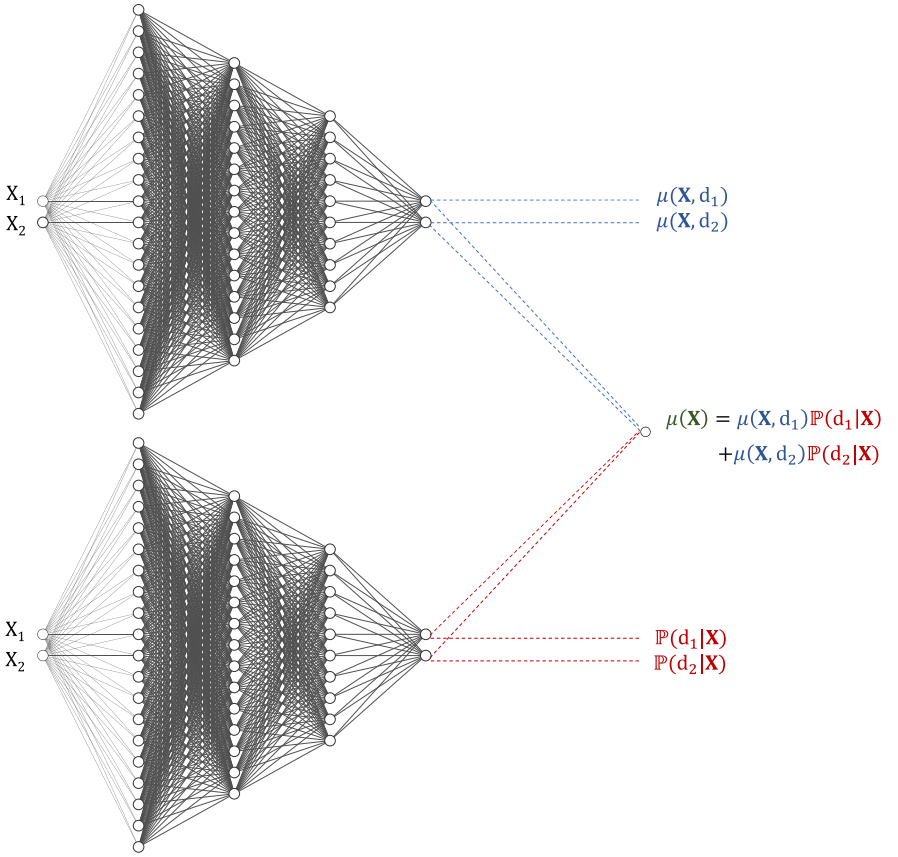

Fig. 3 Illustration of multi-task FNN architecture (19). The upper FNN has the same structure as the multi-output FNN (9) depicted in Fig. 2. The lower FNN is a classifier, giving the conditional probabilities of D, given $\mathbf{X} = \mathbf{x}$. These FNNs output $\mu(\mathbf{X}, d)$ and $\mathbb{P}(D = d|\mathbf{X})$ for $d \in \{d_1, d_2\}$, and they internally compute and output the unawareness price $\mu(\mathbf{X})$

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \Bigg[\sum_{k=1}^{K} L_\mu\Big(Y_i,\ \mu(\mathbf{X}_i, \mathbf{d}_k)\Big)\, \mathbb{1}_{\{\mathbf{D}_i = \mathbf{d}_k\}} \tag{20}$$
$$+ L_p\Big(\mathbf{D}_i,\ (p_k(\mathbf{X}_i))_{1\le k\le K}\Big)\, \mathbb{1}_{\{\mathbf{D}_i \neq \mathrm{NA}\}} + L_\mu\Big(Y_i,\ \mu(\mathbf{X}_i)\Big)\Bigg],$$

for given loss functions L_μ and L_p. This network fitting problem (20) provides the interaction between the two FNNs, and it only considers the parts of the loss function where the corresponding information is available; we discuss this in more detail in the following remarks.

Remark 4 • In comparison to (9), the multi-task FNN architecture (19) adds extra components to model the discriminatory information $\mathbf{D}$, given the non-discriminatory $\mathbf{X}$, and it directly computes the unawareness price $\mu(\mathbf{X})$, defined

through (18), see last term of (19). This model can be fitted by solving the optimization problem (20), which has three different sets of terms. The first terms $L_\mu(Y_i, \mu(\mathbf{X}_i, \mathbf{d}_k))$ are identical to the multi-output FNN fitting (13). The second terms $L_p(\mathbf{D}_i, (p_k(\mathbf{X}_i))_{1\le k\le K})$ fit a categorical classification problem, and we typically choose the multinomial cross-entropy loss for L_p. If we have complete data, and if we only considered these first two terms, we would fit two independent FNN architectures, modeling the best-estimates $\mu(\mathbf{X}, \mathbf{d}_k)$ and the classification probabilities $p_k(\mathbf{X})$, $1 \le k \le K$. The third loss $L_\mu(Y_i, \mu(\mathbf{X}_i))$ in (20) connects best-estimate fitting and categorical probability fitting. This connection should be understood as regularization, i.e., each of the three loss functions acts as a regularizer for the other estimation problems. This regularization takes place on all policies with $\mathbf{D}_i \neq \texttt{NA}$, and on the policies with missing discriminatory information only the (internally computed) unawareness price $\mu(\mathbf{X})$ can be fitted. This leads to robustified results compared to just dropping policies where discriminatory information is not available.

- Intuitively, one should scale the loss functions L_μ and L_p in (20) such that they live on a comparable scale. In our numerical examples, the results have shown little sensitivity in such a scaling. Therefore, we will just use the standard form of the deviance loss L_μ for mean modeling and the cross-entropy loss L_p for categorical probability modeling in our examples below, i.e., without any additional scaling.
- The last component of the objective function in (20) compares the (internally) calculated unawareness price $\mu(\mathbf{X})$ to the response Y. Alternatively, we could fit another regression model $\widetilde{\mu}(\mathbf{X})$ for the unawareness price. This can be done because it does not involve any protected information $\mathbf{D}$. A variant of (20) then replaces the last objective function in (20) as follows

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \Bigg[\sum_{k=1}^{K} L_\mu\Big(Y_i,\ \mu(\mathbf{X}_i, \mathbf{d}_k)\Big)\, \mathbb{1}_{\{\mathbf{D}_i = \mathbf{d}_k\}} + L_p\Big(\mathbf{D}_i,\ (p_k(\mathbf{X}_i))_{1\le k\le K}\Big)\, \mathbb{1}_{\{\mathbf{D}_i \neq \texttt{NA}\}} + L_{\widetilde{\mu}}\Big(\widetilde{\mu}(\mathbf{X}_i),\ \mu(\mathbf{X}_i)\Big)\Bigg], \tag{21}$$

  for given loss functions L_μ, $L_{\widetilde{\mu}}$ and L_p. This approach can be useful in certain situations of over-fitting, but it requires a high-quality model for $\widetilde{\mu}$ in order to outperform the fitting procedure (20).
- In our multi-task FNN we consider two parallel FNNs, see (15), and the connection is only considered by the joint parameter estimation in (20). We could also consider other network architectures where, e.g., we learn a common representation $\mathbf{z}^{(m:1)}(\mathbf{x})$ which serves to jointly construct the readouts of $\mu(\mathbf{x}, \mathbf{d}_k)$ and $p_k(\mathbf{x})$, $1 \le k \le K$. In our numerical experiments this latter approach was less competitive in terms of predictive power compared to the first one, but more work is required to come to a conclusive answer about the 'best' network architecture for this multi-task learning problem.
- The multi-task FNN (19) is solely used to model the discrimination-free insurance price through (5). There are many different notions and definitions of fairness that may complement discrimination-free insurance prices; see, e.g., Grari et al. [11]

and Lindholm et al. [15]. Multi-task learning (19)–(20) can be extended by such complementary fairness notions. This requires that the notion of the chosen fairness criterion can be encoded into a sensible score that can be added to the optimization (20), and depending on the quantities needed in this additional score, we may need to add corresponding outputs to the multi-task learning (19). As a result the multi-output network will be regularized by the corresponding scoring part that accounts for the selected notion of fairness. We emphasize that the selection of a reasonable fairness criterion needs careful consideration, as it may contradict other modeling constraints; for a broader discussion we refer to Lindholm et al. [15].

4 Synthetic health insurance example

We design a synthetic health insurance example that is similar to Lindholm et al. [14], but with a slightly more complicated underlying regression function. Working with a synthetic example, that is, knowing the true data generating model, has the advantage of being able to benchmark the estimated models to the ground truth.

4.1 Data generation

Let the discriminatory information $\mathbf{D} \in \{\text{female}, \text{male}\}$ be the gender of the policyholder. The non-discriminatory information $\mathbf{X} = (X_1, X_2)^\top$ is assumed to have two components, with $X_1 \in \{15, 16, \ldots, 80\}$ denoting the age of the policyholder and $X_2 \in \{\text{non-smoker}, \text{smoker}\}$ the smoking status of the policyholder. There are different claim types: claims that mainly affect females between ages 20 and 40 and males after age 60 (type 1), claims with a higher frequency for smokers and also for females (type 2), and general claims due to other disabilities (type 3). The logged expected frequencies of these claim types are given by

$$\begin{aligned}
\log \lambda_1(\mathbf{X}, \mathbf{D}) &= \alpha_0 + \alpha_1 \mathbb{1}_{\{X_1 \in [20,40],\, \mathbf{D}=\text{female}\}} + \alpha_2 \mathbb{1}_{\{X_1 \geq 60,\, \mathbf{D}=\text{male}\}}, \\
\log \lambda_2(\mathbf{X}, \mathbf{D}) &= \gamma_0 + \gamma_1 X_1 + \gamma_2 \mathbb{1}_{\{X_2=\text{smoker}\}} + \gamma_3 \mathbb{1}_{\{\mathbf{D}=\text{female}\}}, \\
\log \lambda_3(\mathbf{X}, \mathbf{D}) &= \delta_0 + \delta_1 X_1,
\end{aligned}$$

with the following parameters: $(\alpha_0, \alpha_1, \alpha_2) = (-40, 38.5, 38.5)$, $(\gamma_0, \gamma_1, \gamma_2, \gamma_3) = (-2, 0.004, 0.1, 0.2)$, and $(\delta_0, \delta_1) = (-2, 0.01)$. We set for the (true) total expected claim frequency

$$\lambda(\mathbf{X}, \mathbf{D}) = \lambda_1(\mathbf{X}, \mathbf{D}) + \lambda_2(\mathbf{X}, \mathbf{D}) + \lambda_3(\mathbf{X}, \mathbf{D}).$$

If we assume that the number of claims Y of an insurance policyholder with covariates $(\mathbf{X}, \mathbf{D}) = (\mathbf{x}, \mathbf{d})$ is Poisson distributed with expected value $\lambda(\mathbf{x}, \mathbf{d})$, then we obtain the (true) best-estimate price

$$\lambda(\mathbf{x}, \mathbf{d}) = \mathbb{E}\left[Y \mid \mathbf{X} = \mathbf{x}, \mathbf{D} = \mathbf{d}\right], \tag{22}$$

Table 1 Table of considered models fitting approaches

Label	Regression model
(a)	True model with regression function $\lambda(\mathbf{x}, \mathbf{d})$ given by (22)
(b)	Plain-vanilla FNN regression function $\mu(\mathbf{x}, \mathbf{d})$ given by (8)
(c)	Multi-output FNN regression function $\mu(\mathbf{x}, \mathbf{d}_k)$ given by (9)
(d)	Multi-task FNN (Y) regression function given by (19) and (20)
(e)	Multi-task FNN ($\widehat{\mu}$) regression function given by (19) and (21)
(f)	Regression tree boosting (best-estimate benchmark)

where, for simplicity, here we only focus on claim counts. Since in practice this true best-estimate price $\lambda(\mathbf{x}, \mathbf{d})$ is not known, we estimate it from the available data using a regression function denoted by $\mu(\mathbf{x}, \mathbf{d})$. We therefore first use a plain-vanilla FNN to model $\mu(\mathbf{x}, \mathbf{d})$ based on the full input $(\mathbf{X}, \mathbf{D}) = (\mathbf{x}, \mathbf{d})$, see (8). In the next steps we study the multi-output FNN (9) and the multi-task FNN (19), which only use the non-discriminatory information $\mathbf{X} = \mathbf{x}$ as a network input, while the available discriminatory information $\mathbf{D}$ is only used in the loss functions for model fitting; see (13), (20) and (21), respectively. Table 1 illustrates all models that we are going to consider, the labeling (a)–(f) will be kept throughout this example.

To fit these FNNs we first need to generate i.i.d. data $(Y_i, \mathbf{X}_i, \mathbf{D}_i)_{1 \le i \le n}$. We select a portfolio of sample size $n = 100{,}000$ as follows. The age variable X_1 is assumed to be independent of the smoking habits X_2 and the gender $\mathbf{D}$, and we choose the age distribution as given in Figure 4 of Lindholm et al. [14]. Moreover, we choose $\mathbb{P}[\mathbf{D} = \text{female}] = 0.45$, $\mathbb{P}[X_2 = \text{smoker}] = 0.3$ and $\mathbb{P}[\mathbf{D} = \text{female} \mid X_2 = \text{smoker}] = 0.8$. This fully specifies the distribution of the covariates $(\mathbf{X}, \mathbf{D})$, making smoking more common among females compared to males in this population. We simulate $n = 100{,}000$ independent insurance policies from this covariate distribution. This provides us with an empirical proportion of females in the simulated data of 0.4505, fairly close to the true ratio of 0.45. Later, we use this proportion for obtaining the pricing distribution, i.e., we set $\mathbb{P}^*[\mathbf{D} = \text{female}] = 0.4505$; as explained in Lindholm et al. [14], this is the canonical choice for $\mathbb{P}^*$. Finally, we simulate independent observations $Y_i|_{\{\mathbf{X}_i, \mathbf{D}_i\}} \sim \text{Poi}(\lambda(\mathbf{X}_i, \mathbf{D}_i))$, $1 \le i \le n$, giving the (pseudo-)sample $(Y_i, \mathbf{X}_i, \mathbf{D}_i)_{1 \le i \le n}$, representing the portfolio of policies. In the next section we assume the availability of full covariate information for the protected characteristics $\mathbf{D}_i$, whereas in Sect. 4.3 we will assume only partial access to such information.

4.2 Full availability of discriminatory information

4.2.1 Plain-vanilla feed-forward neural network

We start with the plain-vanilla FNN (8), which takes as input the covariates $(\mathbf{X}, \mathbf{D}) \in \mathbb{R}^{q_0}$ with $q_0 = 3$; we use dummy coding for both the gender variable $\mathbf{D}$ and the smoking habits X_2. We choose a network of depth $m = 3$ with $(q_1, q_2, q_3) = (20, 15, 10)$

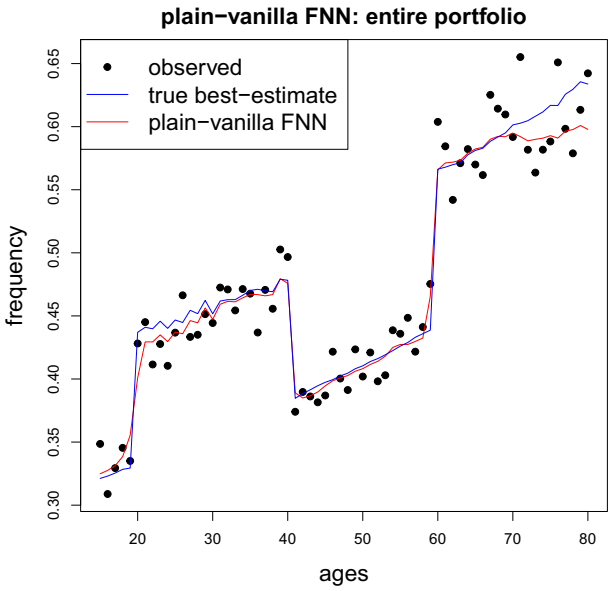

Fig. 4 Best-estimate price $\mu(\mathbf{x}, \mathbf{d})$ of the fitted plain-vanilla FNN (8) (red) compared to the true best-estimate price $\lambda(\mathbf{x}, \mathbf{d})$ (blue) as a function of the age variable x_1 and averaged over smoking habits x_2 and gender $\mathbf{d}$; these results are based on the complete knowledge of discriminatory information. The empirical estimates of the observations are represented by the black dots (color figure online)

hidden neurons in the three hidden layers, the ReLU activation function, and the log-link $g(\cdot) = \log(\cdot)$, which is the canonical link of the Poisson regression model. This network is illustrated in Fig. 1, and has a parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{w})$ of dimension 566, see (10). To implement this FNN we use the library `keras` [7] within the statistical computing software R [18].

We fit this FNN to the simulated data $(Y_i, \mathbf{X}_i, \mathbf{D}_i)_{1 \le i \le n}$; assuming for now access to the full policyholder information, including $\mathbf{D}$. We use the Poisson deviance loss function for L in (12), the `nadam` version of stochastic gradient descent, a batch size of 50 policies, and we explore early stopping based on a 80/20 training-validation split. This is similar to Section 7.3.2 in Wüthrich–Merz [24]; for more details we refer to that source. Since network fitting involves several elements of randomness, see Remark 7.7 in Wüthrich–Merz [24], we average over 10 different FNN calibrations, resulting in the nagging predictor of Richman–Wüthrich [19].

The results are given in Fig. 4. The blue line shows the true best-estimate price $\lambda(\mathbf{x}, \mathbf{d})$ as a function of the age variable $15 \le x_1 \le 80$, averaged over the smoking habits x_2 and the gender variable $\mathbf{d}$ w.r.t. the empirical population density. The black dots show the empirical estimates from the observations Y_i and the red line the plain-vanilla FNN fitted best-estimate price $\mu(\mathbf{x}, \mathbf{d})$ using the full inputs $(\mathbf{X}_i, \mathbf{D}_i)$. Overall, Fig. 4 suggests a rather accurate fit; only at the age boundaries there are some differences which are caused by the noise in the observations Y_i.

Since we know the true regression function $\lambda(\mathbf{x}, \mathbf{d})$, we can explicitly quantify the accuracy of the estimated FNN regression function $\mu(\mathbf{x}, \mathbf{d})$. That is, we do not need

Table 2 Model accuracy of the fitted FNNs and a regression tree boosting model serving as a benchmark; the KL divergences are stated in 10^{-3}

	KL divergence (24) to $\lambda(\mathbf{x}, \mathbf{d})$
(b) Plain-vanilla FNN: full data	0.2204
(c) Multi-output FNN: full data	0.2567
(d) Multi-task FNN (Y): full data	0.2823
(e) Multi-task FNN ($\widehat{\mu}$): full data	0.3070
(f) Regression tree boosting: full data	0.5170

to validate the estimated model on a test data sample, but we can directly compare it to the true model. We use the Kullback–Leibler (KL) divergence to compare the estimated model to the true one. In the case of the Poisson model the KL divergence for given covariate values $(\mathbf{x}, \mathbf{d})$ is given by

$$\begin{aligned} D_{\mathrm{KL}}\big(\lambda(\mathbf{x}, \mathbf{d})\big|\big|\mu(\mathbf{x}, \mathbf{d})\big) &= \sum_{y \in \mathbb{N}_0} e^{-\lambda(\mathbf{x},\mathbf{d})} \frac{\lambda(\mathbf{x}, \mathbf{d})^y}{y!} \log\left(\frac{e^{-\lambda(\mathbf{x},\mathbf{d})} \frac{\lambda(\mathbf{x},\mathbf{d})^y}{y!}}{e^{-\mu(\mathbf{x},\mathbf{d})} \frac{\mu(\mathbf{x},\mathbf{d})^y}{y!}} \right) \\ &= \mu(\mathbf{x}, \mathbf{d}) - \lambda(\mathbf{x}, \mathbf{d}) - \lambda(\mathbf{x}, \mathbf{d}) \log\left(\frac{\mu(\mathbf{x}, \mathbf{d})}{\lambda(\mathbf{x}, \mathbf{d})} \right). \end{aligned} \tag{23}$$

We average the KL divergence of a single instance $(\mathbf{x}, \mathbf{d})$ over the empirical population distribution, which gives us the KL divergence from the estimated model to the true model on our portfolio

$$D_{\mathrm{KL}}\left(\lambda||\mu\right) := \frac{1}{n} \sum_{i=1}^{n} D_{\mathrm{KL}}\big(\lambda(\mathbf{x}_i, \mathbf{d}_i)\big|\big|\mu(\mathbf{x}_i, \mathbf{d}_i)\big). \tag{24}$$

This gives us a measure of model accuracy for the different estimated models; the KL divergence is zero if and only if the estimated model is identical to the true model on the selected portfolio; see Section 2.3 in Wüthrich–Merz [24].

Row (b) of Table 2 shows the KL divergence (24) of the fitted plain-vanilla FNN best-estimate price $\mu(\mathbf{x}, \mathbf{d})$ to the true best-estimate price $\lambda(\mathbf{x}, \mathbf{d})$. The resulting KL divergence is 0.2204×10^{-3}, which is much smaller than a comparable regression tree boosting model that results in a KL divergence of 0.5170×10^{-3}, shown in row (f) of Table 2.

This fitted FNN can now be used for discrimination-free insurance pricing according to formula (5) for the given selected measure $\mathbb{P}^*$. However, we cannot calculate the unawareness price $\mu(\mathbf{x})$ from (3) because this requires the knowledge of the probabilities $(p_k(\mathbf{x}))_{1 \le k \le K}$, see (4). Alternatively, we could directly fit a plain-vanilla FNN for estimating the unawareness price or we could fit a conditional expectation network as proposed in Richman–Wüthrich [20], routes we do not pursue here. We come back to this topic when discussing the multi-task FNN.

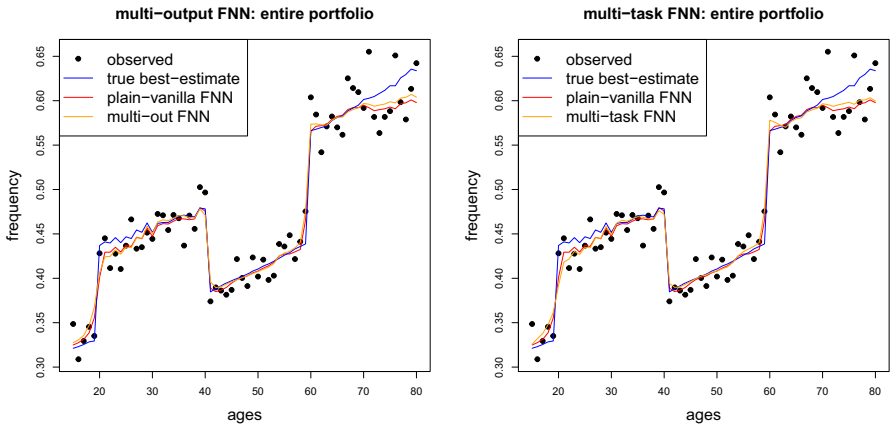

Fig. 5 Best-estimate prices $(\mu(\mathbf{x}, \mathbf{d}_k))_{1\leq k\leq K}$ of the fitted multi-output FNN (9) (lhs) and the fitted multi-task FNN (19) with loss (20) (rhs) (orange); compared to the estimated prices of the plain-vanilla FNN (red) the true best-estimates (blue) and the empirical means (black). The estimates are based on complete knowledge of discriminatory information (color figure online)

4.2.2 Multi-output feed-forward neural network

Next, we fit the multi-output FNN (9) to the same data, only using the non-discriminatory covariates $\mathbf{x}$ as input to the network. This reduces the input dimension to $q_0 = 2$, but on the other hand we have two outputs $\mu(\mathbf{x}, \mathbf{d} = \text{female})$ and $\mu(\mathbf{x}, \mathbf{d} = \text{male})$ in the multi-output FNN. The former reduces the dimension of the network parameter and the latter increases the dimension of the network parameter, resulting in a network parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}_{\text{female}}, \boldsymbol{\beta}_{\text{male}}, \mathbf{w})$ of dimension 557, see (11); this network is illustrated in Fig. 2. We fit this multi-output FNN using exactly the same fitting strategy as above. The results are presented in orange color in Fig. 5 (lhs), and they are compared to the plain-vanilla FNN best-estimates in red color, the true best-estimates in blue color and the empirical observations in black color. We conclude that the two networks (orange and red) provide rather similar results.

Table 2 shows a slightly higher accuracy of the plain-vanilla FNN (row (b)) compared to the multi-output FNN (row (c)), with respective KL divergence values of 0.2204×10^{-3} vs. 0.2567×10^{-3}. In general, these numbers are quite small and the models are rather accurate, which is in support of both FNN models.

4.2.3 Multi-task feed-forward neural network

We now fit the multi-task FNN (19) to the data, still assuming full knowledge of the discriminatory information $\mathbf{D}_i$. The multi-task FNN additionally models the discriminatory covariates $\mathbf{D}_i$ which are then (internally) used to calculate the unawareness price $\mu(\mathbf{x})$, see (18)–(19). That is, in contrast to the plain-vanilla FNN and the multi-output FNN, this is the only one of the three approaches that allows us to directly calculate the unawareness prices within the *same* model as the best-estimate prices.

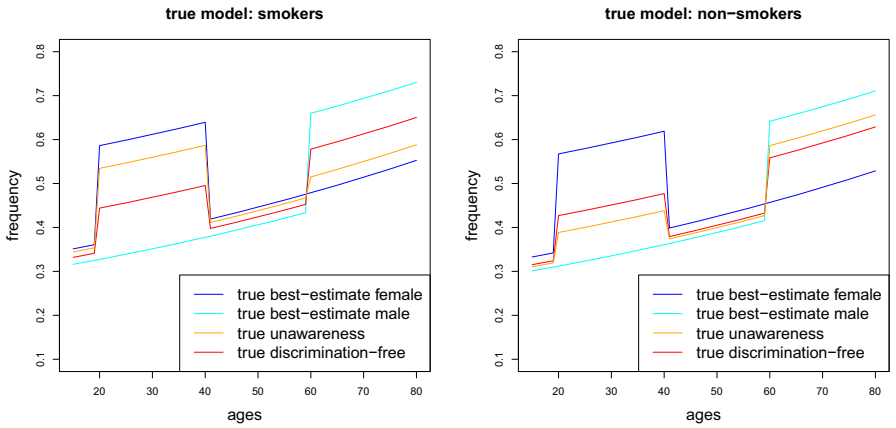

Fig. 6 True model: best-estimate prices $\lambda(\mathbf{x}, \mathbf{d})$, unawareness prices $\lambda(\mathbf{x})$ and discrimination-free insurance prices $\lambda^*(\mathbf{x})$ with (lhs) smokers, and (rhs) non-smokers

We start with objective function (20) considering the response Y for assessing the unawareness price $\mu(\mathbf{x})$.

We use exactly the same fitting strategy as in the previous two modeling approaches. Figure 5 (rhs) shows the resulting best-estimate prices (in orange color) compared to the ones of the plain-vanilla FNN (in red color). Row (d) of Table 2 provides a resulting KL divergence to the true model of 0.2823×10^{-3}. This is slightly higher than in the other two approaches, but still gives a very competitive result. The full advantage of the multi-task FNN approach will become clear once we start to work with incomplete discriminatory information in Sect. 4.3 below.

Finally, we present the fitting results of the multi-task FNN (19) when using objective function (21). For this we first fit a plain-vanilla FNN $\widetilde{\mu}(\mathbf{x})$ to the unawareness price (only considering $\mathbf{x}$). This is done completely analogously to Sect. 4.2.1 except that we drop the discriminatory information from the input. We then use this fitted FNN $\widetilde{\mu}(\mathbf{x})$ in objective function (21), and we use the KL divergence (23) for $L_{\widetilde{\mu}}$ to measure the divergence from $\mu(\mathbf{x})$ to $\widetilde{\mu}(\mathbf{x})$. The results are presented on row (e) of Table 2. We observe that this is the least accurate of all FNN models, the main issue probably being that the regression function $\widetilde{\mu}(\mathbf{x})$ is not sufficiently accurate, and we should rather directly compare the unawareness price $\mu(\mathbf{x})$ to the observations Y as done in (20). For this reason, we will not further pursue this approach below.

4.2.4 Discrimination-free insurance pricing

Having fitted the three FNNs, we can calculate the discrimination-free insurance prices $\mu^*(\mathbf{x})$ by (14), using the empirical gender distribution $\mathbb{P}^*[\mathbf{D} = \text{female}] = 0.4505$ as the pricing distribution. We start by considering the true model $\lambda(\mathbf{x}, \mathbf{d})$. We calculate the discrimination-free insurance prices $\lambda^*(\mathbf{x})$ and the unawareness prices $\lambda(\mathbf{x})$ in the true model. This can be done because all necessary information is available. The corresponding graphs are shown in Fig. 6, and they will serve as a benchmark for the

Table 3 Model accuracy of the unawareness prices $\mu(\mathbf{x})$ and the discrimination-free insurance prices $\mu^*(\mathbf{x})$; the KL divergences to the best-estimate $\lambda(\mathbf{x}, \mathbf{d})$ are stated in 10^{-3}

	KL divergence (24) to $\lambda(\mathbf{x}, \mathbf{d})$
(a0) True unawareness price $\lambda(\mathbf{x})$	6.3174
(d0) Multi-task FNN (Y) unawareness price $\mu(\mathbf{x})$	6.4932
(a1) True discrimination-free price $\lambda^*(\mathbf{x})$	7.8857
(b1) Plain-vanilla FNN discrimination-free price $\mu^*(\mathbf{x})$	8.3222
(c1) Multi-output FNN discrimination-free price $\mu^*(\mathbf{x})$	8.2669
(d1) Multi-task FNN (Y) discrimination-free price $\mu^*(\mathbf{x})$	8.2915

The figures are based on the full knowledge of discriminatory information

estimated FNNs. It is seen that the unawareness prices $\lambda(\mathbf{x})$ closely follow the best-estimate prices $\lambda(\mathbf{x}, \mathbf{d})$ of females for smokers, and the best-estimate prices of males for non-smokers. This reflects the fact that, in our example, smoking habits are rather informative for predicting gender. The discrimination-free insurance prices $\lambda^*(\mathbf{x})$ exactly correct for this inference potential: as seen from Fig. 6, while smokers of either gender have higher predicted claims frequencies than non-smokers, discrimination-free insurance prices lie between the best-estimate prices for males and females, following the same pattern regardless of smoking status.

Table 3 presents the corresponding numerical results. The unawareness price $\lambda(\mathbf{x})$ of the true model has a KL divergence to the best-estimate price $\lambda(\mathbf{x}, \mathbf{d})$ of 6.3174×10^{-3}, see row (a0). That is, we sacrifice quite some predictive accuracy by ignoring the protected information $\mathbf{D}_i$ in the unawareness price $\lambda(\mathbf{x})$. This approach still internally infers the gender from the non-discriminatory information, see (4). Breaking this link further deteriorates the predictive accuracy, resulting in a KL divergence from the discrimination-free insurance price $\lambda^*(\mathbf{x})$ to the best-estimate price $\lambda(\mathbf{x}, \mathbf{d})$ of 7.8857×10^{-3}, see Table 3, row (a1).

Furthermore, Table 3 provides all KL divergences to the true best-estimate price $\lambda(\mathbf{x}, \mathbf{d})$ that can be calculated from the three fitted FNNs: row (b1) considers the discrimination-free insurance price $\mu^*(\mathbf{x})$ in the plain-vanilla FNN (6), row (c1) the discrimination-free insurance price $\mu^*(\mathbf{x})$ in the multi-output FNN (9) and rows (d0) and (d1) the unawareness price $\mu(\mathbf{x})$ and the discrimination-free insurance price $\mu^*(\mathbf{x})$ in the multi-task FNN (19) using objective function (20). The last FNN is the only one that directly provides the unawareness price $\mu(\mathbf{x})$. The accuracy of the resulting discrimination-free insurance prices $\mu^*(\mathbf{x})$ is rather similar between the three network approaches (KL divergences on rows (b1)–(d1) of Table 3). Clearly we sacrifice quite some predictive power by not being allowed to use the protected information $\mathbf{D}_i$, such that the KL divergences increase from 0.25×10^{-3} in Table 2 to roughly 8×10^{-3} in Table 3.

Table 4 Model accuracy of the discrimination-free insurance prices $\mu^*(\mathbf{x})$ relative to the true discrimination-free insurance price $\lambda^*(\mathbf{x})$; the KL divergences are stated in 10^{-3} and these figures are based on the full knowledge of discriminatory information

	KL divergence (24) to $\lambda^*(\mathbf{x})$
(b2) Plain-vanilla FNN discrimination-free price $\mu^*(\mathbf{x})$	0.1748
(c2) Multi-output FNN discrimination-free price $\mu^*(\mathbf{x})$	0.1885
(d2) Multi-task FNN discrimination-free price $\mu^*(\mathbf{x})$	0.2323

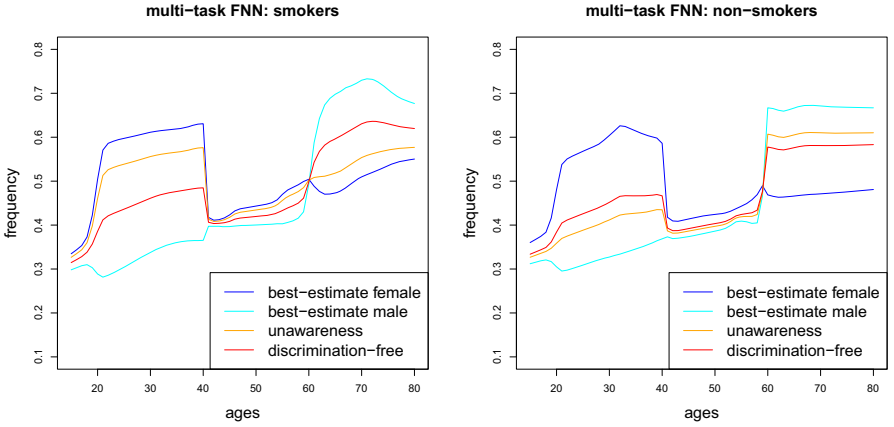

Fig. 7 Multi-task FNN: best-estimate prices $\mu(\mathbf{x}, \mathbf{d})$, unawareness prices $\mu(\mathbf{x})$ and discrimination-free insurance prices $\mu^*(\mathbf{x})$ with (lhs) smokers, and (rhs) non-smokers

Table 4 compares in KL divergence the discrimination-free insurance prices $\mu^*(\mathbf{x})$ of the three fitted FNNs to the true discrimination-free insurance price $\lambda^*(\mathbf{x})$. The plain-vanilla FNN provides slightly more accurate results compared to the multi-output and the multi-task FNNs. Note that the different rankings in Tables 3 and 4 may be caused by the randomness in the data and by potentially different over- or under-fitting to the data. A verification of the explicit reasons for these different rankings is difficult, and different samples may also change this order.

Finally, Fig. 7 illustrates the resulting prices from the multi-task FNN. They should be compared to the true ones in Fig. 6. The interpretation is the same in the two figures. Comparing the two plots we can also clearly see the impact of model uncertainty in Fig. 7, which can only be mitigated by having larger sample sizes.

Table 5 Losses of model accuracy using the unawareness price and the discrimination-free insurance price instead of the best-estimate price; the KL divergences are stated in 10^{-3}

	KL divergences to the best-estimates:	
	Unawareness	Discrimination-free
(a) True model	6.3174	7.8857
	100%	125%
(d) Multi-task FNN	6.7980	8.5339
	100%	126%

4.2.5 Quantification of direct and indirect discrimination

The results of Tables 2, 3 and 4 give motivation to quantify the potential for direct and indirect discrimination. In contrast to these former results, we now compare the unawareness price $\mu(\mathbf{x})$ and the discrimination-free insurance price $\mu^*(\mathbf{x})$ to the best-estimate price $\mu(\mathbf{x}, \mathbf{d})$ *within* a given model. The step from the best-estimate price $\mu(\mathbf{x}, \mathbf{d})$ to the unawareness price $\mu(\mathbf{x})$ accounts for direct discrimination by quantifying the effect of simply being blind w.r.t. the protected information $\mathbf{D}$. Going from the unawareness price $\mu(\mathbf{x})$ to the discrimination-free insurance price $\mu^*(\mathbf{x})$ accounts for indirect discrimination. However, these steps are more subtle for several reasons. First, the KL divergence does not satisfy the triangle inequality and, hence, divergences cannot simply be decomposed along a certain path. Second, in some models we cannot simultaneously calculate the best-estimate, the unawareness and the discrimination-free insurance prices, but need to use different models to estimate these quantities. This applies, e.g., to the multi-output FNN where we do not receive the unawareness price within that network model, but we have to explore another (separate) model to estimate this unawareness price. This critical point does not apply to the multi-task FNN, where we consistently calculate all terms within the same model. Third, interpreting the potential for indirect discrimination more broadly, there are two ingredients, namely, the inference part $\mathbb{P}[\mathbf{D} = \mathbf{d}|\mathbf{X}]$ and the best-estimate prices $\mu(\mathbf{x}, \mathbf{d})$. Only if both of them are sufficiently imbalanced, indirect discrimination becomes relevant (and visible). If we think of different insurance companies selling the same product (with the same underwriting standards and the same claim costs), these companies should use the same best-estimate prices $\mu(\mathbf{x}, \mathbf{d})$. Indirect discrimination will typically differ between these companies, because they will generally have different portfolio distributions $\mathbb{P}(\mathbf{X}, \mathbf{D})$, resulting in different inference potentials. Thus, the statistical dependence (association) between $\mathbf{X}$ and $\mathbf{D}$ is company-specific, and so is the amount of indirect discrimination.

Table 5 reflects the loss of model accuracy if we deviate from the best-estimate price. These losses can be interpreted as the quantification of direct and indirect discrimination. If we normalize the KL divergence of the unawareness price to 100%, then the discrimination-free insurance price adds another 25% to the loss of model accuracy compared to the unawareness price. Thus, direct discrimination is clearly the dominant term, here. Nonetheless, we note that the numbers in Table 5 reflect portfolio

Table 6 Empirical female ratio $\mathbb{P}(\mathbf{d})$ and multi-task FNN estimated female ratio $\widehat{\mathbb{P}}(\mathbf{d})$ under the chosen drop-out rates

Drop-out rate	90%	80%	70%	60%	50%	40%	30%	20%	10%	0%
Available $\mathbf{D}_i$	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Empirical $\mathbb{P}(\mathbf{d})$	44.8%	44.5%	44.3%	44.5%	44.6%	45.5%	45.4%	45.2%	45.1%	45.1%
Multi-task $\widehat{\mathbb{P}}(\mathbf{d})$	45.8%	44.8%	44.3%	44.5%	44.7%	45.5%	45.4%	45.2%	45.0%	45.1%

considerations, and the impact of indirect discrimination on particular sub-populations or individual policies may be bigger.

4.3 Partial availability of discriminatory information

4.3.1 Missing completely at random

So far, all numerical results have been based on the full knowledge of the data $(Y_i, \mathbf{X}_i, \mathbf{D}_i)_{1\le i\le n}$. Next, we turn our attention to the problem of having incomplete discriminatory information, and we analyze how well we can fit our FNNs under this partial information setting. We therefore randomly remove $\mathbf{D}_i$ from the information set. This is done by independently (across the entire portfolio) setting $\mathbf{D}_i = \texttt{NA}$ with increasing (drop-out) rates of 10%, 20%, ..., 90%. That is, in the last case only roughly 10% of the discriminatory labels $\mathbf{D}_i$ are available, and in the first case roughly 90% of all discriminatory labels are available. Working with these drop-outs also changes the empirical female ratio (that can only be calculated on the policies with full information). We state these in Table 6 (row 'empirical $\mathbb{P}(\mathbf{d})$') as we use them for the pricing measure $\mathbb{P}^*(\mathbf{d})$.

We use these datasets with drop-outs (incomplete protected information) to perform two different model fittings. Firstly, in the more naive approach, we just fit a plain-vanilla FNN (6) only using those observations for which the discriminatory information is available and we discard all insurance policies with incomplete information (case deletion). Thus, if, e.g., the drop-out probability is 80% we only use the remaining 20% of the data for model fitting, for which the discriminatory information $\mathbf{D}_i$ is available. Secondly, this naive approach is challenged by a multi-task FNN (19) fitted with the loss function (20), which accounts for partial availability of discriminatory information, but uses the entire portfolio for model fitting. Note that at this stage we do not perform any hyperparameter tuning and we use the identical network architectures across all chosen drop-out rates.

The results are presented in Fig. 8. This figure shows the KL divergences to the fitted best-estimate prices $\mu(\mathbf{x}, \mathbf{d})$ to the true best-estimate price $\lambda(\mathbf{x}, \mathbf{d})$; the dotted horizontal lines illustrate the results of Table 2 reflecting the case of full discriminatory information. We observe that if sufficient discriminatory information $\mathbf{D}_i$ is available, then we have a similar performance between the plain-vanilla FNN and the multi-task FNN. However, below a critical amount of discriminatory information (the vertical dotted line in Fig. 8 is at 50%) we give a clear preference to the multi-task FNN. In

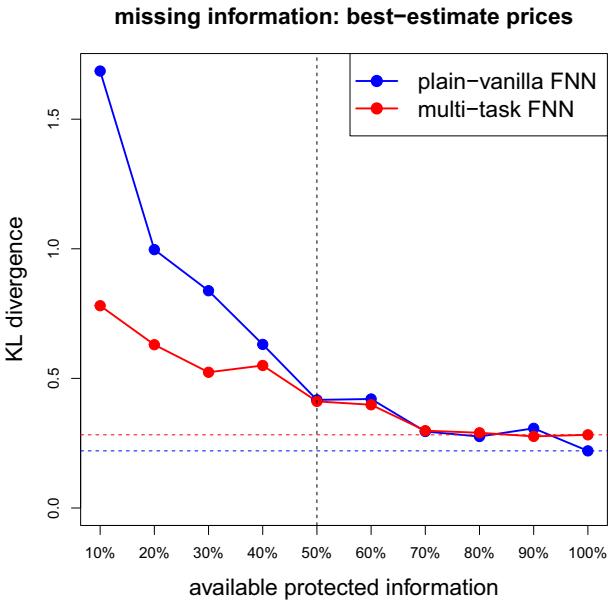

Fig. 8 Comparison of the (naive) plain-vanilla FNN and the multi-task FNN under missing discriminatory information: KL divergences from the fitted best-estimate prices $\mu(\mathbf{x}, \mathbf{d})$ to the true best-estimates $\lambda(\mathbf{x}, \mathbf{d})$; scale on y-axis is in 10^{-3}

fact, if only 10% or 20% of the protected information is available, the KL divergence from the true model is much smaller for the multi-task FNN compared to the naive plain-vanilla FNN. On the other hand, for drop-outs of less than 40% of the protected information we would probably prefer the simpler (naive) FNN architecture because it is comparably good to the more sophisticated multi-task FNN. These conclusions concern the case of best-estimate estimation, and not performing special hyperparameter tuning.

This preference for the multi-task FNN carries over to the discrimination-free insurance prices. In Fig. 9 we illustrate the KL divergences from the estimated discrimination-free insurance prices $\mu^*(\mathbf{x})$ to the true discrimination-free insurance price $\lambda^*(\mathbf{x})$; the case of full discriminatory information is denoted by the horizontal dotted lines and corresponds to rows (b2) and (d2) of Table 4. We observe smaller KL divergences of the multi-task FNN approach if we have discriminatory information on less than 50% of the policies (vertical black dotted line). Furthermore, we notice that in the multi-task FNN case, the KL divergence deteriorates only mildly, when the availability of discriminatory information decreases beyond the 50% point.

The discrimination-free insurance prices of Fig. 9 have simply used the empirical estimates for $\mathbb{P}(\mathbf{d})$ from the insurance policies where full information is available, see row 'empirical $\mathbb{P}(\mathbf{d})$' of Table 6. Having the fitted multi-task FNN we can also use the estimated categorical probabilities $p_k(\mathbf{X}_i)$, $1 \leq k \leq K$, from (19) to estimate the distribution of the protected covariates. Namely, we get an estimate

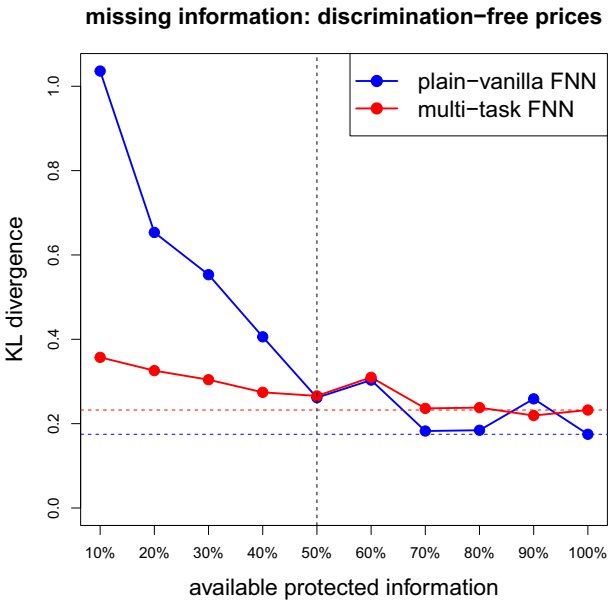

Fig. 9 Comparison of the (naive) plain-vanilla FNN and the multi-task FNN under missing discriminatory information: KL divergences from the fitted discrimination-free insurance prices $\mu^*(\mathbf{x})$ to the true discrimination-free insurance prices $\lambda^*(\mathbf{x})$; scale on y-axis is in 10^{-3}

$$\widehat{\mathbb{P}}(\mathbf{d}_k) = \frac{1}{n}\sum_{i=1}^{n} p_k(\mathbf{X}_i) \qquad \text{for } 1 \le k \le K. \tag{25}$$

These estimates (in the binary gender case $K = 2$) are presented on the last row of Table 6. We observe that they match the empirical estimates, and only for high drop-out rates there are some differences to the empirical estimates. The specific choice only has a marginal influence on the prices.

One may argue that the above comparison between the multi-task FNN and the naive plain-vanilla FNN in case of missing information is not completely fair. Since the naive plain-vanilla FNN uses less data, we should also fit a smaller FNN architecture in this latter case, i.e., we should decrease the depth m of the FNN architecture and/or the numbers of neurons q_j in the hidden layers. This would indeed be the classical view from statistics. However, for complex machine learning models it is not so clear that the above reasoning holds. Networks are fitted with stochastic gradient descent using early stopping to prevent over-fitting. Typically, the effective dimension of such fitted networks is much smaller than the number of parameters involved; the effective dimension is a way of quantifying the number of active parameters in a predictive model, see Abbas et al. [1]. Therefore, one could also argue that making the architecture a priori smaller for the plain-vanilla FNN, disadvantages this approach because it decreases the space of finding an optimal model.

For computational reasons, we have restricted our analysis to a drop-out rate of 80%, i.e., we assume availability of discriminatory information $\mathbf{D}_i$ on only 20% of

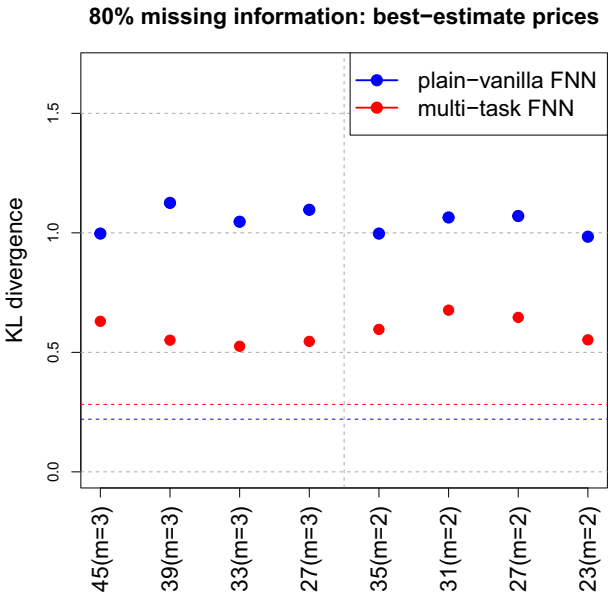

Fig. 10 Comparison of the (naive) plain-vanilla FNN and the multi-task FNN under 80% of missing discriminatory information: best-estimate prices for different FNN architectures, the x-axis shows the total number of chosen hidden neurons $q_1 + \cdots + q_m$ and the depth $m \in \{2, 3\}$ is given in brackets; the dotted lines show the results under full information, scale on y-axis is in 10^{-3}

the portfolio. We have fitted smaller networks than the one considered in Fig. 8 to this incomplete data. First, we have kept the depth $m = 3$ of the networks fixed, and we have decreased the numbers of neurons (q_1, q_2, q_3) in these three hidden layers. We have fitted these reduced architectures both to the naive plain-vanilla FNN case, see Fig. 1, and also to the multi-task FNN case, see Fig. 3. Second, we have decreased the depth to $m = 2$ and starting from $(q_1, q_2) = (20, 15)$ hidden neurons, we have chosen this and smaller architectures in the hidden neurons. The results are shown in Fig. 10. The first result with $q_1 + q_2 + q_3 = 45$ and $m = 3$ corresponds to Fig. 8 (with drop-out rate 80%). We observe robustness of the results across different (smaller) network architectures. From this we conclude that the comparison in Fig. 8 cannot be considered unfair, and the magnitude of outperformance of the multi-task FNN against the plain-vanilla FNN under high drop-out rates is rather stable across different architectures.

4.3.2 Not missing completely at random

In the last example, illustrated by Figs. 8 and 9, discriminatory information $\mathbf{D}_i$ was removed completely at random, i.e., $\mathbf{D}_i$ was set to `NA` by an i.i.d. Bernoulli random variable with a fixed drop-out rate. However, it might be that the missingness of protected information is not completely independent from the remaining covariates $\mathbf{X}_i$. Of course, there are many different ways in which this could happen, and we just provide here one particular example. We take as a baseline the i.i.d. Bernoulli case

Table 7 Not completely missing at random: resulting overall drop-out rates, and empirical female ratios $\mathbb{P}(\mathbf{d})$ and multi-task FNN estimated female ratio $\widehat{\mathbb{P}}(\mathbf{d})$

	Missing at random	Not missing at random	
Drop-out rate on $\mathcal{M}$	70%	80%	90%
Overall drop-out rate	70%	78%	86%
Empirical $\mathbb{P}(\mathbf{d})$	44.3%	42.4%	40.3%
Multi-task $\widehat{\mathbb{P}}(\mathbf{d})$	44.3%	45.0%	44.9%

Table 8 Model accuracy of the discrimination-free insurance prices $\mu^*(\mathbf{x})$ relative to the true discrimination-free insurance price $\lambda^*(\mathbf{x})$ where the drop-out rate of the gender information is not missing completely at random; the KL divergences are stated in 10^{-3}; the resulting overall drop-out rates are given in Table 7

	KL divergence (24) to $\lambda^*(\mathbf{x})$		
	Missing at random	Not missing at random	
Drop-out rate on $\mathcal{M}$	70%	80%	90%
(b3) Plain-vanilla FNN discrimination-free	0.5532	0.8153	0.9306
(d3) Multi-task FNN discrimination-free	0.3042	0.3698	0.3750

$\mathbf{D}_i = \texttt{NA}$ with a drop-out probability of 70%. We then modify this case by choosing a higher drop-out rate for the gender information on the policies $\mathcal{M} = \{X_1 \leq 45, X_2 = \text{smoker}\}$. Besides the base case of 70%, we choose the drop-out rates of 80% and 90% on $\mathcal{M}$. Note that $\mathcal{M}$ collects the smokers with ages below 45, and by the choice of our population distribution, 79.9% on this sub-portfolio are female.

Table 7 shows the resulting overall drop-out rates. These drop-out rates are no longer missing completely at random, because we have higher drop-out rates on $\mathcal{M}$. In our case, this results in a biased empirical gender estimate. This can be seen from the row 'empirical $\mathbb{P}(\mathbf{d})$' which simply calculates the empirical female ratio on the policies where full information is available. Having the fitted multi-task FNN, we can also estimate the female ratio using the categorical probability estimates of $p_k(\mathbf{X}_i)$, see (25). This provides us with the results on the last row of Table 7. We observe that these estimates are close to unbiased, the true value being 45%. We can use this multi-task FNN to estimate $\widehat{\mathbb{P}}(\mathbf{d})$ as pricing measure for the discrimination-free insurance price in the case of data not missing completely at random.

Table 8 and Fig. 11 present the results. The base case is the i.i.d. Bernoulli drop-out case with a drop-out rate of 70%, taken from Fig. 9. This base case is modified to a higher drop-out rate of 80% and 90%, respectively, on sub-portfolio $\mathcal{M}$; see Table 7 for the resulting overall drop-out rates. For the pricing measure $\mathbb{P}^*(\mathbf{d})$ we choose

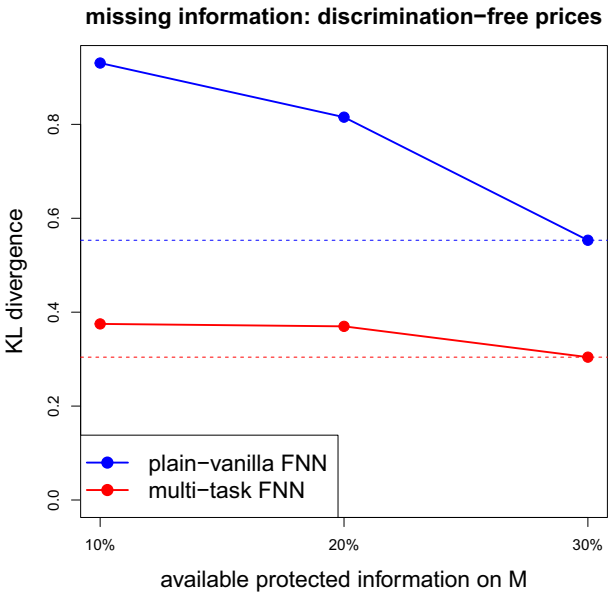

Fig. 11 Comparison of the (naive) plain-vanilla FNN and the multi-task FNN where the drop-out rate is not missing completely at random: KL divergences from the fitted discrimination-free insurance prices $\mu^*(\mathbf{x})$ to the true discrimination-free insurance prices $\lambda^*(\mathbf{x})$; scale on y-axis is in 10^{-3}; the x-axis gives the available discriminatory information $\mathbf{D}_i$ on $\mathcal{M}$ and the resulting overall drop-out rates are given in Table 7

the empirical probability $\mathbb{P}(\mathbf{d})$ in the plain-vanilla FNN case and the multi-task FNN estimate $\widehat{\mathbb{P}}(\mathbf{d})$ in the multi-task FNN case, see Table 7. At first sight, this does not seem to be an entirely fair comparison because the former estimates are biased. However, there is no simple way in the plain-vanilla FNN case to receive better gender estimates, whereas in the multi-task FNN case we obtain these better estimates as an integral part of the prediction model. In this not missing completely at random example we arrive at the the same conclusion, namely, that the multi-task FNN shows superior performance. In fact, in our example, the performance of the multi-task FNN in the situation of not missing completely at random is comparable to the situation of missing completely at random. We remark that even if we would use the same biased gender estimates also in the multi-task FNN to calculate the discrimination-free insurance prices we would come to the same conclusion.

Remark 5 In Remark 1 we have discussed that discrimination may also result from the fact that a certain sub-population is under-represented in the data and, hence, we may have a poorly fitted model on this part of the covariate space. This form of discrimination is directly related to incomplete data not missing completely at random, which is relevant when fitting the best-estimate price $\mu(\mathbf{x}, \mathbf{d})$. The present section has shown that the multi-task FNN can help to improve model accuracy if under-representation is caused by missing discriminatory information. However, if the sub-population is under-represented per se, e.g., there are only few elderly female

smokers in the portfolio, then this multi-task FNN cannot resolve the fundamental class imbalance problem.

5 Non-life insurance pricing example

In this section, we apply the multi-task approach to a real-world non-life motor insurance dataset. It consists of policyholder, vehicle and claims information for a large portfolio observed over a single year of exposure. These historical data, relating to a period close to the turn of the century, were contributed by a large multi-national insurance company. In this case, the true underlying data generating process is unknown. Thus, we follow a slightly different approach to Sect. 4, by comparing the best-estimate $\mu(\mathbf{x}, \mathbf{d})$, unawareness $\mu(\mathbf{x})$ and discrimination-free $\mu^*(\mathbf{x})$ insurance prices derived using the plain-vanilla and multi-task FNN approaches to the actual observed claim frequencies Y_i.

5.1 Data description

The historical non-life data contain 165,511 years of exposure of comprehensive motor insurance policies, with 41,608 claims arising in the same period of exposure. The claims can be assumed to be fully run-off, i.e., no incurred but not reported adjustments need to be made, and they relate to property (motor-hull), liability (third-party property and/or bodily injury) and other associated coverages. The protected information $\mathbf{D} \in D_5 = \{1, 2, 3, 4, 5\}$ refers to the ethnicity code of the policyholder, which was recorded by the insurer at the time of underwriting the policy in one of the five different ethnicity categories (as defined in the jurisdiction in which the policies were written). This information is not used by the insurer for pricing, but it is collected to enable monitoring of insurance penetration.

We make some remarks regarding protecting confidentiality and not disclosing commercially sensitive information. To preserve confidentiality, we do not provide the ethnicity categories that relate to the levels of $\mathbf{D}$, but only provide an integer code which has been assigned to each category. Moreover, the jurisdiction in which this data was observed is not disclosed. The exact coverages and excesses (which influence the frequency) in this portfolio have also not been disclosed, thus, the information shown is not commercially useful. Furthermore, we show only the predictions for a single component in $\mathbf{X}$, the policyholder age X_1 (although we fit the models using all the components of $\mathbf{X}$), whereas for commercial pricing, knowledge of all the variables in $\mathbf{X}$ is needed to predict meaningfully the frequency in a different portfolio. Finally, the year in which the experience was observed is not disclosed; since frequency changes quite rapidly over time based on the portfolio composition and, moreover, the jurisdiction in which the policies are written, the experience shown is not representative of a portfolio with similar policies at the current time or another jurisdiction, thus, the experience is not useful for commercial pricing.

Table 9 shows a summary of the claims, exposures and frequencies for each ethnicity code $\mathbf{D} \in D_5$ in the data.

Table 9 Summary of real-world non-life insurance claims, exposures and frequencies of claims according to the ethnicity codes $\mathbf{D} \in D_5$

Ethnicity code	Number of claims	Exposure	Frequency
1	5223	14,317	36.48%
2	965	3925	24.59%
3	3354	14,363	23.35%
4	5249	20,240	25.93%
5	26,817	112,667	23.80%

The largest group of policyholders consists of those with $\mathbf{D} = 5$, whereas the other groups range from very small ($\mathbf{D} = 2$) to just under a fifth of the size of the largest group ($\mathbf{D} = 4$). The observed claims frequencies are similar for most of the groups, with the exception being those policyholders with $\mathbf{D} = 1$, who have a noticeably higher frequency. The non-discriminatory information $\mathbf{X} = (X_1, X_2, \ldots, X_q)^\top$ consists of typical variables used for non-life insurance pricing, such as policyholder age and gender, driver age, and usage and location of the vehicle. Similar to the previous section, we will use policyholder age when illustrating the results, which we assign to the variable X_1, with $X_1 \in \{18, 19, \ldots, 90\}$. The policies may not be active for the entire calendar year of observation and the exposure v_i measures the time the policy was active during that year; thus, for each policy $1 \leq i \leq n$ the claims, policy and exposure data $(Y_i, \mathbf{X}_i, \mathbf{D}_i, v_i)$ are available.

5.2 Estimating best-estimate, unawareness and discrimination-free insurance prices

The real-world non-life insurance data are substantially more complex than the synthetic health insurance example, thus, a more complex neural network is used, compared to the previous section. Among the variables $\mathbf{X}$, there are three numerical and 16 categorical covariates; we convert the three numerical variables to categorical variables by binning them into 20 groups defined by the 5% quantiles (so that we can use network embeddings for these, which are most easily applied to categorical data, as described in what follows). To obtain the best-estimate, unawareness and discrimination-free insurance prices we use a plain-vanilla FNN (8) with embedding layers that map each level of the categorical variables to a 5-dimensional vector; see Section 7.4 in Wüthrich–Merz [24] for more details. The FNN thus takes as an input the covariates $(\mathbf{X}, \mathbf{D})$ mapped into $\mathbb{R}^{19\times 5}$. For the rest of the network, we choose a depth $m = 2$ with $(q_1, q_2) = (128, 128)$ hidden neurons in the two hidden layers, the hyperbolic tangent activation function, and the logit-link $g(\cdot) = \text{logit}(\cdot)$ (since we want claims frequencies not to exceed one, which would be highly unusual for this insurance portfolio). The final layer of the network multiplies the predicted claims frequency with the exposure v_i to produce the best-estimate prices $\mu(\mathbf{x}, \mathbf{d})$. To regularize the network, which is required due to the large number of variables in $\mathbf{X}$ and $\mathbf{D}$, both the batch normalization technique of Ioffe–Szegedy [12] and the drop-out technique

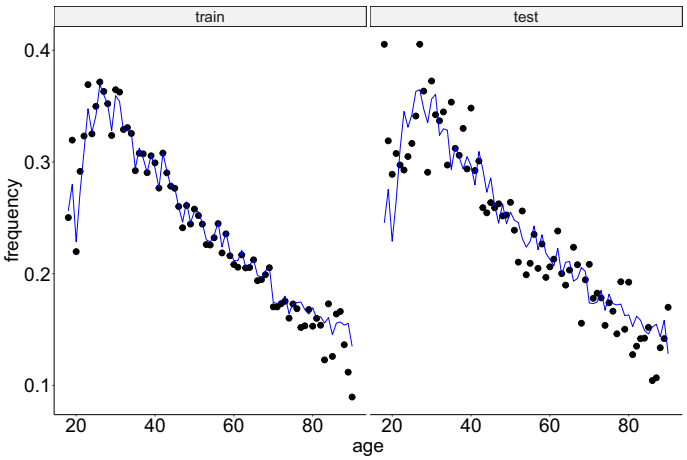

Fig. 12 Best-estimate frequency $\mu(\mathbf{x}, \mathbf{d})$ of the fitted plain-vanilla FNN (8) (blue line) compared to the observed (empirical) frequency (black dots) as a function of the age variable X_1 (note the FNN was fit on all components of $\mathbf{X}$ although we only show X_1) and averaged over ethnicity $\mathbf{d}$; these results are based on the complete knowledge of discriminatory information (color figure online)

of Srivastava et al. [21] are applied between each fully connected FNN layer. This network has a parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{w})$ of dimension 34,676. To implement this FNN we again use the library `keras` [7].

The FNN is fit to the full policyholder data $(Y_i, \mathbf{X}_i, \mathbf{D}_i, v_i)_{1 \le i \le n}$, i.e., including the protected information $\mathbf{D}$. We split the policy data into a training set consisting of 80% of the data and a test set consisting of the remaining 20%; a further 5% sample of the training set is then used as a validation set for exploring early stopping. As above, we use the Poisson deviance loss function for L in (12), and we apply stochastic gradient descent on a batch size of 128 policies. This is similar to what was done above. We average over 20 different FNN calibrations and derive the nagging predictor of Richman–Wüthrich [19] from these, using more calibrations than before due to the greater complexity of the dataset.

The results are given in Fig. 12, for both the training and test sets. The blue line shows the estimates from the plain-vanilla FNN fitted best-estimate price $\mu(\mathbf{x}, \mathbf{d})$ using all covariates $(\mathbf{X}_i, \mathbf{D}_i)$ as a function of the age variable $18 \le X_1 \le 90$, averaged over the rest of the covariates in $\mathbf{X}$ as well as the ethnicity variable $\mathbf{D}$ w.r.t. the empirical population density. The black dots show the corresponding observed frequency. The figure shows an accurate fit to the training set and a reasonable fit to the test set, with the largest errors in both sets made at the extreme ages where there is only little exposure available.

The discrimination-free insurance price $\mu^*(\mathbf{x})$, calculated w.r.t. the empirical probability $\widehat{\mathbb{P}}(\mathbf{d})$, and the unawareness price $\mu(\mathbf{x})$ are shown in Fig. 13 relative to the best-estimate price $\mu(\mathbf{x}, \mathbf{d})$. Both the discrimination-free insurance price and the unawareness price lie within a narrow range of about 1.5% around the best-estimate price. The discrimination-free insurance price at the younger ages X_1 is the most different one from the best-estimate price, whereas the unawareness price follows the

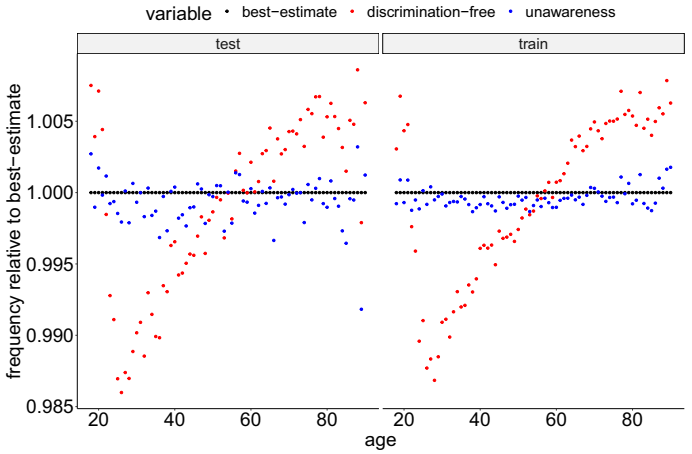

Fig. 13 Discrimination-free insurance price $\mu^*(\mathbf{x})$ (red) and unawareness price $\mu(\mathbf{x})$ (blue) relative to the best-estimate frequency $\mu(\mathbf{x}, \mathbf{d})$ of the fitted plain-vanilla FNN (8) (black) as a function of the age variable X_1 and averaged over ethnicity $\mathbf{d}$ (color figure online)

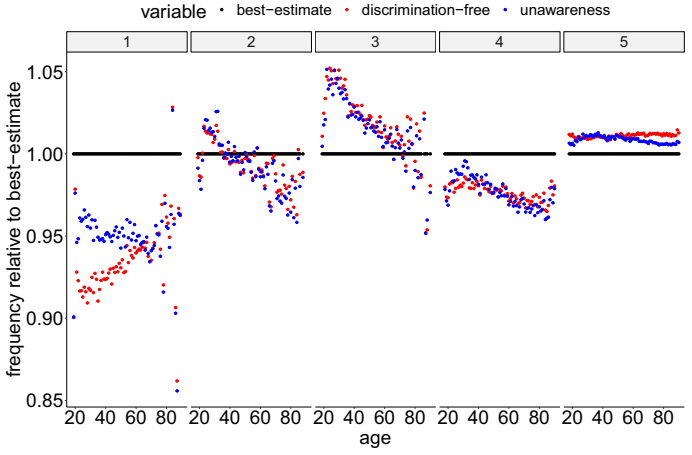

Fig. 14 Discrimination-free insurance price $\mu^*(\mathbf{x})$ (red) and unawareness price $\mu(\mathbf{x})$ (blue) relative to the best-estimate price $\mu(\mathbf{x}, \mathbf{d})$ of the fitted plain-vanilla FNN (8) (black) as a function of the age variable X_1 for the ethnicity codes $\mathbf{D} \in D_5$, test set only (color figure online)

best-estimate price more closely than the discrimination-free insurance price, with some variation in how close the approximation is depending on age. Such deviation implies that ethnicity $\mathbf{D}$ is used by the FNN to determine the best-estimate price and, also, that it is possible to infer implicitly $\mathbf{D}$ from $\mathbf{X}$. In other words, the unawareness price indirectly discriminates by ethnicity in this portfolio.

The differences between the discrimination-free, unawareness and best-estimate prices are emphasized in Fig. 14, which shows the first two of these relative to the best-estimate price for each ethnicity code $\mathbf{D} \in D_5$ for the test set only. For ethnicity code $\mathbf{D} = 5$ with the biggest total exposure (see Table 9, where it can also be seen that

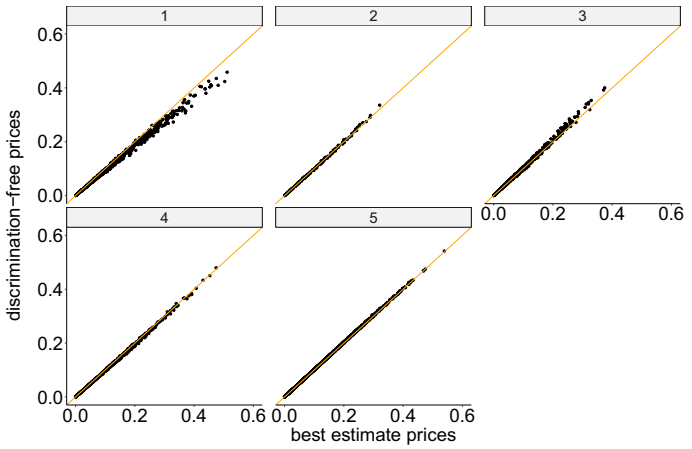

Fig. 15 Discrimination-free insurance price $\mu^*(\mathbf{x})$ from the plain-vanilla FNN compared to the best-estimate price $\mu(\mathbf{x}, \mathbf{d})$ for the ethnicity codes $\mathbf{D} \in D_5$

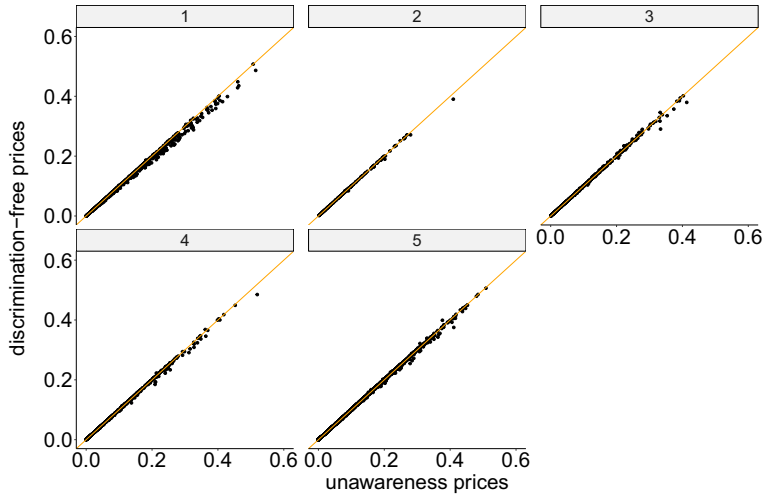

Fig. 16 Discrimination-free insurance price $\mu^*(\mathbf{x})$ from the plain-vanilla FNN compared to the unawareness price $\mu(\mathbf{x})$ for the ethnicity codes $\mathbf{D} \in D_5$

this code has the lowest observed frequency), there is not much difference between the discrimination-free insurance price and the unawareness price, which are relatively close to the best-estimate price, but higher due to the frequencies for this subset of the portfolio being low compared to the rest of the portfolio. The most noticeable divergences between the best-estimate and the unawareness prices are for younger drivers with ethnicity code $\mathbf{D} = 1$, where price differences of more than 5% between these prices occur. In Figs. 15 and 16 we plot the discrimination-free insurance prices against each of the best-estimate and unawareness prices, respectively. We draw a

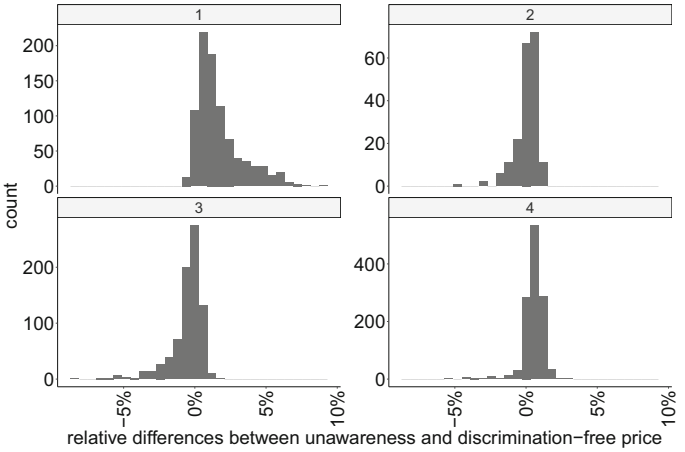

Fig. 17 Relative differences between the unawareness prices $\mu(\mathbf{x})$ and the discrimination-free insurance prices $\mu^*(\mathbf{x})$ of the fitted plain-vanilla FNN (8) for the ethnicity codes $\mathbf{D} \in \{1, \ldots, 4\}$, test set only

Table 10 Model accuracy measured using Poisson deviance losses of the best-estimate, unawareness and discrimination-free insurance prices against the observed claims Y_i in the training and test sets; these figures are based on the full knowledge of protected information

	Training set	Test set
Plain-vanilla FNN best-estimate price $\mu(\mathbf{x}, \mathbf{d})$	0.40056	0.40590
Plain-vanilla FNN unawareness price $\mu(\mathbf{x})$	0.40110	0.40605
Plain-vanilla FNN discrimination-free price $\mu^*(\mathbf{x})$	0.40125	0.40608

similar conclusion, which is that the most significant differences between the three prices occur for drivers with ethnicity code $\mathbf{D} = 1$.

The distributions of the differences between the discrimination-free and unawareness prices are shown in Fig. 17 for the ethnicity codes with less exposure. It can be seen that, while the differences are less than 5% for most policies, there is nonetheless a substantial number of policies for which this is not the case.

We compute the Poisson deviance losses of these three prices w.r.t. the observed claims Y_i in the training and test sets in Table 10. On the training set, a lower Poisson deviance loss is produced by the best-estimate price compared to the other two prices. However, the differences between the prices are significantly smaller on the test set and, in both cases, there is only a negligible loss of predictive accuracy when using the discrimination-free insurance price compared to the other two prices.

5.3 Applying the multi-task network

In the previous section, we have estimated the best-estimate, unawareness and discrimination-free insurance prices using FNNs estimated with full knowledge of the discriminatory information $\mathbf{D}$. We now apply the multi-task FNN approach (19)

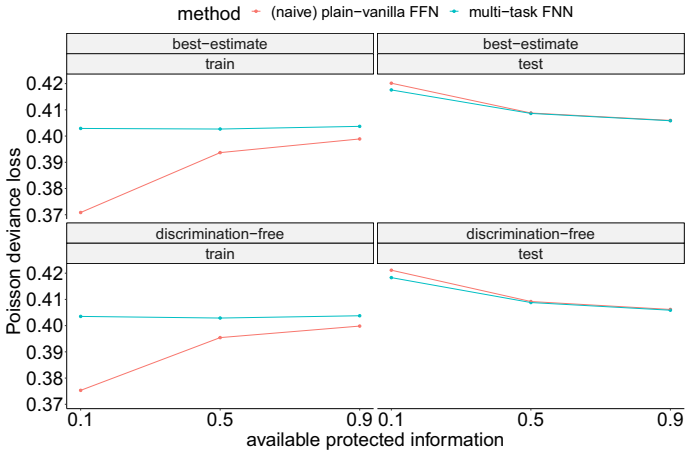

Fig. 18 Comparison of the (naive) plain-vanilla FNN and the multi-task FNN with drop-outs missing completely at random: Poisson deviance losses of the fitted best-estimate prices $\mu(\mathbf{x}, \mathbf{d})$ and discrimination-free insurance prices $\mu^*(\mathbf{x})$ for the observed claims Y_i; the x-axis gives the available discriminatory information $\mathbf{D}_i$

to the non-life insurance data to analyse the performance of this approach in the case of records missing in the discriminatory covariates $\mathbf{D}$. We focus on the case of discriminatory information missing completely at random.

We proceed analogously to Sect. 4.3 by independently setting $\mathbf{D}_i = \texttt{NA}$ with increasing (drop-out) rates of 10%, 50% and 90%; since the computational burden of fitting the models (19) is quite high, we focus only on these three cases. These datasets with drop-outs (incomplete protected information) are then used to perform two different model fittings: the more naive approach, in which we just fit a plain-vanilla FNN (6), using only those observations for which the discriminatory information is available and we drop all insurance policies with incomplete information, and the multi-task FNN (19) fitted with the loss function (20), which accounts for partial availability of discriminatory information, but uses the entire portfolio for model fitting. For both the naive and multi-task approaches, we fit the networks 20 times and average over the 20 different calibrations to derive the best-estimate and discrimination-free insurance prices. These can be compared both to the actual claims Y_i, as well as the corresponding estimates of the best-estimate and discrimination-free insurance prices derived using data with full discriminatory information, as shown in the previous section.

A comparison of the best-estimate and discrimination-free insurance prices estimated using each of the naive and multi-task FNN approaches to the observed frequencies Y_i, for each of the training and test sets, is shown in Fig. 18. On the training set, the multi-task approach performs worse than the naive approach for all three of the drop-out rates used. However, on the test set, the situation is reversed, with the multi-task network performing as well or better than the naive approach for all drop-out rates, with the most noticeable outperformance occurring when the available discriminatory information is only 10% of the total portfolio. This holds for both the

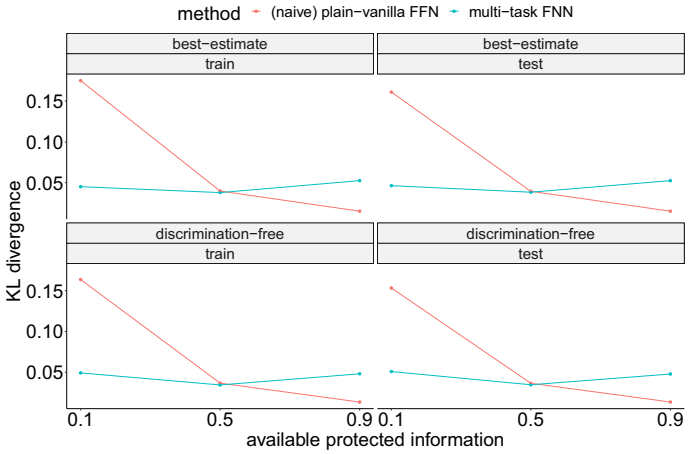

Fig. 19 Comparison of the (naive) plain-vanilla FNN and the multi-task FNN with drop-outs missing completely at random: KL divergences from the fitted best-estimate prices $\mu(\mathbf{x}, \mathbf{d})$ and discrimination-free insurance prices $\mu^*(\mathbf{x})$ to the same prices estimated using all the discriminatory information without drop-out; the x-axis gives the available discriminatory information $\mathbf{D}_i$; scale on y-axis is in 10^{-2}

best-estimate and discrimination-free insurance prices. Since there is a large difference between the performance of the naive method between the training and test sets, particularly at the higher drop-out rates, whereas there is a smaller difference for the multi-task method, it appears that the naive method is over-fitting to the training set due to the smaller amount of data available once a significant number of observations has been dropped-out (has been disregarded). On the other hand, the multi-task FNN over-fits to a smaller extent and provides better models out-of-sample; note that we apply early stopping in both cases.

More pronounced differences emerge when comparing the best-estimate and discrimination-free insurance prices estimated using each of the naive and multi-task FNN approaches to the best-estimate and discrimination-free insurance prices estimated with full knowledge of $\mathbf{D}$ (using the same model described in Sect. 5.2), as shown in Fig. 19.[1] Here, the naive approach is superior when 90% of the protected information is available for both best-estimate and discrimination-free insurance prices, and, when 50% of the information is available, the performance of the multi-task approach is marginally better than the naive approach. However, when only 10% of the protected information is available, the multi-task approach is clearly superior to the naive method. This indicates that the naive approach over-fits in case of small portfolios (with only few insurance policies with full information), even when carefully fine-tuning the stopping rule for gradient descent. In fact, a common experience is that neural networks are difficult to fit on small data having many categorical covariates with many levels, which is the case here, because in this situation gradient descent

[1] Note that Fig. 19 is comparable to Figs. 8 and 9; these figures show the KL divergence between the prices estimated using the naive and multi-task approaches to the same prices estimated using all the discriminatory information without drop-outs. On the other hand, Fig. 18 shows the Poisson deviance loss of the prices estimated using the naive and multi-task approaches w.r.t. the observed claims.

is rather sensitive on these categorical labels. For these reasons, we give preference to the multi-task approach here where we try to use all information (in a suitable way). Despite the worse performance of the multi-task approach in approximating the the best-estimate and discrimination-free insurance prices, compared to the naive approach when 90% of the data is available, nonetheless, the test set performance of the multi-task approach in predicting the claims experience is as good as the naive approach, as shown in Fig. 18, reinforcing that the naive approach seemingly over-fits even when only 10% of the protected data is missing.

Our case studies have shown that the more complex multi-task approach has clearly outperformed the naive FFN approach in situations of low availability of discriminatory information. As a result, we suggest that if the discriminatory information is available for most instances, it is sufficient to use a simple (plain-vanilla) FNN for calculating discrimination-free insurance prices. If the discriminatory information is available on fewer than 50% of insurance policies, then preference should be given to the multi-task approach, which uses all available data and presents superior predictive performance. Furthermore, our empirical results indicate that the multi-task approach is robust to the situation of data on **D** not missing completely at random.

6 Concluding remarks

Addressing the problem of indirect or proxy discrimination involves an apparent paradox: in order to compensate for the potentially discriminatory effect of implicitly inferring policyholders' protected characteristics, information on these very characteristics must be available for regression modeling. Resolving this tension poses clear legal, regulatory and technical challenges. Here, focusing on the latter, we provided a multi-task neural network learning framework, which can generate insurance prices that are free from indirect discrimination. We demonstrated that this multi-task architecture is competitive to conventional approaches when full information is available, while clearly outperforming them in the case of less than 50% of discriminatory information. This was shown both in the case of synthetic data, as well as for a complex real-world non-life insurance dataset. Moreover, we have obtained comparable results for the two cases of protected information missing completely at random and not missing completely at random. When using the real-world insurance dataset, we observed that some minority classes in this particular portfolio may experience relatively significant indirect discrimination if unawareness prices are used. This demonstrates the need for practically applicable discrimination-free pricing methods, which we aimed to address in this paper.

Nonetheless, there is an aspect of the technical challenge that we have not yet fully addressed. Practically, we still need discriminatory information **D** for a part of the portfolio in order to fit our model. Hence, a scheme needs to be in place that allows insurers to access such protected information for a subset of policies. Such a scheme may be constructed commercially, e.g., by offering special discounts to customers who are willing to disclose information on protected characteristics. Even if only a smaller proportion of customers would be willing to disclose this information, nonetheless, the multi-task method presented here shows promising performance even in this case.

Besides addressing privacy concerns, a difficulty with such an approach is to ensure (or mitigate) the potential selection bias that such a commercial promotion will generally have. While our case study illustrated good performance of our model when data are not missing at random, further work on this topic is required.

A related issue is that, to go from best-estimate prices to discrimination-free insurance prices, we need to choose the pricing measure $\mathbb{P}^*(\mathbf{d})$. A natural candidate is to use the empirical version of $\mathbb{P}(\mathbf{d})$, but since this choice will be based on a subset of the portfolio, the question again arises as to whether this subset is representative of the entire portfolio. In the multi-task network approach we receive an estimate for $\mathbb{P}(\mathbf{d})$ as an integral part of the prediction model by the averaging in (25) of the estimated categorical probabilities $p_k(\mathbf{X}_i)$. Alternatively, techniques from survey sampling could be used in order to obtain an estimate of $\mathbb{P}(\mathbf{d})$, using so-called indirect questioning. These techniques were constructed in order to obtain unbiased estimates of population proportions of a single sensitive dichotomous characteristic, such as drug use and sexual preference, based on open answer questionnaires, see the seminal paper by Warner [23]. For more general categorical sensitive characteristics, alternative techniques can be used; see Lagerås–Lindholm [13] and the survey of Chaudhuri–Christofides [5]. Regardless of the specific technique employed, by obtaining a suitable total population estimate for $\mathbb{P}(\mathbf{d})$ it is possible to assess whether the sub-portfolio has been sampled with data missing completely at random or not.

Data Availability R code generating synthetic data and performing the analysis of Section 4 is available at https://github.com/RonRichman/multi_task_dfip. The data used in Section 5 are confidential and cannot be shared.

Appendix A: Feed-forward neural network architecture

To define a FNN architecture we start by choosing the depth $m \in \mathbb{N}$ of this architecture, meaning that we are going to consider m hidden FNN layers $\mathbf{z}^{(j)}$, $1 \leq j \leq m$. For each of these hidden layers $\mathbf{z}^{(j)}$, we need to specify the number of hidden neurons $q_j \in \mathbb{N}$. We initialize the dimension of the input to the network by q_0 being the dimension of the covariates $(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^{q_0}$. Furthermore, we choose an activation function $\phi : \mathbb{R} \to \mathbb{R}$. A hidden FNN layer is then given by the mapping

$$\mathbf{z}^{(j)} : \mathbb{R}^{q_{j-1}} \to \mathbb{R}^{q_j}, \qquad \mathbf{a} \mapsto \mathbf{z}^{(j)}(\mathbf{a}) = \left(z_1^{(j)}(\mathbf{a}), \ldots, z_{q_j}^{(j)}(\mathbf{a})\right)^\top,$$

where we define the neurons $\mathbf{a} \mapsto z_k^{(j)}(\mathbf{a})$, $1 \le k \le q_j$, in the j-th hidden layer by

$$z_k^{(j)}(\mathbf{a}) = \phi\left(w_{0,k}^{(j)} + \sum_{l=1}^{q_{j-1}} w_{l,k}^{(j)} a_j\right) = \phi\left\langle \mathbf{w}_k^{(j)}, \mathbf{a} \right\rangle, \tag{26}$$

with network weights $\mathbf{w}_k^{(j)} = (w_{0,k}^{(j)}, \ldots, w_{q_{j-1},k}^{(j)})^\top \in \mathbb{R}^{q_{j-1}+1}$ in the k-th neuron of the j-th hidden FNN layer, and where $\langle \cdot, \cdot \rangle$ denotes the scalar product between $\mathbf{w}_k^{(j)}$ and $\mathbf{a}$, the latter (by a slight abuse of notation) being extended by a 0-th component being equal to 1. This implies that the j-th hidden layer $\mathbf{z}^{(j)}$ has a network weight $\mathbf{w}^{(j)} = (\mathbf{w}_1^{(j)}, \ldots, \mathbf{w}_{q_j}^{(j)})$ of dimension $q_j(q_{j-1}+1)$, this is the size of the parameter in j-th hidden layer.

The FNN architecture of depth m is received by composing the hidden FNN layers $\mathbf{z}^{(j)}$, $1 \le j \le m$. This yields the mapping

$$\mathbf{z}^{(m:1)} : \mathbb{R}^{q_0} \to \mathbb{R}^{q_m}, \qquad (\mathbf{x}, \mathbf{d}) \mapsto \mathbf{z}^{(m:1)}(\mathbf{x}, \mathbf{d}) = \left(\mathbf{z}^{(m)} \circ \cdots \circ \mathbf{z}^{(1)}\right)(\mathbf{x}, \mathbf{d}). \tag{27}$$

This mapping has network weights (parameter) $\mathbf{w} = (\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{m)})$ of dimension $\sum_{j=1}^m q_j(q_{j-1}+1)$. This mapping (27) transforms the q_0-dimensional vector-valued input $(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^{q_0}$ to a new q_m-dimensional representation $\mathbf{z}^{(m:1)}(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^{q_m}$ of the original covariates $(\mathbf{x}, \mathbf{d})$.

The final step of the FNN architecture is the readout of the q_m-dimensional representation $\mathbf{z}^{(m:1)}(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^{q_m}$. For this we choose a strictly monotone and smooth link function $g : \mathbb{R} \to \mathbb{R}$ and a readout parameter $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{q_m})^\top \in \mathbb{R}^{q_m+1}$. This allows us to define the following FNN regression function, see (6),

$$(\mathbf{x}, \mathbf{d}) \;\mapsto\; g\left(\mu(\mathbf{x}, \mathbf{d})\right) = \beta_0 + \sum_{j=1}^{q_m} \beta_j z_j^{(m:1)}(\mathbf{x}, \mathbf{d}) = \left\langle \boldsymbol{\beta}, \mathbf{z}^{(m:1)}(\mathbf{x}, \mathbf{d}) \right\rangle, \tag{28}$$

where the last term expresses the middle term as a scalar product, similar to (26).

Fig. 1 illustrates a FNN architecture of depth $m = 3$, with input dimension $q_0 = 3$ and numbers of neurons in the hidden layers $(q_1, q_2, q_3) = (20, 15, 10)$. The depth $m \in \mathbb{N}$, the numbers of hidden neurons $q_j \in \mathbb{N}$, $1 \le j \le m$, the activation function ϕ and the link function g are hyperparameters that need to be chosen by the modeler. The network weights $\mathbf{w}$ of dimension $\sum_{j=1}^m q_j(q_{j-1}+1)$ and the readout parameter $\boldsymbol{\beta} \in \mathbb{R}^{q_m+1}$ are the parameters that are fitted to the data for the given network architecture.

References

1. Abbas A, Sutter D, Zoufal C, Lucchi A, Figalli A, Woerner S (2021) The power of quantum neural networks. Nat Comput Sci 1:403–409
2. Araiza Iturria CA, Hardy M, Marriott P (2022) A discrimination-free premium under a causal framework. SSRN Manuscript ID 4079068

3. Batista GEAPA, Monard MC (2002) A study of k-nearest neighbour as an imputation method. In: Abraham A, Ruiz-del-Solar J, Köppen M (eds) Soft computing systems—design, management and applications, Frontiers in Artificial Intelligence and Applications, vol 87. IOS Press, Amsterdam, pp 251–260
4. 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, vol 81, pp 77–91
5. Chaudhuri A, Christofides TC (2013) Indirect questioning in sample surveys. Springer, Berlin
6. Chen J, Shao J (2000) Nearest neighbor imputation for survey data. J Off Stat 16(2):113–131
7. Chollet F, Allaire JJ et al (2017) R interface to Keras. https://github.com/rstudio/keras
8. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O (2021) A survey on missing data in machine learning. J Big Data 8:140
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. Off J Eur Union L 373:37–43
10. Frees EWJ, Huang F (2023) The discriminating (pricing) actuary. North Am Actuar J 27(1):2–24
11. Grari V, Charpentier A, Lamprier S, Detyniecki M (2022) A fair pricing model via adversarial learning. arXiv:2202.12008v2
12. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: 32nd International conference on machine learning, vol 1, pp 448–456
13. Lagerås A, Lindholm M (2020) How to ask sensitive multiple-choice questions. Scand J Stat 47(2):397–424
14. Lindholm M, Richman R, Tsanakas A, Wüthrich MV (2022) Discrimination-free insurance pricing. ASTIN Bull 52(2):55–89
15. Lindholm M, Richman R, Tsanakas A, Wüthrich MV (2023) What is fair? Proxy discrimination vs. demographic disparities in insurance pricing. SSRN Manuscript ID 4436409
16. Pope DV, Sydnor JR (2011) Implementing anti-discrimination policies in statistical profiling models. Am Econ J 3(3):206–231
17. Prince AER, Schwarcz D (2020) Proxy discrimination in the age of artificial intelligence and big data. Iowa Law Rev 105(3):1257–1318
18. R Core Team (2021) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/
19. Richman R, Wüthrich MV (2020) Nagging predictors. Risks 8(3): 83
20. Richman R, Wüthrich MV (2023) Conditional expectation network for SHAP. SSRN Manuscript ID 4514891
21. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958
22. van Buuren, S, Groothuis-Oudshoorn K (2011) `mice`: multivariate imputation by chained equations in R. J Stat Softw 45(3)
23. Warner SL (1965) Randomized response: a survey technique for eliminating evasive answer bias. J Am Stat Assoc 60(309):3–69
24. Wüthrich MV, Merz M (2023) Statistical foundations of actuarial learning and its applications. Springer Actuarial
25. Xin X, Huang F (2023) Anti-discrimination insurance pricing: regulations, fairness criteria, and models. North Am Actuar J **(in press)**
26. Yoon J, Jordon J, van der Schaar M (2018) GAIN: missing data imputation using generative adversarial nets. Proc Mach Learn Res 80:5689–5698