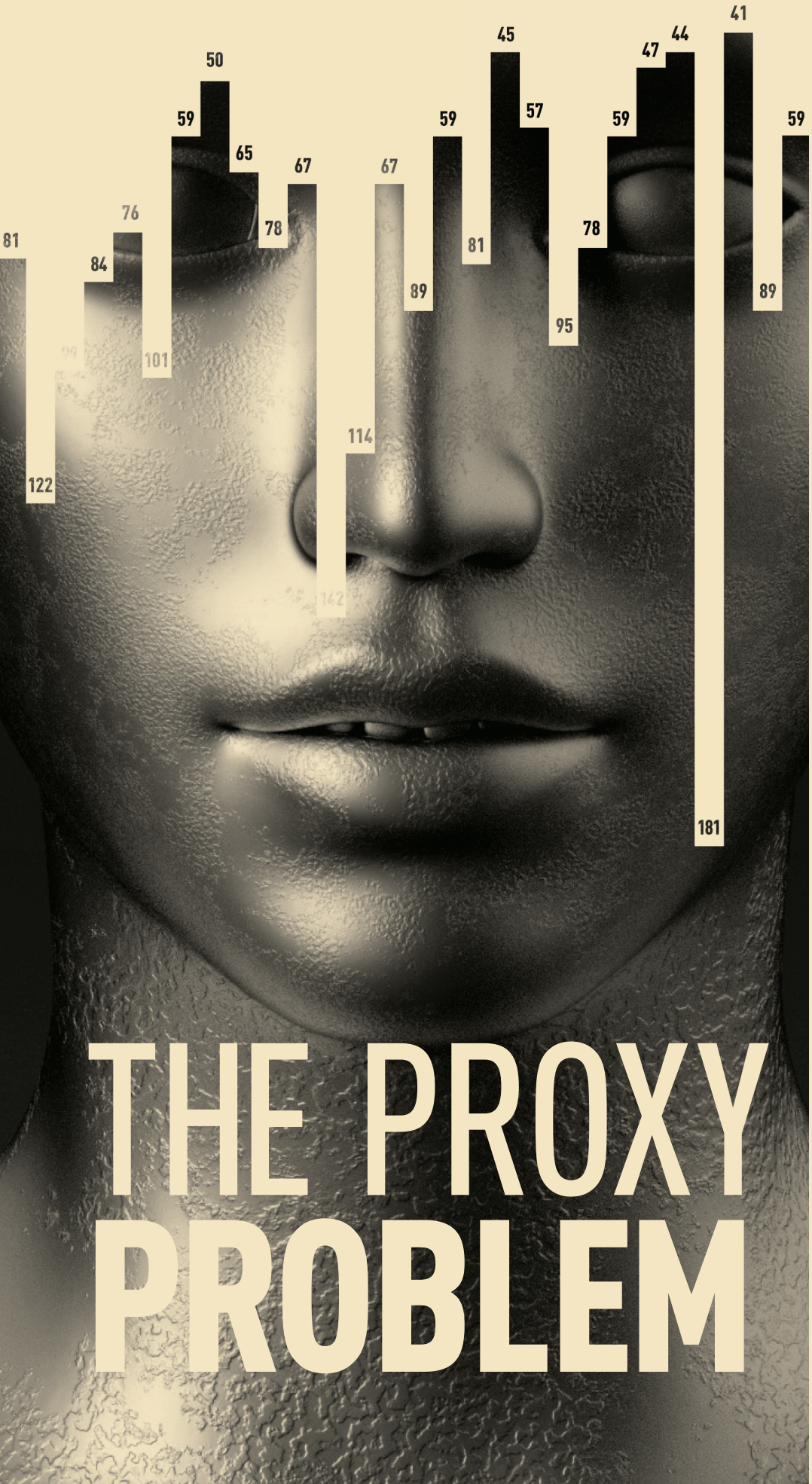


THE PROXY PROBLEM

If we want to avoid discrimination in insurance pricing, we need to be able to take protected characteristics into account, say **Mathias Lindholm**, **Ronald Richman**, **Andreas Tsanakas** and **Mario Wüthrich**

As advanced modelling methodologies become widely available to actuaries, the way models are used within financial services is increasingly constrained by legal developments and regulatory scrutiny. Two examples in the UK are the Financial Conduct Authority's review of general insurance pricing practices and the Information Commissioner's Office consultation on an AI auditing framework.

A more longstanding and familiar regulation is the EU gender discrimination directive, which requires that pricing models do not discriminate by gender. The risks of inadvertent discrimination with respect to protected characteristics seem to be higher in complex models than in simple ones, as complex models may exploit intricate patterns in data to derive proxies for, say, gender. In addition to the legal and regulatory risks, ethical concerns could arise if models were found to be using unacceptable proxies.

Defining discrimination in such an intuitive way may appear straightforward, but without a

rigorous definition of discrimination, it becomes difficult to guarantee that pricing models are free of it. How can we make sure that illegal or unwanted discriminatory factors are not influencing the results of a model?

Our recent research paper 'Discrimination-Free Insurance Pricing' (**bit.ly/2KLG5CK**) proposes an approach to ensuring that the results of actuarial models are not influenced by protected characteristics. This proposed discrimination-free pricing method is a simple add-on to existing pricing methodologies and does not require major changes to insurers' predictive models. It can remove discriminatory effects from all categories of pricing techniques currently in use, from generalised linear models (GLMs) to gradient boosting machines and deep neural networks.

Method

We start by taking it as given that protected characteristics such as gender are not used within pricing models as rating factors – meaning direct discrimination is avoided. What do we mean, then, when we say that a price may still be discriminatory? We illustrate our ideas with a simple stylised example; a full mathematical definition of discrimination in pricing can be found in our paper.

We consider the case of a simple pricing model for a health insurance portfolio. The two relevant covariates are the policyholder's gender and age class. The portfolio population is split 50/50 between women and men – shown in *Figure 1*, together with the split across ages. In this example, 90% of policyholders in the younger age classes are female, with the reverse happening for older age classes.

The expected claims costs by age class and gender are shown by the grey and dark yellow lines in *Figure 2*; we can view these as 'best-estimate prices'. We can see that, as age increases, claims costs for females and males diverge – in particular, claims costs for females become progressively higher. The question is: how should insurance be priced?

A common method for avoiding discrimination is simply ignoring gender. Then, the insurance rate for a policyholder of any gender at age 50 is nothing but the average cost of the corresponding age class. We call a price calculated in this way an 'unawareness price', shown by the black line in *Figure 2*. It is striking that the unawareness price is very close to the best-estimate price for women at lower ages, and then drops to nearly the best-estimate price for men at higher ages. This is due to the much higher prevalence of women within the lower age classes (90%), as we saw in *Figure 1*. The unawareness price uses age as a proxy for gender – to be precise, the calculation of unawareness prices implicitly relies on the conditional probability of gender, given age. In summary, ignoring gender in price calculation did not remove its impact on prices. This is indirect discrimination.

What should the price be for, say, a policyholder aged 50? We know that gender must somehow be allowed for in the calculation, since ignoring it leads to indirect discrimination. Furthermore, prices should lie somewhere between the extremes given by the grey and dark yellow lines in *Figure 2*; in particular, the price at age 50 should be a weighted average of the corresponding

FIGURE 1 Portfolio proportions – distribution of gender across age classes and population average

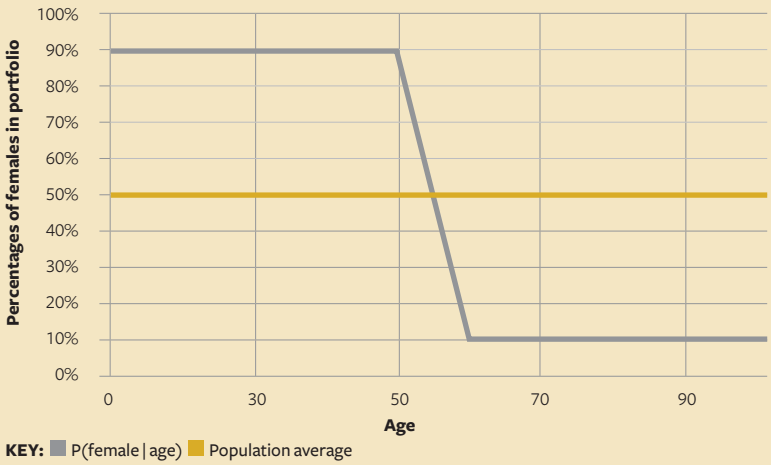


FIGURE 2 Different types of insurance prices

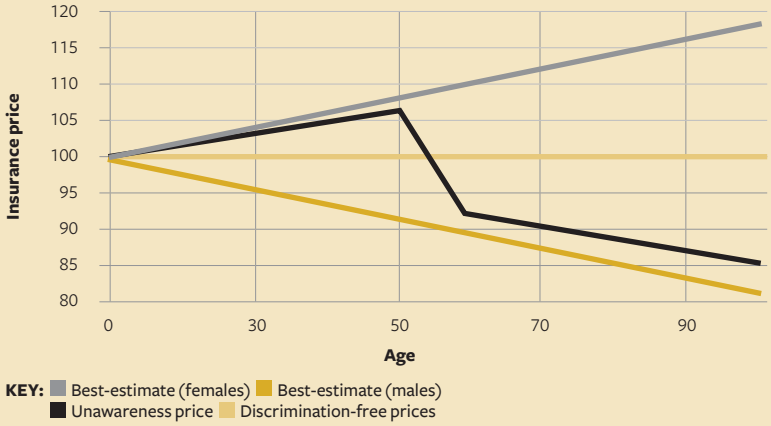


IMAGE: SHUTTERSTOCK

best-estimate prices for men and women. For age not to be a proxy for gender, these weights should not depend on the proportion of women in each age group. This means that 'discrimination-free prices' must be represented by a straight line in *Figure 2*. Finally, note that the overall population is split equally between men and women. This implies that a horizontal line, as depicted in pale yellow, is a suitable choice for a discrimination-free price.

This stylised example has been constructed with some care in order to clearly illustrate what can go wrong when unawareness prices are used. It shows that, in order to account for discriminatory characteristics, one needs to actually use the very same characteristics as part of the pricing procedure – recall that the intuitively discrimination-free prices we derived were based on best-estimate prices.

In many realistic situations, the differences between unawareness prices and discrimination-free prices may be smaller. Still, to be certain that no indirect discrimination takes place, we need a practical alternative to unawareness prices.

Interpretation

Our discrimination-free pricing formula can be derived by arguing from two distinct directions; we only give a summary of the technical arguments here. First, recall that insurance prices are generally calculated as conditional expectations of claims costs (given the rating factors available). These expectations are sometimes re-weighted, for example assigning a higher probability to some scenarios than the data would imply, in order to derive a profit-loaded premium. Our approach utilises a similar trick. However, for us the aim of the re-weighting is different: the statistical decoupling of discriminatory from non-discriminatory factors, without changing the structure of the predictive model underlying best-estimate prices. Specifically, if $u(x,d)$ is the best-estimate price, depending on both the rating factors x and the discriminatory characteristics d, discrimination-free prices arise from 'averaging out' the discriminatory characteristics d: $\sum_d u(x,d)P(d)$

A second justification relies on causal inference, a branch of statistics that is attracting increasing public interest, partly due to Dana Mackenzie and Judea Pearl's 2018 publication *The Book of Why*. Causal inference uses graphs to represent not just correlations in the data, but also the actual direction of causal effects, which are subsequently estimated from observational data. This allows users to assess the impact of changes in the values of chosen variables while stripping out confounding effects. The pricing formula we propose can (in some circumstances) be interpreted within the framework of causal inference – as representing the direct causal effect of the rating factors on the insurance experience, without confounding by other discriminatory characteristics such as gender.

FIGURE 3 Assumed claims costs underlying the simulated data used in the example. The higher costs at ages 20-40 for women are for birth-related costs. Costs for smokers are higher than for non-smokers due to costs associated with cancer.

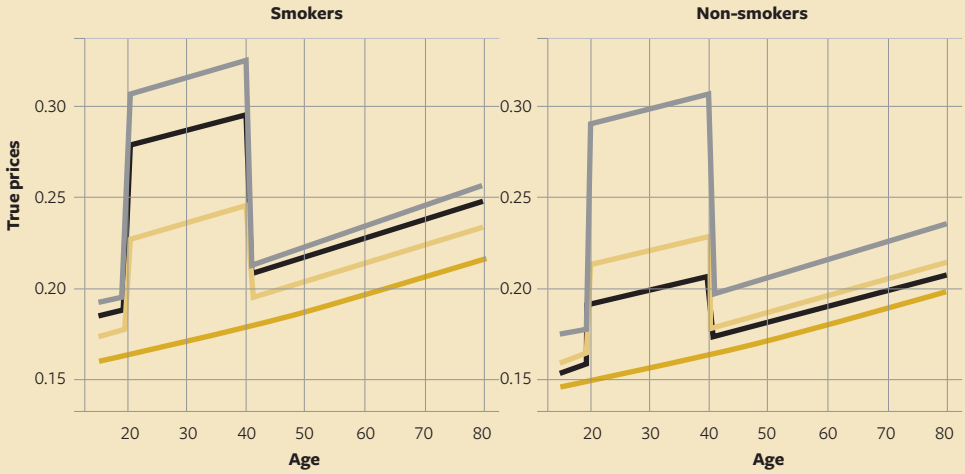
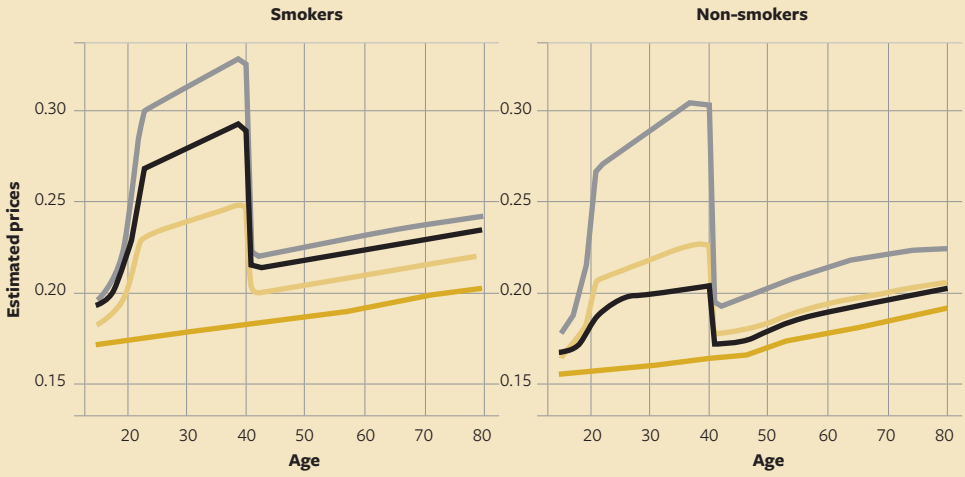


FIGURE 4 Predicted claims using deep neural networks calibrated on claims data simulated for 100,000 policyholders.



KEY: Best-estimate (females) Best-estimate (males) Unawareness price Discrimination-free prices

Applications

We now consider the application of discrimination-free pricing in a more complex model of health insurance claims. We simulated data for three types of healthcare costs, based on the two rating factors of gender and smoking status: costs of birth-related injuries, only applying to females aged 20-40; cancer-related costs, which are higher for smokers; and all other healthcare costs. We also assumed that women are more likely to smoke than men. For the remaining assumptions used in this example, we refer to our paper.

We show the true claims costs (grey and dark yellow lines) based on the model underlying the simulated data in *Figure 3*, as well as the unawareness and discrimination-free prices. The best-estimate claims costs are consistently higher for women than for men. The unawareness prices for smokers are closer to the best-estimate prices for women, since, in our example, being a smoker is predictive of being a woman. Likewise, the unawareness prices for non-smokers are closer to the prices for men. On the other hand, the discrimination-free prices do not reflect the gender-based information contained in the smoking status – they only capture the direct effect of smoking on the (higher) level of claims produced.

Having applied the method to the true claims costs, we now investigate how well the method works on noisy simulated data. For this purpose, the claims of 100,000 policyholders were simulated using the claims cost model discussed, on the assumption that claims costs are distributed according to a Poisson distribution. We then fit a deep neural network to the simulated data to act as our pricing model, considering both gender and smoking status (to derive best-estimate prices) and subsequently estimate discrimination-free prices.

Furthermore, we recalibrate the network using only the smoking rating factor, to derive unawareness prices. The predictions from these models are shown in *Figure 4*.

It can be seen that the deep neural networks successfully approximate the true claims costs, and that both the discrimination-free and unawareness prices are similar to the true values shown in *Figure 3*. This leads us to conclude that the method of producing discrimination-free prices works well in the given model.

Avoiding bias (when avoiding bias)

A basic requirement of a good pricing model is that the total costs predicted by the model should be equal to the expected total costs from the portfolio under consideration. Most actuarial models (such as GLMs) fulfil this requirement, but it can be shown that the discrimination-free prices introduced in this article do not. A correction to these prices for this bias is therefore required, the simplest option being pro-rata adjustment.

Conclusions

We have proposed an easily implementable method for removing the effects of discrimination from pricing models by removing the proxying of characteristics such as gender by other covariates. We have provided examples showing that ignoring discriminatory characteristics does not lead to discrimination-free prices, meaning that unawareness prices ignore the wrong thing. Instead, we should include discriminatory characteristics in a model and remove their effect afterwards. Our proposal works for any kind of predictive model – from GLMs to neural networks – and can thus be applied as an add-on to existing pricing models used by actuaries. Mathematical details can be found in our paper.

"In order to account for discriminatory characteristics, one needs to actually use the very same characteristics as part of the pricing procedure"

What our method requires is data on characteristics, whose use may be considered discriminatory. Many such characteristics are not recorded by companies, so development of this work must consider how to overcome this problem. Our claim is that information on discriminatory characteristics is necessary to remove discrimination from pricing. While the technical foundation for this idea is solid, communicating it may not be easy, particularly in view of concerns around privacy.

We have not tried to define which factors should be treated as discriminatory – a societal question beyond our analysis. We recommend that companies should assess whether any rating factors currently used in pricing models might be functioning as problematic proxies from a legal, regulatory or ethical perspective. An example is the use of postal code information within models, since postal codes can correlate highly with ethnicity – by applying our method, it might be possible to provide insurance at a more reasonable cost to groups that may have been disadvantaged in the past. This indicates that the broader implications of discrimination-free pricing within specific markets should be considered.

DR MATHIAS LINDHOLM
is an associate professor in mathematical statistics at Stockholm University

RONALD RICHMAN
is an associate director (R&D and Special Projects) at QED Actuaries and Consultants

PROFESSOR ANDREAS TSANAKAS
is professor of risk management at Cass Business School, City, University of London

PROFESSOR MARIO WÜTHRICH
is professor for actuarial science in the Department of Mathematics at ETH Zürich