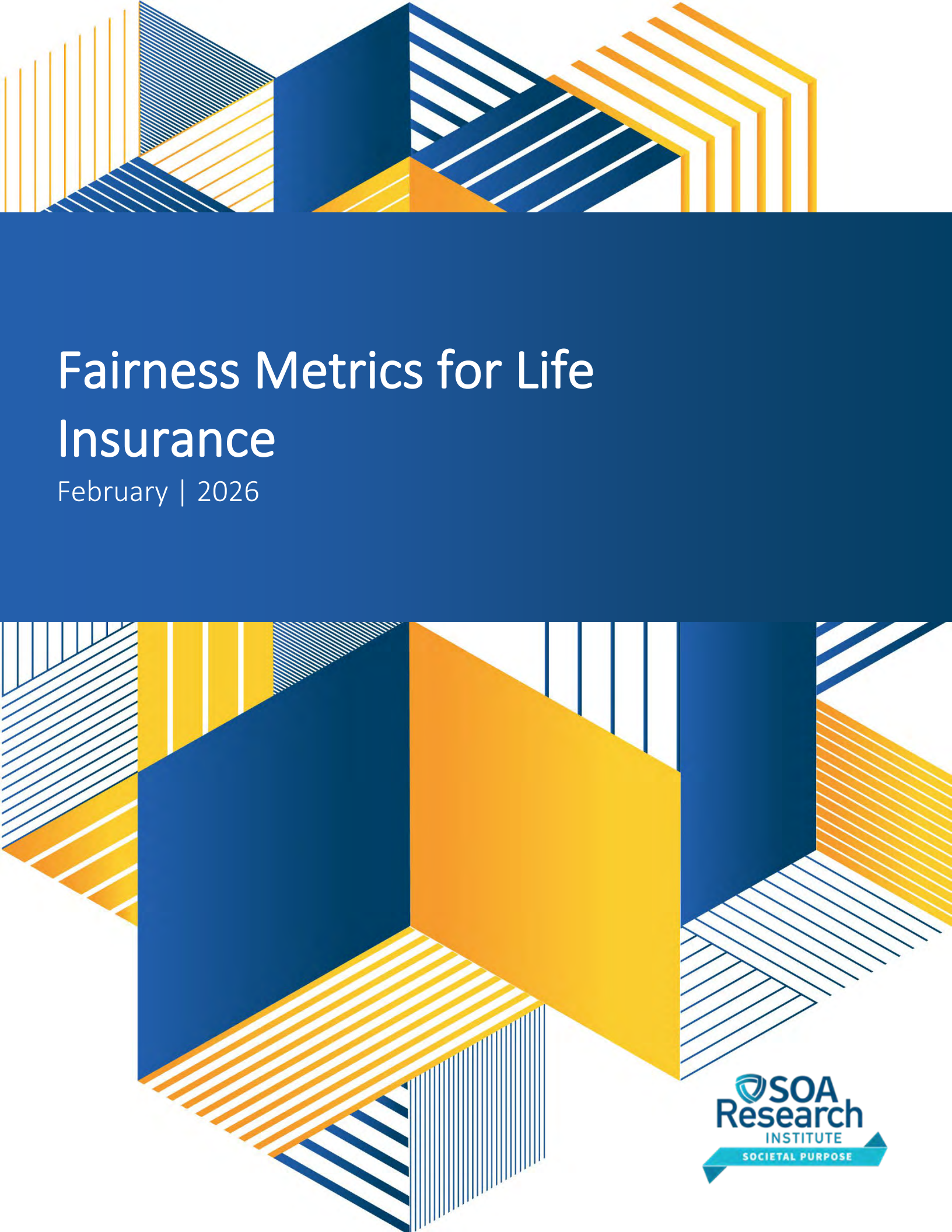
Fairness Metrics for Life Insurance

February | 2026

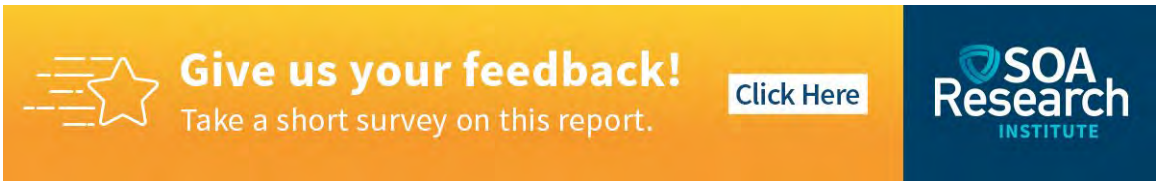


Fairness Metrics for Life Insurance

AUTHORS Eric Krafcheck, FCAS, MAAA
Milliman Inc

Igor Balnozan, PhD
University of New South Wales

Fei Huang, PhD
University of New South Wales



Contents

Contents **3**

Executive Summary **4**

Background and Introduction **5**

Section 1: Defining Fairness **7**
1.1 Individual vs Group Fairness 7
1.2 Actuarial Fairness 10

Section 2: Individual Fairness Criteria **11**
2.1 Fairness Through Unawareness 11
2.2 Fairness Through Awareness 12
2.3 No Omitted-Variable Bias 14

Section 3: Group Fairness Criteria **16**
3.1 Independence 16
3.2 Sufficiency 20
3.3 Separation 22

Section 4: Conclusion **23**

Section 5: Acknowledgments **25**

References **26**

About The Society of Actuaries Research Institute **28**

Fairness Metrics for Life Insurance

Executive Summary

This paper offers a critical, practical framework for evaluating and implementing fairness in life insurance, recognizing that fairness is not a one-size-fits-all concept. By systematically analyzing a range of fairness criteria—both individual and group-based—the paper reveals that different approaches carry ethical, operational, and financial implications for insurers.

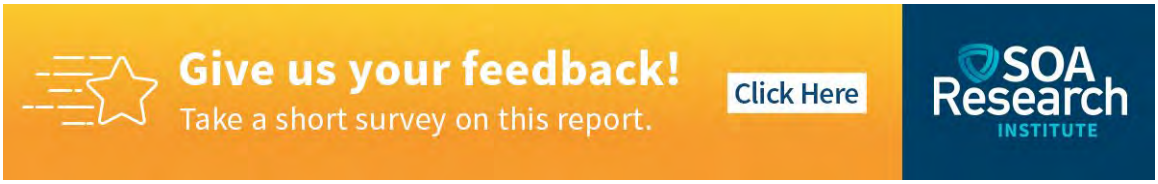
Measuring fairness in life insurance is inherently complex. Individual fairness, closely aligned with actuarial principles, ensures that premiums and outcomes are commensurate with each applicant's risk. However, this may result in unequal treatment across groups, especially when underlying social disparities are present. Group fairness, conversely, seeks parity in outcomes across demographic or other cohorts, but can lead to inconsistent treatment of individuals and introduce challenges such as adverse selection and cross-subsidization.

The paper critically examines methods and metrics for evaluating fairness while highlighting their strengths and limitations in real-world insurance contexts. It demonstrates that technical compliance with a fairness metric does not guarantee that ethical or social objectives are met, and that the choice of metric must be tailored to specific business practices, regulatory constraints, and stakeholder expectations.

Importantly, this paper does not advocate for a single fairness definition or metric. Instead, it equips practitioners with tools to articulate their chosen definition of fairness, calculate relevant metrics, and understand the trade-offs involved. The framework encourages readers to:

- Carefully select fairness metrics that align with their strategic objectives and regulatory environment.
- Recognize that achieving fairness at both the individual and group level is often impossible, necessitating transparent trade-offs.
- Conduct ongoing analysis to identify unintended consequences, such as loss of predictive accuracy, increased operational costs, or shifts in market dynamics.
- Engage with stakeholders to ensure that fairness initiatives meet the needs of the relevant parties.

As the life insurance industry continues to evolve with advancements in technology and shifts in societal expectations, further research and ongoing dialogue among actuaries, regulators, and industry stakeholders could contribute to refining these fairness metrics and their real-world application. Such efforts may help develop balanced life insurance practices that not only comply with ethical and regulatory standards but also foster trust in the broader societal context.



Background and Introduction

With technological and societal advancements, there continues to be much discussion and effort aimed toward rethinking whether processes involved in the development and management of life and other insurance products are fair. Regulators, underwriters, actuaries, other insurance professionals, academics, consumers, advocates, and other stakeholders have varying perspectives on what fairness means. Defining fairness for life insurance processes is complex and challenging as well as subjective.

The Society of Actuaries Research Institute submits to the public discourse this summary of fairness criteria and metrics for life insurance. The insurance industry is well regulated, with regulations focused on two objectives: financial solvency of insurance companies and fair treatment of policyholders and claimants. The U.S. legal framework offers various definitions that one might consider definitions of fairness. This report intends to objectively inform readers about defining and quantifying fairness in the context of processes involved in developing and managing life insurance, especially risk classification, without recommending specific solutions. The contents are intended to help the reader understand various definitions of fairness, metrics aligned to those definitions, strengths and weaknesses of both the definitions and metrics, as well as various considerations when making fairness determinations or assessments.

Questions of fairness and discrimination have been core issues for the life insurance industry for decades. As life insurance serves a societal role in various ways, including providing financial security for families, efforts have focused on broadly ensuring fair access to it.

The issue of bias has gained attention in recent years. Anecdotal or experimental evidence of seemingly unfair outcomes, particularly in relation to the use of predictive models (e.g. machine learning) and artificial intelligence, has led to much emerging research. The topics pursued include fairness and bias, as well as explainability and possible risks of increasingly opaque complex models. As a regulated industry, concerns of data privacy, accountability, and oversight are already within the scope of insurers' stated responsibilities; however, various stakeholders seek to further understand whether certain data-driven decisions could produce unfair or biased outcomes—whether by amplifying and entrenching potential historical biases in the data these models are trained on or from algorithmic inaccuracy.[1]

This paper addresses some key issues currently faced by the insurance industry, in particular fairness definitions and metrics that can potentially be used in the life insurance industry. A variety of fairness definitions and criteria are introduced, discussed, and critically analyzed, including a discussion about the ethical implications and trade-offs in adopting each criterion. Concepts presented within the paper are not intended to only apply to certain applications within life insurance, though some readers may find applications in underwriting (e.g. acceptance of risks, pricing, etc.) and claims (e.g. fraud models) of particular interest.

This paper presents a framework that can be used to articulate a definition of fairness and describes various metrics that can be calculated and used to quantify a given definition. Questions of fairness can be significantly more complicated than they initially seem, as there is no single definition of what it means for something to be fair. Rather, there are a plethora of competing and mutually incompatible definitions that have been proposed, and each takes a different perspective on fairness.

Notably, this paper does not recommend any single definition or metric. A determination of the most appropriate fairness metric or metrics to use ultimately requires an understanding of the application and the relevant circumstances—there is no "one size fits all" metric that is appropriate for all situations. As can often be the case

[1] (Xin, Hooker, & Huang, 2024), (O'Neil, 2016)

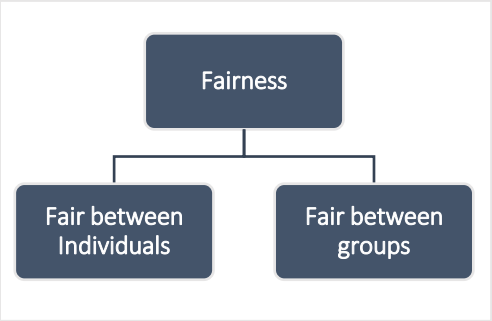
with ethical issues, conflicts can arise between different but reasonable positions, and often there is no single "right"' answer. This paper seeks to equip the reader with the understanding of those concepts and the established literature on the topic.

Section 1: Defining Fairness

1.1 INDIVIDUAL VS GROUP FAIRNESS

Fairness is a nuanced concept. To define fairness, a fundamental question must first be answered – fair for whom? **The various fairness definitions found in the literature[2] can be organized into two categories: fairness defined as being fair between individuals (individual fairness), or fairness defined as being fair between groups (group fairness).[3]**

Figure 1
FAIRNESS DEFINITION TREE



In this paper, "group" is intended to broadly mean a cohort of individuals that are assigned to a single classification based on one or more attributes for the purposes of evaluating fairness, regardless of whether any of the defining attributes are sensitive in nature. For example, while much of the fairness literature discusses groups defined based on sensitive attributes such as sex or race, insurers may also have interest in applying fairness concepts to groups defined based on other attributes, such as whether the insured is a smoker or their occupation.

"Individual," on the other hand, is intended to refer to the most granular risk classification that is used in the insurance practice being evaluated. For instance, if age, sex, and smoking status are used to determine life insurance premiums, then "individuals" in this process would be each unique combination of age, sex, and smoking status.

It is important to understand that the two classes of fairness definitions—individual fairness and group fairness—are unrelated to individual insurance and group insurance. An individual fairness criterion could be tested for group insurance, and a group fairness criterion could be tested for individual insurance. For example, if one wished to test group fairness for pricing on an individual term insurance product with some underwriting, the groups could be defined based on issue state and the test could compare the average price for insureds from Hawaii (with a mean life expectancy of 80.7) against insureds from Mississippi (mean life expectancy of 71.9).[4] To test individual fairness in pricing for a group term insurance product that includes some underwriting, each policy—which covers a group of individuals and has a unique set of attributes used for pricing—would be treated as a single "individual' during the fairness testing process.

The distinction between individual and group fairness is fundamental, as each approach leads to different ways of evaluating equity in insurance practices. **Individual fairness seeks to achieve equality of treatment or outcome among individuals, at the expense of potentially different distributions of outcomes across groups, whereas group**

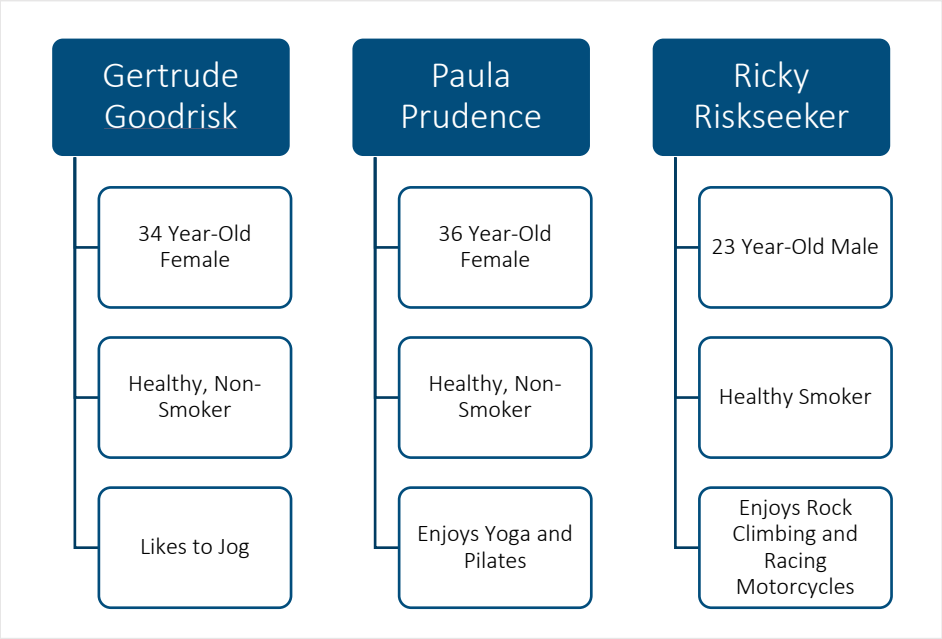
[2] (Charpentier, 2024), (Xin & Huang, 2023)
[3] (Xin & Huang, 2023)
[4] National Vital Statistics Reports Volume 71, Number 2 August 23, 2022, https://www.cdc.gov/nchs/data/nvsr/nvsr71/nvsr71-02.pdf

fairness seeks to achieve equality of the distributions of outcomes across groups, at the expense of inconsistent treatment or outcome of individuals.

Under an individual fairness framework, for instance, one might evaluate whether each insured's premium is directly commensurate with their mortality risk (i.e., life insurance premiums are solely a function of a mortality risk score). In other words, this means two identical mortality risks[5] would receive the same price, whereas an insured classified as having a lower mortality risk would receive a lower price than an insured classified as having a higher mortality risk.

Individual fairness is determined, then, by how specific individuals are treated relative to one another. For a practical example, consider the individuals described below in Figure 2. Paula Prudence and Gertrude Goodrisk have many characteristics of low relative mortality risk with their healthy habits, good health, and relatively lower baseline mortality rates. On the other hand, Ricky Riskseeker has several characteristics of higher relative mortality risk with risky habits (smoking and high-risk hobbies) as well as a higher baseline mortality rate.[6]

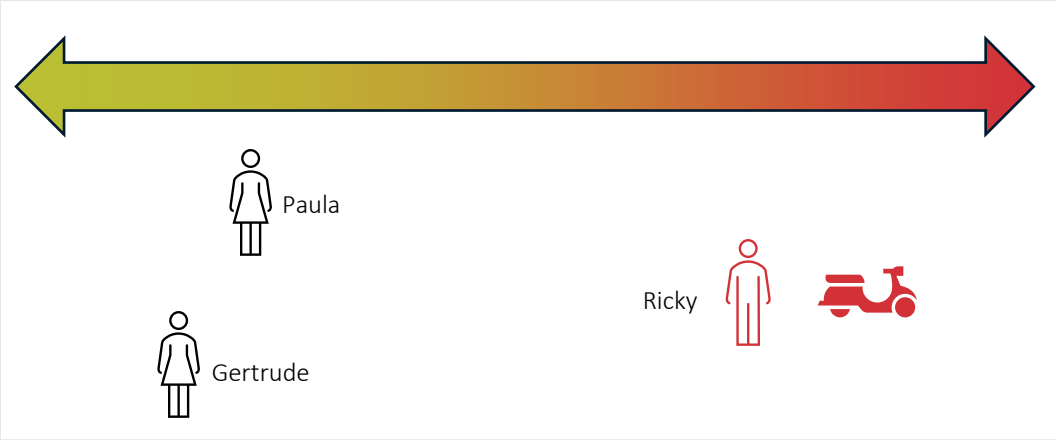
Figure 2
CHARACTERISTICS OF HYPOTHETICAL INSURED INDIVIDUALS



To meet an individual fairness standard, Paula and Gertude would pay similar premiums since they represent similar mortality risks, and Ricky would pay a higher premium since his mortality risk is higher. Said differently, Paula and Gertrude have similar and relatively lower expected life insurance claim costs, whereas Ricky has relatively higher expected claim costs. All else being equal, this again means that Ricky will have a higher life insurance premium.

[5] In practice, the individual evaluation of mortality typically maps the individual lives to an underwriting class, where the mortality risk evaluation happens, and which itself is a statistical estimate. An insurer can't precisely compute a single individual's risk contribution to a pool, only the corresponding risk estimate for a typical member from a group of individuals that share one or more observable characteristics.
[6] Using the 2017 Commissioners Standard Ordinary Tables, using age last birthday, male/female, smoker/non-smoker ultimate rates.

Figure 3
RELATIVE LIFE INSURANCE PREMIUMS FOR HYPOTHETICAL INSURED INDIVIDUALS



In contrast, **group fairness analyzes fairness of outcomes between groups, not between the individuals, and seeks to ensure that the treatment or decision is fair between the groups. In this way, group fairness evaluates whether outcomes are the same across different groups.** The characteristics used to define those groups could be anything that can be used for classifying populations—occupational classes, levels of education, recreational activities, and so forth. For example, one might evaluate average premium (rates) for life insurance policies across occupation classes.

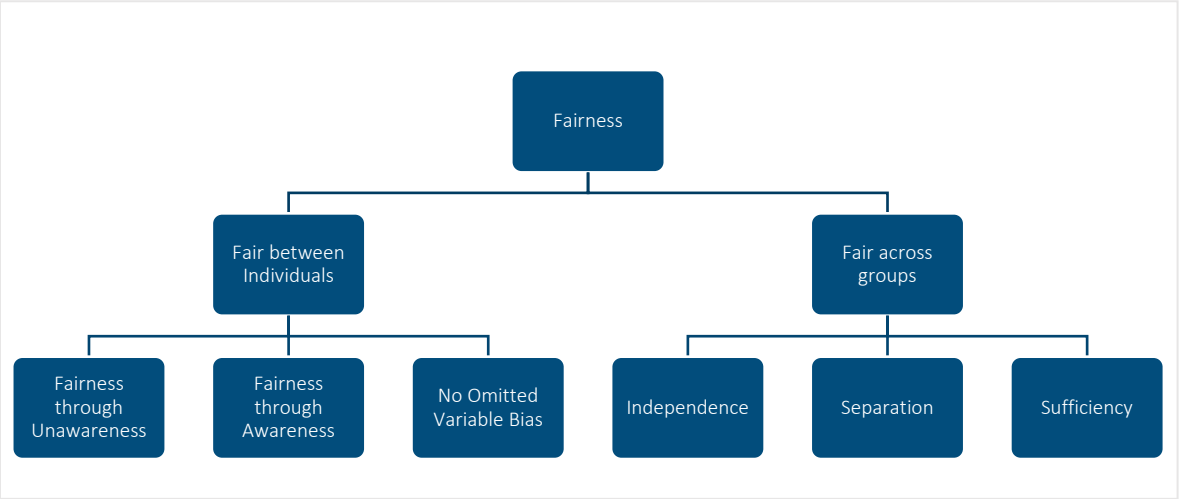
Group fairness metrics, in general, consist of calculating and comparing statistics for each group, such as the group average (e.g., average premium for males versus the average premium for females). Conversely, individual fairness metrics generally focus on quantifying the similarity or differences of inputs between various individuals.

Researchers have argued that an individual–group criteria split is somewhat illusory, because in practice even methods for implementing "individual" fairness criteria must ironically treat an individual person as a typical member of some "group", where the group is defined as all members of the population that share the same combination of attributes for the finite set of observed characteristics considered by the model or insurance practice.[7] This is particularly relevant in the context of insurance, where risks are generally categorized into classes in response to the inherent randomness associated with the coverage provided or other considerations such as volume and volatility of the data.

Ultimately, in discussing and measuring "fairness" in a given application, it is important to specify a precise definition that applies to the context at hand. Actuaries, with their quantitative expertise, are well-suited to express many fairness criteria in technical terms. Examples of some of the various definitions of fairness in use are summarized in the tree below. These definitions, which are by no means exhaustive, are discussed in further detail in Section 2.

[7] (Binns R. , 2020), (Côté, Côté, & Charpentier, 2024)

Figure 4
FAIRNESS CRITERIA TREE[8]



1.2 ACTUARIAL FAIRNESS

As the insurance industry has discussed the topic of fairness, the term "actuarial fairness" has frequently been referenced. Actuarial fairness generally refers to achieving equality across risks in the ratio of the outcome to actual or expected claim costs, such as consistency in loss ratios (ratio of premium to claim costs). In other words, differences in the outcome (e.g., differences in premiums, acceptance rates, etc.) across risks are acceptable if those differences are commensurate with differences in claim costs.

Actuarial fairness can be evaluated at both an individual and group level. In the example above, if Ricky's expected claim costs are twice that of Gertrude's, Ricky's premium would need to be twice that of Gertrude's to achieve actuarial fairness (assuming expenses are proportional to claim costs). This would achieve actuarial fairness at an individual level. Alternatively, if one were interested in evaluating fairness by sex, one could evaluate whether the ratio of claim costs to premium is consistent between males and females. Although achieving actuarial fairness at the individual level generally implies actuarial fairness at the group level, the reverse is not always true.

While actuarial fairness has traditionally been used by insurance companies as justification for many business practices, other stakeholders have raised questions as to whether the strict use of this fairness criteria is sufficient, or if other criteria should be used.

Please note that the fairness criteria presented in this paper vary in regard to the degree which they align with actuarial fairness.

[8] Note that this is distinct from the Fairness Tree published by (Center for Data Science and Public Policy, University of Chicago, 2024)

Section 2: Individual Fairness Criteria

There are a variety of criteria that have been defined on an individual fairness basis. The individual approach to fairness is straightforward and can be thought of as a concept of "equality" across individuals. The table below summarizes several individual fairness criteria.

Table 1
INDIVIDUAL FAIRNESS CRITERIA

Criterion	Principle	Example
Fairness through unawareness	Leave certain variables, such as those containing sensitive class information, out of risk calculations.[9]	Ensuring no sensitive rating variables are used in underwriting.
Fairness through awareness	Similar individuals—where similarity is defined by an assumed distance metric—receive similar treatment (e.g. prediction), where the similarity metric has been adjusted to account for known disparities among sensitive characteristics.[10]	Grouping insurance applicants according to a chosen similarity metric—such as a risk score that is normalized for known disparities in mortality rates by race or ethnicity due to access to healthcare—then assigning the same premium to individuals within each group who have the same risk score, while allowing for different premiums across groups with meaningfully different risk profiles.
No omitted-variable bias	Prevent indirect adverse impact arising by diagnosing any proxy relationship between sensitive characteristics and other variables and adjusting for the effect of the sensitive characteristics.[11]	Removing the effect of race or ethnicity on model outputs.

2.1 FAIRNESS THROUGH UNAWARENESS

The underlying principle of fairness through unawareness is to intentionally omit some data, such as an insured's sensitive class information, from the model. In practice, many insurers by default follow fairness through unawareness since they do not collect or the law restricts them from using data on statutorily "protected" characteristics.[12] This criterion is also known as "fairness through blindness."[13]

Proponents of this approach argue that by avoiding any direct incorporation of sensitive class information, the result is inherently fair. This argument is strengthened when the other inputs used are unbiased (i.e., uncorrelated) with respect to that sensitive class information.

[9] (Baumann & Loi, 2023) (Xin & Huang, 2023)
[10] (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012) (Xin & Huang, 2023)
[11] (Xin & Huang, 2023)
[12] For example, see Emerson, Jackob. 2022 April 15. "States that restrict payers from collecting race and ethnicity data from members." *Becker's Payer Issues*. https://www.beckerspayer.com/payer/states-that-restrict-payers-from-collecting-race-and-ethnicity-data-from-members.html
[13] (Baumann & Loi, 2023)

The weakness of this approach, however, is that it only avoids *directly* incorporating the information. **If sensitive class attributes are correlated with other attributes incorporated in the model, information about the sensitive class may still enter the analysis and be used indirectly.** As a hypothetical example, suppose that an insurer uses medical claims data in calculating a mortality risk score, and that medical claims data includes a medical diagnosis. If the diagnosis is one that is predictive of race or ethnicity, predictor values assigned to those medical claim codes may capture the direct impact of the disease or condition itself and also the impact of exogenous information that would be included in the estimated parameter if the mortality risk score model were built to predict all-cause mortality—and so would be a proxy variable for the blinded sensitive attribute. Thus, while fairness through unawareness would be satisfied, attributes that indirectly included sensitive information were nonetheless included.

Insurance redlining is an example of deliberately using proxy variables (National Commission on Urban Problems, 1968). Under this practice, within urban areas deemed to be "high risk", carriers would either decline to issue auto and home coverage or they would explicitly charge a higher rate. The urban areas marked with these red lines on maps were populated with a significant number of racial and ethnic minorities. Since the disparate treatment was driven by ostensibly neutral factors like zip code, no protected class information was used when declining the applications, or charging a higher premium, but the outcome was a disparate impact on those racial and ethnic minorities. This is an example where Fairness through Unawareness would have been satisfied, but the outcome was nonetheless biased.

Additional analysis regarding the relationship between the sensitive class and the data being used by the insurer may be required for the insurer to be certain that there is no occurrence of an unintentional, indirect impact. Blindness does not guarantee that there is not a disproportionate impact on the outcome, nor does it guarantee that the result could not be characterized as unfair.

Thus, the central paradox of fairness through unawareness is that ensuring a fair outcome requires that sensitive class information be collected, posing potential challenges in this criterion's selection and use.

Like other individual fairness criteria, measurability is a challenge **and there is no proposed metric for fairness through unawareness**. As such, it would be difficult to quantitatively test whether fairness through unawareness is satisfied. **Verification that fairness through unawareness is satisfied would require auditing the process and underlying data used in the pricing or underwriting process of interest to validate the lack of any direct incorporation of sensitive characteristics. Fairness through unawareness is satisfied when it can be verified that there is no direct inclusion of the sensitive characteristics.**

2.2 FAIRNESS THROUGH AWARENESS

The aim of fairness through awareness is to ensure that individuals who are similar, according to a specific measure, are treated in a similar manner.[14] In this context, "awareness" refers to knowing how to measure similarity between individuals, which is defined formally using a selected similarity metric. If relevant, this similarity metric can be selected to allow the user to reflect information about known disparities between groups. An example of this for life insurance is a mortality risk score that is used to assess similarity between individuals, where the mortality risk score

[14] (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012), (Kusner, Loftus, Russell, & Silva, 2018)

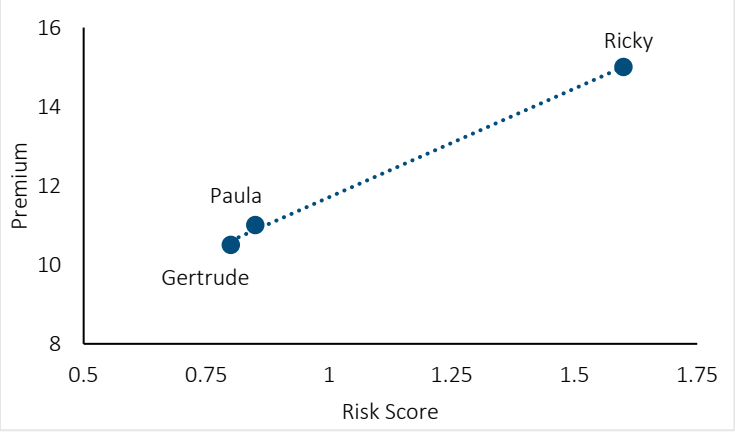
is determined to be appropriate for accepting risks and is normalized to account for the fact that different groups have different access to care (and that differences in access to care have an impact on mortality).

Fairness through awareness was originally formulated in response to the perceived shortcomings in some group fairness definitions, particularly demographic parity (discussed in Section 3.1 below),[15, 16] which requires that the classes of the group variable being evaluated are not predictive of which outcome an individual receives. For example, if underwriting classes in insurance followed demographic parity, then knowing an individual's demographic information would not provide any indication of which class they are assigned to—each class would have the same demographic makeup. In contrast, fairness through awareness evaluates whether individuals who receive the same outcome are actually similar based on a measurement specific to the task, such as a risk score, rather than their demographic characteristics. In this way, for many insurance applications, fairness through awareness more so resembles actuarial fairness compared to demographic parity.

To expand on the hypothetical example from Section 1, recall the three individual insureds: Gertrude Goodrisk, Paula Prudence, and Ricky Riskseeker, who have mortality risk scores of 0.8, 0.85, and 1.6, respectively. Assume that the risk scores have been adjusted for known disparities that exist in access to care by race and ethnicity. If the normalized risk scores are relative measures of mortality risk and are the only aspect being considered in this fairness analysis, satisfying fairness through awareness would mean that Gertrude and Paula, with risk scores of 0.8 and 0.85, should be treated similarly, while Ricky, with a risk score of 1.6, would be treated differently. If assessing life insurance prices, Gertrude would be expected to pay the lowest premium, Paula to pay a slightly higher premium, and Ricky to pay the highest premium. Figure 5 shows how a simple model could demonstrate fairness through awareness, assuming that all control variables are equivalent for these three insureds—same policy type, same underwriting done, and so forth.

Figure 5

HYPOTHETICAL MODELED LIFE INSURANCE PREMIUMS AGAINST RISK SCORES SATISFYING FAIRNESS THROUGH AWARENESS

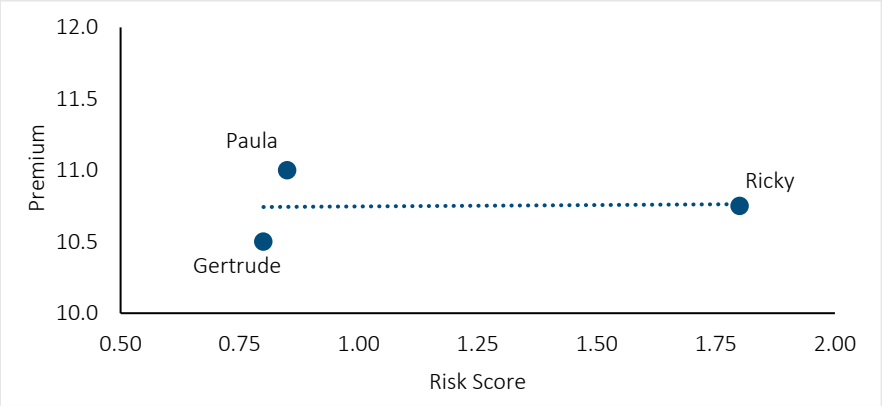


If Ricky, with his mortality risk score of 1.6, paid a lower premium than Paula or Gertrude—or even the same premium—this would not qualify as a fair outcome under fairness through awareness, as seen in Figure 6.

[15] Demographic parity is also known as Independence and discussed in section 3.1.
[16] (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012)

Figure 6

HYPOTHETICAL MODELED LIFE INSURANCE PREMIUMS AGAINST RISK SCORES NOT SATISFYING FAIRNESS THROUGH AWARENESS



The main challenge with this approach is choosing the right similarity metric for each situation, as the approach generally requires a measure that is tailored to both the problem (the particular scenario, task, or decision-making process at hand such as accepting or rejecting risks) and the desired outcome (the intended result or objective that the fairness assessment seeks to achieve). Additionally, it is important to consider circumstances specific to the analysis when defining the metric used and deciding which inputs to use in calculating similarity between individuals.[17] If the similarity metric is "blind" to group differences and disparities (that is, in the example above, the mortality risk scores are not adjusted for known differences that exist in access to care across different groups), then fairness through awareness provides no more fairness than fairness through unawareness. However, the causes of disparities are not always known, so it may not be possible to account for all disparities.

2.3 NO OMITTED-VARIABLE BIAS

Fairness through awareness improves upon fairness through unawareness by letting users consider disparities when assessing similarity between individuals. However, proxy effects for sensitive characteristics can still appear if the similarity metric or other variables are correlated with the sensitive characteristics. As a result, leaving out the sensitive attribute can still cause bias (i.e., indirect adverse impact). To address this, several bias mitigation techniques have been discussed in the literature which directly include the sensitive characteristic in the modeling process.

In this paper, we use the term "no omitted-variable bias" to mean a fairness criterion that requires removing the proxy effect of sensitive characteristics from the modelled outcome by directly including the sensitive characteristics in the analysis. **That is, by including the sensitive characteristic in the analysis (i.e. not omitting it), the goal of no omitted-variable bias is to ensure there is also no indirect effect of those attributes on outcomes.**

To return to the example of Gertrude, Paula and Ricky, suppose that the insurer wants to ensure that its mortality risk scores do not serve as a proxy for sex. If smoking is more prevalent among males[18] and females are more than

[17] As an example, many insurers will offer multiple life insurance product options—they may vary based on the death benefits offered, the build-up of a cash value over time, the manner in which the product is sold, the target market for the product, or innumerable other dimensions. Depending on all of those facts and the risk appetite of the insurer, different amounts of information may be collected from the applicant. The selected similarity metric may need to incorporate the degree to which a policy is underwritten (e.g. whether a guaranteed issue policy or a fully underwritten policy with fluids drawn and medical records reviewed), the size of the policy sought, the type of life insurance product purchased (as permanent insurance products incorporate a build-up of surrender value over time, potentially reducing the insurance risk), all in addition to the specifics for each insured life.

[18] (Centers for Disease Control and Prevention, 2024)

twice as likely to participate in yoga,[19] then any algorithm that does not directly use sex as a predictor but uses smoking and yoga participation would inherently also be capturing the effect of sex in the model due to the correlation between these attributes and sex. Accordingly, if one controlled for the effect of sex by using it directly in the analysis, the remaining effect of smoking and yoga participation would capture the pure effect of these attributes (i.e., not a combination of their effect and the effect of sex).

There are several ways to diagnose whether bias exists due to the omission of the sensitive characteristic. One way is to incorporate the sensitive characteristic directly into the model as an explanatory variable to act as a control variable. The coefficients of this model can then be compared to coefficients from a version of the model that excludes the sensitive characteristic. If the coefficients for certain variables change significantly when the sensitive characteristic is added, this would suggest those variables are correlated with the sensitive characteristic.[20]

Alternatively, this kind of indirect adverse impact may be diagnosed by taking the explanatory variables of the model (excluding the sensitive characteristic) and evaluating whether they are predictive of the sensitive characteristic. This would be done by using the explanatory variables in a separate model, where the target variable is the sensitive characteristic. If the explanatory variables are predictive of the sensitive characteristic, this would suggest the explanatory variables act as proxies for the sensitive characteristic when they are included in the model without the sensitive characteristic present.[21]

Once such indirect adverse impact has been diagnosed, there are several options for mitigation. One option is to include the sensitive characteristic as a control variable in the model. The disadvantage of this approach is that when multicollinearity exists, it can be difficult to control how the model parses out the effects of each variable. Alternatively, one could take an iterative modeling approach, wherein the modeler includes the sensitive characteristic in an initial model and excludes any explanatory variables that have been identified as having proxy effects. After, the target variable can be adjusted for the effect of the sensitive characteristic, and a secondary model using the remaining explanatory variables (the variables excluded from the initial model) can be run on the adjusted target variable. This approach ensures the effect of the sensitive variable is not captured in the parameter estimates of the explanatory variables identified as having proxy effects.

A third approach is to control for the sensitive characteristics in the modeling stage and then adjust the modeling outputs via a statistical procedure that removes or "neutralizes" the effect of the sensitive characteristic. Lindholm refers to this approach as "discrimination-free pricing. "[22]

A major drawback of any of these no omitted-variable bias approaches is that they require the direct use of the sensitive characteristics in the modeling process. The explicit use of sensitive characteristics in model construction or adjustment could be perceived as controversial to some stakeholders, as some may view it as inappropriate or even discriminatory in itself, regardless of the intention to promote fairness.

[19] (Elgaddal & Weeks, 2022)
[20] (Pope & Sydnor, 2011)
[21] (du Preez, et al., 2024)
[22] (Lindholm, Richman, Tsanakas, & Wuthrich, 2022)

Section 3: Group Fairness Criteria

To evaluate group fairness, several criteria may be considered which are outlined and discussed below. Note that these criteria fundamentally differ in how fairness is measured.[23]

Table 2
GROUP FAIRNESS CRITERIA

Criterion	Principle	Example
Independence	The modeled risk score or predicted value of the outcome variable is statistically independent from any sensitive characteristics.[24]	Testing whether the rejection rate for life insurance applications is equal across each race and ethnicity class.
Sufficiency	After conditioning on the predicted value of the outcome variable, the observed/true value of the outcome variable is statistically independent from any sensitive characteristics.[25]	The actual distribution of claims incurred by the insurer is equal across groups, once controlling for the estimated mortality risk score within each group.
Separation	After conditioning on the observed, true value of the outcome variable, the modeled risk score or predicted outcome is statistically independent from any sensitive characteristics.[26]	The likelihood of a submitted claim being selected for fraud investigation or recission is equal for race and ethnic groups, once controlling for the ultimate detection of fraud or recission of the policy.

3.1 INDEPENDENCE

Under the independence group fairness criteria, fairness is achieved if outcomes are equal across groups (i.e., the outcome variable is statistically *independent* of the levels of the group variable). This is also known as "demographic parity" or "statistical parity."[27]

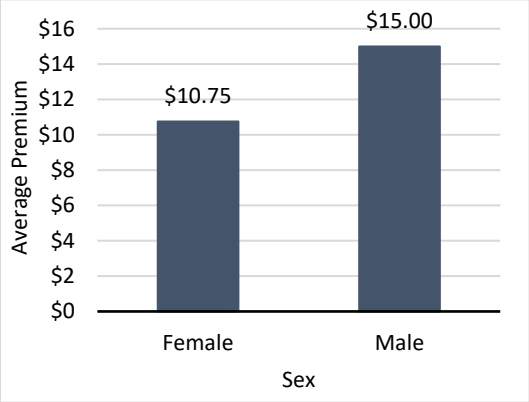
Calculating a demographic parity metric begins with segmenting the population into groups. Once individuals are assigned to groups, analysis is done at the group level. For each group, the outcome or variable of interest, such as underwriting rejection rate, is calculated, and then compared across groups.

The comparison can be stated as an absolute difference (e.g. Group A - Group B), which is referred to as "mean differences", or expressed on a relative basis (Group A / Group B), which is referred to as "impact ratios" or "adverse impact ratios."[28][29] Z-tests or t-tests can be used to test whether differences observed would be considered statistically significant, if appropriate for the analysis. The appropriate comparison to use depends on the nature of the outcome variable of interest.

Continuing the hypothetical example from Section 1, assume Gertrude, Paula, and Ricky had premiums of $10.50, $11.00, and $15.00, respectively, and that one was interested in testing whether the premiums are fair across sexes. Under this example, the average premium for females is $10.75 (average of the premiums for Gertrude and Paula), whereas the average premium for males is $15.00, as outlined in Figure 7:

[23] (Barocas, Hardt, & Narayanan, 2023) (Baumann & Loi, 2023) (Lindholm, Richman, Tsanakas, & Wuthrich, What is fair? Proxy discrimination vs. demographic disparities in insurance pricing, 2024)
[24] (Barocas, Hardt, & Narayanan, 2023) (Côté, Côté, & Charpentier, 2024) (Xin & Huang, 2023)
[25] (Barocas, Hardt, & Narayanan, 2023) (Côté, Côté, & Charpentier, 2024) (Lindholm, Richman, Tsanakas, & Wuthrich, What is fair? Proxy discrimination vs. demographic disparities in insurance pricing, 2024)
[26] (Barocas, Hardt, & Narayanan, 2023) (Côté, Côté, & Charpentier, 2024) (Lindholm, Richman, Tsanakas, & Wuthrich, What is fair? Proxy discrimination vs. demographic disparities in insurance pricing, 2024)
[27] (Barocas, Hardt, & Narayanan, 2023) (Lindholm, Richman, Tsanakas, & Wuthrich, What is fair? Proxy discrimination vs. demographic disparities in insurance pricing, 2024)
[28] (Gailey, 2023)
[29] (Irving, 2024)

Figure 7
HYPOTHETICAL AVERAGE LIFE INSURANCE PREMIUMS BY SEX



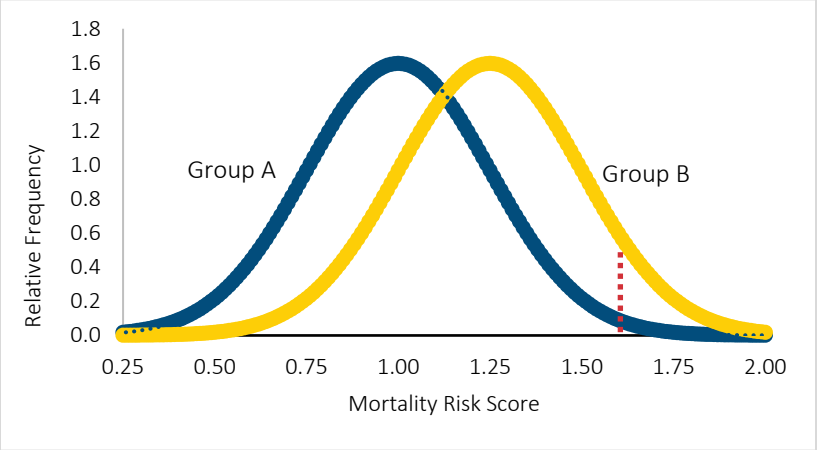
In this example, the average premiums by sex differ by $4.25, or 39.5%. The premiums in this situation do not meet the independence criteria for fairness by sex (i.e., do not achieve demographic parity), assuming that this measured difference is practically (and to the extent relevant, statistically) significant.

A clear strength for assessing demographic parity is that it is a criterion that has several metrics (whether differences or ratios) that are straightforward to calculate and simple to explain. Once individuals are assigned to groups, the mean differences or impact ratio metrics are directly measurable from the data.

The appeal of its simplicity does not mean that an approach under the independence criterion is not without its drawbacks. **For one, independence alone may be viewed as unfair to individuals. Since it does not require that similar cases are treated similarly, an approach that is deemed fair under independence may treat two individuals who are similar in all respects aside from what group they are assigned to differently, depending on the remediation practices employed.** [30] As a simple example, consider a case where an insurer calculates a mortality risk score based on prescription data and uses that information alone to accept or reject insurance applications, and from quantitative studies the insurer can demonstrate that their risk score is predictive of mortality outcomes. In this example, assume that the distribution of those mortality risk scores for two groups differs, as shown in Figure 8. If the insurer sets only a single risk score cutoff for accepting applications as indicated by the red dotted line below, a higher proportion of Group B will be rejected because of that decision. Rejected applications are those with risk scores higher than (to the right-hand side of) the red dotted line in Figure 8. This would fail a strict definition of fairness under independence.

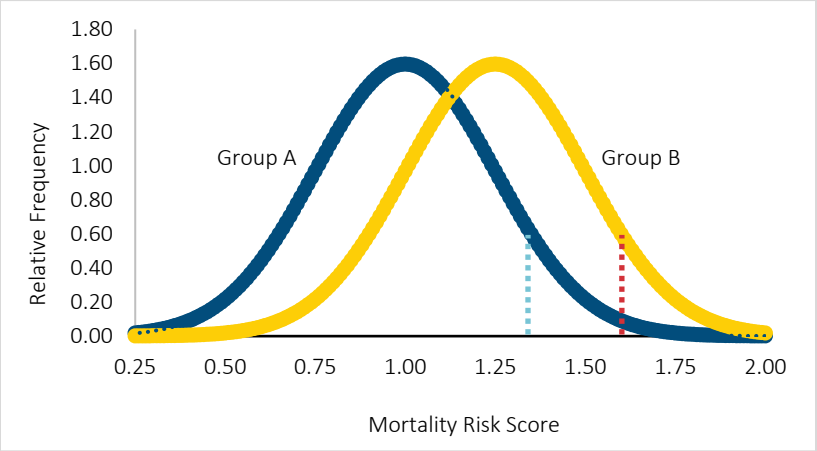
[30] (Baumann & Loi, 2023)

Figure 8
MORTALITY RISK SCORE DISTRIBUTION BY GROUP



However, if the insurer seeks to achieve group fairness through independence and so wants to equalize the proportion of applications rejected from both groups, a simple remediation would be to set different risk score cut-offs for the two groups to equalize the proportion of each that are rejected (this remediation is simple only in a paper; in practice, it would be highly complex). This new underwriting cut-off (Figure 9) is shown below as a light blue dotted line, and is set to reject the same proportion of Group A as is rejected in underwriting Group B. Using this new cutoff, the insurer's underwriting would now satisfy independence. However, an individual in Group A who received a mortality risk score of 1.5 would be rejected by the insurer, while an otherwise identical individual with a mortality risk score of 1.5 in Group B would be accepted.

Figure 9
MORTALITY RISK SCORE DISTRIBUTION BY GROUP—DISPARATE TREATMENT



This illustrates a key shortcoming of strict independence in a life insurance context, since many people would expect that an insurer acting fairly would treat these otherwise similar cases in a similar manner, and explicit adjustments based on the group membership violate that notion of individual fairness. Further, the direct use of an individual's group classification in the business practice (such as defining underwriting cut-offs based on an individual's group classification) may not even be allowed based on state statutes.

An alternative remedy is to remove variables from the model or process that contribute to the differences in the outcomes across groups so that the independence criteria is met (in other words, calculating a new mortality risk

score that is fully independent of the characteristics defines the groups.[31] Practically, this means sacrificing some degree of accuracy in predictions, which ultimately introduces a level of cross-subsidization. For example, in the context of pricing, individuals with lower risk would pay more than their commensurate risk level in order to subsidize individuals with higher risk. This can have further consequences when it comes to anti-selection. If one company, for instance, uses the approach in Figure 7 and a second company calculated a mortality risk-score that satisfies independence, individuals would fall into one of three categories, those who would:

1. pass underwriting for both companies,
2. be declined in underwriting by both companies, and
3. be accepted by one company and declined by the other.

Depending on the accuracy of the two approaches and the degree of "shopping" that applicants do, the resulting mortality rates may differ. For this reason, insurance carriers as economic actors tend to benefit from individual fairness definitions since they result in more precise risk classification and reduce adverse selection. Thus, there is a natural tension that emerges when adopting independence as a fairness criterion.[32]

In the extreme, mandated community rating can resolve this conflict between individual fairness and independence and anti-selection issues.[33] In community rating, insurers are precluded from varying premiums based on certain characteristics, such as age, sex, health status or other factors. Under this approach, some or all of the variation in premiums is eliminated and, as a result, the differences that are otherwise present between individuals and between groups are removed. For community rating to be successful, lower-risk individuals must be incentivized or required to participate in the risk pool at the higher, community-rated premium, because otherwise a rational, lower-risk individual will determine that the community-rated premium is overpriced.[34] At the margin, as some proportion of lower-risk individuals exit the insurance pool, the community-rated premium for the remaining participants will increase and additional remaining lower-risk individuals will opt out of purchasing insurance. This self-reinforcing phenomenon is colloquially called a "death spiral" and a potential consequence of community-rating in the absence of non-market incentives or mandated requirements to produce a stable market structure.[35]

Given the drawbacks listed here and elsewhere in the literature that exist under strict independence, several "relaxations" of the requirements have been developed. Some of these are discussed in Table 3.

[31] (Côté, Côté, & Charpentier, 2024) (Frees & Huang, 2023) (Komiyama, Takeda, Honda, & Shimao, 2018)
[32] (Baumann & Loi, 2023)
[33] (Baumann & Loi, 2023)
[34] (Baumann & Loi, 2023), for an implemented example see (Commonwealth of Australia, 2021)
[35] For contrasting health insurance examples, see (Frech & Smith, 2015) and (Buchmueller & DiNardo, 2002)

Table 3
RELAXED GROUP FAIRNESS CRITERIA

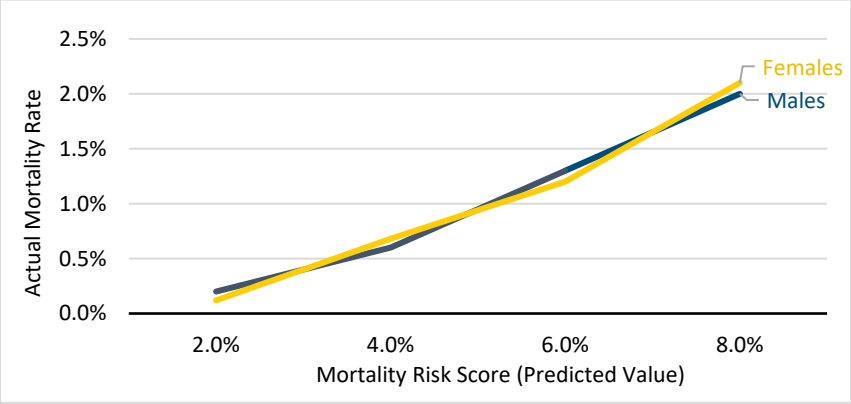
Criterion	Principle	Example
Relaxed Demographic Parity	A relaxation of "independence" where deviations in the distributions for different values of sensitive variables are constrained.[36]	The price of insurance for a non-sensitive group can't be less than e.g. 80% of the price of insurance in the sensitive group.
Conditional Demographic Parity	A relaxation of "independence" where some legitimate non-sensitive attributes are allowed to enter the pricing model even if they correlate with sensitive group status, assuming they are actuarially relevant.[37]	An underwriting model is permitted to use a mortality risk score based on prescription data and medical claims data that has some correlation to sensitive classes, so long as the prescription and medical claims data can be shown to be actuarially relevant.
Conditional Disparate Impact	A further relaxation of conditional demographic parity where some defined difference is permitted between groups, conditional on having the same non-sensitive attributes.[38]	The expected mortality risk score between two sensitive groups, conditional on prescription data and medical claims data, are permitted to differ by up to 10%.

Another motivation for using "relaxed" criteria is that they can act as compromises between two conflicting criteria. For instance, "fairness through unawareness" and "independence" are both special cases of "conditional demographic parity," in which all available non-sensitive attributes are either deemed legitimate or nonlegitimate, respectively.[39]

3.2 SUFFICIENCY

A drawback of the independence fairness criteria is that it does not at all consider the model's accuracy. That is, it does not account for how well the model's predictions align with actual outcomes. For instance, in continuing the Gertrude, Paula, and Ricky example above, assume that the premiums were based on the mortality risk score, the model for which produces actual-to-expected results when reviewing actual mortality rates by sex (Figure 10):

Figure 10
MALE AND FEMALE ACTUAL MORTALITY RATES BY MORTALITY RISK SCORE



[36] (Xin & Huang, 2023)
[37] (Xin & Huang, 2023; Corbett-Davies, 2017)
[38] (Xin & Huang, 2023)
[39] (Xin & Huang, 2023)

In this example, males and females with similar mortality risk scores (i.e., similar model predictions) have similar actual mortality rates. Or, in another way, the observed mortality rates do not vary by sex for any given mortality risk score.

The example above demonstrates the concept of sufficiency, which pertains to the principle that, after conditioning on the predicted outcome (e.g., modeled risk score), the observed outcome variable (i.e., true value) should be independent of the defined group characteristics.[40] That is, the model's accuracy is the same across groups for individuals with similar model predictions. In this way, sufficiency ensures that the predictive power of a model is consistent across different groups, which is why it is also referred to as "predictive parity."[41] For this reason, sufficiency is inherently related to the concept of actuarial fairness.

Potential metrics that align with the sufficiency criterion focus on how often the model's predictions turn out to be correct for each group. Examples include the positive predictive value (PPV), also known as precision, and the negative predictive value (NPV). PPV measures the proportion of positive predictions that are true positives (i.e., out of all the times the model predicted something would happen, how often was it actually right?), while NPV measures the proportion of negative predictions that are true negatives (i.e., out of all the times the model predicted something would not happen, how often was it actually right?). To apply these metrics in practice, one could calculate the PPV and NPV for each group and compare them across groups. For example, if an insurance company uses a model to predict who is high risk, it would compare the PPV of these predictions across different groups to ensure that the accuracy of these predictions does not favor one group over another.

The general strengths of the sufficiency criterion include its direct applicability and interpretability in practical settings. Since sufficiency is often a by-product of standard machine learning techniques aimed at minimizing prediction errors, it can be straightforward to implement.[42] However, a potential weakness is that achieving sufficiency does not necessarily imply fairness in other aspects, such as equal opportunity or demographic parity. Additionally, sufficiency may inadvertently sustain historical biases if the underlying data reflects those prejudices.[43]

Within the context of insurance, given the significant element of randomness in death claims and long lag between pricing or underwriting and the actual claim, there also can be a fundamental challenge in applying a strict sufficiency criterion to pricing or underwriting outcome for most long-duration life insurance products. For example, it may not be possible to credibly evaluate whether a model's predictions are consistently accurate across groups for every possible prediction value due to the historical claims experience. Accordingly, some degree of clustering of the predicted values is likely needed to evaluate this fairness criteria in practice in the insurance industry.

Determining thresholds of acceptability for sufficiency metrics involves setting acceptable ranges for differences in PPV and NPV across groups. This can be challenging, as it requires balancing statistical significance with practical. For instance, a small but statistically significant difference in PPV between groups might not be practically significant, whereas a large difference could indicate substantial unfairness.

Potential unintended consequences of using sufficiency as a fairness criterion include the possibility of neglecting other fairness aspects. For example, while sufficiency ensures that predictions are equally accurate across groups, it does not address whether the model inadvertently perpetuates existing inequalities. Additionally, focusing solely on sufficiency might lead to models that are technically fair but fail to address broader ethical concerns about equity and justice in insurance practices. Since the sufficiency criterion requires the absence of a cross-subsidization

[40] (Barocas, Hardt, & Narayanan, 2023) (Baumann & Loi, 2023) (Binns R. , 2020)
[41] (Lindholm, Richman, Tsanakas, & Wuthrich, What is fair? Proxy discrimination vs. demographic disparities in insurance pricing, 2024)
[42] (Barocas, Hardt, & Narayanan, 2023), Chapter 3.
[43] (Barocas, Hardt, & Narayanan, 2023)

mechanism in the risk pool, sufficiency seems unsuitable for "social-good" insurance products that seek to leverage those cross-subsidies.[44]

3.3 SEPARATION

Separation is similar in concept to sufficiency, but the conditioning element is flipped. **That is, separation in insurance fairness pertains to the principle that, after conditioning on the observed (i.e., true value) of an outcome variable, the predicted outcome (e.g., modelled risk score) should be statistically independent from the defined group characteristics.** [45]

Consider, for example, a model used to detect instances of claim fraud. To test the separation fairness criteria in the context of sex, one could take the data used to train the model and segment it between claims that were fraudulent and those that were not (i.e., segment by their actual value). Then, within each cohort, one could compare by sex the percentage of claims the model correctly predicted as being fraudulent to test whether this percentage varied by sex (Figure 11):

Figure 11
MALE AND FEMALE PREDICTED FRAUD RATE BY ACTUAL FRAUD EXPERIENCE

The example in Figure 11 demonstrates that the hypothetical model's predicted rate of fraud does not materially vary by sex, regardless of whether there actually was fraud, which would be considered fair under the separation criteria.

More formally, for classification models where the outcome of the model places an individual into a distinct group (e.g., identifying fraudulent claims), Separation corresponds to evaluating whether the error rates are consistent across groups.[46] One way to do this is to compare the true positive rates (TPR) and false positive rates (FPR) across groups (also known as equalized odds[47]). That is, across groups, one would compare the percentage of the time the model correctly identifies something as positive (e.g. fraudulent claim) and the percentage of the time the model incorrectly identifies something as positive. **If the TPR and FPR are equal across groups, the model satisfies separation**.

[44] (Baumann & Loi, 2023)
[45] (Barocas, Hardt, & Narayanan, 2023) (Baumann & Loi, 2023) (Binns R. , 2020)
[46] (Lindholm, Richman, Tsanakas, & Wuthrich, What is fair? Proxy discrimination vs. demographic disparities in insurance pricing, 2024)
[47] (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021)

Another approach is to average the discrepancies in TPR and FPR across groups to obtain a single measure of fairness.[48] Additionally, one might compare the total counts of true positives and false positives between groups, rather than their frequencies, to assess disparity.

Consider an example where an insurer uses a binary classification model to identify fraud claims. To apply the TPR and FPR metrics, the insurer would calculate the true positive and false positive rates for each group (e.g., different racial or sex groups). If Group A has a TPR of 82% and Group B has a TPR of 88%, while Group A has an FPR of 22% and Group B has an FPR of 16%, the insurer would recognize a disparity potentially indicating the separation criterion isn't satisfied. Alternatively, if the insurer finds that the average discrepancies in TPR and FPR between groups are minimal (or statistically insignificant), this suggests closer adherence to separation.

The strengths of using separation as a fairness metric include its direct measurability and clear interpretability (e.g., equality of error rates across groups). However, a significant weakness is that it evaluates fairness based on actual outcomes rather than uncertain future outcomes. This can be particularly challenging for many insurance applications where the intended purpose of many models is to predict expected outcomes. For example, a model used for pricing insurance is intended to predict the expected costs of a particular risk, not the actual cost, as insurance is inherently based on a risk-pooling principle where individuals share the cost of risk, regardless of which individuals incur the actual losses.[49] For life insurance coverages that may require years or decades for the outcome to be observed, collecting sufficient observations to conduct a fairness analysis may be protracted. Similar discussions on the drawbacks of separation have also been discussed for general insurance.

Section 4: Conclusion

In summary, this paper has explored the multifaceted concept of fairness within the life insurance industry, delving into both individual and group fairness criteria, and the associated metrics for assessing and ensuring fairness. By examining the operational and financial implications of implementing fairness requirements, this paper highlighted the inherent trade-offs between achieving social equity and maintaining actuarial soundness.

While the focus of this paper is on fairness metrics, there are a variety of other factors practitioners will need to consider when implementing bias analyses in practice. These considerations range from how to obtain sensitive data when it has not been collected (see *Statistical Methods for Imputing Race and Ethnicity*[50] for more information on how to infer sensitive data), to weighing between statistical significance (the likelihood that a given result is not due to chance) versus practical significance (whether the observed result is large enough to be meaningful).

Practitioners must also consider whether the evaluation of a model itself is appropriate for assessing whether there is bias in the business practice at hand. For example, an insurer may use a model to inform pricing decisions but may implement manual overrides to the model due to business decisions at implementation. Alternatively, a pricing model may be developed to predict loss, when the premium charged also includes provisions for expenses that are determined outside of the model. In these cases, the practitioner may want to evaluate whether solely testing fairness on the model and not the premium actually charged is appropriate.

Finally, stakeholders must consider the practical and financial trade-offs of fairness requirements. These may include reduced model accuracy (e.g. unbalanced premiums, anti-selection etc.), increased operational costs (cost

[48] (Pfisterer, Siyi, & Lang, 2024)
[49] (Baumann & Loi, 2023)
[50] Society of Actuaries Research Institute. (2024). *Statistical methods for imputing race and ethnicity* (Research Report). Authors: Larry Baeder, Erica Baird, Peggy Brinkmann, Joe Long, Caleb Stracke, Kweweli Togba-Doya, Gabriele Usan, Natalie Weaver, & Meseret Woldeyes. Society of Actuaries Research Institute. https://www.soa.org/resources/research-reports/2024/stat-methods-imputing-race-ethnicity/.

to conduct and audit testing, cost of reporting, and potential legal defense costs), and shifts in how costs and benefits are distributed among insurers, policyholders, and other stakeholders. The real-world impact of these trade-offs is not yet fully understood.

It is also important to note that fairness metrics come with limitations. For instance, while fairness metrics can highlight disparities, they do not always explain their underlying causes, and as such, often require additional due diligence to understand underlying causes. Furthermore, satisfying a technical fairness criterion does not guarantee that the intended ethical outcome has been achieved, since ethical perspectives can vary among stakeholders and often extend beyond what can be captured by quantitative measures. Finally, it is often not possible to simultaneously satisfy multiple definitions of fairness. **In many cases it may not be possible to achieve fairness at both an individual and group level.** The extent to which these fairness criteria agree depends on whether correlation exists between the attributes used to define the groups and other variables in the data (i.e., outcome variable and explanatory variables). For example, it is well known that male average life expectancy is less than female average life expectancy.[51] Thus, an insurer will not be able to define a premium scale with the same average premiums across males and females while also reflecting the differing risks and costs. As such, a tradeoff must be made in deciding what it means to be fair, which underscores why different stakeholders may have different perspectives on fairness definitions. While there are some fairness approaches that seek to combine elements of both individual and group fairness criteria, there is often no single "right" answer.

All of this should not be taken to mean that individul and group fairness are completely incompatible with one another. Different definitions of fairness within individual and group fairness relate to precisely how one defines equality. For a concrete pricing example, fairness between groups may be stated in terms of demographic parity, where an example of a fair outcome could require average premiums to be equal across groups. Fairness between groups could alternatively be defined as an equal accuracy across groups of estimating the expected present value of future claims (in other words, equal error rates). In this case, individual and group fairness could be achieved at the same time, if the distribution of risk were similar between the groups. The applicability of the fairness criterion also depends on the choice of target variables.

As the life insurance industry continues to evolve with advancements in technology and shifts in societal expectations, it is imperative for actuaries, regulators, and industry stakeholders to engage in ongoing dialogue and research to refine these fairness metrics and their real-world application. This will help in developing balanced life insurance practices that not only comply with ethical and regulatory standards but also foster trust and inclusivity in the broader societal context. Actuarial organizations whose missions include informing public policy, such as the American Academy of Actuaries, will be vital in these efforts. The insights provided in this paper aim to equip practitioners with a comprehensive understanding of the complexities surrounding fairness in life insurance, guiding them in making informed decisions that balance ethical considerations with practical feasibility.

[51] For example, by referring to the 2017 Commissioners Standard Ordinary Tables, and noting that male mortality rates are generally higher than the comparable female mortality rates.

Section 5: Acknowledgments

The researchers' deepest gratitude goes to those without whose efforts this project could not have come to fruition: the Project Oversight Group and others generously shared their wisdom, insights, advice, guidance, and review of this study prior to publication.

Project Oversight Group members:

Dorothy L. Andrews, ASA, CSPA, MAAA, Ph.D.

Brian Bayerle, FSA, MAAA

Andrew Clark, PhD

Bruce A. Friedland, FSA, MAAA, CLU, ChFC, MBA,

Thomas P. Hinrichs, FSA, MAAA

Hezhong (Mark) Ma, FSA, MAAA

Shisheng (Rose) Qian, FSA, CERA

David Sandberg, FSA, CERA, FCA, MAAA

Mark A. Sayre, FSA, CERA

David Schaub, FSA, CERA, MAAA, AQ

Matthew S. Wolf, FSA, CERA, MAAA

At the Society of Actuaries Research Institute:

Lisa S. Schilling, FSA, EA, FCA, MAAA

Barbara Scott, Senior Research Administrator

References

Barocas, S., Hardt, M., Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities.* MIT Press. Retrieved from https://fairmlbook.org/

Baumann, J., Loi, M. (2023). Fairness and risk: an ethical argument for a group fairness definition insurers can use. *Philosophy & Technology, 36*(3), 45.

Binns, R. (2020). On the apparent conflict between individual and group fairness. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, (pp. 514-524).

Buchmueller, T., Dinardo, J. (2002). Did Community Rating Induce an Adverse Selection Death Spiral? Evidence from New York, Pennsylvania, and Connecticut. *American Economic Review, 92*(1), 280-294. doi:10.1257/000282802760015720

Center for Data Science and Public Policy, University of Chicago. (2024). *Aequitas*. Retrieved May 6, 2024, from http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/

Centers for Disease Control and Prevention. (2024, September 10). Tobacco product use among adults—United States, 2022. *2022 National Health Interview Survey (NHIS) Highlights*.

Charpentier, A. (2024). *Insurance, biases, discrimination and fairness* (1st ed.). Cham: Springer Nature Switzerland. doi:10.1007/978-3-031-49783-4

Commonwealth of Australia. (2021). *About private health insurance*. Retrieved December 5, 2023, from https://www.health.gov.au/topics/private-health-insurance/about-private-health-insurance

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A. (2017). Algorithmic decision and the cost of fairness. *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797–806). New York, NY: USA: Association for Computing Machinery. doi:10.1145/3097983.3098095

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

du Preez, V., Bennet, S., Byrne, M., Couloumy, A., Das, A., Dessain, J., Galbraith, R., King, P., Mutanga, V., Schiller, F., Zaaiman, S., Moehrke, P., van Heerden, L. (2024). From bias to black boxes: understanding and managing the risks of ai–an actuarial perspective. *British Actuarial Journal, 29*(6).

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science* (pp. 214-226). New York, NY: Association for Computing Machinery. doi:10.1145/2090236.2090255

Elgaddal, N., Weeks, J. D. (2022). Yoga among adults age 18 and older: United States, 2022. *NCHS Data Brief, No. 501*. Retrieved from https://www.cdc.gov/nchs/data/databriefs/db501.pdf

Frech, H. E., Smith, M. P. (2015). Anatomy of a Slow-Motion Health Insurance Death Spiral. *North American Actuarial Journal, 19*(1), 60-72. doi:10.1080/10920277.2014.982871

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

Gailey, A. (2023). *What Is disparate impact testing?* Retrieved October 14, 2025, from https://media.crai.com/wp-content/uploads/2023/01/30101848/FE-Insights-What-is-Disparate-Impact-Testing.pdf

Irving, J. M. (2024, July 30). *Blog Post Series: AI Fairness 360- Mitigating Bias in Machine Learning Models*. Retrieved October 14, 2025, from Medium: https://medium.com/@james.irving.phd/blog-post-series-ai-fairness-360-mitigating-bias-in-machine-learning-models-c1ec744c91c4

Komiyama, J., Takeda, A., Honda, J., Shimao, H. (2018). Nonconvex optimization for regression with fairness constraints. *International conference on machine learning*, (pp. 2737–2746).

Kusner, M., Loftus, J., Russell, C., Silva, R. (2018). *Counterfactual fairness.* Retrieved from https://arxiv.org/abs/1703.06856

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

Lindholm, M., Richman, R., Tsanakas, A., Wüthrich, M.V. (2023). What is fair? Proxy discrimination vs. demographic disparities in insurance pricing. *Scandinavian Actuarial Journal*. doi:10.1080/03461238.2024.2364741

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

National Commission on Urban Problems. (1968). *Building the American City: Report of the National Commission on Urban Problems to the Congress and to the President of the United States.* Washington DC: U.S. Government Printing Office.

O'Neil, C. (2016). *Weapons ofMath Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York, NY: Crown Publishing Group.

Pfisterer, F., Siyi, W., Lang, M. (2024). *Fairness metrics.* Retrieved May 06, 2024, from Vignette for R package 'mlr3fairness', version 0.3.2.

Pope, D. G., Sydnor, J. R. (2011). Implementing anti-discrimination policies in statistical. *American Economic Journal: Economic Policy, 3*(3), 206-231.

Xin, X., Huang, F. (2023). Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. *North American Actuarial Journal*, 1-35.

Xin, X., Hooker, G., Huang, F. (2024). *Why you should not trust interpretations in machine learning: adversarial attacks on partial dependence plots.* arXiv preprint arXiv:2404.18702.

About The Society of Actuaries Research Institute

Serving as the research arm of the Society of Actuaries (SOA), the SOA Research Institute provides objective, data-driven research bringing together tried and true practices and future-focused approaches to address societal challenges and your business needs. The Institute provides trusted knowledge, extensive experience and new technologies to help effectively identify, predict and manage risks.

Representing the thousands of actuaries who help conduct critical research, the SOA Research Institute provides clarity and solutions on risks and societal challenges. The Institute connects actuaries, academics, employers, the insurance industry, regulators, research partners, foundations and research institutions, sponsors and non-governmental organizations, building an effective network which provides support, knowledge and expertise regarding the management of risk to benefit the industry and the public.

Managed by experienced actuaries and research experts from a broad range of industries, the SOA Research Institute creates, funds, develops and distributes research to elevate actuaries as leaders in measuring and managing risk. These efforts include studies, essay collections, webcasts, research papers, survey reports, and original research on topics impacting society.

Harnessing its peer-reviewed research, leading-edge technologies, new data tools and innovative practices, the Institute seeks to understand the underlying causes of risk and the possible outcomes. The Institute develops objective research spanning a variety of topics with its strategic research programs: aging and retirement; actuarial innovation and technology; mortality and longevity; diversity, equity and inclusion; health care cost trends; and catastrophe and climate risk. The Institute has a large volume of topical research available, including an expanding collection of international and market-specific research, experience studies, models and timely research.

Society of Actuaries Research Institute
8770 W Bryn Mawr Ave, Suite 1000
Chicago, IL 60631
www.SOA.org