PartnerRe

When Testing Life Insurance Models, What's Fair About Fairness Metrics and Methods?

Contents

Abstract

Is this model fair? It's a question that actuaries, underwriters, data scientists, compliance officers, and regulators in life insurance are increasingly being asked – and the answer is rarely straightforward.

Fairness is contextual. While the distinction between equality and equity is often blurred, in life insurance it carries meaningful implications. An equality-based approach would prohibit the risk differentiation that underwriting relies on, pushing the industry toward identical outcomes rather than accurate risk-based pricing.

This paper offers a primer on fairness metrics and methods in the context of life insurance. When the question arises: "Is this model fair?", the answer is rarely straightforward. Whether you are an actuary, underwriter, or data scientist, this paper provides a practical foundation for navigating fairness in life insurance. The central message is clear: metrics alone are insufficient, and not all definitions of fairness are compatible with a risk-based system.

Fairness Concepts

What does it mean to treat someone fairly? There are two main ways to define fairness – and as we will see, they can lead to very different outcomes. **Equality** is based on distributing rewards or treating all individuals equally. This would mean, regardless of individual circumstances, premiums would be otherwise equal – one rate for all. **Equity,** on the other hand, is about proportionality – those who carry more risk pay more, and those who carry less, pay less.

Now that we have defined what it means to be fair, the question remains – which approach is right? The answer depends on how society, and regulators, define fairness. If insurance is seen as a social necessity, where those with less risk offset those with more risk, then equality is the goal. If, however, fairness means pricing based on individual risk – where those carrying more risk (e.g., medical impairments in life insurance, reckless behavior in property & casualty) pay more – then equity is the answer. It is important to note that most legislation fails to be explicit on these points.

Fairness can also be examined through the lens of justice theory, which offers three additional perspectives worth considering:

- **Distributive justice** refers to equality of outcomes. This would parallel the idea of equality above.
- **Procedural justice** refers to equality of processes. In areas where equity is the norm, having fair process and procedures is important.
- **Interactional justice** – equality of interpersonal treatment – how do insurance companies treat the individuals (with respect, professionalism and recognition of differences).

These three perspectives align with the framework outlined by Fairlearn's[1] – a widely used toolkit for assessing fairness in AI – particularly its notions of allocation harms and quality of service harms.Fairness discourse is global – regulators and legislators around the world have begun addressing these concepts, including through the NAIC Model Bulletin on Use of AI Systems by Insurers (2023)[2], CO SB21-169 (2021)[3], NY Circular Letter No. 1 (2019)[4] and No. 7 (2024), Canada's proposed Bill C-27 (AIDA, lapsed 2025)[5], the EU AI Act (2024)[6]. Regardless of source or locale, most of this legislation tends to address one or more of the following concepts.

- **Direct discrimination** based on any protected class (e.g., race, gender identity) is widely prohibited.
- **Proxy discrimination** – where a variable is more closely associated with a protected class than with the outcome being predicted – should also be prohibited.
- **Technical biases** – such as differences in baseline predictions, how factors are weighted, or accuracy rates across populations – can cause a model to perform differently for different groups.
- **Disproportionate outcomes that are not otherwise justified** – this is where it gets complicated. Without that qualifier, we risk ignoring legitimate risk-based differences and focusing solely on adverse impact to a group, regardless of the underlying circumstances.

As a result, authors of bills and regulations often lack clarity on what type of fairness is required – and in some cases, inadvertently supply a list of potentially contradictory measures.

1 Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI (MSR-TR-2020-32). Microsoft.
2 National Association of Insurance Commissioners. (2023). Model bulletin: Use of artificial intelligence systems by insurers. NAIC.
3 Colorado General Assembly. (2021). Senate Bill 21-169: Protecting consumers from unfair discrimination in insurance practices. State of Colorado.
4 New York State Department of Financial Services. (2019). Insurance circular letter no. 1: Use of external consumer data and information sources in underwriting for life insurance. NYDFS.
5 Parliament of Canada. (2025). Bill C-27: An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act [Legislative summary].
6 European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

A Precursor to Fairness is Validity

Before examining fairness metrics, we must first address validity – a precursor to any meaningful fairness assessment – and why it matters in an insurance context. Validity – broadly referred to as construct validity – refers to the notion that the model/tool/data assesses what it is intended to assess. Evidence to support validity can come from many sources: expert opinion (content validity) and empirical data (criterion-related validity) are among the most common.

Validity is important for a number of reasons. It is difficult to imagine a model that "does not work" as being fair to any group. While some techniques to establish validity might not be universally applied to every situation, at least some evidence would be necessary to establish the "reasonable justification" mentioned above.

If differences do exist among protected classes, it is necessary to establish why – and to the extent possible, demonstrate that a sensible rationale exists for those differences. The more indirect the relationship between the outcome of interest (e.g., mortality) and the variables being used, the more likely fairness issues could arise. At a minimum, variables should be justifiable (a clear, defensible reason for their inclusion), transparent (a clear explanation of how they are used and how they affect outcomes) – and ideally, a direct causal link can be established.

In life insurance, however, a simple, direct causal link is rarely achievable – and the degree to which a variable meets the justifiable and transparent criteria matters enormously. BMI, for example, is statistically related to mortality, but its impact is indirect, operating through other health conditions rather than causing death directly. The pathway, while indirect, is traceable and coherent – and its relationship to mortality can be articulated and explained. Location presents a different challenge. While location may correlate with mortality, the pathway is diffuse and cannot be clearly attributed to any single underlying factor – making it difficult to justify its inclusion or explain its effect on any individual applicant in a transparent and defensible way. The more a variable relies on correlation alone, without a coherent and explainable mechanism, the more carefully its use must be scrutinized – particularly where it affects protected groups differently.

Primer on Fairness Metrics

With fairness concepts and validity established, we now examine the metrics commonly used to measure fairness – and their limitations in an insurance context. Drawing from numerous sources, we have compiled a representative – though not exhaustive – set of commonly used metrics (see Appendix 1). Some metrics only apply when a threshold is used to turn scores into decisions (e.g., decline vs. accept), while others work directly with continuous predictions. An important caution is the frequent misuse of the 4/5ths rule – one of the threshold-based metrics listed in Appendix 1.

Although originally established as a compliance benchmark in U.S. employment law, it has often been oversimplified or misapplied in other contexts – a problem well documented by Fairlearn[7]. Interestingly, the application of the 4/5ths rule has been recommended to life insurers in Singapore, where guidance on fairness assessments in predictive underwriting suggests a ratio between 0.8 and 1.2 as an acceptable threshold[8]. The misuse of the 4/5ths rule and the challenge of selecting an appropriate fairness framework both illustrate why interpreting fairness metrics is not straightforward; we expand on these below:

1 **Metrics are sensitive to directionality**. A metric may suggest fair outcomes when focused on declines but suggest otherwise if the focus shifts to acceptances. Thus, results can reverse depending on which outcome is treated as the "positive" class, and there is no agreed-upon standard for which direction should be preferred.

2 **They cannot determine whether group-level differences are justified**. Observed group differences in outcomes may result from real differences in underlying risk of the set of individuals or from bias – but fairness metrics cannot tell which is which. They highlight disparities without explaining whether those disparities are appropriate or problematic.

3 **Threshold-based metrics are highly sensitive to where the cutoff is set**. Fairness results can change significantly depending on where the cutoff is placed to convert risk measures into binary decisions (e.g., decline vs. accept). This makes it harder to interpret whether observed disparities reflect differences in the risk-outcome relationship or are simply artifacts of threshold choice.

In short, fairness metrics can help flag disparities but interpreting them requires understanding the structure of the underlying data as well as what definition of fairness you are actually trying to achieve – equity or equality.

7 Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI (MSR-TR

8 Monetary Authority of Singapore & Veritas Consortium. (2022). FEAT Principles Assessment Case Studies (Veritas Document 4). Monetary Authority of Singapore. https://www.mas.gov.sg/-/media/mas-media-library/news/media-releases/2022/veritas-document-4---feat-principles-assessment-case-studies.pdf

Simulated Example – When Fairness Metrics Mislead

To illustrate this critical limitation – that metrics cannot distinguish justified from unjustified disparities – we present a simulated example comparing three scenarios where the underlying data differs fundamentally, yet the metrics produce similar values.

Figure 1: Risk–outcome relationships versus fairness metrics. The top panels show how risk scores relate to outcomes for each group. Ellipses capture each group's pattern – their angle shows whether the score predicts outcomes within that group. The dashed line marks the decision threshold. The bottom panels show fairness metrics at that threshold. Group-Based Risk Differences and Proxy Discrimination produce nearly identical metric values despite fundamentally different underlying relationships. Metrics detect disparities but cannot distinguish legitimate risk measurement from problematic bias. Example uses PartnerRe simulated data.

In **Figure 1**, we use ellipse plots to show the differences among three scenarios. Each ellipse (an oval shape representing a scatterplot of x-y pairs) shows where 95% of the data falls. The x-axis shows the risk measure, while the y-axis shows the outcome. The ellipse shape reflects the spread of the data – when the spread in outcomes is greater, the ellipse becomes taller – and when the spread in risk is greater, the ellipse becomes wider. The lines show the direction and magnitude of the relationship between the outcome and risk – as risk increases, so does the outcome.

Below, we describe how the underlying relationships differ across the three scenarios and how key fairness metrics behave in each case. Because this example uses a threshold, we focus on equality-oriented metrics – equity-oriented metrics, such as calibration, require continuous scores and are not applicable here. The equality-oriented metrics we evaluated include whether:

- Groups are selected at the same rate (demographic parity)
- The selection rate gap signals discrimination (4/5ths rule)
- Groups are incorrectly flagged at the same rate (false positive rate balance)
- Groups are missed at the same rate (false negative rate balance)
- Both error rates are equal across groups (equalized odds)

We also include one mixed metric, which measures whether flagged individuals face the same true risk across groups (predictive parity).

- In the **No Group Differences** case, the two groups have similar averages and spread for the risk measure, and both follow the same relationship (slope) between the risk measure and the outcome. Equality largely holds – groups are selected at nearly identical rates (demographic parity, 4/5ths rule), error rates are nearly balanced (false positive rate balance, false negative rate balance, equalized odds), and positive predictions are slightly less accurate for one group (predictive parity). Notably, even in this **ideal scenario**, small differences can emerge – a reminder that metrics can flag minor variation even when **no meaningful disparity exists.**
- In both the **Group-Based Risk Differences** and **Proxy Discrimination** cases, the groups have the same averages and spread. In the **Group-Based Risk Differences** case, the score meaningfully predicts the outcome within both groups – the slope is the same for each group. In **the Proxy Discrimination** case, there is no within-group relationship between the outcome and risk – this is why the yellow and blue lines are nearly flat and parallel. The risk measure is related **to group identity** and not the outcome – any apparent relationship with the outcome disappears once group identity is accounted for. Despite these fundamental differences, both cases show similar metric patterns – selection rates diverge between groups (demographic parity, four fifths ratio), error rates differ (false positive rate balance, false negative rate balance, equalized odds), and positive predictions are less accurate for one group (predictive parity).

The similarity in metric patterns across the Group-Based Risk Differences and Proxy Discrimination cases reveals a core limitation – metrics can detect when groups differ but cannot evaluate whether those differences are appropriate. The Fairlearn user guide makes the same point, noting that metrics alone "measure differences in predictions across sensitive groups"[9] but do not establish whether those differences are warranted. Because of this limitation, we next turn to methods designed to go beyond simple disparity measures and assess whether observed group differences can be explained by legitimate factors or may reflect unjustified bias requiring further investigation.

9 Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI (MSR-TR-2020-32). Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

Assessing Whether Group Differences Are Justified

We have demonstrated the limitations of fairness metrics, now we turn to methods that go further – investigating why group differences exist and whether they are justified by underlying risk factors. A metric can show that groups have different outcomes (e.g., adverse impact ratio, false-positive rate parity), yet it does not explain why the difference exists or whether it is supported by underlying risk.

The approaches below go further, and interested readers can examine Appendix 2 for more details. These are methods – not single-number scores – that investigate why differences appear. In practice, these methods fall into three broad styles:

- **Exploratory comparisons:** Simple checks that look directly at past decisions or score performance (e.g., matching individuals with similar risk factors, or plotting predicted risk vs. actual outcomes by group).
- **Model-based tests:** Approaches that use a predictive model to account for accepted risk drivers, then test whether group membership still explains differences (e.g., adding a group variable to residuals, interaction/moderation tests, counterfactual checks when the attribute is in the model).
- **Causal analyses:** Deeper techniques that require a defined cause-and-effect model of how predictors relate to the outcome and then separate acceptable vs. unacceptable pathways of group influence (e.g., direct versus indirect effects analysis).

Worked Example – Bringing Exploratory and Model-Based Checks Together

We now turn to a worked example that brings these methods together – applying exploratory and model-based checks to two populations across two scenarios to illustrate how justified disparities can be distinguished from potential bias. Fairness metrics can flag disparities but cannot determine whether those disparities reflect real health differences that justify different outcomes (an equity-aligned view) or whether disparate outcomes themselves are problematic regardless of their cause (an equality-oriented view). The simulation showed that metrics produce similar values whether group differences stem from legitimate risk factors or from proxy discrimination – they cannot distinguish justified disparities from unjustified bias.

Here, we compare two groups, where the risk scores reflect relative mortality risk compared to a reference profile:

- **Group A** the reference group, with age profile average above 51 years, spread about 20 years above or below the mean. Risk scores average 130%.
- **Group B** the protected group under review, averaging about 46 years of age, with most people falling roughly 19 years above or below that. Risk scores average 170%.

These population characteristics stay the same across both scenarios – the scenarios differ in analytical scope, not in the underlying populations.

We first focus on equity-aligned metrics, since in life insurance, we aim for equity. These include whether flagged individuals face the same true risk (predictive parity), whether predicted scores reflect actual risk similarly across groups (calibration), and whether those who experienced the outcome had similar scores across groups (balance for positive class) – all of which work with continuous scores.

We see that across both scenarios, Group A consistently shows higher average scores (calibration), is more likely to die among those flagged as high risk (predictive parity) and has higher average scores among those who died (balance for positive class). The specific values differ between scenarios, but the trend is consistent – metrics suggest there is no equity in either scenario.

Next, we move on exploratory analyses – with the understanding that in our application of the metrics, neither scenario appears equitable.

Figure 2 reveals patterns the metrics cannot:

- **Panel A – Score distribution:** The density plot shows where each group's predicted risk scores cluster – a higher curve indicates more individuals at that score level. In both scenarios, Group A has a higher proportion of individuals in the high-risk range than Group B. In Scenario II, Group B develops a longer tail into the high-risk range compared to Scenario I, though Group A still dominates that range in both scenarios.
- **Panel B – Binned averages with ellipses:** The ellipses summarize the relationship between predicted risk and observed mortality for each group – the angle of the ellipse reflects how strongly the score predicts the outcome within that group. In Scenario I, both ellipses run at similar angles and overlap substantially, suggesting the score predicts mortality consistently across groups. In Scenario II, the ellipses tilt at noticeably different angles – a visual signal that the score may function differently by group.

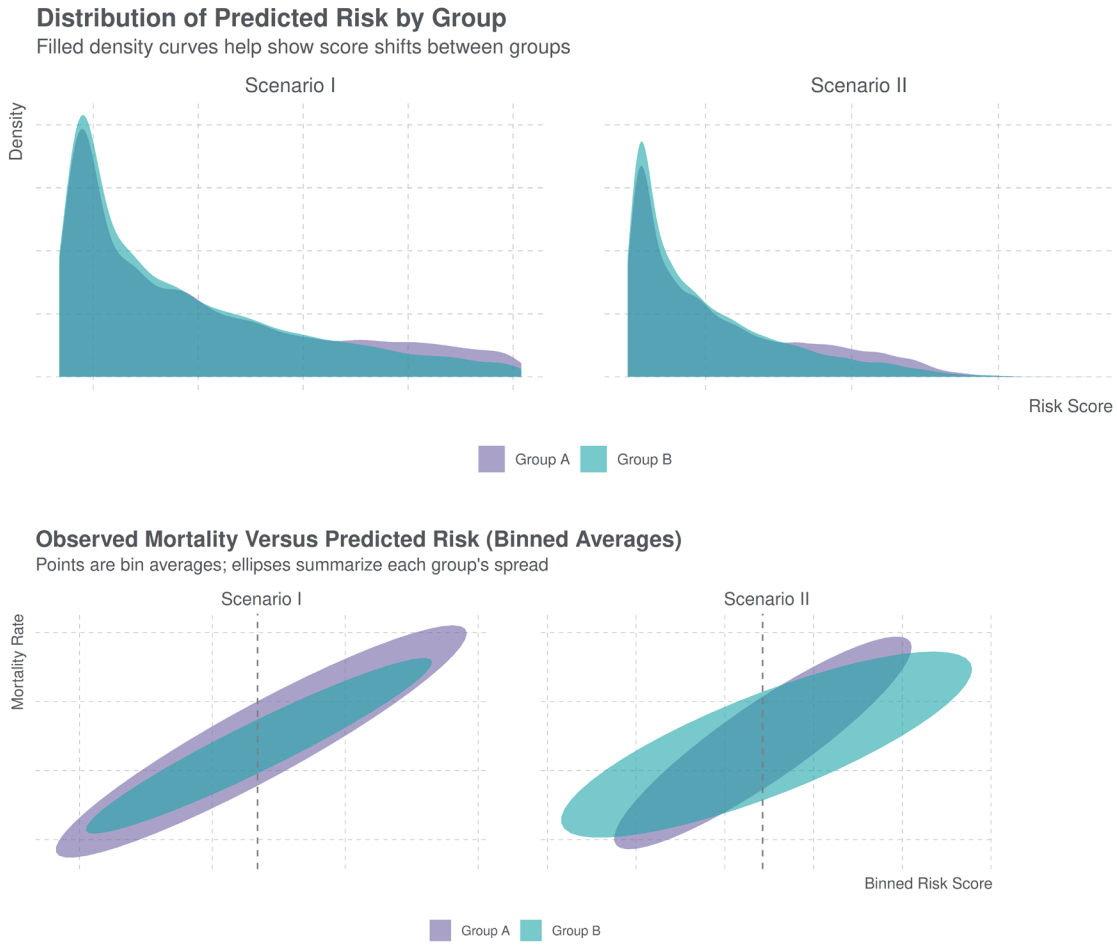

Figure 2: Visual diagnostics and metrics for assessing score behavior across groups. Two visual diagnostics combined with continuous score metrics examine whether the risk score functions equivalently across groups. Panel A shows score distributions. Panel B (within-group calibration using binned averages) shows whether the score-outcome relationship is consistent – ellipse angles reveal this pattern. Scenario I shows modest differences with aligned ellipses. Scenario II shows larger disparities with differently-angled ellipses, suggesting the score may function differently by group. These diagnostics reveal patterns but cannot determine whether differences are justified or problematic. Example uses PartnerRe simulated data.

While metrics suggested neither scenario was equitable, the exploratory analysis points to Scenario II as the more concerning case – where the ellipses suggest the score may not be functioning consistently across groups. Exploratory analyses alone cannot confirm this so we turn to model-based tests to determine whether the metrics are accurately reflecting the underlying relationships.

These models assess two critical questions**:** Does any meaningful group difference remain after we account for the risk score (i.e., Within-Group Calibration; Appendix 2)? And does the score predict mortality the same way for both groups (i.e., Moderation / Interaction Testing; Appendix)?

Table 1: Sequential tests to determine whether the risk score functions fairly across groups.

WHAT THIS ANSWERS	MODEL	SCENARIO I	SCENARIO II
Baseline: Is there a disparity?	1: Group only	Group B has 26% lower mortality	Group B has 32% lower mortality
Does the score predict mortality?	2: Risk only	21% higher mortality per 1-point increase in score	1% higher mortality per 1-point increase in score
↳ Does this effect hold when we add group?	3: Risk + Group	Effect unchanged (21% per point)	Effect unchanged (1% per point)
CRITICAL: Does the score work the same way for both groups?	4: Risk × Group + Interaction term	Yes – no significant difference Effect remains 21% for both groups	No – works differently Group B: 2% per point vs. Group A: 3% per point)

Each model adds complexity to isolate where differences arise. All models control for age and gender identity. The interaction term in Model 4 tests whether the relationship between risk score and mortality differs by group – a significant interaction signals the score may not measure risk consistently across groups.

While neither scenario was equitable and the exploratory analysis pointed to Scenario II as more concerning, the model-based tests tell a more nuanced story. In Scenario I, outcomes are equitable, where Group B has lower mortality risk at the baseline, and differences persist even after accounting for the risk score. The differences in outcomes reflect real health differences that the score captures fairly. In Scenario II, outcomes are not equitable, where Group B also shows lower baseline mortality, but the score's predictive power differs by group – people with identical risk scores face different mortality risks depending on their group. This pattern could reflect several possibilities from real health differences concentrated at the extremes (Group A has substantially worse health profiles in the top 5% of risk in this dataset), different disease pathways leading to the same score but different outcomes, or the score inadvertently captured group identity. The significant interaction shows that we need to dig deeper to understand what's driving this and whether the model is making meaningful distinctions or showing bias.

The equity-equality distinction matters. Only model-based tests like these can distinguish justified disparities (unequal but equitable – real health differences captured consistently) from problematic ones (neither equal nor equitable – score may not be measuring risk consistently). Scenario I demonstrates that unequal outcomes can be fair when the measurement tool works consistently. Scenario II's significant interaction is a red flag requiring further investigation, not a conclusive diagnosis.

Conclusion

Fairness assessment requires structured analytical approaches that combine validity testing, appropriate statistical methods, and contextual interpretation. The goal is not eliminating all group differences, but ensuring differences reflect legitimate factors rather than discriminatory processes – a task requiring both a technical approach and contextual judgment that metrics alone cannot provide. For life insurance specifically, this means preserving the ability to differentiate risk, where justified, while preventing discrimination – a balance that equality-focused metrics fundamentally cannot achieve.

Contributors

Jody Daniel, Senior Data Scientist, Life & Health
Tom Fletcher, SVP, Global Head of Data Science Consulting, Life & Health

Appendix

Common fairness metrics for predictive models.

Table 2 summarizes metrics used to evaluate models that produce binary decisions (threshold-based) or continuous scores. Metrics are grouped by what aspect of fairness they examine: Access Parity examines whether groups are selected at similar rates; Error Burden Parity examines whether groups experience similar rates of correct and incorrect predictions; Predictive Value & Reliability examines whether predictions have similar meaning across groups; and Calibration & Outcome Alignment examines whether predicted scores match actual outcomes within each group. Equality-oriented metrics focus on whether groups experience similar outcomes. Equity-aligned metrics focus on whether the model measures risk accurately within each group.

Table 2: Common fairness metrics for predictive models.

Metric	Alignment	What it measures & why it's used	How it's measured
Access Parity (Selection Rates)			
Demographic Parity (Statistical Parity, Group Fairness)	**Equality-oriented**	Checks whether individuals from different groups are selected or flagged at similar rates, regardless of whether those selections are correct. Detects broad selection bias.	Calculate the percentage of individuals flagged in each group and compare.
Conditional Demographic Parity	**Equality-oriented**	Determines whether differences in selection rates remain after adjusting for valid risk factors (e.g., age, sex, health). Helps separate explainable risk differences from unexplained group effects.	Group people into levels of risk (or other valid factors), then compare selection rates between groups within each level.
4/5ths Rule (80% Rule, Disparate Impact Ratio)	**Equality-oriented**	A widely used compliance benchmark: signals possible discrimination if one group's selection rate is <80% of the majority group's.	Divide the minority group's selection rate by the majority group's selection rate.
Error Burden Parity (Error Distribution)			
Equal Opportunity	**Equality-oriented**	Evaluates whether people who truly experience the event are correctly identified at similar rates across groups. Highlights unequal benefit from correct identification.	For each group, compute the proportion of individuals with the event who were flagged. Compare these proportions.
False Positive Rate Balance	**Equality-oriented**	Evaluates whether people who do not experience the event are wrongly flagged at similar rates across groups. Highlights unfair over-flagging.	For each group, compute the proportion of individuals without the event who were flagged incorrectly. Compare these proportions.
False Negative Rate Balance	**Equality-oriented**	Evaluates whether people who do experience the event are missed at similar rates across groups. Highlights unequal harm from under-flagging.	For each group, compute the proportion of individuals with the event who were not flagged. Compare these proportions.
Equalized Odds	**Equality-oriented**	Combines the above: requires that both correct identification rates and false alarm rates be similar across groups. Considered a stronger fairness condition.	For each group, compare both the proportion of true events flagged and the proportion of non-events incorrectly flagged; smaller differences indicate better parity.

Metric	Alignment	What it measures & why it's used	How it's measured
Predictive Value & Reliability			
Predictive Parity (Positive Predictive Value / Precision)	**Mixed**	Ensures that when the model flags someone, the flagged individuals have similar true risk across groups. Addresses fairness in the meaning of a positive result.	For each group, calculate the proportion of flagged individuals who actually experience the event and compare.
Calibration & Outcome Alignment			
Calibration (Test-Fairness)	**Equity-aligned**	Assesses whether predicted risk scores match actual event rates equally across groups. Indicates whether risk predictions are trustworthy for everyone.	Divide predictions into score ranges ("risk bins"). For each bin and group, calculate the average predicted risk and compare it with the observed event rate.
Balance for Positive / Negative Classes	**Equity-aligned**	Checks whether predicted scores themselves are distributed similarly for people who did and did not experience the event across groups. Identifies if one group tends to receive systematically higher or lower scores for the same outcomes.	For each group, compute the average predicted score among those who experienced the event and among those who did not. Compare these averages across groups.

Unlike the fairness metrics described earlier (which only measure differences), the approaches below are methods – they are used to investigate why differences exist and whether they can be explained by accepted risk factors or instead signal potential bias. Each method makes its own assumptions and requires judgment calls (e.g., which risk factors are legitimate to adjust for, or how to define a valid "what if" scenario). These are not single-number scores; they are analyses that help determine if observed disparities in predictions or outcomes are reasonable. ★ Methods marked with a star are summarized from the Fairlearn User Guide (v0.12); other methods are referenced individually in the table.

Table 3: Methods for Assessing Whether Group Differences Are Justified or Reflect Bias.

What it measures & why it's used	How it's applied	Key concerns & assumptions
Fairness Through Unawareness[10]		
Basic input check: confirms that a protected attribute (e.g., race, ethnicity) is not explicitly included as a predictor.	Inspect the model's input variables or rating factors to verify that the protected attribute itself (or an obvious renamed version) is not present.	Does not ensure fairness — other predictors may still act as proxies (e.g., ZIP code can correlate with race).
Individual Fairness Test[11] **(pairwise comparison)**		
Looks at actual decisions: asks whether two applicants with the same accepted risk factors but different group membership were treated the same.	Audit past underwriting or claim decisions: match pairs of similar individuals (same accepted risk factors) who differ only in the protected attribute; compare outcomes.	Needs reliable historical data on both risk factors and the protected attribute. Requires agreement on which factors are acceptable to match on.

10 Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning (Chapter 2: Fairness Through Unawareness; Chapter 3: Individual Fairness). Retrieved from https://fairmlbook.org

11 Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214–226). ACM. https://doi.org/10.1145/2090236.2090255

What it measures & why it's used	How it's applied	Key concerns & assumptions
Counterfactual Fairness[12]		
Tests how a trained model responds when you hypothetically change a protected attribute that is included in the model (e.g., gender identity, if legally allowed). Used to see if the model's use of that attribute seems appropriate after holding other accepted risk factors fixed.	Take the model and an applicant's data. Change the protected attribute (e.g.,female → male) while keeping the other accepted risk factors the same; observe how the prediction changes.	Only meaningful if the attribute is actually a model input. You must decide which other variables to hold constant. Interpretation depends on legal/regulatory context — e.g., gender identity may be allowed, but you may still want to confirm its impact is reasonable.
Residual / Conditional Parity Analysis[13]		
After accounting for accepted risk drivers, checks if group membership still helps explain remaining differences in predictions or outcomes.	Build your usual model using accepted factors; then add the protected attribute and see if it still explains leftover variation in predictions or errors.	Results depend on having the right predictors — leaving out important risk drivers can make the test misleading. Needs enough data per group.
Moderation / Interaction Testing (Cleary-type)[14],[15]		
Checks whether the relationship between the model's score and actual outcomes differs by group (i.e., the model works differently for some groups).	Run a regression of outcome on the score and add an interaction term (score × group). If the interaction is meaningful, the score–risk relationship differs by group.	Needs outcome data and group labels; can be unreliable if groups are small or have few outcomes.
Within-Group Calibration Testing[16]		
Visual check of whether predicted risk rates match observed event rates within each group. Helps see if the score's risk signal works consistently.	For each group, create a diagnostic plot: X-axis = predicted risk/score; Y-axis = observed event rate. Color by group to compare how well predictions align with outcomes.	Needs enough data per group to form a stable curve; results can be noisy with small samples. This is a diagnostic — it shows differences but not why they exist.
Path-Specific Effect Analysis[17],[18]		
Breaks down how a protected attribute influences predictions — separating effects that travel through accepted risk factors vs. effects that act directly or through other unwanted paths.	Build and fit a cause-and-effect statistical model (e.g., causal mediation or structural equation model) describing how predictors influence each other and the outcome. Estimate how much of the group difference flows through acceptable vs. unacceptable paths.	Requires a defendable understanding of how predictors cause or influence each other and the outcome. Complex to build and explain; results depend on whether your assumed cause-and-effect structure is correct.

12 Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. In Advances in Neural Information Processing Systems (Vol. 30). https://arxiv.org/abs/1703.06856
13 Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (Vol. 29, pp. 3323-3331). https://arxiv.org/abs/1610.02413
14 Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 5(2), 115–124. https://doi.org/10.1111/j.1745-3984.1968.tb00613.x
15 De Ayala, R. J. (2009). The theory and practice of item response theory (Chapter 9: Differential prediction and test bias). New York, NY: Guilford Press.
16 Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230. https://doi.org/10.1186/s12916-019-1466-7
17 Nabi, R., & Shpitser, I. (2018). Fair inference on outcomes. In Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI). https://arxiv.org/abs/1802.06309
18 Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1706.02744