São Paulo 2025

An Empirical Fairness Evaluation Framework for AI Models in Insurance: Balancing Fairness and Business Value

Yun-Tuan Chang

SPEAKER Cathay Life Insurance, Taiwan / Data analyst

國泰人壽
Cathay Life Insurance

Contents

01. About Cathay Life Insurance and the Speakers

About Cathay Life Insurance

Company Snapshot

- Largest insurance company in Taiwan
- Offers individual life, health, unit-linked, and group insurance products

Number of Customers	Num. of In-force Policies
8 million	20 million

Total Assets (US$)	Number of Tied Agents
280 billion	24,000

Awards & Recognition

IIC ASIA 2025

Insur-Innovator Connect Awards 2025

Digital Transformation Trailblazer Award

2025 & 2024

ITC Asia Insurer Awards

(2025) Digital Transformation Trailblazer Award
(2024)Data and Analytics Master Award

2024 & 2023

Celent Model Insurer Awards

Data, Analytics and AI

AIIA 2022

Digital Insurer of the Year

AI-powered Insurance and Actuarial Work

The demand for AI is growing day by day, and Cathay Life currently has over 40 AI products.

Core insurance process

Internal operational support

Product design

- Risk scoring model

Marketing

- Marketing scoring model
- Targeted marketing list
- Product Recommendation

Underwriting

- Precise health screenings
- Underwriting robot

Actuarial

- Policy loan model
- Lapse rate model

Claim

- Claim risk scoring model
- Claim automation

Service

- NPS management
- Potential user prediction
- GenAI Q&A
- Customer lifetime value

HR & Channel management

- QA robot
- Agent recruitment model

Risk management

- Money laundering risk model
- Risk prediction for applications

Investment

- Property pricing model
- AI-generated report

About the Speakers

Yun-Tuan Chang

Data Analyst,
Data and AI Development Department

I-Hsiu Kao

Data Scientist,
Data and AI Development Department

Chia-An Wang

Senior Business Analyst,
Data and AI Development Department

02. Fairness AI In Insurance

Key Question of Fairness AI in Insurance

Rethinking Fairness for AI Models in Insurance: Why Sufficiency Matters

Fairness Evaluation Framework and Methodology

Key Question of Fairness AI in Insurance

Importance of Fairness AI in Insurance Industry

Fairness in insurance is increasingly driven by regulatory pressure.

Evaluating fairness is not just about metrics — it's key to model deployment, compliance risk management, and long-term business viability.

The Practical Challenge

However, there is often a trade-off between **fairness** and **model performance**. Improving fairness may reduce risk differentiation, while focusing solely on accuracy may overlook unequal impacts on certain groups. This presents a key challenge: how to strike a balance that ensures both **regulatory compliance** and **business value**.

Key Questions

Can we achieve a **balance between model accuracy and fairness** to meet the application needs of AI models in the insurance sector?

Rethinking Fairness for AI Models in Insurance: Why Sufficiency Matters

Aligning Fairness with Insurance Modeling Context

Fairness standard in machine learning is commonly categorized **into independence, separation, and sufficiency** (Barocas et al., 2019). In insurance AI, choosing the right fairness standard requires aligning with domain needs—**models must estimate risk accurately, respect individual differences, avoid systemic bias, and adhere to actuarial principles, where premiums reflect actual risk**.

Sufficiency supports fairness without sacrificing the core goal of insurance AI—accurate, risk-aligned prediction

- **Sufficiency** also known as well-calibrated prediction, evaluates whether predicted risk scores reflect actual risk equally across groups. It enables consistent, comparable predictions, supports bias detection, and aligns with actuarial fairness—making it well-suited for both risk assessment and potential pricing applications in insurance.
- **Independence** evaluates outcome parity across groups, ignoring true risk. This conflicts with risk-based modeling and may misrepresent high-risk individuals as low-risk.
- **Separation** evaluates error rate parity across groups. This reduces precision in identifying individual risk, which is critical in actuarial decision-making.

Rethinking Fairness for AI Models in Insurance: Sufficiency Fairness Standard

What Counts as Fair Under the Sufficiency Standard?

We assess fairness using calibration-based metrics

- **Calibration Curves** – show how well predicted risk matches actual outcomes
- **ECE (Expected Calibration Error)**– quantify calibration quality

$$\mathrm{ECE} = \sum_{i=1}^{K} \frac{n_i}{n} \cdot \left|\mathrm{acc}(B_i) - \mathrm{conf}(B_i)\right|$$

K: Number of bins used to partition the predicted probabilities (e.g., 10 bins).

B_i: The i^{th} bin, containing predictions with confidence scores falling into that bin's interval.

n_i: Number of samples in bin B_i.

n: Total number of samples.

$\mathrm{acc}(B_i)$: Accuracy in bin B_i, i.e., the proportion of correct predictions.

$\mathrm{conf}(B_i)$: Average predicted confidence in bin B_i.

Fairness under sufficiency means predictions are well-calibrated and consistent across all groups

- Lower ECE = better fairness
- ECE < 1% is considered well-calibrated (Guo et al., 2017)
- Group ECE differences should be small to ensure fairness across subpopulations

Fairness Evaluation Framework for AI Models in Insurance

To assess the fairness of AI models applied in the insurance domain, we developed **a Fairness Evaluation Framework,** consisting of two main stages, followed by a Fairness-Performance Trade-off assessment:

01. Fairness evaluation

- Apply the sufficiency criterion to specific subgroups (e.g., age, region, category) identified by regulations or business concerns.
- Evaluate calibration within each group to determine whether the model treats them fairly.

02. Fairness improvement

- **Data adjustment:** Debias the dataset to ensure group balance
- **Post-prediction processing:** Adjust model outputs to improve group-level fairness

Fairness-Performance Trade-off

Measure the impact of fairness improvements on model performance using:

- **Model metrics:** AUC, RMSE
- **Business metrics:** LIFT Value for high/low risk groups

03. Fairness Evaluation Framework and Methodology

Fairness evaluation:

Key Fairness Metrics

Fairness evaluation on Mortality Risk Identification Model

Fairness evaluation: Key Fairness Metrics

Sufficiency Fairness Criterion (Primary Indicators of Fairness)

- **Calibration Test:** Ensures that predicted risk scores align with actual event rates within each group.

Input Data Check (Ensure the data itself doesn't introduce hidden bias)

- **Chi-Square Test:** Checks if the distribution of protected attributes is balanced, identifying potential data bias.
- **Proxy Discrimination:** Assesses the association between input variables (X) and model predictions (Y) to detect indirect bias.

Dimension	Sufficiency Fairness Criterion	
Method	ECE（Expected Calibration Error）	**Fair:** ECE < 1%, small group differences **Unfair:** ECE > 4–10%, large group gaps
	Calibration Curve	**Fair:** Lines close to 45° across groups **Unfair:** Systematic deviation from perfect calibration

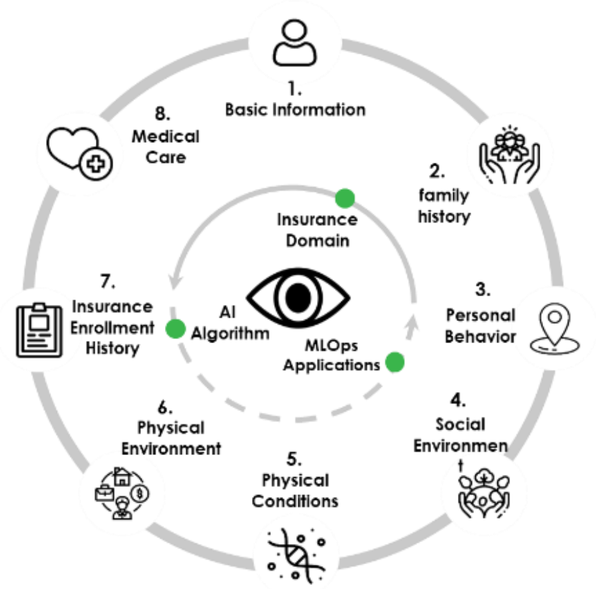
Dimension	Input Data (Supplementary Checks)	
Method	Chi-Square test	< 0.05 → Suggests potential unfairness
	Proxy Discrimination	Min Category Ratio: < **0.05** → Indicates severe imbalance Max Proxy Discrimination P-Value: < **0.05** → Indicates proxy discrimination risk Max Proxy Coefficient: < **0.1** → Considered low association

Fairness evaluation: Mortality Risk Identification Model Overview

This model predicts mortality to segment health risks, identifying a 500% difference between high- and low-risk customers—enhancing both risk assessment and underwriting. Fairness is evaluated across two key group dimensions:

- **Age Groups:** Five-year age bands for individuals under 75
- **Region:** Taiwan's seven administrative regions

Cathay Eye Model



Model Structure

- **Model Type**: LightGBM-based risk assessment model
- **Prediction Target**: 9-year mortality risk
- **Data Foundation**:
  - **X Variables**: 367 features extracted from 8 variable categories
  - **Sample Data**: Historical records from 6.5 million policyholders

Model Output

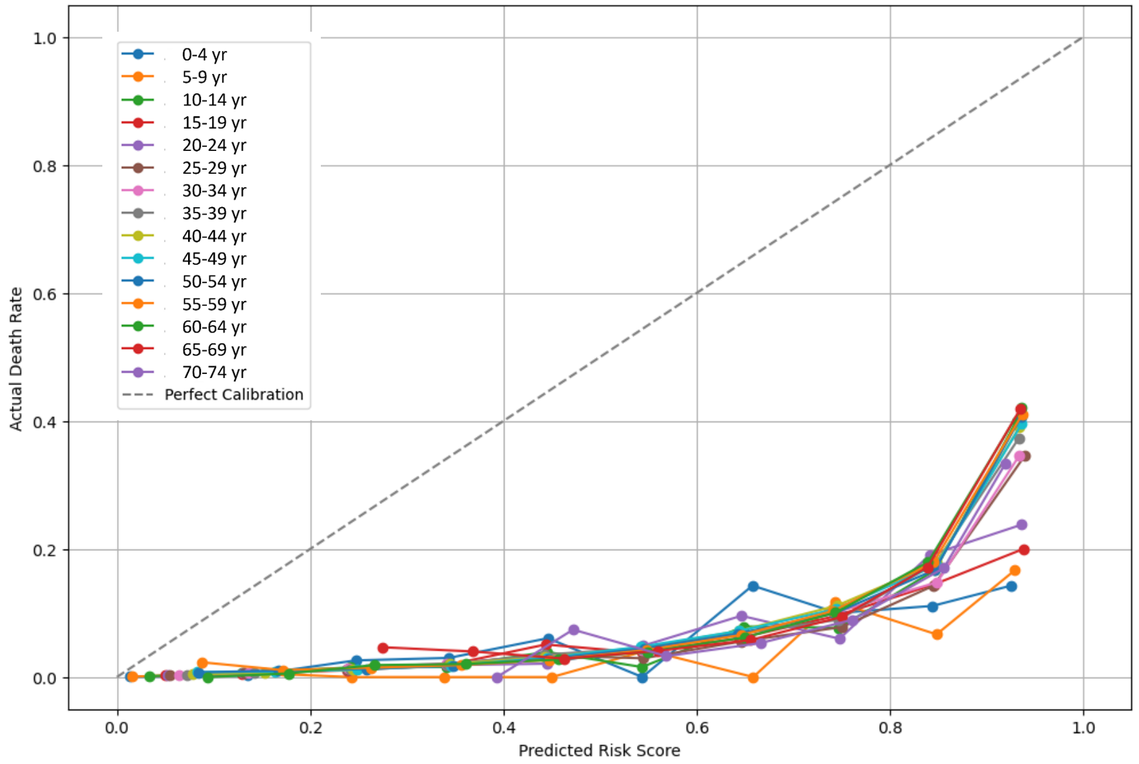
- **Mortality Probability (Risk Score, 0-1)**
- **Labeling Method**: Ranked by age and gender percentiles, assigned labels from 1 to 1000 (1 = low risk, 1000 = high risk)

Application

- **Risk Segmentation**: Tailored underwriting strategies for high- and low-risk groups
- **Underwriting Support**: High-risk customers flagged for additional medical screening

Fairness evaluation: Age Groups

Calibration Curve



Calibration Error (ECE)

Group	Sample Size	Mortality Rate	ECE
0-4yr	179k	0.0007	0.0118
5-9yr	225k	0.0009	0.0165
10-14yr	287k	0.0021	0.0359
15-19yr	422k	0.0032	0.0515
20-24yr	473k	0.0036	0.0586
25-29yr	522k	0.0042	0.0655
30-34yr	666k	0.0069	0.0985
35-39yr	702k	0.0119	0.1512
40-44yr	593k	0.0191	0.2165
45-49yr	587k	0.0273	0.2832
50-54yr	610k	0.0377	0.3462
55-59yr	574k	0.0525	0.4163
60-64yr	479k	0.0784	0.5002
65-69yr	250k	0.1256	0.5875
70-74yr	146k	0.2110	0.6300

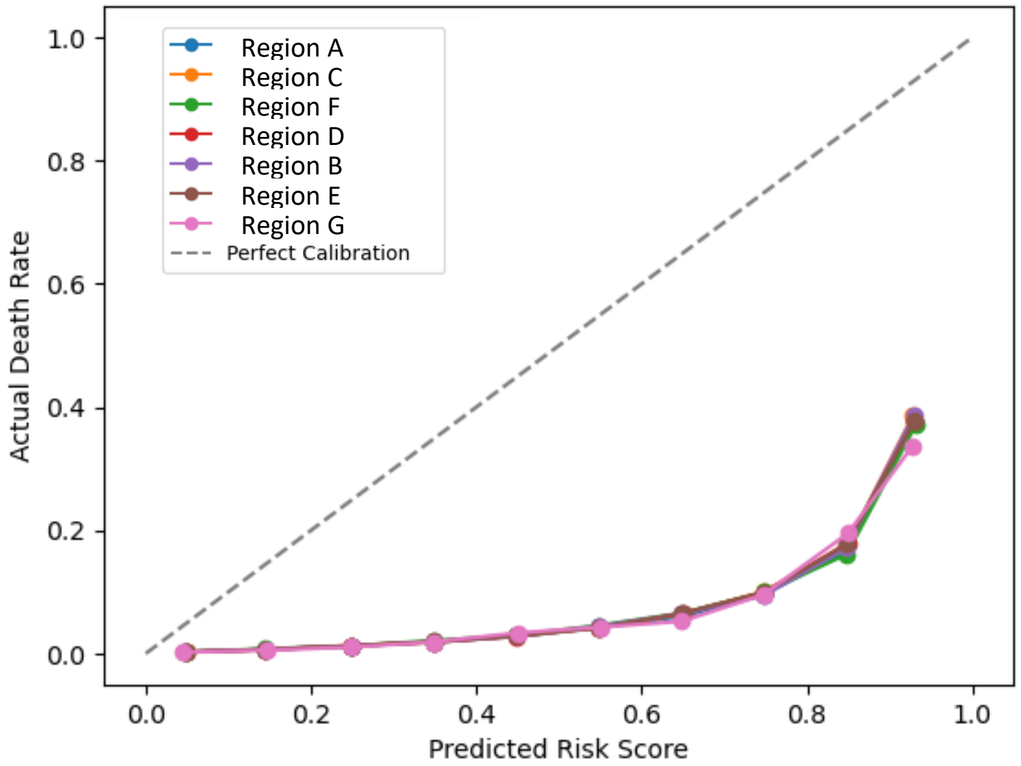
Overall Calibration Error (ECE): 0.2226

Examining Fairness in Input Data

Chi-Square P-Value	Min Category Ratio	Max Proxy Discrimination P-Value	Max Proxy Coefficient
0.0218	0.000000	0.000000e+00	0.055450

Fairness evaluation: Regions

Calibration Curve



Calibration Error (ECE)

Group	Region	Sample Size	Mortality Rate	ECE
A	Northern Region	1,790k	0.0270	0.2212
B	North-Central Region	1,103k	0.0260	0.2061
C	Central Region	1,324k	0.0286	0.2162
D	South-Central Region	980k	0.0334	0.2301
E	Southern Region	1,184k	0.0351	0.2421
F	Eastern Region	314k	0.0320	0.2214
G	Offshore Islands	22k	0.0253	0.1952

`Overall Calibration Error (ECE): 0.2226`

Examining Fairness in Input Data

Chi-Square P-Value	Min Category Ratio	Max Proxy Discrimination P-Value	Max Proxy Coefficient
0.000000	0.0033	6.393852e-30	0.029141

Fairness evaluation: Conclusion

Fairness Evaluation Results of the Model

Evaluated Groups	Sufficiency Metric (Calibration Error /Curve)	Examining Input Data	Fairness Conclusion
Age Groups	High-age groups show elevated ECE (ECE ≥ 0.6), with clear deviation in calibration curves	Chi-Square P = 0.0 Proxy P = 0.0 Proxy Coef ≈ 0.055 → Significant proxy effect	**Unfair**
Regions	ECE ranges from 0.19 to 0.24, low variation and well-aligned calibration curves	Chi-Square P = 0.0 Proxy P = 6e-30 Proxy Coef ≈ 0.029 → Proxy effect present	**Unfair**

03. Fairness Evaluation Framework and Methodology

Fairness improvement:

Experimental Design

Experimental Results

Why Systematic Fairness Methods Matter

Emerging Fairness Challenges

- Traditional variables like age are manageable via business rules.
- Variables like region or income are harder to control.
- We explore scalable methods to support broader fairness needs.

Age as a Starting Point

- Well-understood in insurance practice.
- Provides a simple test case for evaluating fairness techniques.
- Methods tested on age may extend to more complex group variables.

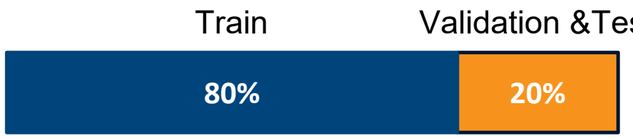
Fairness improvement: Experimental Design

Experiments List

Processing Stage	Approach	Experiment ID	Description	Main Objective
Baseline		M0	Original model	Raw model without any pre- or post-processing
Data Preprocessing	Feature Filtering	MCF	Correlation Filtering	Remove features highly correlated with evaluated groups to reduce proxy bias
	Feature Smoothing	MDIR	DIR (all numeric variables)	Smooth numeric variables within groups for better statistical fairness
Post-processing	Model Calibration	MC-0	Baseline calibration model using M0 architecture	Trains on 16% less data than the original model to reserve enough samples for calibration
		MC-1	Group-based sigmoid calibration	Calibrate sigmoid curves for each group to improve ECE and reliability
		MC-2	Global sigmoid + group normalization	Apply global sigmoid calibration followed by group normalization for joint fairness
		MC-3	Group normalization + group sigmoid calibration	Normalize by group first, then apply group sigmoid for better stability and fairness

Fairness improvement: Data Preprocessing

Experiment Data Set

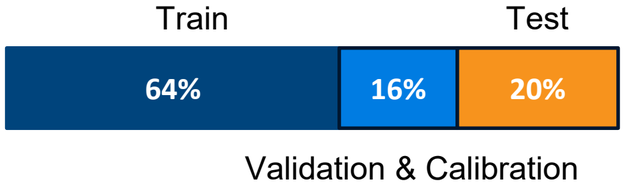


Fairness-Oriented Data Preprocessing: MCF vs. MDIR

Aspect	MCF – Correlation Filtering	MDIR – Group-based Feature Smoothing
Core Strategy	Removes features highly correlated with evaluated groups	Smooths **numeric features(294)** within each group
Target Bias Type	Proxy bias	Statistical distributional bias
Operation	Feature selection (filtering based on **correlation threshold >0.3**)	Feature transformation (group-level normalization: **repair level=1**)
Feature Handling	Risky features are removed (variables: 337)	All features are retained and adjusted (variables: 367)

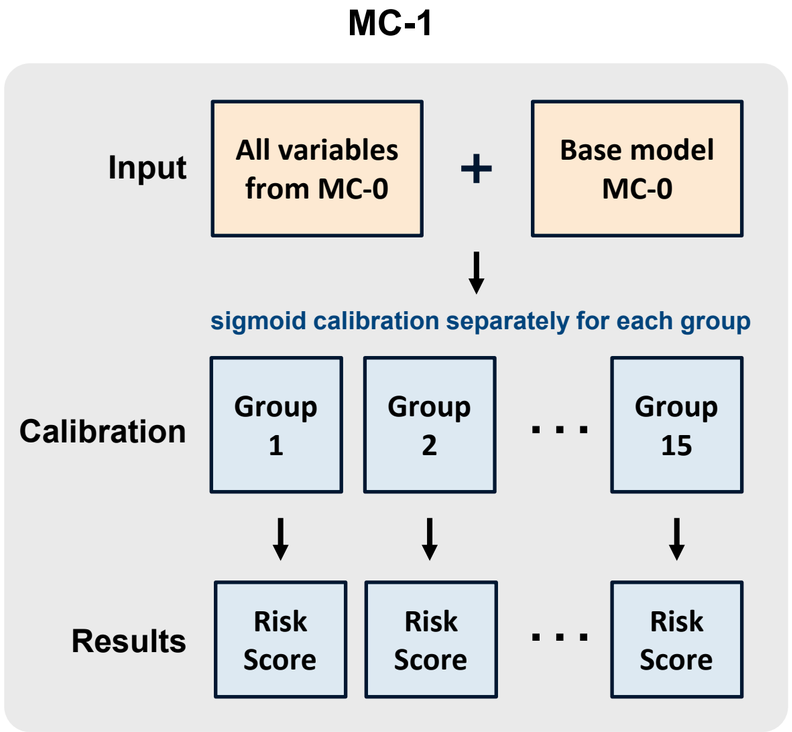
Fairness improvement: Post-processing

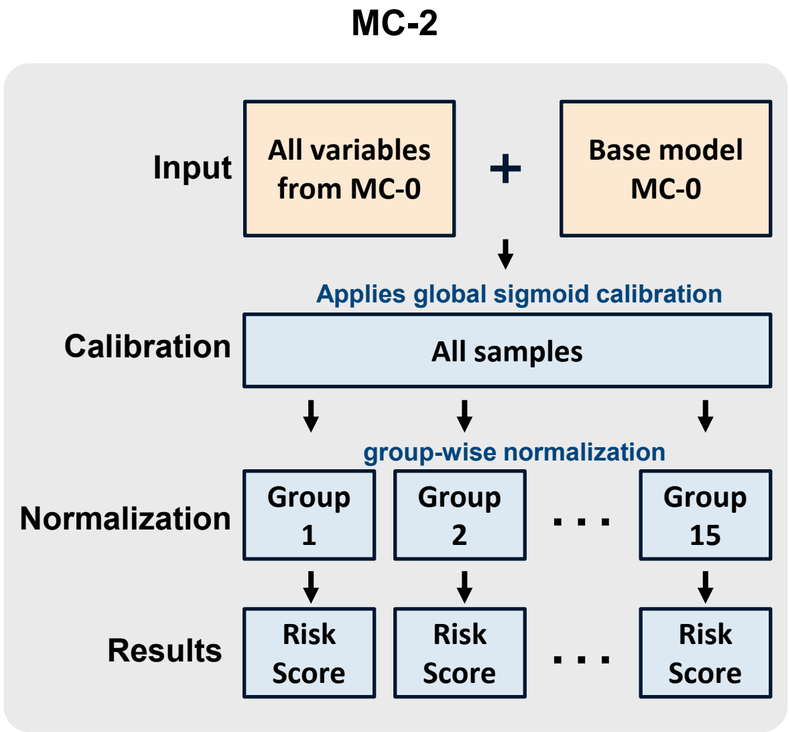
Experiment Data Set (MC-0~3)

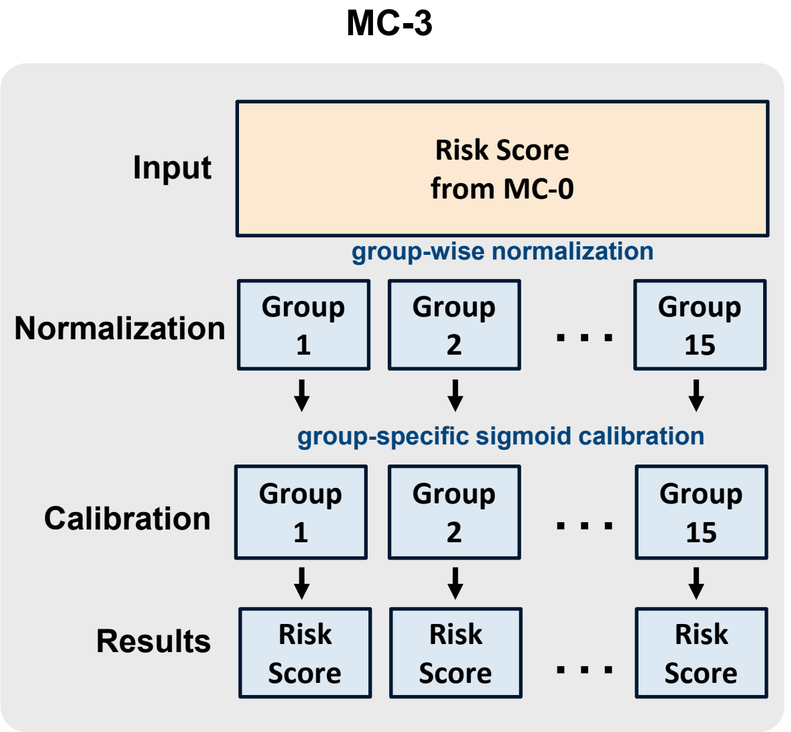


To enable post-hoc calibration, 20% of the original training set (16% of total data) was reserved, reducing the training size for MC-0~3 models compared to M0.

Calibration Experiments



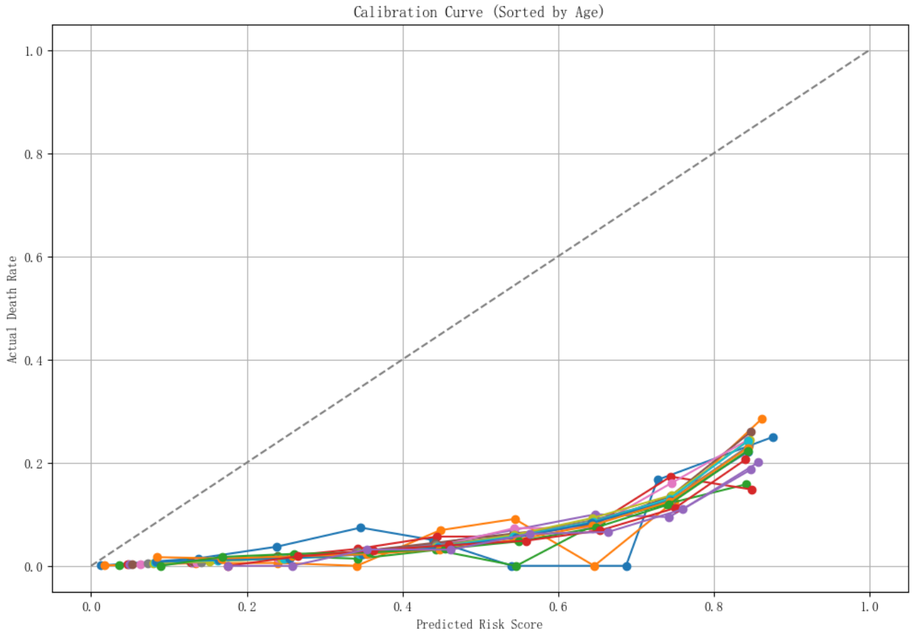




Fairness improvement – Data Preprocessing

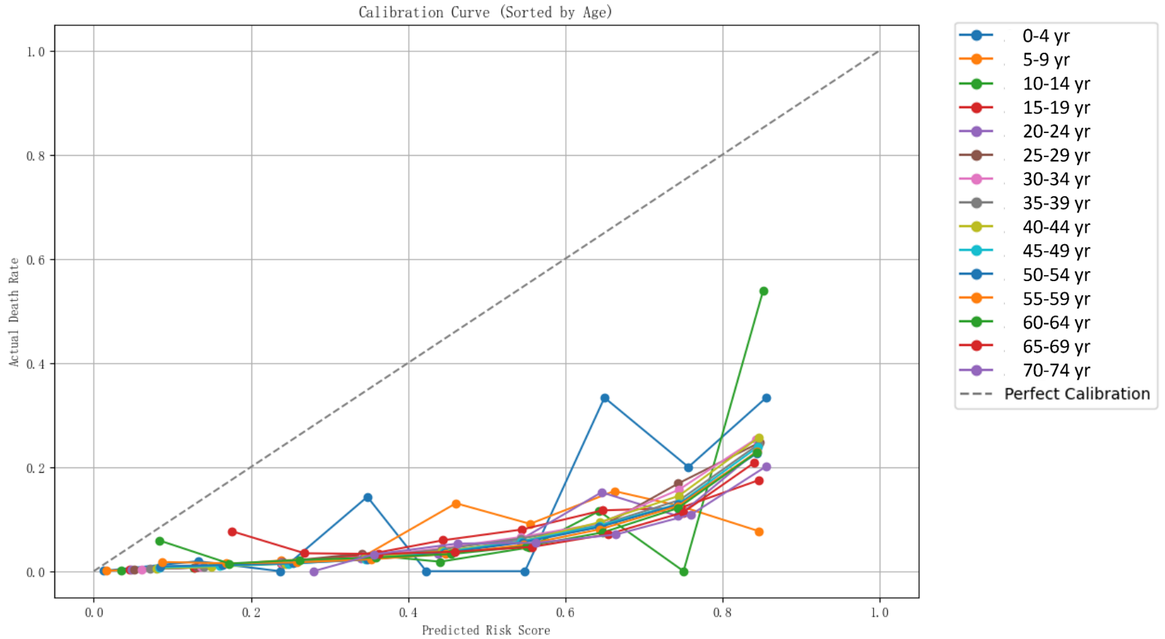
M0
Overall Calibration Error (ECE): 0.2226

MCF



MCF
Overall Calibration Error (ECE): 0.2255

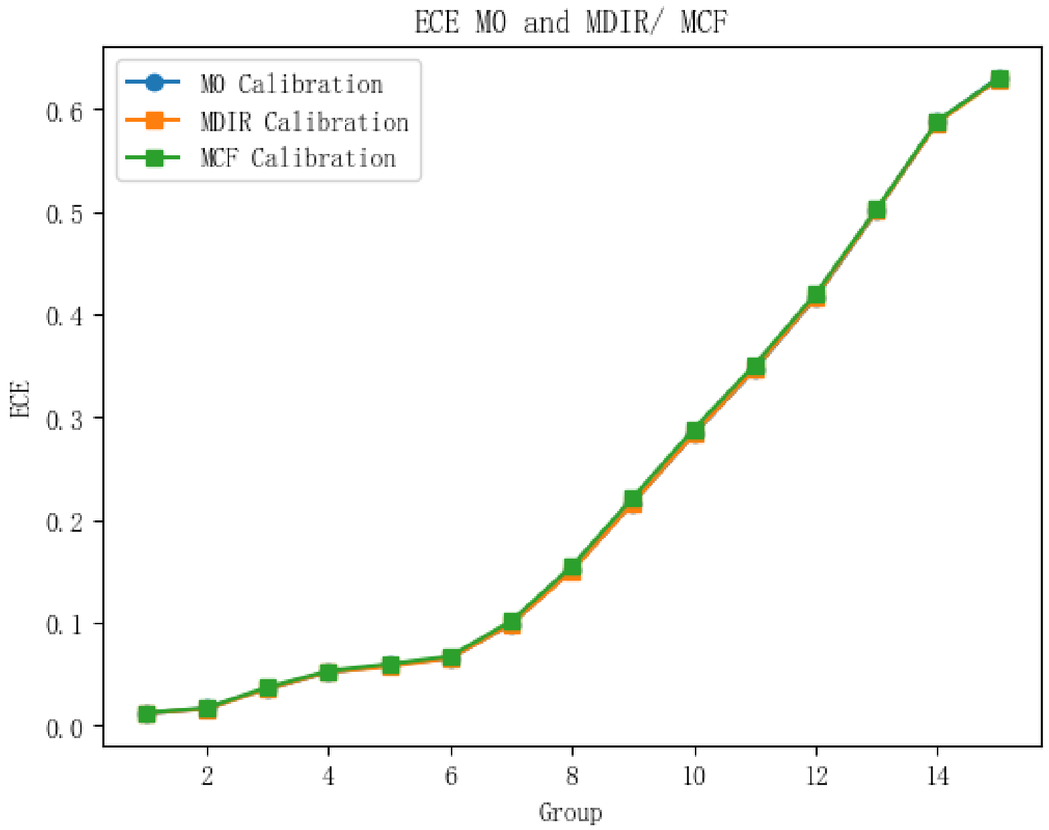
MDIR



MDIR
Overall Calibration Error (ECE): 0.2224

Fairness improvement – Data Preprocessing

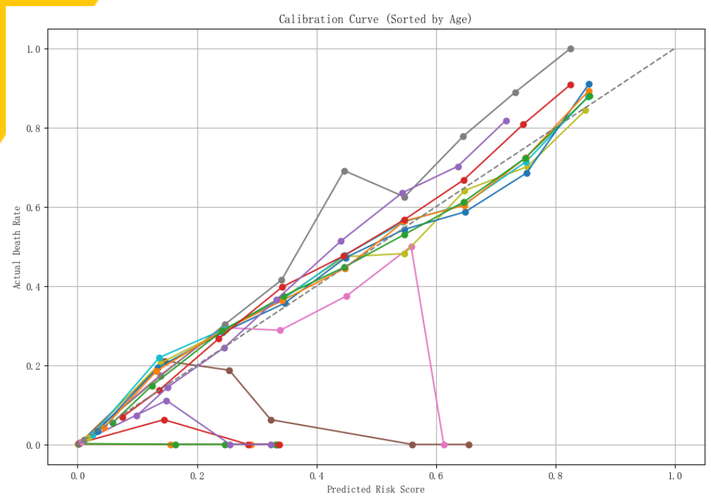
Group	Sample Size	M0	MCF	MDIR
		ECE	ECE	ECE
70-74yr	146k	0.6300	0.6295	0.6283
65-69yr	250k	0.5875	0.5882	0.5866
60-64yr	479k	0.5002	0.5023	0.5008
55-59yr	574k	0.4163	0.4200	0.4169
50-54yr	610k	0.3462	0.3509	0.3466
45-49yr	587k	0.2832	0.2885	0.2833
40-44yr	593k	0.2165	0.2218	0.2165
35-39yr	702k	0.1512	0.1551	0.1502
30-34yr	666k	0.0985	0.1014	0.0984
25-29yr	522k	0.0655	0.0673	0.0653
20-24yr	473k	0.0586	0.0596	0.0573
15-19yr	422k	0.0515	0.0529	0.0518
10-14yr	287k	0.0359	0.0371	0.0356
5-9yr	225k	0.0165	0.0170	0.0162
0-4yr	179k	0.0118	0.0123	0.0120



Fairness improvement – Post-processing

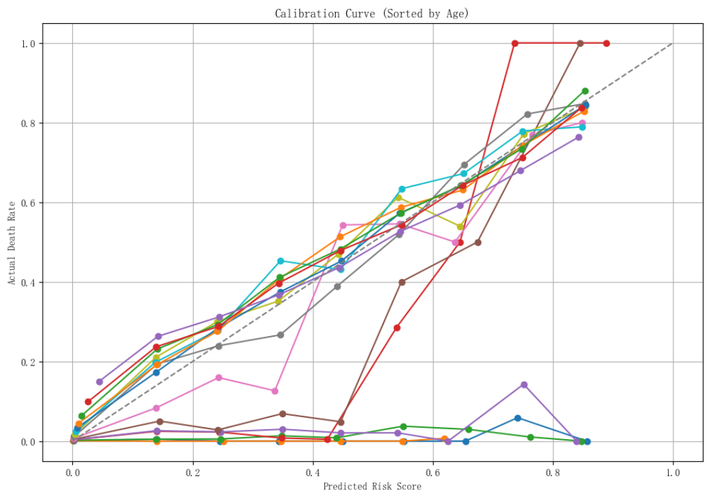
M0
Overall Calibration Error (ECE): 0.2226

MC-1



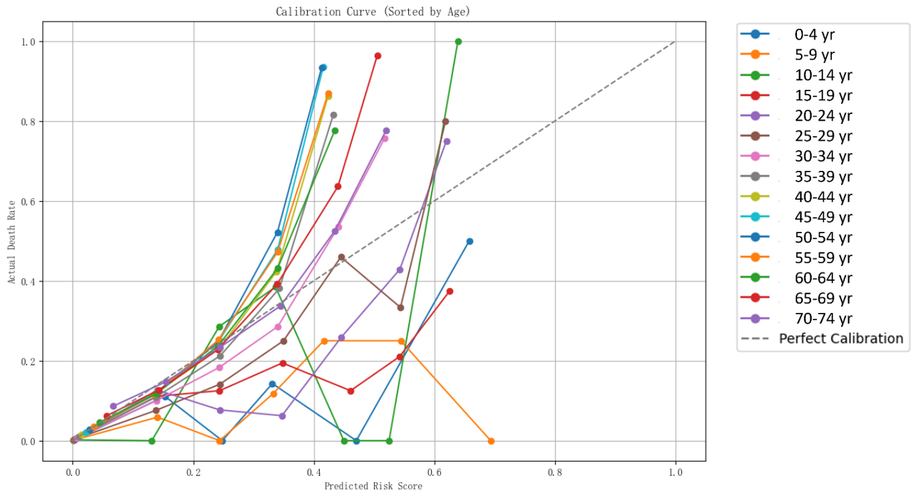
MC-1
Overall Calibration Error (ECE): 0.0017

MC-2



MC-2
Overall Calibration Error (ECE): 0.0182

MC-3

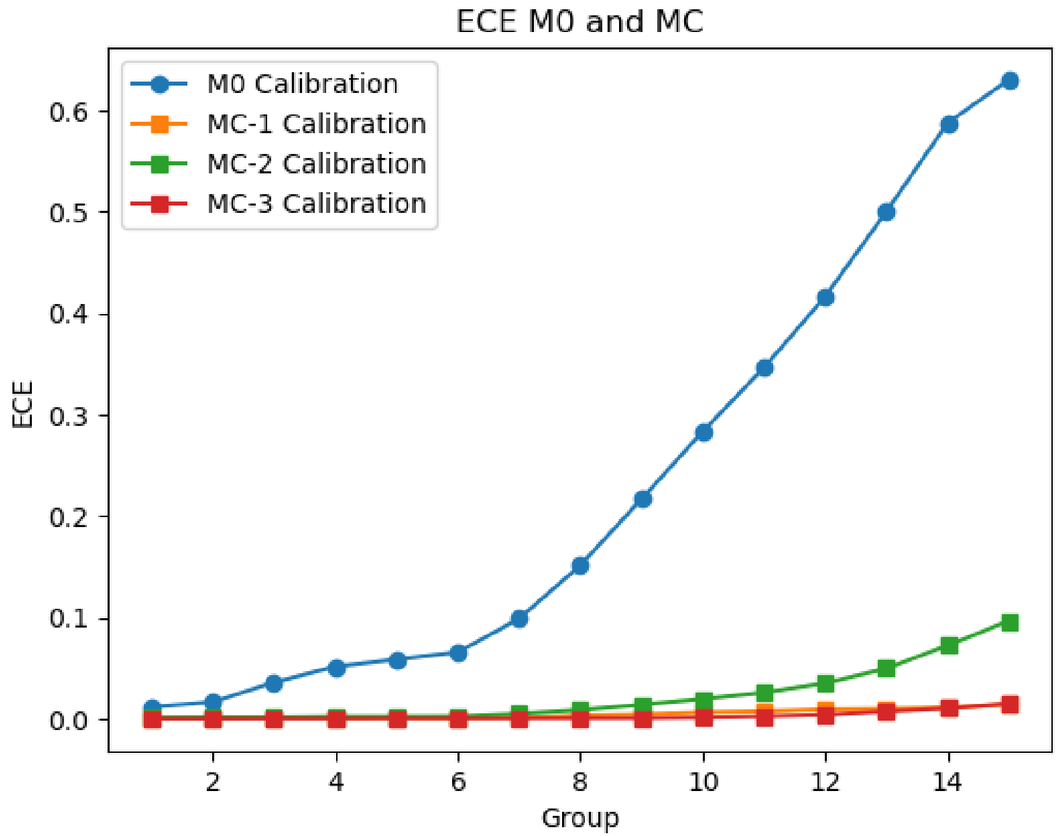


MC-3
Overall Calibration Error (ECE): 0.0013

The calibration curve of the MC-3 model appears visually discontinuous and more volatile, mainly due to the compression of predicted values after group normalization and the smaller sample sizes in some groups.

Fairness improvement – Post-processing

Group	Sample Size	M0	MC-1	MC-2	MC-3
		ECE	ECE	ECE	ECE
70-74yr	146k	0.6300	0.0141	0.0965	0.0151
65-69yr	250k	0.5875	0.0111	0.0723	0.0104
60-64yr	479k	0.5002	0.0099	0.0497	0.0075
55-59yr	574k	0.4163	0.0094	0.0351	0.0041
50-54yr	610k	0.3462	0.0075	0.0259	0.0025
45-49yr	587k	0.2832	0.0065	0.0196	0.0014
40-44yr	593k	0.2165	0.0045	0.0139	0.0007
35-39yr	702k	0.1512	0.0035	0.0088	0.0006
30-34yr	666k	0.0985	0.0011	0.0051	0.0005
25-29yr	522k	0.0655	0.0003	0.0026	0.0005
20-24yr	473k	0.0586	0.0003	0.0024	0.0005
15-19yr	422k	0.0515	0.0003	0.0024	0.0003
10-14yr	287k	0.0359	0.0003	0.0018	0.0003
5-9yr	225k	0.0165	0.0001	0.0018	0.0001
0-4yr	179k	0.0118	0.0001	0.0015	0.0001



Fairness improvement – Result

Fairness optimization for age groups shows that data preprocessing methods provide limited improvement in reducing prediction bias across groups. In contrast, post-processing strategies using group-specific sigmoid calibration—particularly MC-1 and MC-3—effectively reduce prediction bias for older age groups and narrow the calibration error gaps between age segments.

Experiment ID	Max ECE (G15)	Min ECE (G1)	Standard deviation of ECE across all groups	Whether fairness has improved
M0	0.6300	0.0118	0.2089	-
MCF	0.6295	0.0123	0.2090	No significant improvement
MDIR	0.6283	0.0120	0.2088	No significant improvement
MC-0	0.6247	0.0117	0.2141	-
MC-1	0.0141	0.0001	0.0046	Significant improvement
MC-2	0.0965	0.0015	0.0280	Some improvement
MC-3	0.0151	0.0001	0.0193	Significant improvement

04. Empirical Results: Fairness-Performance Trade-off

Evaluating the Fairness-Performance Trade-off:

Model metrics

Business metrics

Evaluating the Fairness-Performance Trade-off

Purpose

- To understand the range of performance shifts that may occur when improving fairness.
- This enables better planning when aligning models with both regulatory and business expectations.

Metrics Used

- Both AUC and LIFT are **compared before and after fairness optimization** to understand how much performance is affected when pursuing fairness.

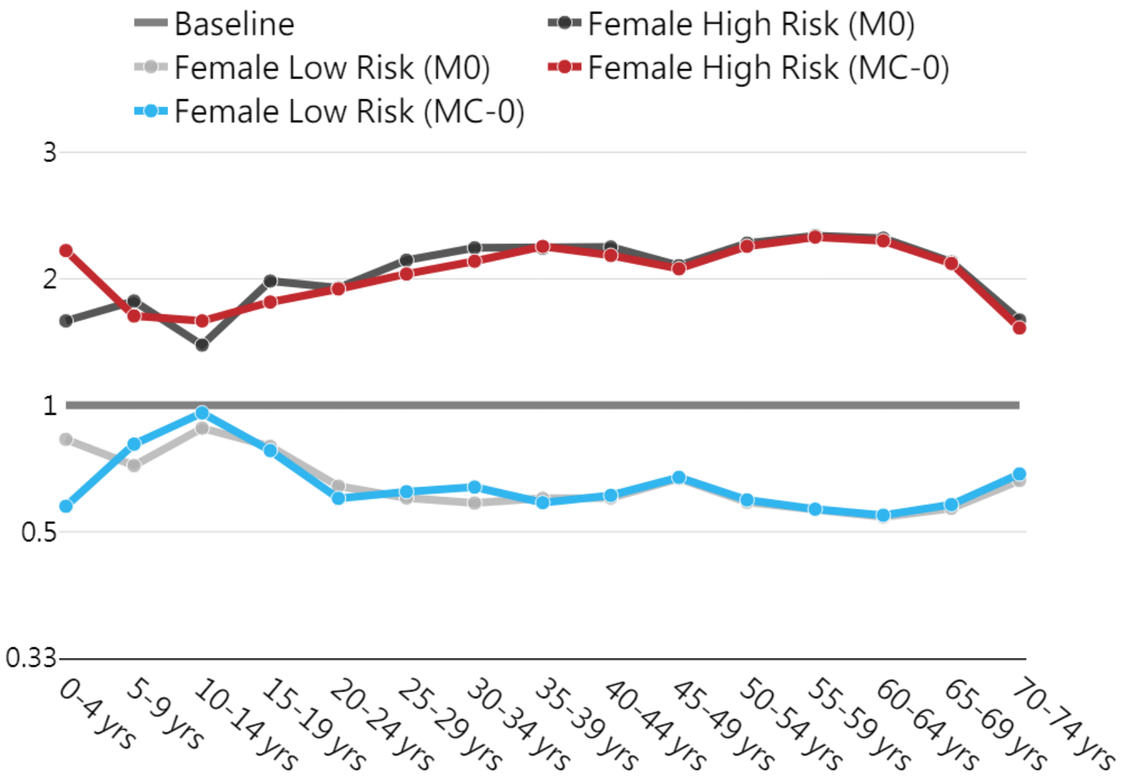
Metric Type	Metric Name	Purpose	Calculation / Interpretation
Model Metric	**AUC** (Area Under the ROC Curve)	Measures overall predictive accuracy	AUC closer to 1 indicates better discrimination between positive and negative cases
Business Metric	**LIFT** (Mortality Segmentation)	Evaluates business relevance of predictions	Compare mortality rate of **high-risk group vs. overall** (↑ is better) and **low-risk vs. overall** (↓ is better) → larger gap = better risk stratification
			Example: High-risk mortality = 10%, overall = 2% → LIFT = 5.0 (well-separated)

Fairness-Performance Trade-off – Baseline Model

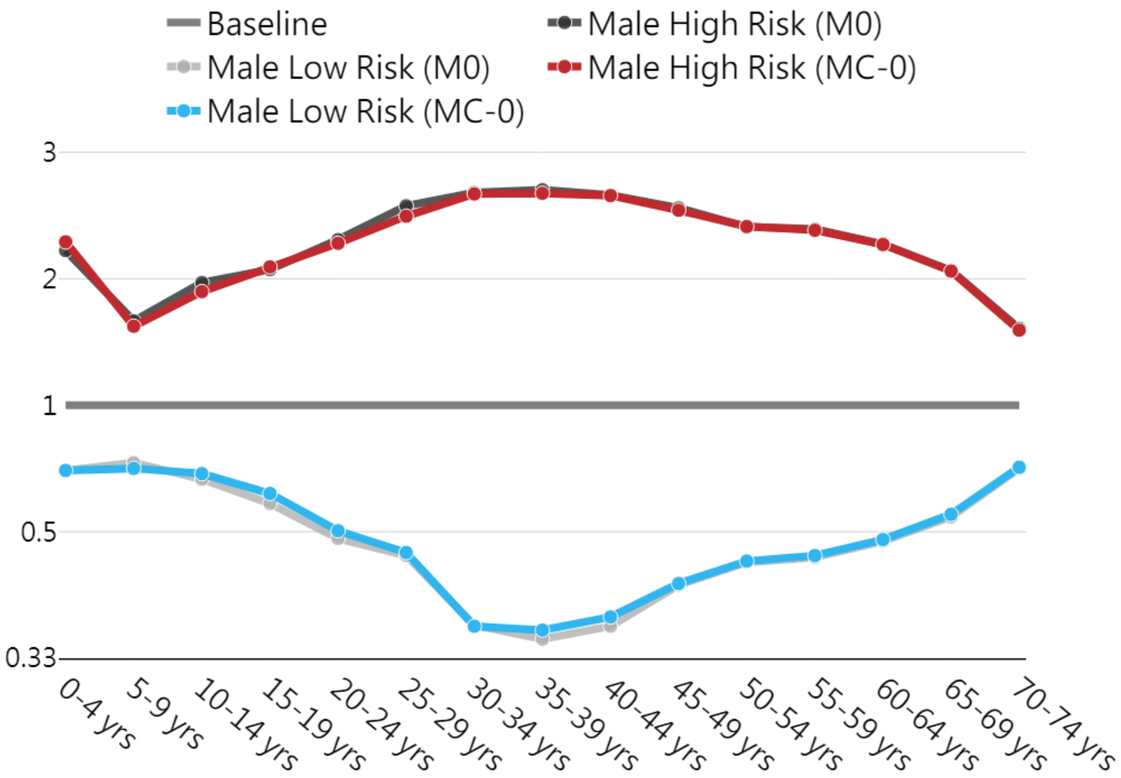
M0 vs MC-0 Comparison

- Assess performance impact when introducing fairness optimization.
- MC-0 reserves data for calibration, reducing training data.
- Goal: Understand trade-off between fairness pipeline and model performance

M0
AUC: 0.8881
MC-0
AUC: 0.8878



Female

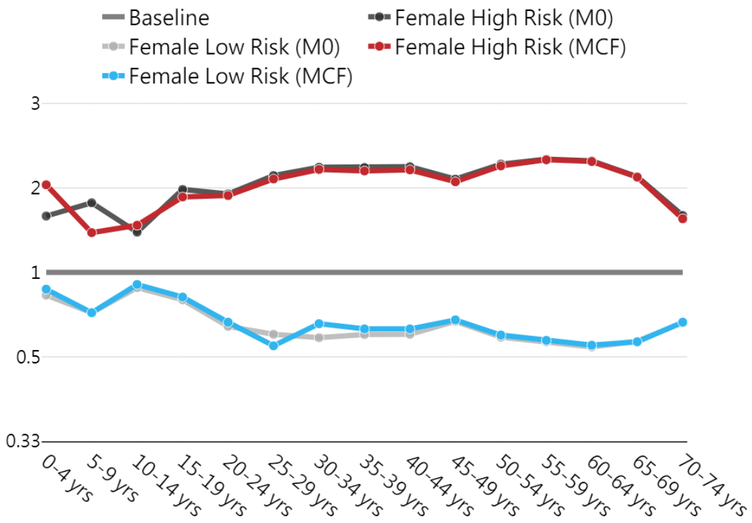


Male

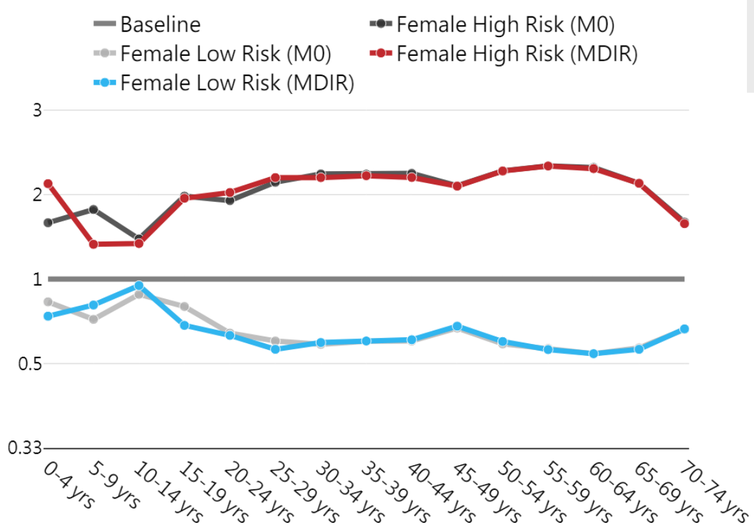
Fairness-Performance Trade-off – Preprocessing

Female

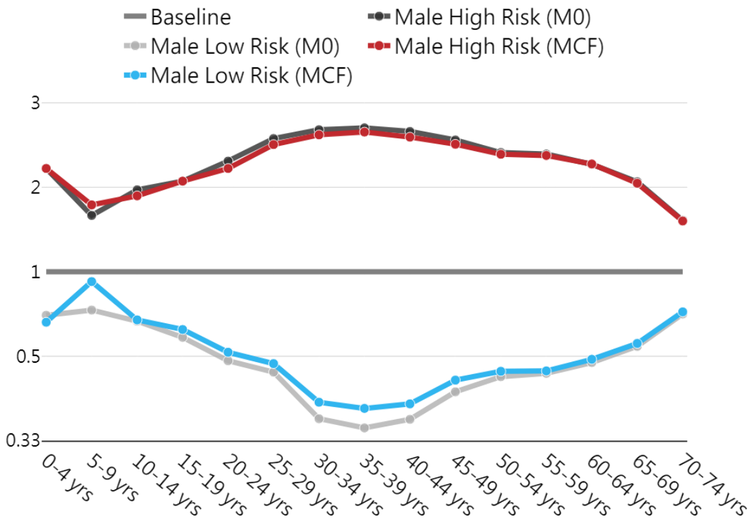
MCF
AUC: 0.8871

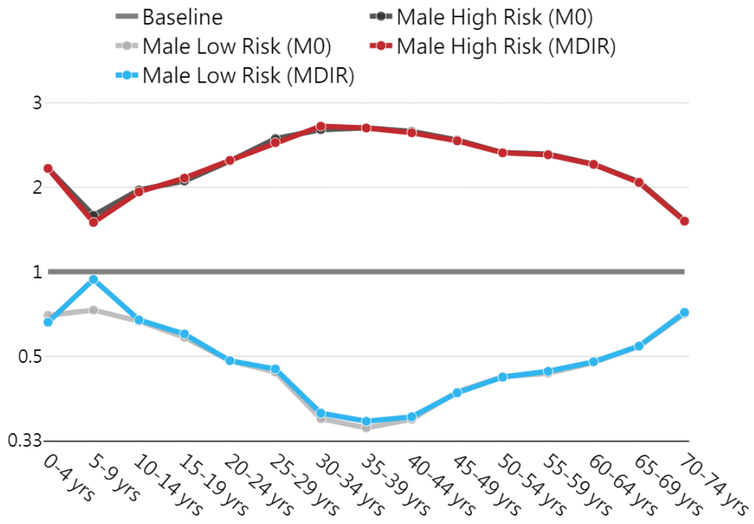


MDIR
AUC: 0.8871



Male

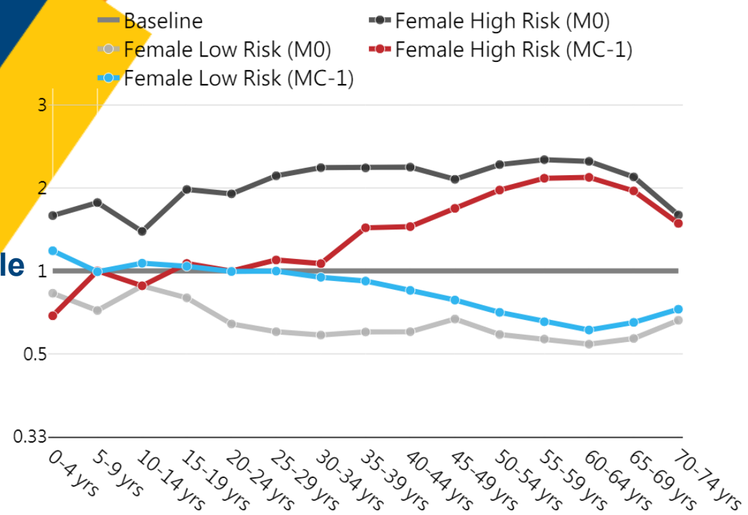




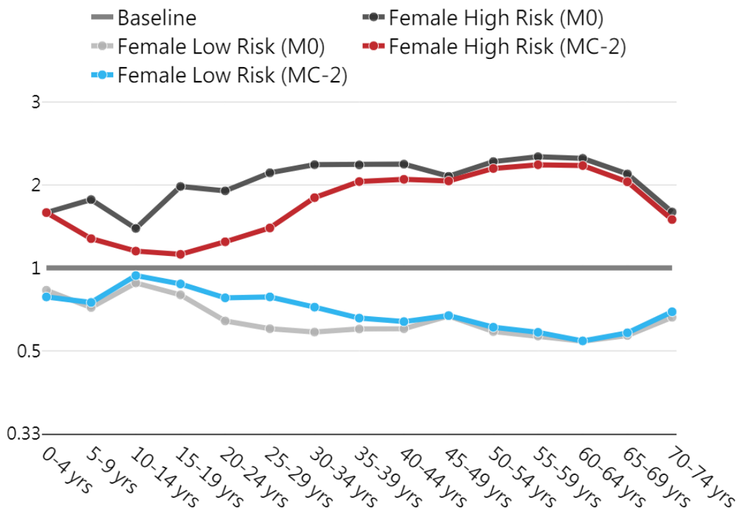
M0
AUC: 0.8881
Male Mortality Rate: 4.63%
Female Mortality Rate: 3.04%

Fairness-Performance Trade-off – Post-processing

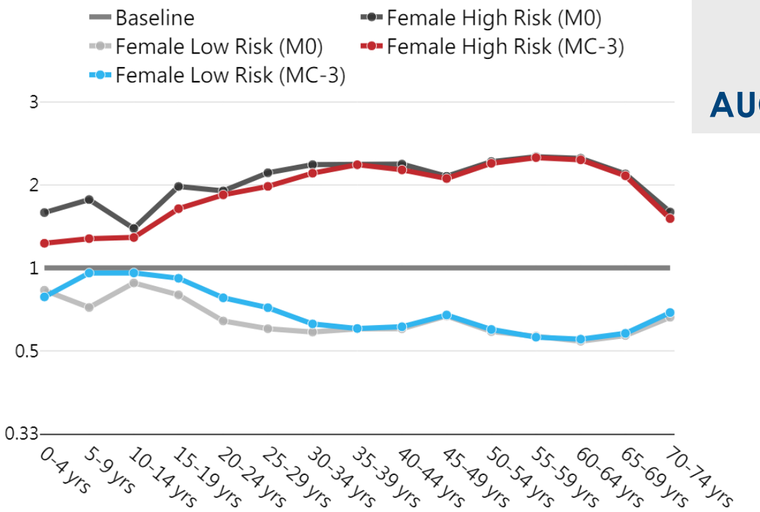
MC-1
AUC: 0.8754



MC-2
AUC: 0.8764



MC-3
AUC: 0.8872

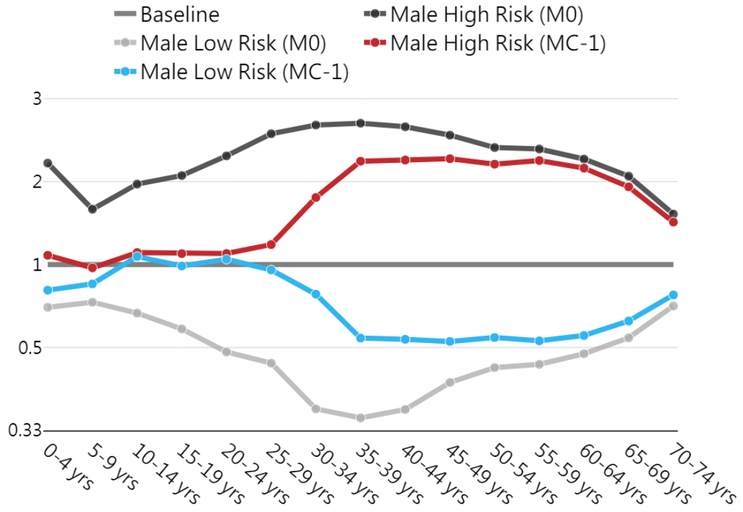


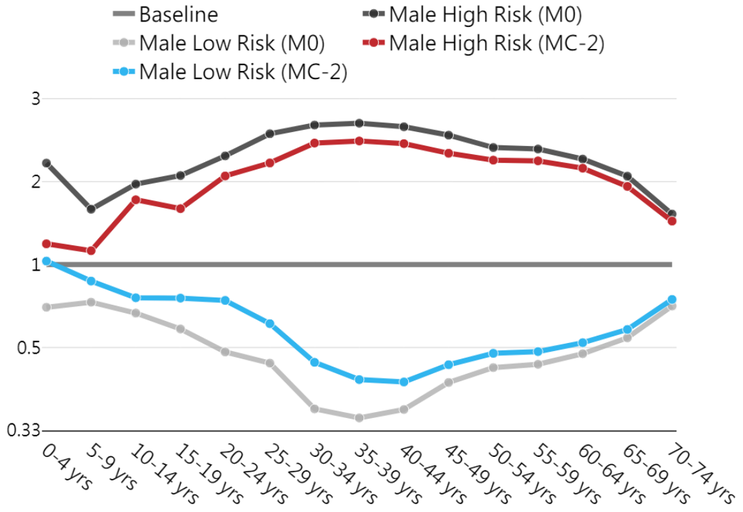
M0
AUC: 0.8881

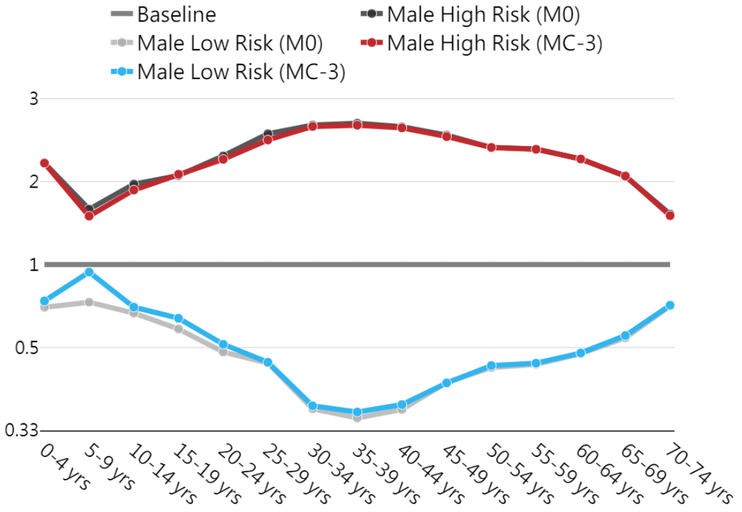
MC-0
AUC: 0.8878

Female

Male







Fairness-Performance Trade-off - Conclusion

- **MC-3 showed the best trade-off,** achieving fairness improvement without sacrificing AUC, making it the most practical post-processing strategy.
- **Performance impact was more evident in younger age groups,** likely due to smaller sample sizes and lower event rates, which reduce calibration stability.
- **Recommendation:** Apply MC-3's combined calibration + normalization approach to younger groups to avoid score compression and improve risk discrimination.

Category	Model	AUC (Model Metric)	LIFT Value (Business Metric)	Fairness (Age Groups)	Trade-off Summary
Baseline	M0	0.8881	–	–	–
Pre-processing	MCF	0.8871	Stable	No significant improvement	Good performance but no fairness gain
	MDIR	0.8871	Stable	No significant improvement	Good performance but no fairness gain
Post-processing	MC-1	0.8754	Noticeable change at young groups	Significant improvement (G15 ECE = 0.0141/ Overall ECE: 0.0017)	Fairness improved with slight performance decrease
	MC-2	0.8764	Slight change	Moderate improvement (G15 ECE = 0.0965/ Overall ECE: 0.0182)	Balanced fairness and performance, minor performance drop
	MC-3	0.8872	Stable	Significant improvement (G15 ECE = 0.0151/ Overall ECE: 0.0013)	Fairness improved with nearly unchanged performance

05. Conclusion and Key Takeaways

Conclusion

Key Takeaways

Conclusion

Can we achieve a balance between model accuracy and fairness to meet the application needs of AI models in the insurance sector?

- Post-processing methods, especially MC-1 and MC-3, effectively reduced ECE in older age groups, validating the sufficiency-based fairness approach.
- MC-3 achieved substantial fairness improvement with minimal impact on model performance, demonstrating that fairness and predictive utility can coexist when appropriate strategies—like group-specific calibration with normalization—are applied.

Insurance AI models can be fair without sacrificing accuracy.
With the right post-processing design, insurers can meet regulatory fairness standards while preserving model reliability.

Key Takeaways

- **Sufficiency is a practical fairness standard**

Well-calibrated predictions align with actuarial principles and support risk-based insurance modeling.

- **Post-processing can effectively improve fairness**

Group-specific calibration (e.g., MC-1, MC-3) reduces disparities without requiring model retraining.

- **Fairness and accuracy are not mutually exclusive**

MC-3 demonstrated that fairness gains can be achieved with minimal performance trade-off.

- **A structured fairness framework aids responsible AI adoption**

Our two-stage evaluation and improvement process helps insurers ensure compliance and trustworthiness.

Thank you! Obrigado!

Questions?

Reference

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. Retrieved from https://fairmlbook.org

Luo, R., Bhatnagar, A., Bai, Y., Zhao, S., Wang, H., Xiong, C., Savarese, S., Ermon, S., Schmerling, E., & Pavone, M. (2023). Local calibration: Metrics and recalibration. *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. Stanford University & Salesforce AI Research.

Xin, X., & Huang, F. (2023). Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. *Journal of Insurance Issues*, 46(2), 285–319. https://doi.org/10.1080/10920277.2023.2190528

Baumann, J., & Loi, M. (2023). Fairness and risk: An ethical argument for a group fairness definition insurers can use. *Ethics and Information Technology*, 36, Article 45. https://doi.org/10.1007/s10676-023-09728-6

Appendix. Fairness improvement over Regions

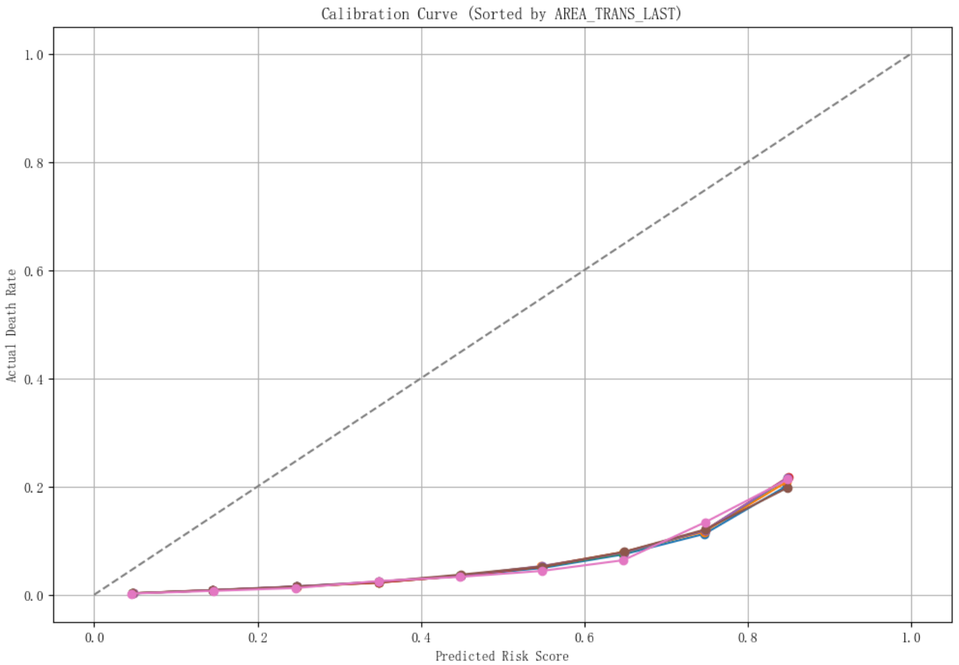
Fairness improvement

Evaluating the Fairness-Performance Trade-off

Fairness improvement – Data Preprocessing

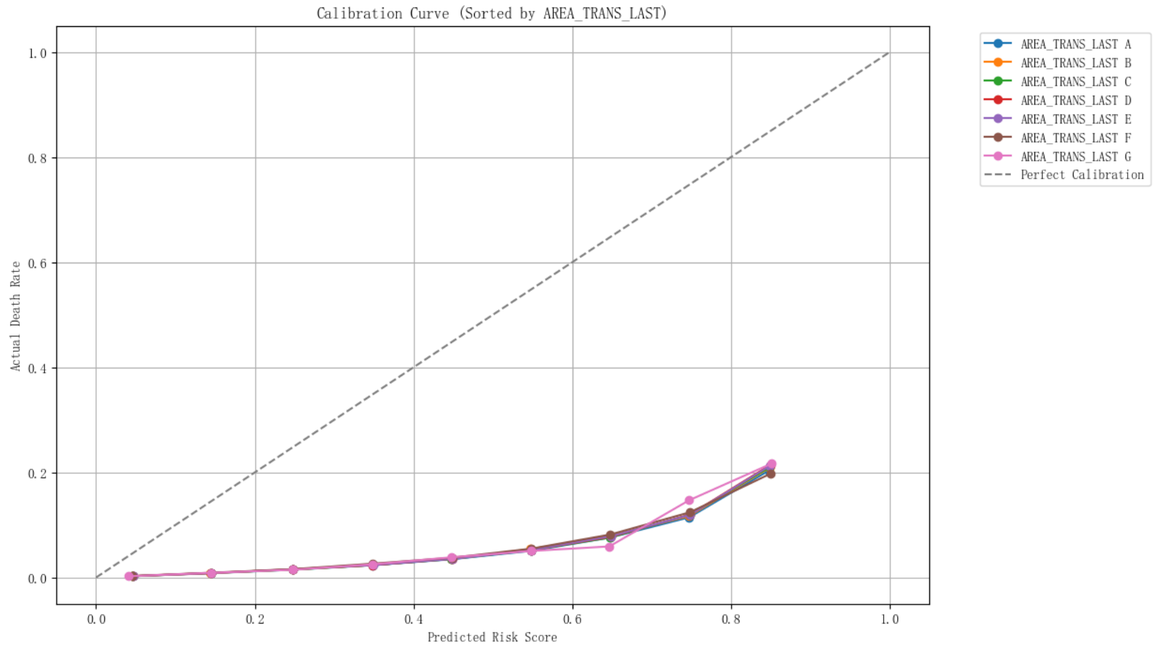
M0
Overall Calibration Error (ECE): 0.2226

MCF



MCF
Overall Calibration Error (ECE): 0.2237

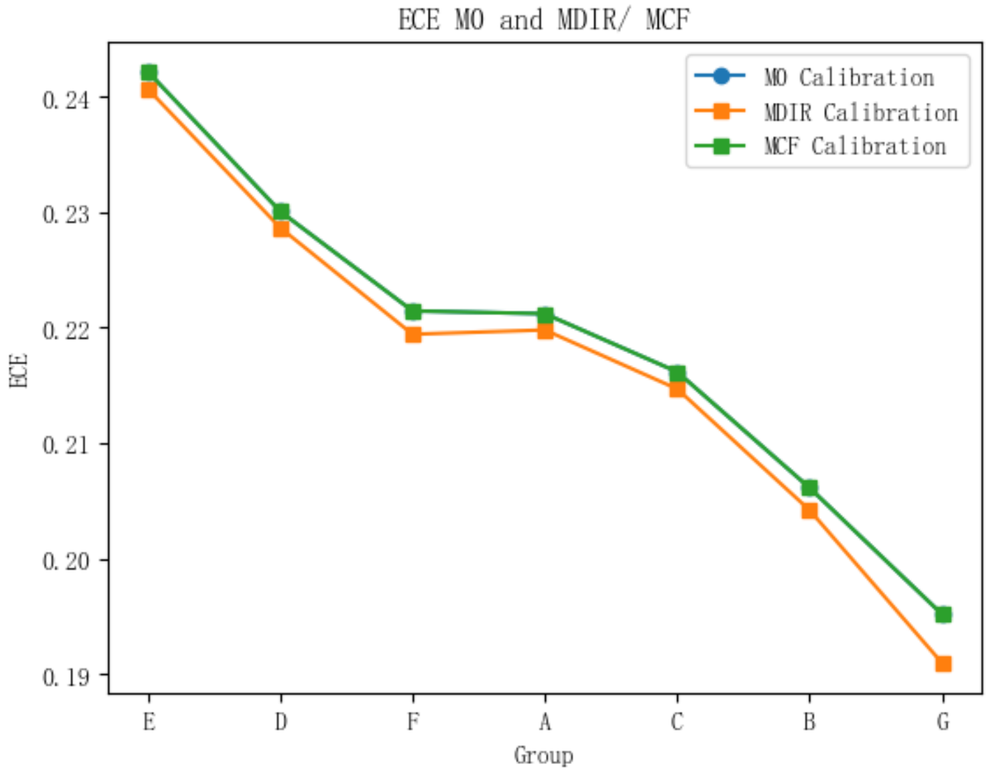
MDIR



MDIR
Overall Calibration Error (ECE): 0.2211

Fairness improvement – Data Preprocessing

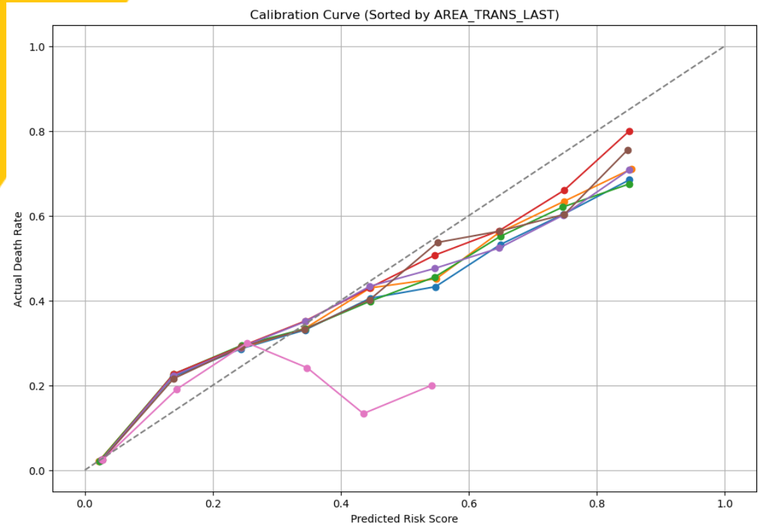
Group	Sample Size	M0	MCF	MDIR
		ECE	ECE	ECE
E	1,184k	0.24215	0.24215	0.24065
D	980k	0.23005	0.23005	0.22864
F	314k	0.22143	0.22143	0.21941
A	1,790k	0.22119	0.22119	0.21978
C	1,324k	0.21615	0.21615	0.21467
B	1,103k	0.20611	0.20611	0.20418
G	22k	0.19518	0.19518	0.19095



Fairness improvement – Post-processing

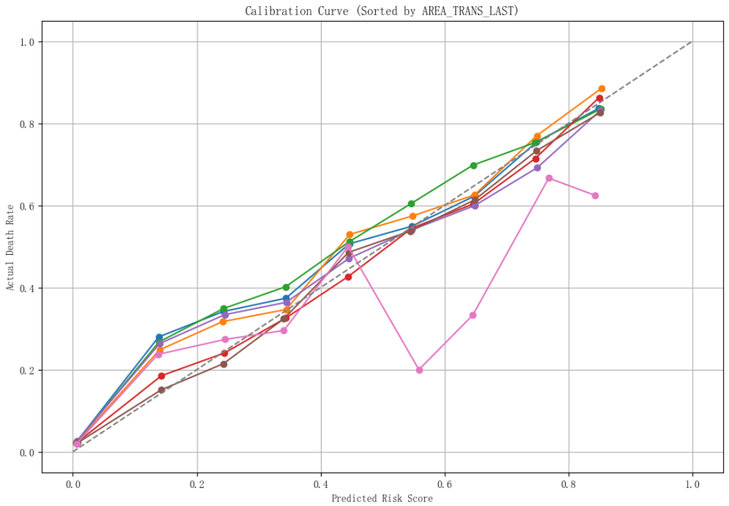
M0
Overall Calibration Error (ECE): 0.2226

MC-1



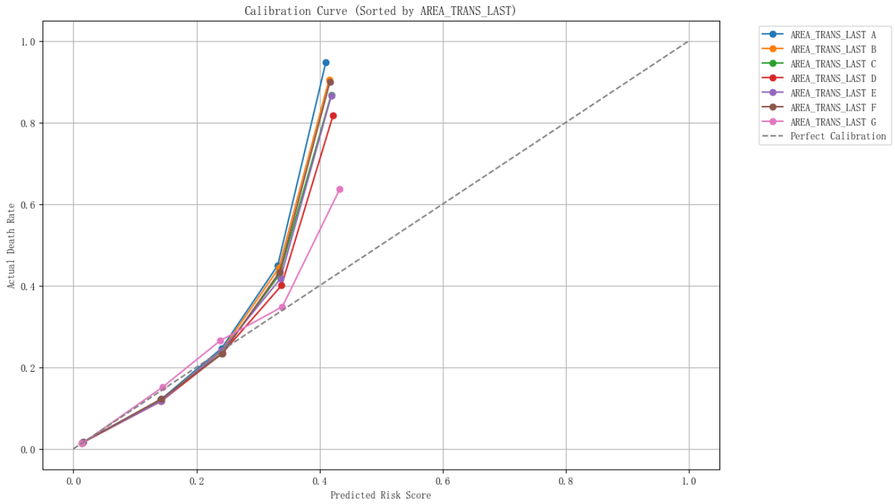
MC-1
Overall Calibration Error (ECE): 0.0135

MC-2



MC-2
Overall Calibration Error (ECE): 0.0181

MC-3

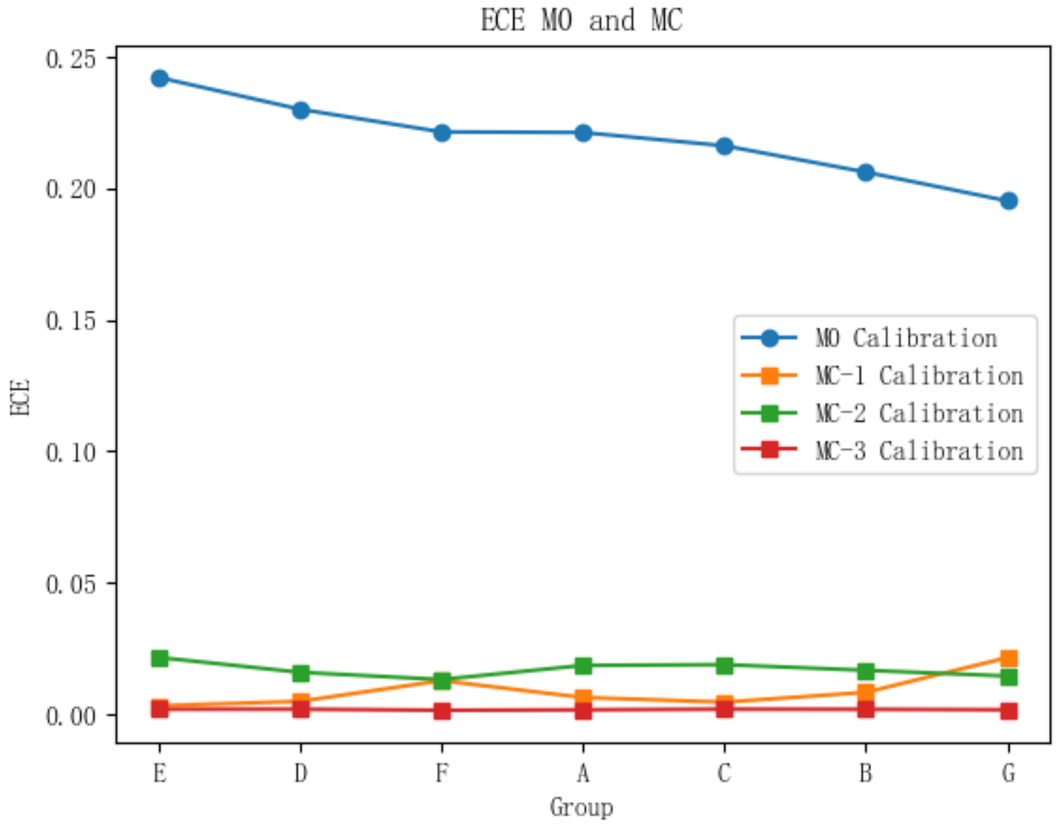


MC-3
Overall Calibration Error (ECE): 0.0026

The calibration curve of the MC-3 model appears visually discontinuous and more volatile, mainly due to the compression of predicted values after group normalization and the smaller sample sizes in some groups.

Fairness improvement – Post-processing

Group	Sample Size	M0	MC-1	MC-2	MC-3
		ECE	ECE	ECE	ECE
E	1,184k	0.24215	0.00312	0.02151	0.00182
D	980k	0.23005	0.00482	0.01587	0.00183
F	314k	0.22143	0.01268	0.01312	0.00143
A	1,790k	0.22119	0.00629	0.01848	0.00158
C	1,324k	0.21615	0.00451	0.01877	0.00184
B	1,103k	0.20611	0.00824	0.01663	0.00178
G	22k	0.19518	0.02147	0.01444	0.00161



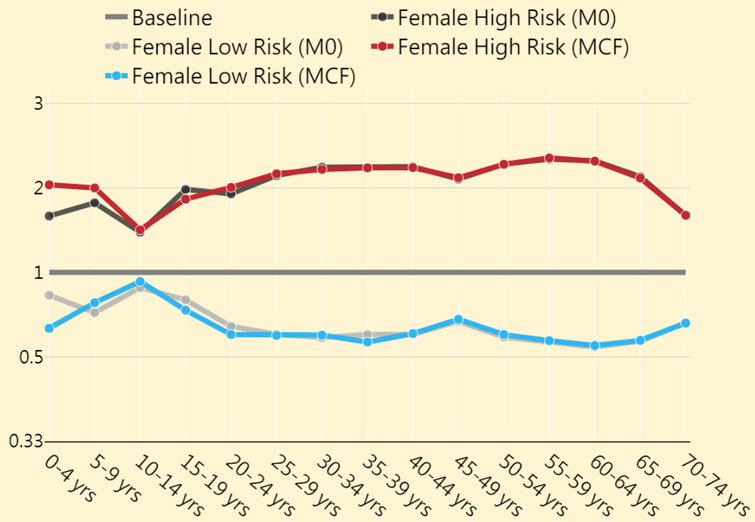
Fairness improvement – Result

Fairness optimization for Regions shows that data preprocessing methods provide limited improvement in reducing prediction bias across groups. In contrast, post-processing strategies using group-specific sigmoid calibration—particularly MC-1 and MC-3—effectively reduce prediction bias.

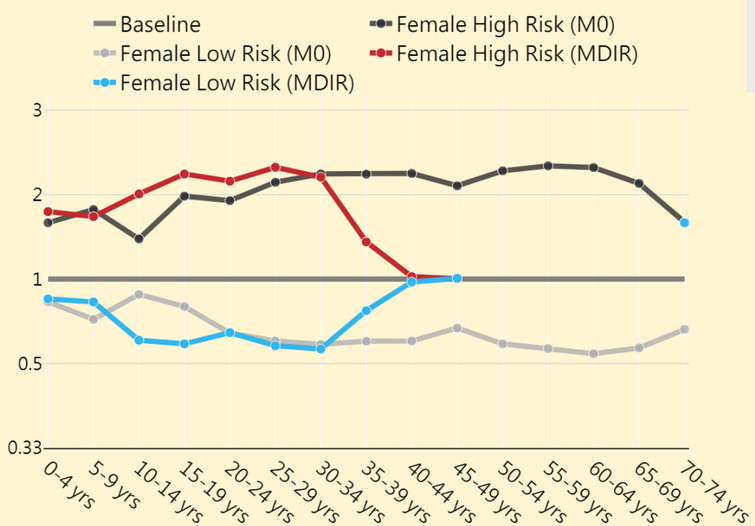
Experiment ID	Max ECE	Min ECE	Standard deviation of ECE across all groups	Whether fairness has improved
M0	0.2421	0.1951	0.0235	-
MCF	0.2421	0.1951	0.0235	No significant improvement
MDIR	0.2406	0.1909	0.0249	No significant improvement
MC-0	0.2421	0.1951	0.0235	-
MC-1	0.0214	0.0031	0.0092	Significant improvement
MC-2	0.0215	0.0131	0.0042	Some improvement
MC-3	0.0018	0.00143	0.0002	Significant improvement

Fairness-Performance Trade-off – Preprocessing

MCF
AUC: 0.8881



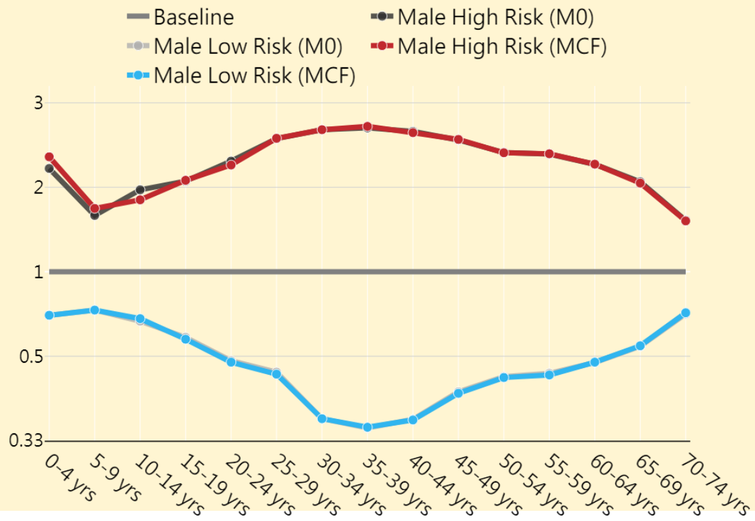
MDIR
AUC: 0.8881

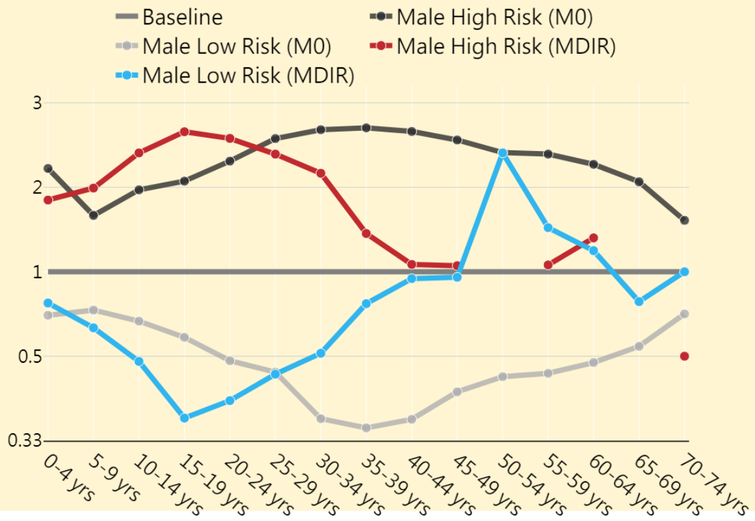


M0
AUC: 0.8881
Male Mortality Rate: 4.63%
Female Mortality Rate: 3.04%

Female

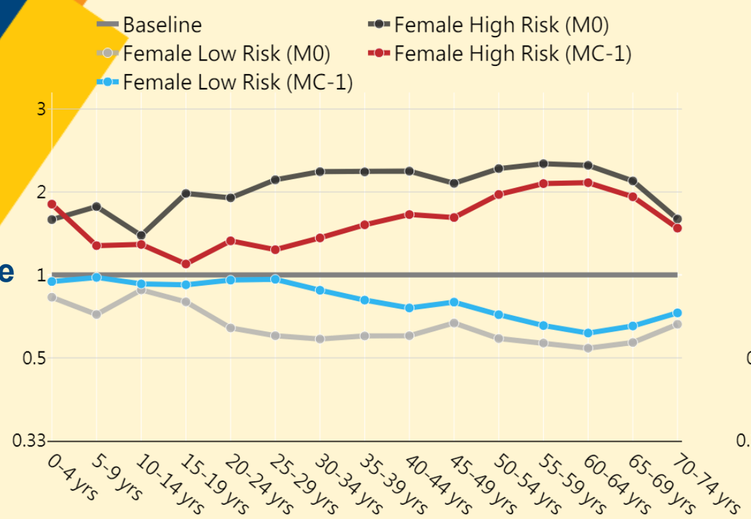
Male



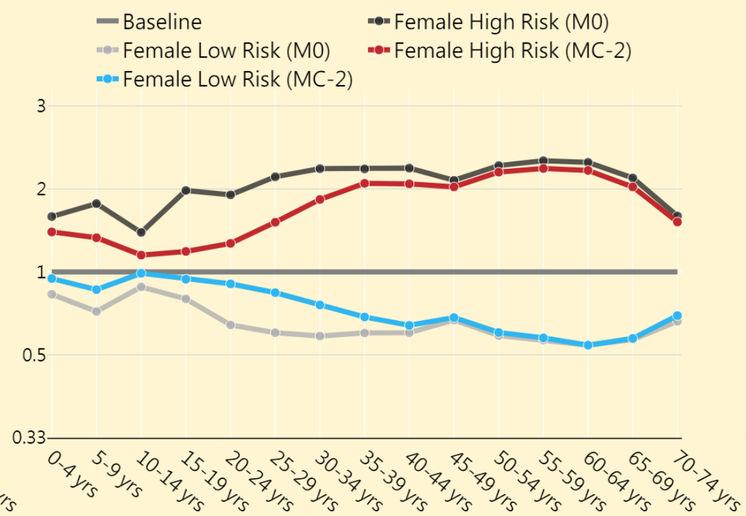


Fairness-Performance Trade-off – Post-processing

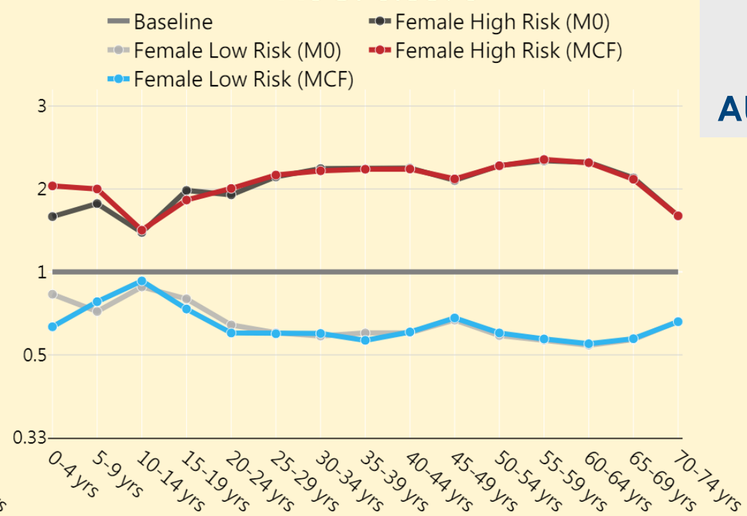
MC-1
AUC: 0.8503



MC-2
AUC: 0.8815



MC-3
AUC: 0.8875

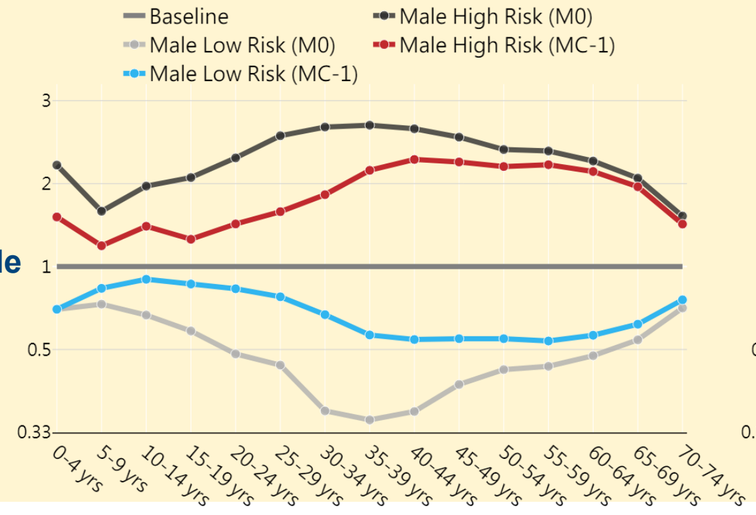


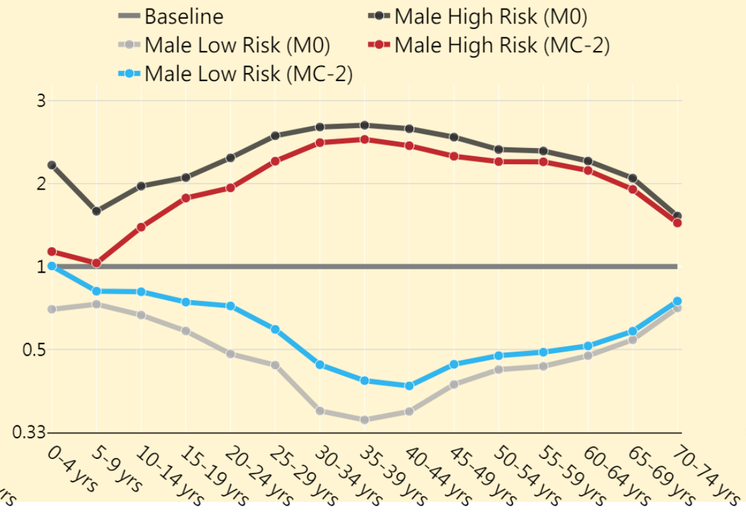
M0
AUC: 0.8881

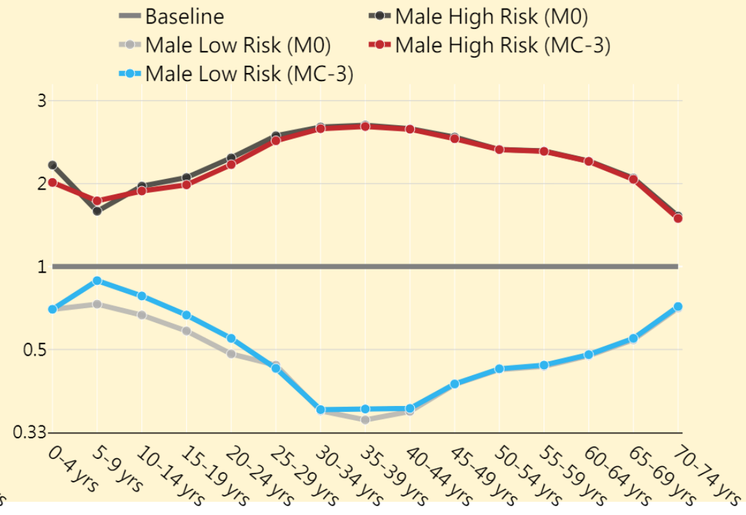
MC-0
AUC: 0.8878

Female

Male







Fairness-Performance Trade-off - Conclusion

- **MC-3 showed the best trade-off,** achieving fairness improvement without sacrificing AUC, making it the most practical post-processing strategy.
- **Recommendation:** Apply MC-3's combined calibration + normalization approach to younger groups to avoid score compression and improve risk discrimination.

Category	Model	AUC (Model Metric)	LIFT Value (Business Metric)	Fairness (Age Groups)	Trade-off Summary
Baseline	M0	0.8881	–	–	–
Pre-processing	MCF	0.8881	Stable	No significant improvement	Good performance but no fairness gain
	MDIR	0.8881	Stable	No significant improvement	Good performance but no fairness gain
Post-processing	MC-1	0.8503	Noticeable change at young groups	Significant improvement (Overall ECE: 0.0135)	Fairness improved with slight performance decrease
	MC-2	0.8815	Slight change	Moderate improvement (Overall ECE: 0.0181)	Balanced fairness and performance, minor performance drop
	MC-3	0.8875	Stable	Significant improvement (Overall ECE: 0.0026)	Fairness improved with nearly unchanged performance