SOA Research Institute
SOCIETAL PURPOSE

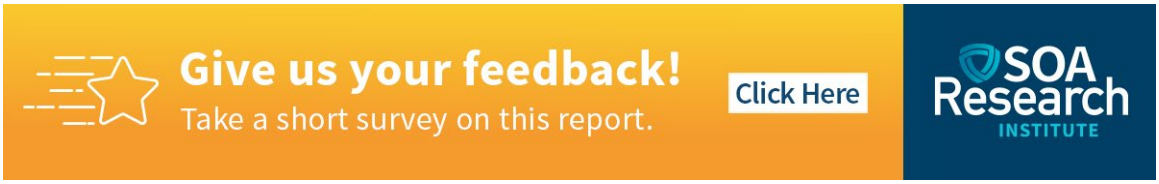
Evaluating Bias Mitigation Strategies in AI Models for Insurance Pricing and Risk Assessment

OCTOBER | 2026

Evaluating Bias Mitigation Strategies in AI Models for Insurance Pricing and Risk Assessment

AUTHORS Hong Li, FSA, ACIA
Jiangang He, FSA, FCIA

SPONSOR Society of Actuaries Research Institute



Caveat and Disclaimer
The opinions expressed and conclusions reached by the authors are their own and do not represent any official position or opinion of the Society of Actuaries Research Institute, the Society of Actuaries, or its members. The Society of Actuaries Research Institute makes no representation or warranty to the accuracy of the information.

Copyright © 2026 by the Society of Actuaries Research Institute. All rights reserved.

CONTENTS

Executive Summary **5**

Section 1 **Key Terms, Scope, and How to Read This Report** **7**

Section 2 **Introduction** **9**

Section 3 **Applications** **11**
3.1 Life Insurance 11
3.1.1 Embedded Statistical Bias Mechanism 11
3.1.2 Data Generation Process 12
3.1.3 Exploratory Data Patterns in the Simulation 13
3.1.4 Model Evaluation Results 15
3.2 Auto Insurance 18
3.2.1 Embedded Statistical Bias Mechanism 18
3.2.2 Data Generation Process 19
3.2.3 Exploratory Data Patterns in the Simulation 20
3.2.4 Model Evaluation Results 21
3.3 Credit Insurance 24
3.3.1 Embedded Statistical Bias Mechanism 24
3.3.2 Data Generation Process 25
3.3.3 Exploratory Data Patterns in the Simulation 26
3.3.4 Model Evaluation Results 27
3.4 Mortgage Insurance 30
3.4.1 Embedded Statistical Bias Mechanism 30
3.4.2 Data Generation Process 31
3.4.3 Exploratory Data Patterns in the Simulation 31
3.4.4 Model Evaluation Results 32
3.5 Health Insurance 35
3.5.1 Embedded Statistical Bias Mechanism 35
3.5.2 Data Generation Process 36
3.5.3 Exploratory Data Patterns in the Simulation 37
3.5.4 Model Evaluation Results 38

Section 4 **Indirectly Associated Variables, Measurement Bias, and Group Disparities** **42**
4.1 Definitions and Notation 42
4.2 Mechanisms that Generate Indirectly Associated Disparities 42
4.3 A Useful Target in Indirect Association and Measurement-Bias Settings 43

Section 5 **Predictive Models and Mitigation Approach** **44**
5.1 Modeling Goal 44
5.2 Model Overview 44
5.2.1 Generalized Linear Model 45
5.2.2 Neural Network 45
5.2.3 Adversarial Neural Network 45

Section 6 **Evaluation Design and Group-Audit Metrics** **47**
6.1 Evaluation Goals 47
6.2 Predictive Performance: ROC Curve and AUC 47
6.3 Fixed Operating Points: Selecting the Top $\boldsymbol{X}\%$ as "High Risk" 47
6.4 Group-audit Metrics Reported 48
6.5 Estimation, Tuning, and Audit Workflow 48

Section 7 **Conclusion** **50**

Acknowledgments **51**

Copyright © 2026 Society of Actuaries Research Institute

Appendix A: Mathematical Specification of Models 52
A.1 Generalized Linear Model 52
A.2 Standard Neural Network (Example with Two Hidden Layers) 52
A.3 Adversarial Neural Network (ADV-NN) 52
A.3.1 Predictor and Observed Indirectly Associated Adversary Variable 52
A.3.2 Loss Functions and Training Objective 53
A.4 Implementation Details 53

Appendix B: Mathematical Specification of Evaluation Metrics 55
B.1 Fixed Operating Point and Decision Indicator 55
B.2 AUC (Ranking Performance) 55
B.3 Demographic Parity (Selection Rate) Metric 55
B.4 Absolute True Positive Rate Gap 55

Appendix C: Data-Generating Process of Life Insurance 56
C.1 Notation and Basic Functions 56
C.2 Group and Indirectly Associated Variable Generation 56
C.3 UNOBSERVED Score and Biased Measurement 56
C.4 Covariates and Additional Observed Risk Factors 56
C.5 Outcome Model and Intercept Calibration 57
C.6 Where Group Disparities Enter 58

Appendix D: Data-Generating Process of Auto Insurance 59
D.1 Notation and Basic Functions 59
D.2 Group and Territory Generation 59
D.3 Age, Mileage, and Transformed Exposure 59
D.4 Unobserved Violations, Recorded Violations, and Prior Claims 59
D.5 Outcome Model and Intercept Calibration 59
D.6 Where Group Disparities Enter 59

Appendix E: Data-Generating Process of Credit Insurance 60
E.1 Notation and Basic Functions 60
E.2 Group and Indirectly Associated Variable Generation 60
E.3 Unobserved Credit Strength and Biased Observation 60
E.4 Other Covariates 60
E.5 Outcome Model and Intercept Calibration 61
E.6 Where Group Disparities Enter 61

Appendix F: Data-Generating Process of Mortgage Insurance 62
F.1 Notation and Basic Functions 62
F.2 Group and Indirectly Associated Variable Generation 62
F.3 Unobserved Credit Score and Biased Measurement 62
F.4 Covariates (LTV and DTI) 62
F.5 Outcome Model and Intercept Calibration 62
F.6 Where Group Disparities Enter 62

Appendix G: Data-Generating Process of Health Insurance 63
G.1 Notation and Basic Functions 63
G.2 Group, Indirectly Associated Variable, Age, Sex, and Frailty Generation 63
G.3 Chronic Condition and Unobserved Morbidity 63
G.4 Observed Morbidity and Measurement Bias 64
G.5 Outcome Model and Intercept Calibration 64
G.6 Where Group Disparities Enter 64

About The Society of Actuaries Research Institute 66

Copyright © 2026 Society of Actuaries Research Institute

Evaluating Bias Mitigation Strategies in AI Models for Insurance Pricing and Risk Assessment

Executive Summary

AI-based predictive models are increasingly used throughout insurance for pricing, underwriting, triage, risk selection, claim management, and related decisions. Their effectiveness depends on the quality of the underlying data, yet observed insurance data may not measure the intended risk cleanly.

This is the central issue studied in the report. Observed variables such as territory, ZIP code, credit-related scores, recorded violations, health-care utilization, or morbidity scores may contain useful predictive information, but they may also reflect measurement differences, administrative processes, access barriers, or social and geographic patterns. If a model treats these observed variables as pure risk signals, it may learn distorted information rather than the underlying risk the insurer intends to measure.

The guiding question here is: when observed insurance data contain group-correlated measurement distortion or indirect association, can model design improve the alignment between model outputs and underlying risk while preserving useful predictive performance? This question concerns model validity as well as group-level audit outcomes: a model that relies on distorted signals may be less aligned with the risk being priced even when its average predictive performance remains strong.

This question matters practically because a model can perform well on average while still relying on a distorted data pathway. The concern is especially important at high-risk thresholds. Many insurance models do not determine an action for every applicant or policyholder. Instead, they identify a small subset for manual review, additional underwriting, requests for documentation, more restrictive terms, surcharges, or declination. A small distortion in the upper tail of predicted risk can therefore have material consequences for who receives additional scrutiny or less favorable treatment.

A model can be accurate on average while still relying on a distorted data pathway.

This report treats biased measurement and indirect association as model-quality issues as well as fairness concerns. If a model relies on an observed variable that is systematically distorted relative to underlying risk, the resulting prediction may be less credible as an estimate of the risk being priced. Similarly, excluding an explicit group label from the model does not necessarily remove group-related information. Other variables may still transmit that information indirectly into model predictions.[1] Please refer to Section 1 for definitions of key terms and scope of this study.

Across the controlled applications studied in the report, standard predictive models can produce material differences in high-risk selection rates and error rates across groups even when the audit label does not enter the true outcome equation directly. These differences arise because the model learns from the observed data structure.

[1] In some contexts, this type of indirect route may be referred to as a proxy pathway. However, to avoid any potential confusion with the term “proxy” as a specifically defined term in certain legal contexts, this report avoids using the term “proxy.”

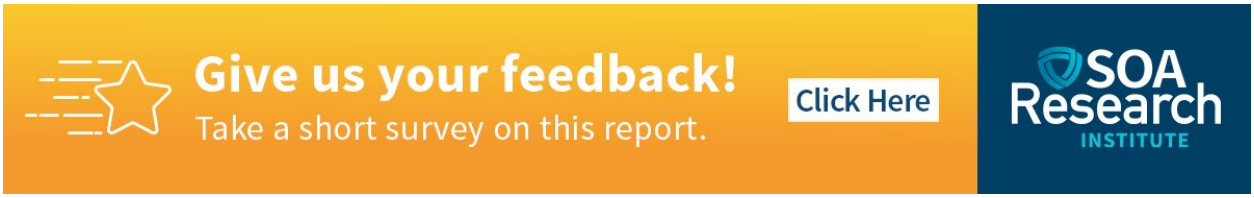
Copyright © 2026 Society of Actuaries Research Institute

If the observed data contain distorted measurements or indirectly associated variables, the fitted model can reproduce those distortions in its predictions.

The report also finds that mitigation can help. In the applications studied, the adversarial neural network generally reduces recoverable information about the designated indirectly associated variable and lowers reported group differences in high-risk flags or error rates. The magnitude of the improvement varies across applications, and stronger adversarial penalties can reduce predictive performance.

The main practical takeaway is that, for insurers, model evaluation extends beyond average predictive accuracy to include whether the model is learning the intended risk signal. Robust model validation generally includes diagnostic testing for biased measurement, review of indirectly associated variables, group-level audit diagnostics, and threshold-specific analysis at operational decision points such as the top 1%, 2%, or 5% of predicted risks.

The report does not conclude that every measured group difference is unfair. It does not estimate the experience of any particular insurer, product, jurisdiction, or market, and it does not treat group-comparison metrics as substitutes for actuarial judgment. Its conclusion is narrower and more practical: when observed inputs are systematically distorted or indirectly associated with group membership, fairness concerns and model-validity concerns become closely connected. Insurers can reduce model risk by identifying distorted data pathways, testing whether those pathways affect high-impact decisions, and comparing model designs that better align predictions with the risk the insurer intends to measure.



Copyright © 2026 Society of Actuaries Research Institute

Section 1 Key Terms, Scope, and How to Read This Report

Several terms used in discussions of insurance modeling, artificial intelligence, and group-based model evaluation may carry legal, regulatory, or policy meanings. This report uses these terms in a statistical, actuarial, and model-audit sense. The group differences reported in the synthetic applications are statistical diagnostics of fitted-model behavior and are not intended to be interpreted as legal findings.

The following definitions apply throughout the report.

- Artificial intelligence (AI) refers in this report primarily to flexible machine-learning models, especially neural networks, which can learn nonlinear relationships and interactions from observed insurance data.
- Underlying or unobserved risk refers to the risk-relevant quantity or outcome-generating process that the insurer intends to measure but does not observe directly without error or distortion. Depending on the application, this may be unobserved credit quality, underlying morbidity, true driving behavior, or another risk-relevant characteristic. In the synthetic simulations, this quantity is known by construction from the data-generating process. In operational insurance data, it would generally not be observed directly.
- Statistical bias means a systematic distortion in observed predictors or model outputs relative to underlying or unobserved risk. The term is used here in a model-quality sense and is not intended as a legal determination.
- Measurement bias occurs when an observed variable does not measure the corresponding unobserved risk-relevant quantity in the same way across audit groups. For example, an observed credit score, violation count, or morbidity measure may contain a systematic group-correlated shift in addition to random measurement noise.
- The audit label, denoted by A, is a group indicator used in the simulation design and in ex post held-out auditing. It is not included among the model predictors, used as the adversarial target, used to select the adversarial penalty parameter, or used in model refitting. The audit label is introduced only after the fitted models have been fixed to calculate group-level audit diagnostics.
- An indirectly associated variable is an observed feature that carries materially recoverable information about the audit label because of a meaningful association in the data-generating process. It can therefore transmit group-related information indirectly into prediction even though the audit label itself is excluded from the model. The term is narrower than “any variable with nonzero correlation” and is used as a statistical description rather than as a legal designation.
- A group-neutral outcome equation means that the audit label does not enter the true simulated outcome equation directly. It does not necessarily mean that every observed or unobserved risk-factor distribution is identical across groups. Group differences may still arise through indirectly associated variables, measurement distortion, or differences in the distributions of other simulated characteristics.
- The model-development sample is the portion of the simulated data used for model fitting, cross-validation, and tuning. Within this sample, outcome-stratified cross-validation divides the observations into fitting and validation folds while preserving the distribution of the binary outcome as closely as possible.
- The held-out evaluation sample is kept separate from model fitting and tuning. It is used only after the models and tuning parameters have been fixed to evaluate predictive performance and calculate ex post group-audit metrics.

Copyright © 2026 Society of Actuaries Research Institute

- The generalized linear model (GLM) is the actuarial benchmark used in the report. For the binary outcomes studied here, it is implemented as a logistic model.

- The standard neural network (NN) is the flexible machine-learning benchmark. It is trained to predict the insurance outcome from the observed predictors without an adversarial penalty.

- The adversarial neural network (ADV-NN) is the mitigation model. It is trained to predict the insurance outcome while making a designated observed indirectly associated variable more difficult to recover from the learned representation. The adversarial target is the observed indirectly associated variable, not the audit label.

- The adversarial penalty parameter, denoted by λ, controls the strength of the adversarial component of ADV-NN training. Smaller values place greater emphasis on the predictive task, while larger values place greater emphasis on reducing recoverable information about the designated indirectly associated variable.

- Area under the receiver operating characteristic curve (ROC AUC) measures predictive ranking performance. It is the probability that a randomly selected observation with the adverse outcome receives a higher predicted score than a randomly selected observation without the adverse outcome. Higher ROC AUC indicates better ranking performance.

- A high-risk operating point is the threshold at which negative action is taken and is defined by an action rate, such as 1%, 2%, or 5%. The model ranks observations by predicted risk, and the specified percentage with the highest predicted scores is selected for the relevant operational action. In underwriting applications, this action may involve additional review, documentation, more restrictive terms, a surcharge, or declination. In the health application, it may instead represent a triage or intervention decision.

- The demographic-parity ratio compares the rates at which the two audit groups are selected into the high-risk category. A ratio closer to one indicates more similar selection rates across groups.

- The absolute true-positive-rate (TPR) gap compares the proportions of truly adverse cases selected into the high-risk category across the two audit groups. A value closer to zero indicates more similar true-positive rates.

ROC AUC, demographic-parity ratios, and absolute TPR gaps reported in the application sections are calculated on the held-out evaluation sample. The audit label is used only for the two group-audit metrics. It is not used to fit the GLM, NN, or ADV-NN, to train the adversarial branch, or to select λ.

The guiding technical question of the report is whether, when observed insurance data contain group-correlated measurement distortion or indirect association, model design can improve the alignment between model outputs and underlying risk while preserving useful predictive performance. Group-comparison metrics are retained as audit diagnostics for that question and are not treated as substitutes for actuarial judgment.

How to Read This Report

Readers interested primarily in the report's high-level implications may read the Executive Summary, this section, the Introduction, and the Conclusion. Readers interested in a particular line of business may proceed from this section directly to the corresponding life, auto, credit, mortgage, or health application. Each application restates the model and figure conventions required to interpret its results, so the other line-of-business applications may be skipped. Readers seeking technical details on notation, model specification, adversarial training, tuning, and evaluation may consult the methodological sections and Appendices A–B.

Copyright © 2026 Society of Actuaries Research Institute

Section 2 Introduction

Predictive models are now used throughout insurance to support pricing, underwriting, triage, risk selection, claim management, and other decisions. In this report, artificial intelligence (AI) refers to flexible machine-learning models, especially neural networks, which can learn complex relationships from observed insurance data. Compared with traditional actuarial models, these methods can identify complex patterns more automatically, but they may also be more sensitive to distortions in the training data (Eling, Nuessle, and Staubli, 2022; Smith, Pirchalski, and Golbin, 2022).

The central issue studied in this report is that the data available to an insurer may not measure underlying risk cleanly. In many applications, the insurer observes variables such as territory, ZIP code, credit-related scores, recorded violations, health-care utilization, or morbidity scores. These variables may be useful for prediction, but they may also reflect measurement differences, administrative processes, access barriers, or social and geographic patterns that are not the same as the underlying insurance risk that the model is intended to estimate. When this occurs, a predictive model may learn from the observed signal as if it were a pure risk signal, even though part of that signal is distorted.

This report refers to this problem as statistical bias. Bias in this context means a systematic distortion in observed predictors or model outputs relative to the underlying risk process. The concern is not that a model is imperfect in general. The concern is that the model may assign different scores, prices, flags, or underwriting actions to otherwise similar risks because the observed data are systematically different across groups. A model can therefore be accurate in aggregate while still relying on a data pathway that weakens the alignment between prediction and underlying risk.

If the data available to an insurer are imperfect, do predictive models learn true risk or distorted signals?

Consider a credit or mortgage example. Two applicants may have similar underlying repayment quality, but one applicant may have a lower observed credit score because the available data understate that applicant's credit history or financial position. A model trained on the observed score may treat that lower value as a true risk difference and assign a higher predicted default risk. A similar issue can arise in auto insurance. Two drivers may have the same underlying driving behavior, but one may have more recorded violations because the driver lives in a territory with more intensive monitoring or enforcement. A model that uses recorded violations may then respond partly to the recording process rather than only to the underlying driving behavior.

A related issue is the indirect route through which group-related information enters a model through another observed variable. The model does not need to observe a subgroup label directly. If geography, neighborhood characteristics, credit-related measures, utilization patterns, or area-deprivation measures are correlated with subgroup membership, then these variables can carry part of the same information into the fitted model. Recent SOA work notes that leaving a variable out of risk calculations does not by itself remove indirect effects when other observed variables carry materially related information about the omitted characteristic (Krafcheck, Balnozan, and Huang, 2026).

Removing an explicit group label does not necessarily remove group-related information from a model.

This report studies the interaction between predictive modeling, biased measurement, and indirectly associated variables. The guiding question is whether a model trained on imperfect observed data is responding to the underlying risk process or to distorted information in the observed inputs. The same question has both model-validity and fairness implications. If a fitted model relies heavily on a distorted pathway, then its outputs may be less credible as

Copyright © 2026 Society of Actuaries Research Institute

estimates of the risk the insurer intends to measure. Reducing that reliance can therefore be understood as a model-governance objective as well as a fairness objective.

The analysis uses controlled synthetic simulations across five insurance settings: life, auto, credit, mortgage, and health. Synthetic data are generated from an explicit simulation design rather than drawn from an insurer's actual portfolio. This design allows the underlying risk process to be specified directly. The simulation can separately define the unobserved risk signal, the observed variables available to the model, the indirectly associated variables, and the measurement distortions. Because the mechanism is known, the analysis can examine whether a fitted model is recovering the intended risk signal or responding to the distorted observable.

Each simulation omits a direct audit-label term from the true outcome equation. This is what this report means by a group-neutral outcome equation; it does not necessarily imply that every risk-factor distribution is identical across groups. Instead, group differences can arise through observed variables. One group may be more likely to be associated with a territory, ZIP code, neighborhood, or area-deprivation measure. One group may also have an observed score that is shifted away from the unobserved score that truly drives risk. If the fitted model produces different high-risk flags or error rates across groups, those differences arise from the observed data structure rather than from a direct group effect in the outcome equation.

The five applications are designed to illustrate this mechanism in settings familiar to insurance practice. In life insurance, the simulation uses an accelerated-underwriting setting in which an observed credit-related underwriting score may be measured with group-correlated distortion. In auto insurance, the simulation separates unobserved driving behavior from recorded violations that may reflect territory-specific monitoring or enforcement. In credit insurance and mortgage insurance, observed credit scores can differ from unobserved credit quality. In health insurance, observed morbidity measures can differ from unobserved health status because administrative data may reflect utilization, coding, access, and documentation patterns.

The study compares three model classes. The first is a generalized linear model (GLM), used as a familiar actuarial benchmark. The second is a standard neural network (NN), used as a flexible machine-learning benchmark. The third is an adversarial neural network (ADV-NN), which is trained to predict the insurance outcome while reducing dependence on a selected observed indirectly associated variable. The adversarial model is not intended to remove all information associated with risk. Its purpose is to preserve useful predictive information while discouraging the model from relying too strongly on a pathway suspected of carrying distorted or group-related signals (Ganin et al., 2016; Zhang, Lemoine, and Mitchell, 2018).

The fitted models are evaluated using both predictive and audit measures. Predictive performance is measured by how well a model ranks higher-risk cases above lower-risk cases. Group-comparison metrics are then calculated after model fitting, using held-out data, to examine whether high-risk selection rates and error rates differ across groups. These metrics are used as model-audit diagnostics. They summarize how the fitted model behaves after training and help identify whether observed distortions or indirect associations appear to affect model outputs.

The simulations pay particular attention to fixed high-risk operating points, such as the top 1%, 2%, or 5% of predicted risks. This reflects the way many insurance models are used operationally. A model may not trigger action for every applicant or policyholder. Instead, it may identify a small group for manual review, additional underwriting, extra documentation, more restrictive terms, surcharge, or declination. In those settings, differences in the upper tail of the score distribution can matter more than differences in the middle of the distribution.

The case studies are controlled, mechanism-based simulations. They do not estimate the experience of any particular insurer, product, jurisdiction, or market. They show how indirect associations and biased measurement can affect model behavior under controlled conditions, and they evaluate whether mitigation can improve alignment between fitted-model outputs and unobserved risk.

Copyright © 2026 Society of Actuaries Research Institute

Section 3 Applications

This section presents five stylized synthetic applications before turning to the more detailed methodological discussion. Each application follows the same structure: it first describes the embedded statistical-bias mechanism, then summarizes the simulated data patterns, and finally reports model-evaluation results. For orientation, the generalized linear model (GLM) is used as the actuarial benchmark, the standard neural network (NN) as the flexible machine-learning benchmark, and the adversarial neural network (ADV-NN) as the mitigation model. The area under the receiver operating characteristic curve (ROC AUC) summarizes ranking performance. A demographic-parity ratio closer to one indicates more similar high-risk selection rates across audit groups, while an absolute true-positive-rate (TPR) gap closer to zero indicates more similar identification rates among cases that experience the adverse outcome. A 2% high-risk operating point means that the 2% of held-out cases with the highest predicted risks are selected for the relevant operational action. The mathematical definitions, model specifications, and tuning workflow are provided in the methodological section that follows and in the appendices.

The adversarial penalty λ is selected within the model-development sample using outcome-stratified cross-validation. Candidate values whose mean validation ROC AUC is within 0.005 of the highest mean cross-validated ROC AUC among the candidate ADV-NN specifications are retained. Among those candidates, the model with the largest validation adversary loss is selected; under the common validation procedure, a larger adversary loss indicates that the designated indirectly associated variable is more difficult to recover from the learned representation. The audit label is not used in this selection. Because the frontier figures are evaluated on the held-out sample, the validation-selected point need not be the highest or leftmost point in the displayed frontier.

3.1 LIFE INSURANCE

3.1.1 EMBEDDED STATISTICAL BIAS MECHANISM

This hypothetical life-insurance case study uses accelerated underwriting as an illustration. In accelerated underwriting, insurers rely more heavily on fast, data-driven screening tools and less on full traditional underwriting, such as extensive medical exams or laboratory evidence. As a result, observable variables drawn from external data sources can play a larger role in underwriting decisions. In this simulation, one such variable is a hypothetical credit-related underwriting score.

The outcome variable is a binary indicator:

$$Y \in \{0,1\},$$

which records whether the adverse event of interest, namely a covered death event, occurs over the stated horizon. The predictive models are trained on the observed input vector:

$$X = (T, Z, \text{Age}, \text{BMI}, S, C, R),$$

where T is an observed territory or neighborhood indicator, Z is the observed credit-related underwriting score available to the insurer, Age and BMI are continuous underwriting covariates, S is a smoking indicator, C is a chronic-condition indicator, and R is a risky-occupation indicator.

The group variable retained for ex post auditing in this case study is the audit label:

$$A \in \{a, b\},$$

Copyright © 2026 Society of Actuaries Research Institute

which is not used in model training and is retained only for ex post evaluation. The observed indirectly associated variable is T, because it is statistically related to A and is part of the insurer's ordinary observed data. In addition, the simulation distinguishes between:

- the unobserved credit-related underwriting score Z^*, which is risk-relevant, group-neutral, and unobserved by the insurer, and
- the observed score Z, which is the quantity actually used in model training.

The key distortion mechanism is that, in an accelerated-underwriting environment, a credit-related variable may carry useful underwriting information but may also be measured imperfectly for some groups. In particular, some disadvantaged groups may have observed values of Z that are systematically lower than their unobserved risk-relevant values Z^*. This can arise because observed credit-related variables may reflect broader socio-economic obstacles, thinner files, differential access to financial services, or other non-risk-related frictions. Because a lower observed score implies higher predicted risk in this application, the shift can increase the likelihood of additional review, less favorable terms, a surcharge, or declination. The point of the simulation is not to model every source of that gap, but to represent the statistical consequence of such a gap in a transparent way.

The measurement process is represented as

$$Z = Z^* - \theta \cdot 1_{\{A=a\}} + \varepsilon, \ with\ \theta > 0.$$

where θ is the magnitude of the downward shift for Group a and ε is random noise. Because higher values of Z correspond to lower underlying risk, this shift makes Group a appear riskier than an otherwise identical policyholder in Group b even when unobserved risk is the same.

3.1.2 DATA GENERATION PROCESS

The synthetic dataset is designed as a controlled accelerated-underwriting environment in which the audit label does not enter the outcome equation directly, while observed underwriting variables can still generate group differences through indirect association and biased measurement.

At a high level, the simulation proceeds as follows:

- generate group membership:
$$A \in \{a, b\};$$
- generate an observed territory or neighborhood indicator, T, with $T = 1$ more prevalent in Group a;
- generate an unobserved credit-related underwriting score Z^*, independently of A;
- construct the observed score Z by shifting Z^*downward for Group a and adding random noise;
- generate additional underwriting covariates, including age, BMI, smoking status, chronic-condition status, and risky occupation;
- generate the binary outcome Y from a logistic risk model driven by $Z^*$and the other covariates, with nonlinear terms and interactions;
- calibrate the intercept so that the simulated event rate matches the desired target level.

Copyright © 2026 Society of Actuaries Research Institute

The interpretation is that accelerated underwriting uses a limited but scalable set of observed variables to make underwriting decisions. In that setting, an insurer may rely on a credit-related underwriting score Z because it is fast, readily available, and operationally useful. However, if Z is a distorted measure of the unobserved signal Z^* for one subgroup, then a predictive model trained on Z may inherit that distortion. The simulation therefore separates the role of the **unobserved signal** from the **observed underwriting variable** so that the effect of measurement bias can be studied directly.

The simulation specifies the underlying risk process before generating observations. Because the true mechanism is known, model predictions can be evaluated against underlying risk rather than observed measurements alone.

An important point is that the outcome equation is part of the simulation design. In actual insurance practice, mortality outcomes are observed first, and the analyst estimates a model from data. Here, the process is reversed: the data-generating mechanism is specified first and then sample outcomes from that known structure. Because the data-generating mechanism is specified explicitly, the underlying risk process is known by construction. The fitted models can therefore be evaluated on whether they recover the intended risk signal or respond to distortions in the observed inputs.

Group a comprises about 35% of the portfolio. The observed territory indicator $T = 1$ is intentionally much more common in Group a than in Group b (about 60% versus 20%), creating a clear indirect pathway from group membership to the observed data. In the simulated sample, the adverse-event rate is approximately 15%, or about 15,000 deaths out of 100,000 policies. The unobserved score Z^* is group-neutral, while the observed score Z is shifted downward for Group a. Details regarding the simulation process are provided in Appendix C.

3.1.3 EXPLORATORY DATA PATTERNS IN THE SIMULATION

The following figures illustrate how the simulated data reflect the intended bias mechanism before model fitting begins. Key distributions are summarized to clarify both the indirect association and the overall class balance in the simulated sample. The portfolio contains 100,000 policies, with 34,805 in Group a and 65,195 in Group b. The indirectly associated territory indicator equals one for 34,102 policies and is substantially more common in Group a than in Group b, approximately 60% versus 20%. The simulated sample contains 14,929 adverse events, corresponding to an event rate of approximately 15%.

Figure 1 and Figure 2 illustrate the distributions of age and BMI. Figure 3 contrasts the unobserved, group-neutral score Z^* with the observed score Z that includes measurement noise and a group-dependent downward shift. That contrast makes the measurement-bias channel visible in the data seen by the model.

Copyright © 2026 Society of Actuaries Research Institute

Figure 1
DISTRIBUTION OF AGE

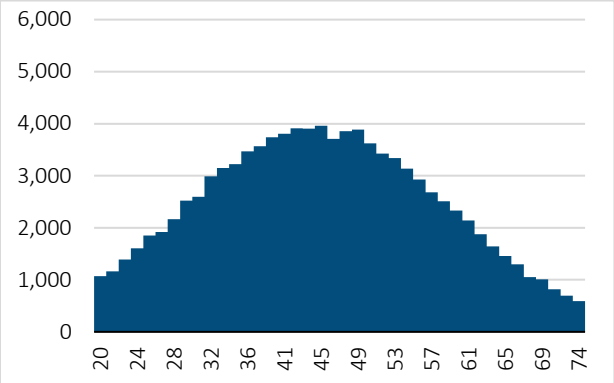


Figure 2
DISTRIBUTION OF BMI

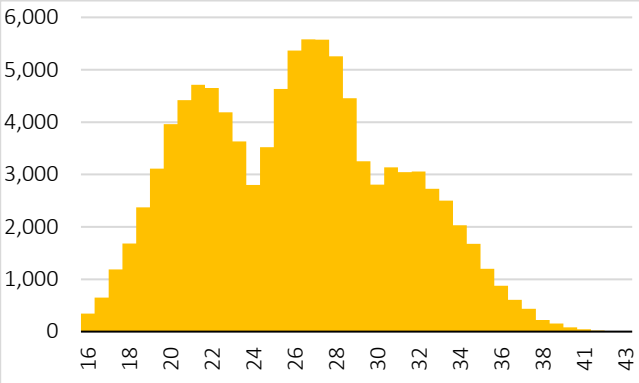
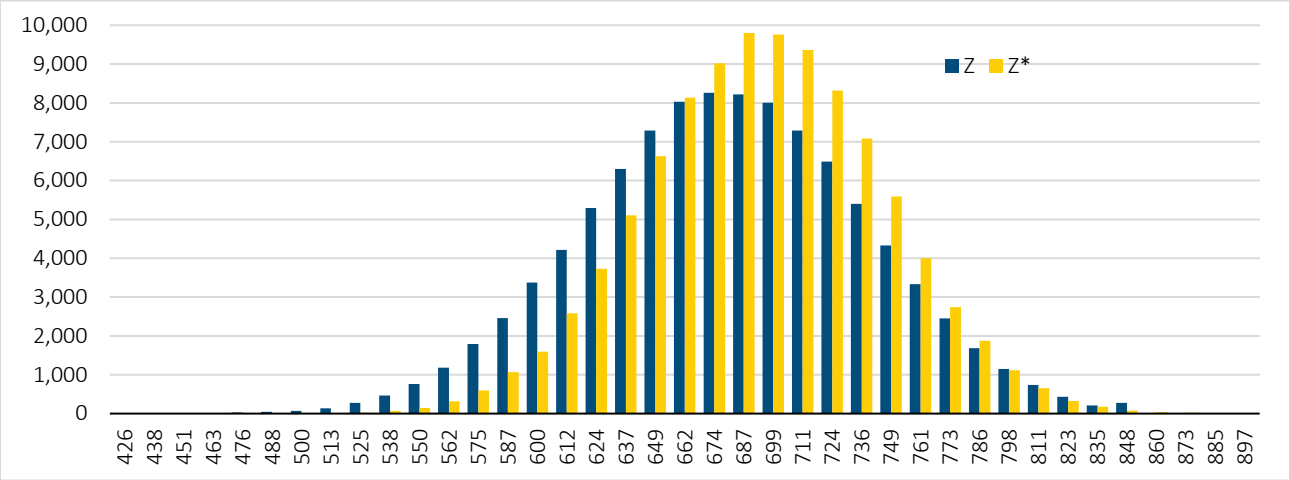


Figure 3
OBSERVED SCORE Z AND UNOBSERVED SCORE Z*



Copyright © 2026 Society of Actuaries Research Institute

3.1.4 MODEL EVALUATION RESULTS

Throughout this subsection, GLM denotes the generalized linear model benchmark, NN denotes the standard neural-network benchmark, and ADV-NN denotes the adversarial neural network. All reported metrics are computed on the held-out evaluation sample. In the frontier figure, higher ROC AUC indicates better ranking performance, while an absolute TPR gap closer to zero indicates more similar true-positive rates across audit groups. The 2% high-risk operating point selects the 2% of cases with the highest predicted risks for a potentially unfavorable action, such as additional review or more restrictive terms. In the 1%, 2%, and 5% figures, a demographic-parity ratio closer to one and an absolute TPR gap closer to zero indicate more similar group outcomes. The highlighted ADV-NN was selected using development-sample validation; the audit label was not used in training, adversarial training, λ selection, or refitting.

Results are intended to be interpreted as illustrative examples of how mitigation changes audited disparities in a controlled life-underwriting setting. Figure 4 and Figure 5 compare the group-specific predicted-risk distributions for the GLM and the selected ADV-NN. Most predicted probabilities remain in a relatively low-risk region, while the operationally relevant differences arise in the right tail because high-risk actions are applied only to a small share of cases. Figure 6 and Figure 7 magnify the right tails of those distributions.

Copyright © 2026 Society of Actuaries Research Institute

Figure 4
PREDICTED RISK DISTRIBUTIONS FOR THE REFERENCE GROUP

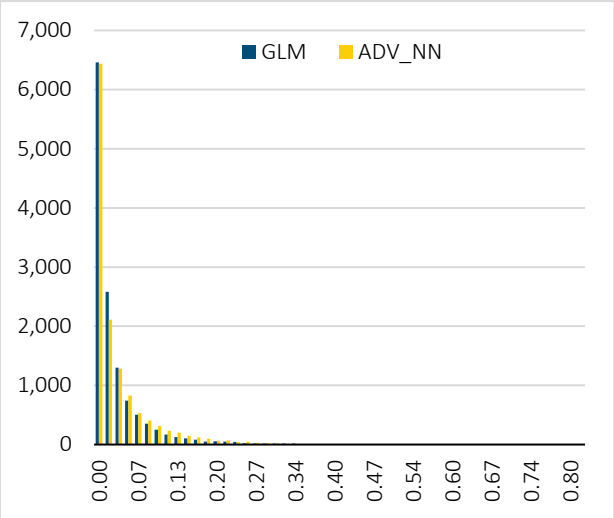


Figure 5
PREDICTED RISK DISTRIBUTIONS FOR THE DISADVANTAGED GROUP

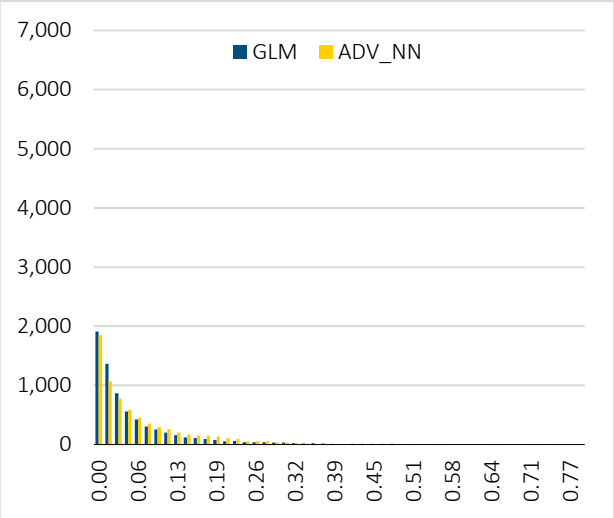


Figure 6
RIGHT TAIL OF PREDICTED RISK DISTRIBUTIONS FOR THE REFERENCE GROUP

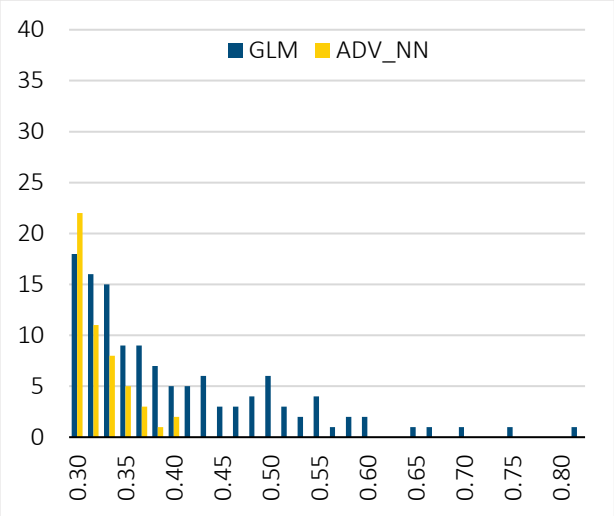


Figure 7
RIGHT TAIL OF PREDICTED RISK DISTRIBUTIONS FOR THE DISADVANTAGED GROUP

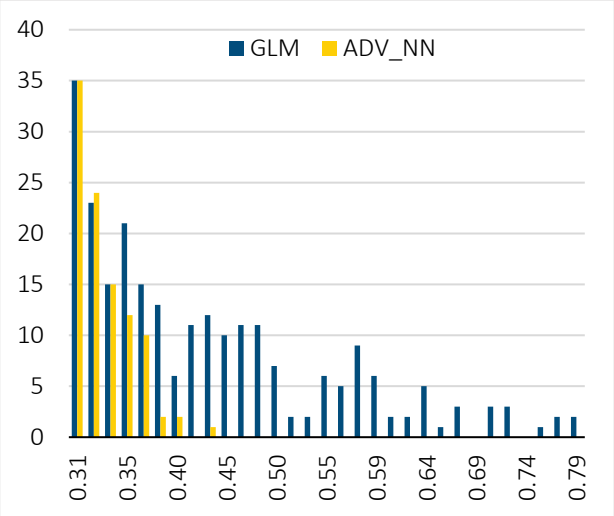
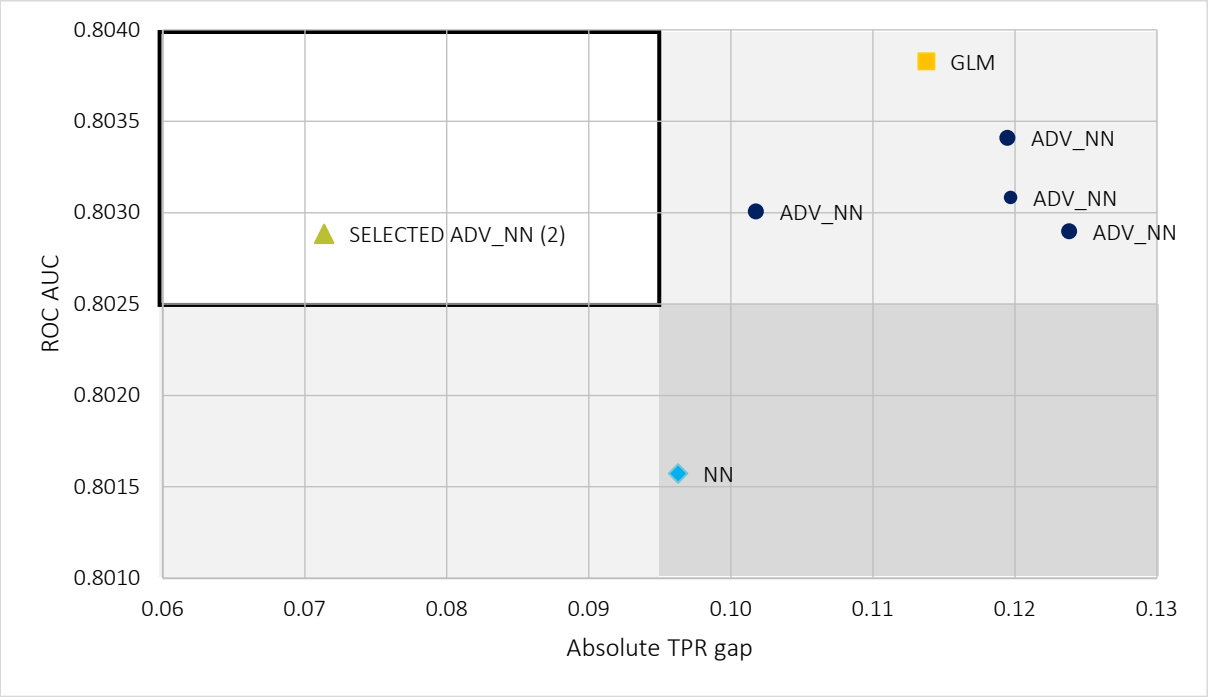


Figure 8 is the main summary chart for the life case study. Each dark-blue point represents one ADV-NN fit at a prespecified value of λ. The vertical axis is ROC AUC, so higher is better; the horizontal axis is the ex post true positive rate (TPR) gap at the 2% high-risk operating point, so smaller is better. Points nearer the upper-left therefore represent a better fairness–accuracy trade-off, and the highlighted selected point is the specification selected by the validation rule.

Copyright © 2026 Society of Actuaries Research Institute

Figure 8

FAIRNESS–ACCURACY FRONTIER FROM AN ADVERSARIAL-PENALTY SWEEP AT THE 2% HIGH-RISK OPERATING POINT



In this simulation, the GLM attains an ROC AUC of 0.8038 with a 2% TPR gap of about 0.114. The standard neural network (NN) lowers the gap to about 0.096 while also lowering AUC to 0.8016. The selected ADV-NN moves further left to a TPR gap of about 0.071, while maintaining an AUC of 0.8028. In this case, the adversarial specification therefore produces a meaningful reduction in the ex post audit gap at only a very small cost in ranking performance.

Figure 9 extends the comparison across 1%, 2%, and 5% high-risk action rates. The selected ADV-NN keeps the demographic-parity ratio closer to one than either benchmark model at each operating point. Figure 10 shows the same pattern for TPR gaps: the selected ADV-NN reduces the gap materially relative to the GLM and also improves on the standard NN, especially at the 2% operating point. These comparisons are ex post audit summaries of fitted-model outputs, not quantities used directly in model training.

To move from a single operating point (2%) to a broader operational perspective, Figure 9 evaluates demographic parity across multiple high-risk selection thresholds. It shows how demographic parity ratios vary as the high-risk selection threshold changes. Across all operating points, the adversarial neural network yields ratios closer to one than the GLM and standard neural network, indicating more similar high-risk selection rates across groups. Disparities are largest at the most restrictive cutoff and decline as the high-risk rate increases.

Complementing the selection-rate comparison in Figure 9, Figure 10 focuses on the alignment of true positive rates across groups at the same operating points. It reports the TPR gap as a function of the high-risk selection rate. For all models, the gap increases as the operating point becomes less restrictive, reflecting greater exposure of group differences at higher coverage levels. Across operating points, the adversarial neural network consistently exhibits absolute TPR gaps that are closer to zero than the GLM and standard neural network, indicating improved alignment of true positive rates across groups.

Copyright © 2026 Society of Actuaries Research Institute

Figure 9
DEMOGRAPHIC PARITY RATIO OF HIGH-RISK SELECTION RATES ACROSS OPERATING POINTS (1%, 2%, 5%)

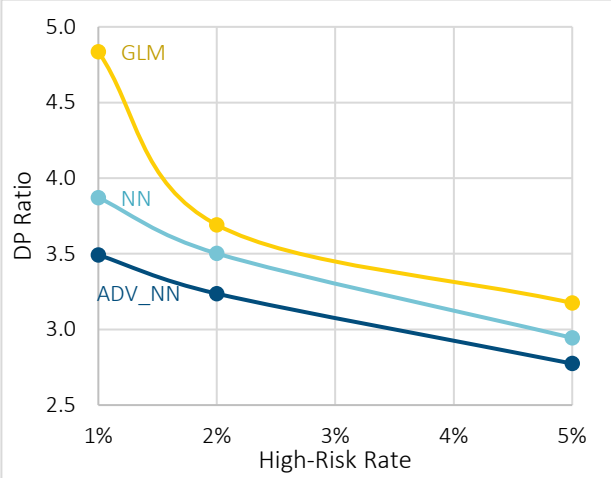
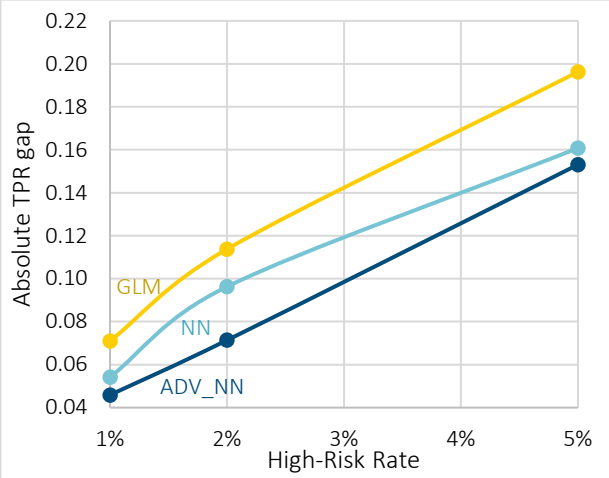


Figure 10
TPR GAP ACROSS OPERATING POINTS (1%, 2%, 5%)



Key Takeaway

In this synthetic life insurance setting, adversarial learning reduced audited disparities substantially while maintaining nearly the same predictive ranking performance.

3.2 AUTO INSURANCE

3.2.1 EMBEDDED STATISTICAL BIAS MECHANISM

This hypothetical auto-insurance case study is intended to be interpreted as a binary underwriting or heightened-review use case illustration. The question is whether an observed behavioral variable may reflect not only underlying driving risk, but also differences in how behavior is monitored and recorded across territories. Two otherwise similar drivers can therefore have the same unobserved violation tendency but different recorded violation counts simply because they are exposed to different surveillance environments.

The outcome variable is a binary next-period claim indicator Y. The predictive models are trained on observed underwriting inputs that include age, annual mileage, the transformed mileage measure $M_h = log(1 + M / 12{,}000)$, recorded moving violations V, a prior at-fault claim flag P_{prior}, and territory T. The group variable retained for ex post audit is A, which is excluded from model training. Territory is the observed indirectly associated variable because it is correlated with the audit label and is part of ordinary insurer data.

$$A \in \{a, b\},$$

A is not used in model fitting and is retained only for ex post audit. The key distortion mechanism is that recorded violations combine true behavior with a territory-specific surveillance uplift. In the simulation, territory does not directly enter the unobserved-risk outcome equation. Instead, it indirectly acts as a recording-environment variable:

Copyright © 2026 Society of Actuaries Research Institute

drivers in $T = 1$ accumulate additional observed violations even when their unobserved violation count V^* is the same.

$$V = V^* + \Delta, \qquad \Delta \sim \text{Poisson}(2.00T).$$

Formally, the observed count is written as $V = V^* + \Delta$, where Δ is a Poisson uplift that appears only in $T = 1$. Because T is correlated with A, the recorded count V can indirectly transmit group information into the fitted model even though the claim equation itself does not use A and does not contain a direct territory effect.

This construction therefore creates two related channels. First, the observed indirectly associated T is statistically related to the audit group. Second, the observed count V is a biased measurement of the unobserved behavioral quantity V^* because it includes territory-linked surveillance intensity. Together, these channels let the fitted model learn subgroup-related distortion even though the audit label itself is excluded from training.

3.2.2 DATA GENERATION PROCESS

The synthetic dataset is designed as a stylized auto-insurance underwriting environment with a binary next-period claim indicator. The goal is not to reproduce any one insurer's proprietary portfolio, but to construct a controlled setting in which true claim risk is driven by familiar exposure and behavior variables while group differences can still emerge through indirect association and measurement distortion.

At a high level, the simulation proceeds as follows:

- generate group membership A, where Group a is the disadvantaged group;
- generate territory T, with $T = 1$ more prevalent in Group a than in Group b;
- generate age and annual mileage, define the youthful-driver indicator, and transform mileage as $M_h = log(1 + M / 12{,}000)$;
- generate unobserved moving violations V^* from a Poisson model driven by transformed mileage and youthful-driver status (but not by territory);
- construct recorded violations V by adding a territory-dependent surveillance uplift to V^*;
- generate a prior at-fault claim flag from a logistic model driven by unobserved violations and youthful-driver status;
- generate the binary claim outcome Y from a logistic model driven by transformed mileage, unobserved violations, youthful-driver status, prior claim history, and a mileage-by-unobserved-violations interaction; and
- calibrate the intercept so that the simulated claim rate matches the target level.

This design is intended to be read as a controlled simulation exercise rather than as a claim that territory is never a legitimate underwriting variable. The point is narrower: if one observed behavioral variable is distorted by the monitoring environment, a flexible model can treat that distortion as if it were a genuine risk signal.

The simulation specifies the underlying risk process before generating observations. Because the true mechanism is known, model predictions can be evaluated against underlying risk rather than observed measurements alone.

An important point is that the outcome equation is part of the simulation design. In an empirical pricing application, claim outcomes are observed first and the analyst estimates the predictive relationship from data. Here, the process is reversed: the data-generating mechanism is specified first and then sample outcomes from that known structure. Because the data-

Copyright © 2026 Society of Actuaries Research Institute

generating mechanism is specified explicitly, the underlying risk process is known by construction. The fitted models can therefore be evaluated on whether they recover the intended risk signal or respond to distortions in the observed inputs.

Group a comprises 35% of the portfolio, and $T = 1$ is much more prevalent in Group a than in Group b (about 65% versus 15%). Annual mileage is centered near 12,000 miles, so the transformed mileage variable still corresponds to realistic annual exposure. The overall claim rate is approximately 15%, and because territory does not enter the unobserved claim equation directly, any small residual differences in raw claim rates by group or territory are sampling noise rather than built-in structural differences. Details regarding the simulation process are provided in Appendix D.

3.2.3 EXPLORATORY DATA PATTERNS IN THE SIMULATION

The following figures illustrate how the simulated data reflect the intended bias mechanism before model fitting begins. The sample contains 100,000 policies, with 35,094 in Group a and 64,906 in Group b. The indirectly associated territory indicator equals one for 32,680 policies. $T = 1$ occurs for approximately 65% of Group a and 15% of Group b. The sample contains 14,901 claims and 85,099 non-claims.

Figure 11 and Figure 12 display the simulated distributions of age and transformed mileage. On the original scale, annual mileage is centered near 12,000 miles, with the middle 50% of the portfolio roughly between 8,300 and 14,900 miles. Figure 13 contrasts the recorded violation count available to the model with the unobserved count used in the claim equation. In the simulated sample, the mean unobserved violation count is approximately 0.72 in both territories, whereas the mean recorded count rises to approximately 2.70 in the higher-surveillance territory, compared with approximately 0.72 in the other territory.

Copyright © 2026 Society of Actuaries Research Institute

Figure 11
DISTRIBUTION OF AGE

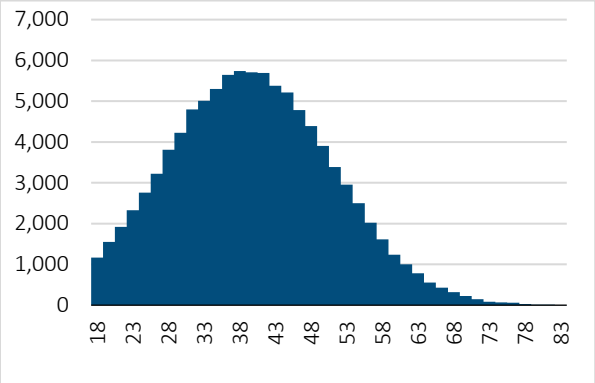


Figure 12
DISTRIBUTION OF LOG(MILEAGE)

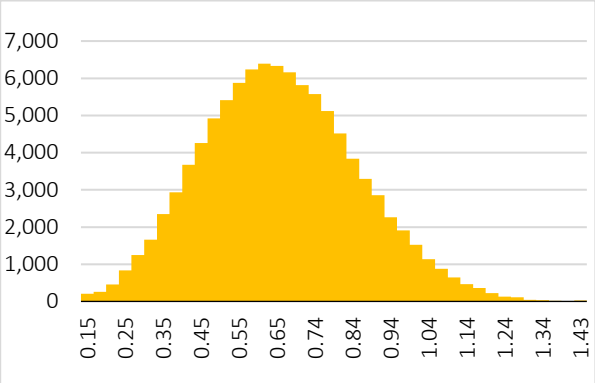
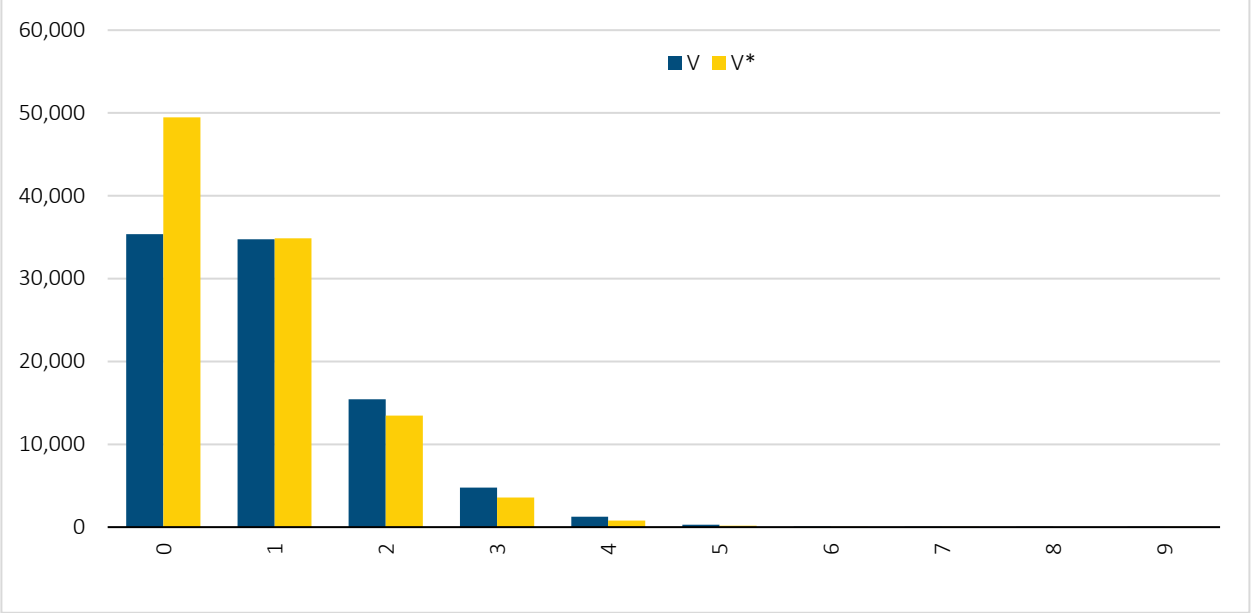


Figure 13
RECORDED VIOLATION COUNT V AND UNOBSERVED VIOLATION COUNT V*



3.2.4 MODEL EVALUATION RESULTS

Throughout this subsection, GLM denotes the generalized linear model benchmark, NN denotes the standard neural-network benchmark, and ADV-NN denotes the adversarial neural network. All reported metrics are computed on the held-out evaluation sample. In the frontier figure, higher ROC AUC indicates better ranking performance, while an absolute TPR gap closer to zero indicates more similar true-positive rates across audit groups. The 2% high-risk operating point selects the 2% of cases with the highest predicted risks for a potentially unfavorable action, such as additional review or more restrictive terms. In the 1%, 2%, and 5% figures, a demographic-parity ratio closer to one and an absolute TPR gap closer to zero indicate more similar group outcomes. The highlighted ADV-NN was selected using development-sample validation; the audit label was not used in training, adversarial training, λ selection, or refitting.

Results are intended to be interpreted as illustrative examples of how mitigation changes audited disparities in a binary underwriting ranking task. Figure 14 through Figure 17 compare predicted claim probabilities across groups for the GLM baseline and the adversarial model. Most policies remain in a low-risk region, but the operational

Copyright © 2026 Society of Actuaries Research Institute

differences arise in the right tail because only a small share of policies is flagged for high-risk action. Relative to the GLM, the adversarial model pulls the two group-specific tails closer together.

Figure 16 and Figure 17 magnify the right tail of the predicted-risk distributions in Figure 14 and Figure 15. This tail region is the part of the portfolio most relevant when the model is used to rank a small fraction of cases for heightened underwriting attention. Compared with the GLM baseline, the adversarial model reduces the separation between groups in that extreme-risk region.

Figure 14
PREDICTED RISK DISTRIBUTIONS FOR THE REFERENCE GROUP

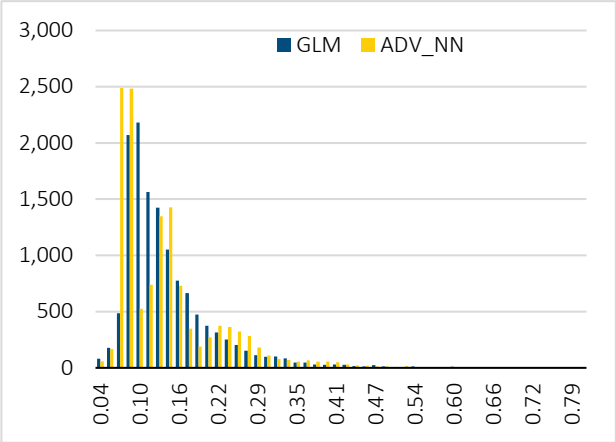


Figure 15
PREDICTED RISK DISTRIBUTIONS FOR THE DISADVANTAGED GROUP

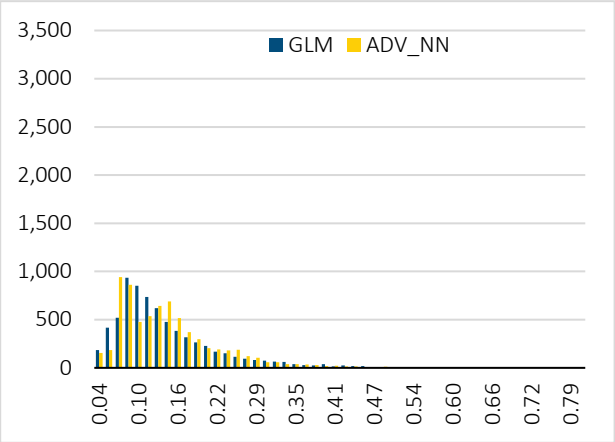


Figure 16
RIGHT TAIL OF PREDICTED RISK DISTRIBUTIONS FOR THE REFERENCE GROUP

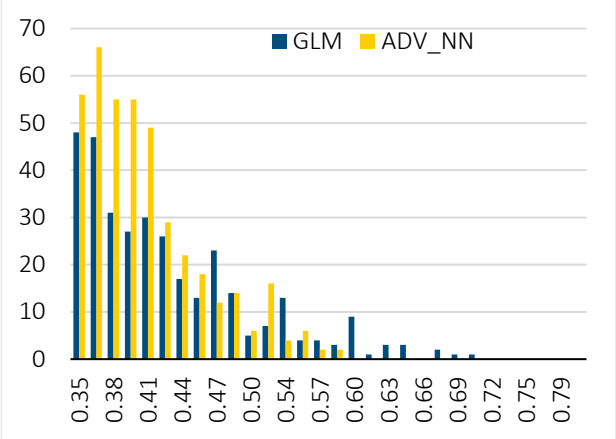
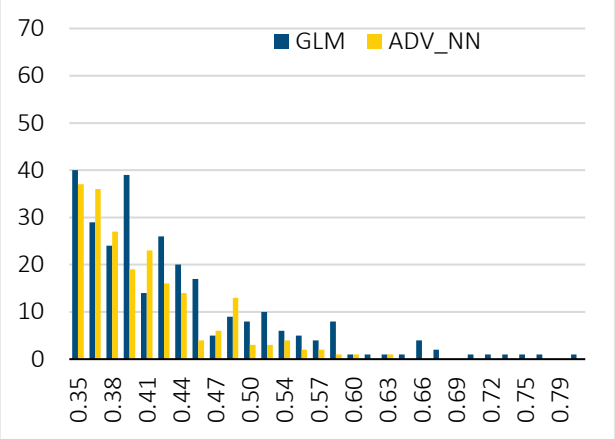


Figure 17
RIGHT TAIL OF PREDICTED RISK DISTRIBUTIONS FOR THE DISADVANTAGED GROUP



After comparing the group-specific distributions in the extreme-risk region, Figure 18 summarizes the fairness–accuracy trade-off across values of the adversarial penalty parameter λ, while Figure 19 and Figure 20 examine how the selected model behaves across different high-risk operating points.

Copyright © 2026 Society of Actuaries Research Institute

Figure 18

FAIRNESS–ACCURACY FRONTIER FROM AN ADVERSARIAL-PENALTY SWEEP AT THE 2% HIGH-RISK OPERATING POINT

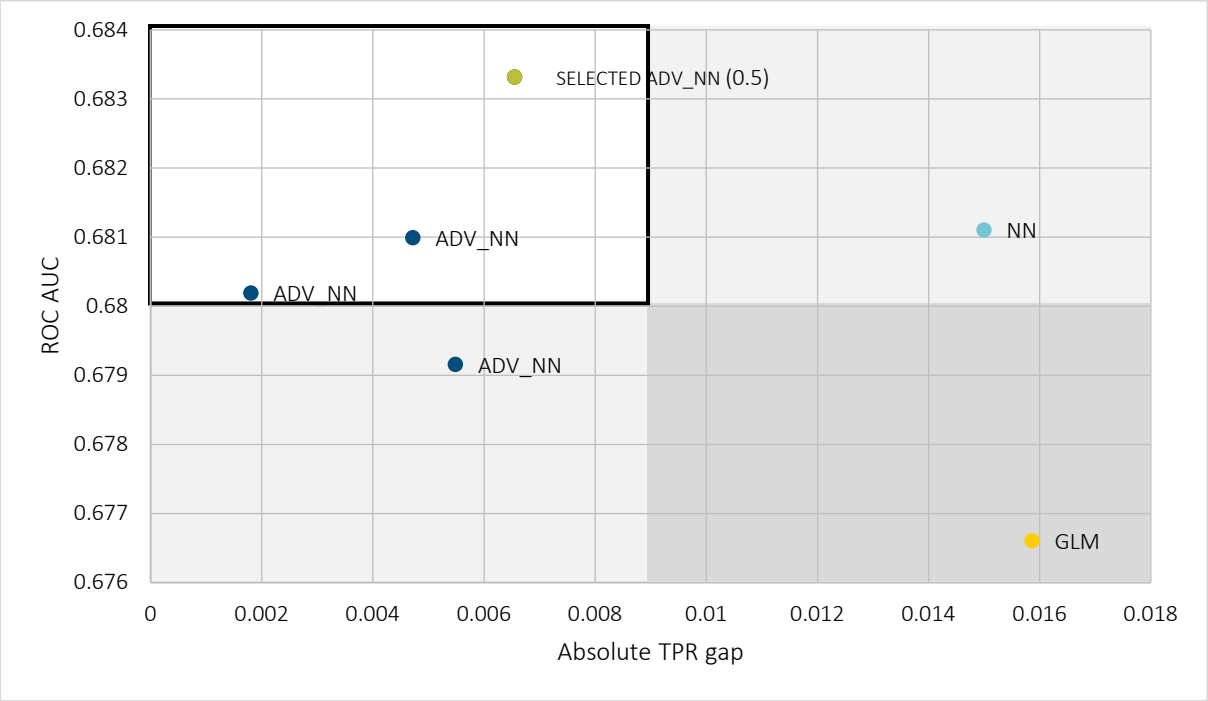


Figure 18 shows that the GLM attains a held-out ROC AUC of 0.6766 with a 2% TPR gap of 0.0159, while the standard NN attains an ROC AUC of 0.6811 with a gap of 0.0150. The validation-selected ADV-NN attains an ROC AUC of 0.6833 and reduces the gap to 0.0065. In this application, the selected ADV-NN therefore improves both predictive ranking and the reported TPR-gap diagnostic relative to the two benchmark models.

Copyright © 2026 Society of Actuaries Research Institute

Figure 19
DEMOGRAPHIC PARITY RATIO OF HIGH-RISK SELECTION RATES ACROSS OPERATING POINTS (1%, 2%, 5%)

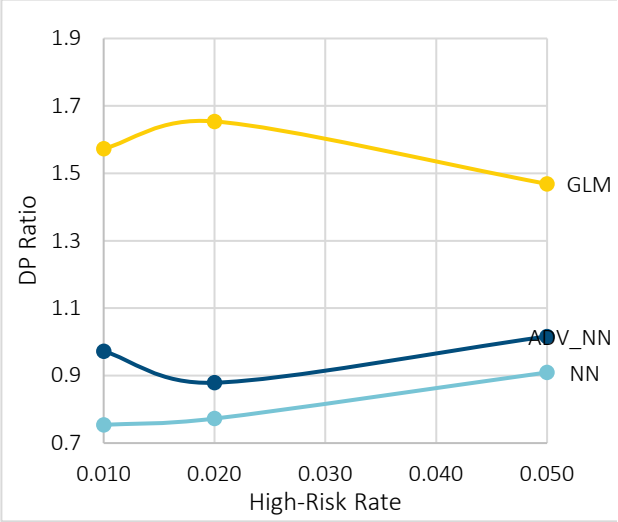


Figure 20
TPR GAP ACROSS OPERATING POINTS (1%, 2%, 5%)

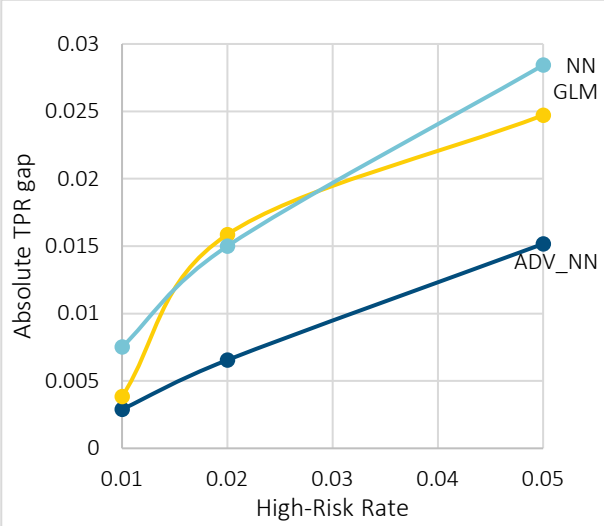


Figure 19 extends the comparison across 1%, 2%, and 5% high-risk action rates; Figure 20 reports the corresponding TPR gaps. At the 2% operating point, the demographic-parity ratio is 1.654 for the GLM, 0.773 for the NN, and 0.879 for the selected ADV-NN. Because values closer to one indicate more similar selection rates, the selected ADV-NN is closer to parity than either benchmark. Its 2% TPR gap is 0.0065, compared with 0.0159 for the GLM and 0.0150 for the NN.

3.3 CREDIT INSURANCE

3.3.1 EMBEDDED STATISTICAL BIAS MECHANISM

The hypothetical credit-insurance case study explores a common feature of consumer-credit data: observed credit scores may reflect not only underlying repayment risk, but also differences in how easily individuals can establish, maintain, and document formal credit histories. In practice, two applicants with similar unobserved repayment capacity may nevertheless have different observed credit scores because one applicant faces greater obstacles in building a conventional credit file. Such obstacles may include thinner credit histories, shorter relationships with mainstream lenders, more fragmented reporting records, or reduced access to standard credit products. As a result, an observed credit score can understate the credit quality of one group even when underlying risk is the same.

The outcome variable in this case study is a binary indicator Y that records whether default occurs over a one-year horizon. The predictive models are trained on the observed input vector $X = (\mathrm{S}, \mathrm{D}, \mathrm{L}, \mathrm{Z})$, where S is the observed credit score available to the insurer, D is debt-to-income, L is the recent delinquency count, and Z is a ZIP-based neighborhood indicator. The audit label in this case study is $A \in \{a, b\}$, where Group a is the disadvantaged group and Group b is the reference group. This audit label is not used in model training and is retained only for ex post evaluation. The observed indirectly associated variable is Z, because it is statistically related to group membership and is part of the insurer's ordinary observed data. In addition, the simulation distinguishes between the unobserved credit score S^*, which represents underlying credit strength and is group-neutral by construction, and the observed credit score S, which is the quantity actually used in model training.

Copyright © 2026 Society of Actuaries Research Institute

The key distortion mechanism is that the observed credit score may be a biased measurement of the unobserved score for one group. In particular, Group a may have observed values of S that are systematically lower than their unobserved values of S^*, even when true repayment risk is the same. The measurement process is represented as

$$S = S^* - \theta 1_{\{A=a\}} + \varepsilon_S,$$

where $\theta > 0$ is the magnitude of the downward shift for Group a and ε_S is measurement noise. Because higher values of S^*correspond to lower underlying risk, this shift makes Group a appear riskier than an otherwise identical applicant in Group b, even when unobserved credit quality is the same.

This construction creates two distinct but related channels. First, the observed indirectly associated Z is statistically related to the audit group. Second, the observed credit score S is a biased measurement of the unobserved, risk-relevant score S^*. Together, these channels allow the model to learn subgroup-related distortion even though the audit label itself is excluded from training.

3.3.2 DATA GENERATION PROCESS

The synthetic credit-insurance dataset is designed as a stylized underwriting environment with a binary one-year default outcome. The goal is not to reproduce any one insurer's proprietary portfolio, but to construct a controlled setting in which default risk is group-neutral at the unobserved level while observed underwriting variables can still generate group differences through indirect association and biased measurement.

At a high level, the simulation proceeds as follows:

- generate group membership, with Group a representing the disadvantaged group and Group b the reference group;
- generate a ZIP-based neighborhood indicator Z, with $Z = 1$ more prevalent in Group a than in Group b;
- generate an unobserved, group-neutral credit score $S^* \sim N(700,60^2)$;
- construct the observed credit score S by shifting S^*downward for Group a and adding random noise;
- generate debt-to-income D from a lognormal distribution and clip it to a plausible range;
- generate recent delinquency count L from a Poisson distribution;
- standardize $S^*$and generate the binary default outcome Y from a logistic model driven by unobserved credit strength, debt burden, and delinquency history; and
- calibrate the intercept so that the overall default rate matches the target rate used in the reported simulation.

The interpretation is that credit-insurance decisions often rely on bureau-based or score-based measures because they are scalable, standardized, and operationally convenient. However, if the observed score S is a distorted measure of the unobserved signal S^* for one subgroup, then a predictive model trained on S may inherit that distortion. The simulation therefore separates the role of the unobserved credit signal from the observed score so that the effect of measurement bias can be studied directly.

The simulation specifies the underlying risk process before generating observations. Because the true mechanism is known, model predictions can be evaluated against underlying risk rather than observed measurements alone.

An important point is that the outcome equation is part of the simulation design. In actual underwriting practice, default outcomes are observed first, and the analyst estimates a predictive model from data. Here, the process is reversed: the data-generating mechanism is specified first and then sample outcomes from that known structure. Because the data-generating mechanism is specified explicitly, the

Copyright © 2026 Society of Actuaries Research Institute

underlying risk process is known by construction. The fitted models can therefore be evaluated on whether they recover the intended risk signal or respond to distortions in the observed inputs.

In the reported simulation, Group a comprises roughly one-third of the portfolio, and the ZIP-based variable is substantially more common in Group a than in Group b. The unobserved credit score $S^*$is group-neutral, whereas the observed score S is shifted downward for Group a. Debt-to-income and delinquency count are generated independently of group and provide additional variation in default risk. The binary outcome Y is then generated from unobserved credit strength and these financial covariates, not from the audit label itself. Additional details regarding the simulation can be found in Appendix E.

3.3.3 EXPLORATORY DATA PATTERNS IN THE SIMULATION

The simulated sample contains 100,000 applicants, including 35,094 in the disadvantaged group and 64,906 in the reference group. The ZIP-based indirectly associated variable equals one for 39,251 applicants and is substantially more common in the disadvantaged group, approximately 65% versus 25%. The sample contains 11,979 defaults, corresponding to a default rate of approximately 12%.

Figure 21
DISTRIBUTION OF DEBT-TO-INCOME RATIO

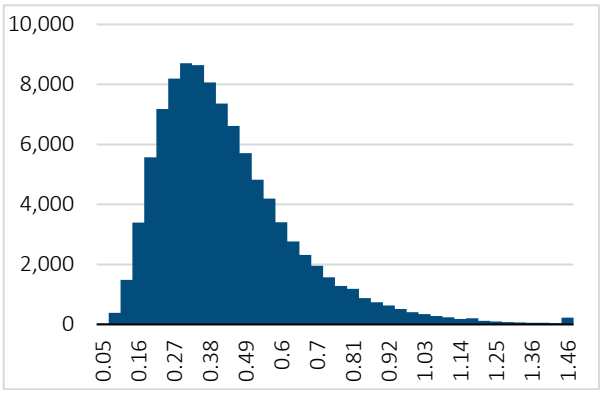


Figure 22
DISTRIBUTION OF DELINQUENCY COUNT

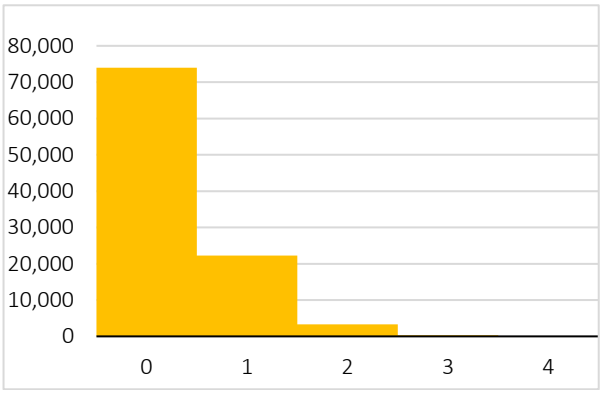


Figure 23
OBSERVED CREDIT SCORE S AND UNOBSERVED CREDIT SCORE S*

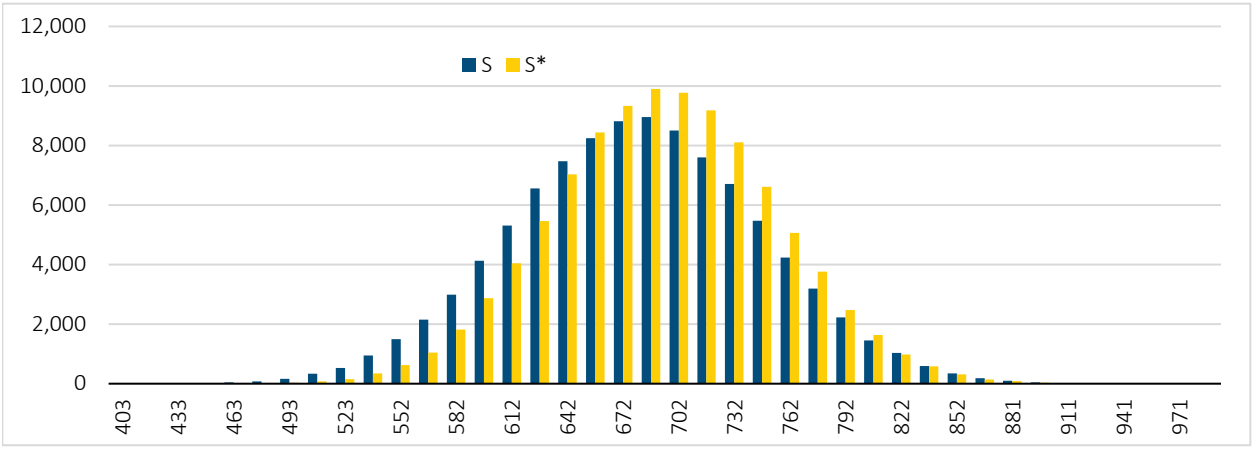


Figure 21 displays the debt-to-income distribution, which is right-skewed in the usual way for leverage measures. Figure 22 shows the delinquency-count distribution, where most applicants have no recent delinquencies and only a small fraction have multiple events. Figure 23 contrasts the observed credit score S with the unobserved, group-

Copyright © 2026 Society of Actuaries Research Institute

neutral score $S^{\star}$. That contrast is the central object of interest: the model observes the biased score S, while the default outcome is generated from $S^{\star}$.

3.3.4 MODEL EVALUATION RESULTS

Throughout this subsection, GLM denotes the generalized linear model benchmark, NN denotes the standard neural-network benchmark, and ADV-NN denotes the adversarial neural network. All reported metrics are computed on the held-out evaluation sample. In the frontier figure, higher ROC AUC indicates better ranking performance, while an absolute TPR gap closer to zero indicates more similar true-positive rates across audit groups. The 2% high-risk operating point selects the 2% of cases with the highest predicted risks for a potentially unfavorable action, such as additional review or more restrictive terms. In the 1%, 2%, and 5% figures, a demographic-parity ratio closer to one and an absolute TPR gap closer to zero indicate more similar group outcomes. The highlighted ADV-NN was selected using development-sample validation; the audit label was not used in training, adversarial training, λ selection, or refitting.

Results in this section are intended to be read as controlled simulation evidence rather than as empirical estimates for any one lender or insurer. Figure 24 through Figure 27 compare the group-specific score distributions and their right tails. As in the other case studies, the fairness-relevant differences are concentrated in the highest-risk region where adverse operational decisions are most likely to occur.

Copyright © 2026 Society of Actuaries Research Institute

Figure 24
PREDICTED RISK DISTRIBUTIONS FOR THE REFERENCE GROUP

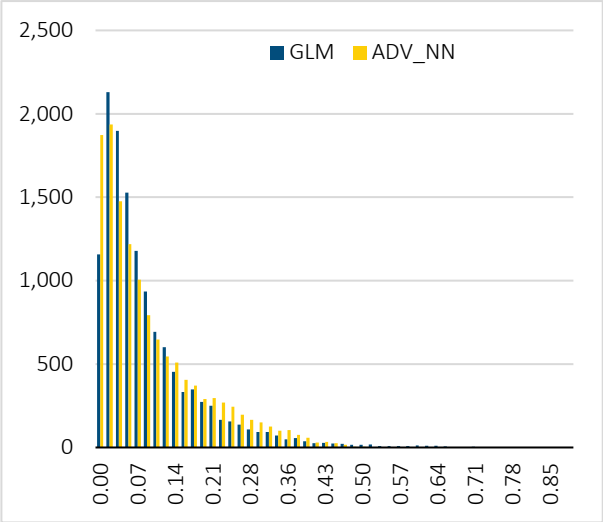


Figure 25
PREDICTED RISK DISTRIBUTIONS FOR THE DISADVANTAGED GROUP

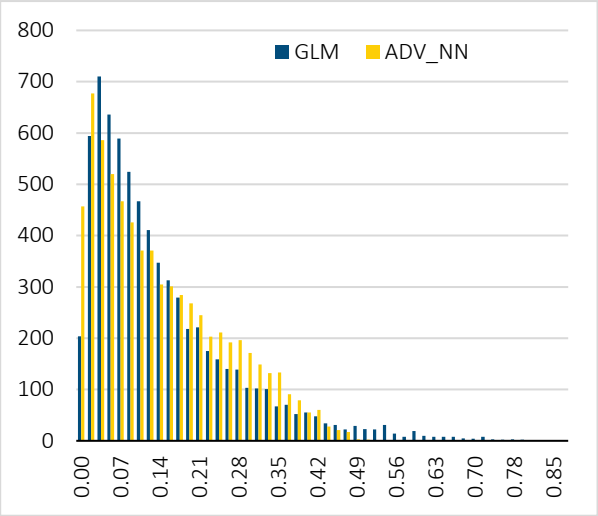


Figure 26
RIGHT TAIL OF PREDICTED RISK DISTRIBUTIONS FOR THE REFERENCE GROUP

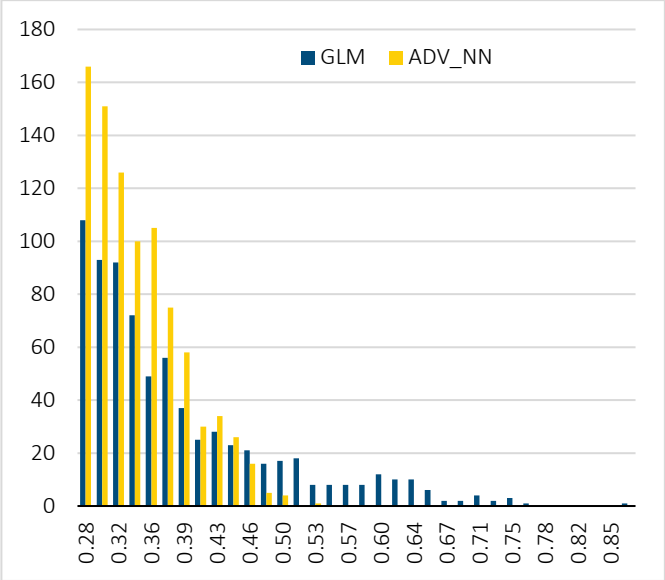


Figure 27
RIGHT TAIL OF PREDICTED RISK DISTRIBUTIONS FOR THE DISADVANTAGED GROUP

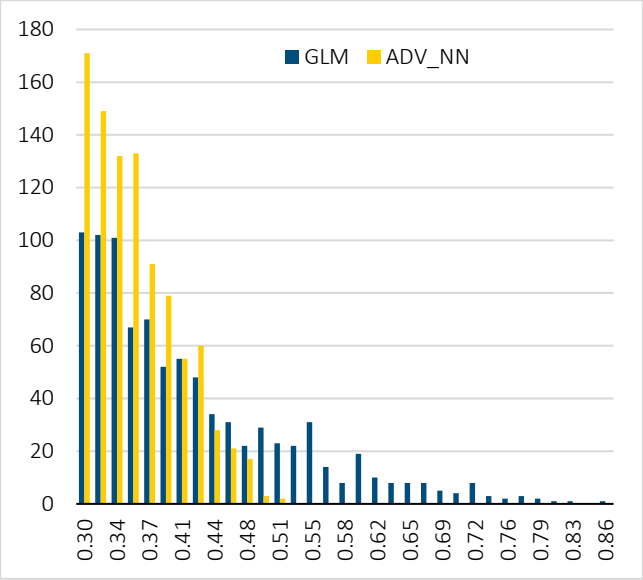


Figure 26 and Figure 27 magnify the right tail of the score distributions shown in Figure 24 and Figure 25. This tail region is the part of the portfolio most likely to be flagged for manual review, denial, surcharge, or tighter underwriting. Compared with the GLM baseline, the adversarial model narrows the gap between groups in this extreme-risk region, which is precisely where the largest group disparities arise when only a small share of applicants is classified as high risk.

Figure 28 is the main summary chart for the credit case study. Each dark-blue point is one ADV-NN fit at a prespecified value of λ. Higher ROC AUC is better, while a smaller 2% TPR gap is better. In the reported held-out

Copyright © 2026 Society of Actuaries Research Institute

results, the GLM attains an AUC of about 0.782 with a 2% TPR gap of about 0.064. The standard NN has a slightly lower AUC, about 0.781, and a slightly larger gap, about 0.069. The selected ADV-NN lowers the gap modestly to about 0.062 at an AUC of about 0.780. The improvement is modest and is accompanied by a small reduction in ROC AUC.

Figure 28

FAIRNESS–ACCURACY FRONTIER FROM AN ADVERSARIAL-PENALTY SWEEP AT THE 2% HIGH-RISK OPERATING POINT

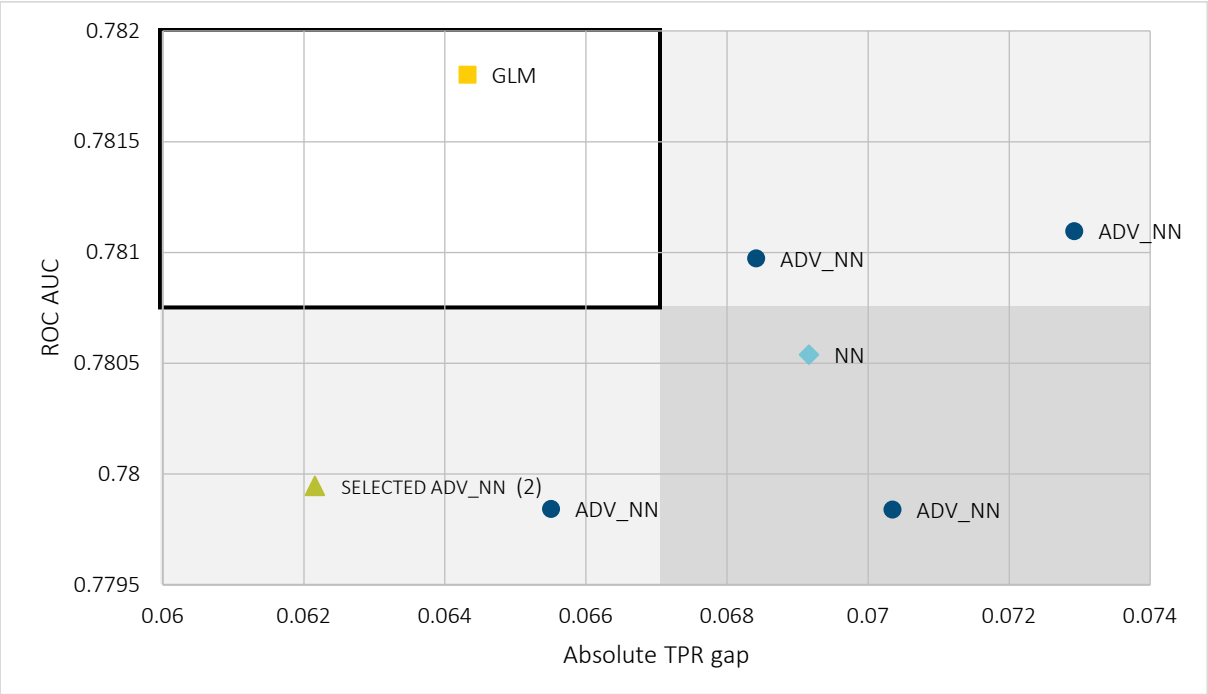


Figure 29 extends the analysis beyond the single 2 percent operating point by reporting demographic-parity ratios at 1 percent, 2 percent, and 5 percent high-risk rates. The GLM and the standard neural network both show strong over-representation of the disadvantaged group in the flagged set, especially at the strictest cutoff. The adversarial model keeps the demographic-parity ratio materially closer to one across all reported operating points, indicating that the fairness improvement is not confined to a single underwriting threshold.

Figure 30 complements the demographic-parity comparison by reporting the TPR gap across operating points. Across all three thresholds, the adversarial neural network exhibits the smallest TPR gap, whereas the GLM and standard neural network maintain noticeably larger disparities. The overall pattern is consistent with the design of the simulation: when the model is trained on a downward-biased observed credit score and a correlated ZIP variable, standard predictors can reproduce substantial group differences, while adversarial learning can attenuate those differences with a small reduction in ROC AUC.

Copyright © 2026 Society of Actuaries Research Institute

Figure 29
DEMOGRAPHIC PARITY RATIO OF HIGH-RISK SELECTION RATES ACROSS OPERATING POINTS (1%, 2%, 5%)

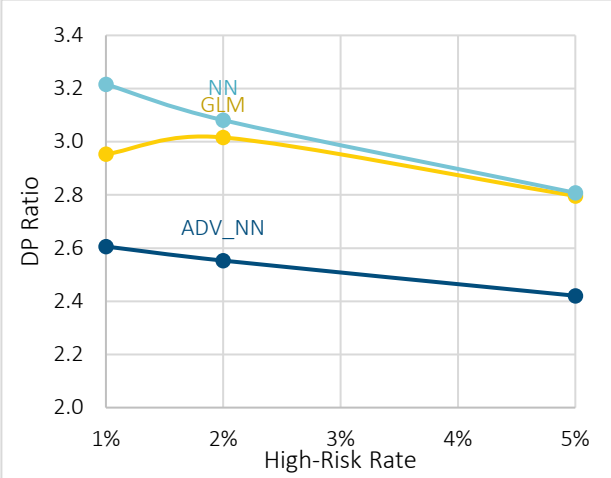
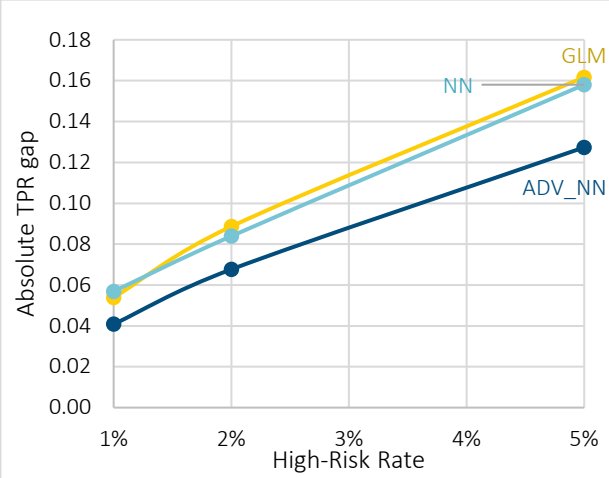


Figure 30
TPR GAP ACROSS OPERATING POINTS (1%, 2%, 5%)



3.4 MORTGAGE INSURANCE

3.4.1 EMBEDDED STATISTICAL BIAS MECHANISM

This hypothetical mortgage-insurance case study is intended to be read as a stylized mortgage-default or mortgage-insurance underwriting screen illustration, rather than as a literal model of FHA pricing or every feature of private mortgage insurance. The question is whether observed credit-score information may reflect not only underlying repayment quality, but also group-linked differences in how easily borrowers can establish, maintain, and document conventional credit histories.

The outcome variable is a binary default indicator Y. The predictive models are trained on the observed input vector $(FICO, LTV, DTI, Z)$, where $FICO$ is the observed credit score, LTV is loan-to-value, DTI is debt-to-income, and Z is a ZIP- or neighborhood-based variable. The group variable used only for ex post audit is A. The observed indirectly associated variable is Z because it is statistically related to the audit group and part of the insurer's ordinary data.

The key distortion mechanism is that the observed $FICO$ score may be a biased measurement of unobserved credit quality for one group. The simulation distinguishes between an unobserved, group-neutral score $FICO^*$ and the observed score $FICO$ that is shifted downward for Group a and then clipped to the standard 300–850 range. Because higher $FICO$ implies lower default risk, this measurement equation makes otherwise similar Group a applicants appear riskier on the observed scale.

$$\text{FICO} = \text{FICO}^* - \theta 1_{\{A=a\}} + \varepsilon,$$

where $\theta > 0$ is the downward measurement shift for Group a, and ε is random measurement noise. This construction creates two related channels. First, the observed indirectly associated Z is statistically related to the audit group. Second, the observed $FICO$ score is a biased measurement of the unobserved, risk-relevant score $FICO^*$. Together, these channels allow the model to learn subgroup-related distortion even though A itself is excluded from training.

Copyright © 2026 Society of Actuaries Research Institute

3.4.2 DATA GENERATION PROCESS

The synthetic mortgage dataset is designed as a stylized underwriting and mortgage-insurance environment with a binary default outcome. The goal is not to reproduce any particular insurer's portfolio, but to construct a controlled setting in which default risk is group-neutral at the unobserved level while observed underwriting variables can still generate group differences through indirect association and biased measurement.

At a high level, the simulation proceeds as follows:

- generate group membership A, where Group a is the disadvantaged group and Group b is the reference group;
- generate a ZIP- or neighborhood-based variable Z, with $Z = 1$ more prevalent in Group a than in Group b;
- generate an unobserved, group-neutral credit score $FICO^*$ from a group-neutral bimodal mixture;
- construct the observed $FICO$ score by shifting downward for Group a, adding random measurement noise, and clipping the result to the standard 300–850 range;
- generate loan-to-value (LTV) and right-skewed debt-to-income (DTI) ratios so that both are negatively related to unobserved credit strength;
- generate the binary default outcome Y from a logistic model driven by unobserved credit quality, centered LTV, centered DTI, and an LTV by DTI interaction; and
- calibrate the intercept so that the simulated default rate matches the target level.

This design captures the empirical idea that stronger unobserved credit tends to be associated with lower leverage and lower repayment burden, without claiming to reproduce a specific insurer's book. It is therefore more appropriate to interpret the case as a mortgage-default or mortgage-insurance underwriting screen than as a literal premium manual.

The simulation specifies the underlying risk process before generating observations. Because the true mechanism is known, model predictions can be evaluated against underlying risk rather than observed measurements alone.

An important point is that the outcome equation is part of the simulation design. In actual underwriting practice, default outcomes are observed first, and the analyst estimates a predictive model from data. Here, the process is reversed: the data-generating mechanism is specified first and then sample outcomes from that known structure. Because the data-generating mechanism is specified explicitly, the underlying risk process is known by construction. The fitted models can therefore be evaluated on whether they recover the intended risk signal or respond to distortions in the observed inputs.

Group a comprises 35% of the portfolio, and the indirectly associated Z is much more prevalent in Group a than in Group b (about 60% versus 20%). The default rate is approximately 20%. Because unobserved credit quality is generated independently of A, the model differences studied here arise through the observed FICO score and the ZIP-based indirectly associated variable rather than through a direct A term in the outcome equation. Additional details regarding the simulation are provided in Appendix F.

3.4.3 EXPLORATORY DATA PATTERNS IN THE SIMULATION

The following figures illustrate how the simulated data reflect the intended bias mechanism before model fitting begins. Key distributions are summarized to clarify the indirect association and the measurement-bias mechanism. The sample contains 100,000 mortgage applications, including 19,930 defaults and 80,070 non-defaults. Mean unobserved $FICO^*$ is essentially the same across groups, whereas mean observed $FICO$ is approximately 701 for Group a and 736 for Group b because of the embedded measurement shift.

Copyright © 2026 Society of Actuaries Research Institute

Figure 31
DISTRIBUTION OF LOAN-TO-VALUE RATIO

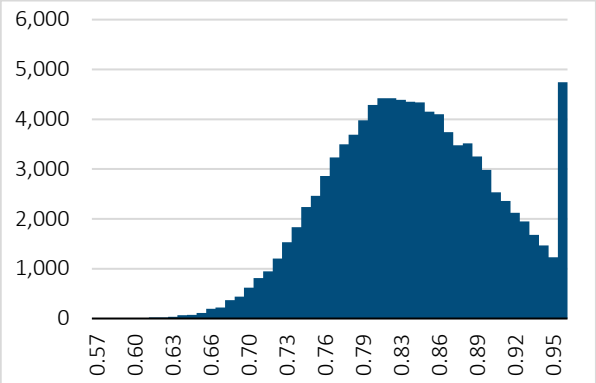


Figure 32
DISTRIBUTION OF DEBT-TO-INCOME RATIO

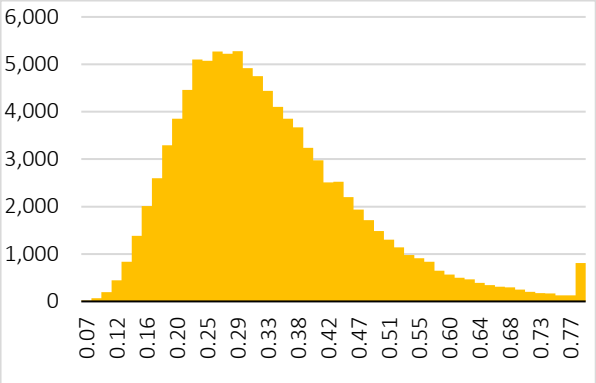


Figure 33
OBSERVED FICO AND UNOBSERVED CREDIT SCORE FICO*

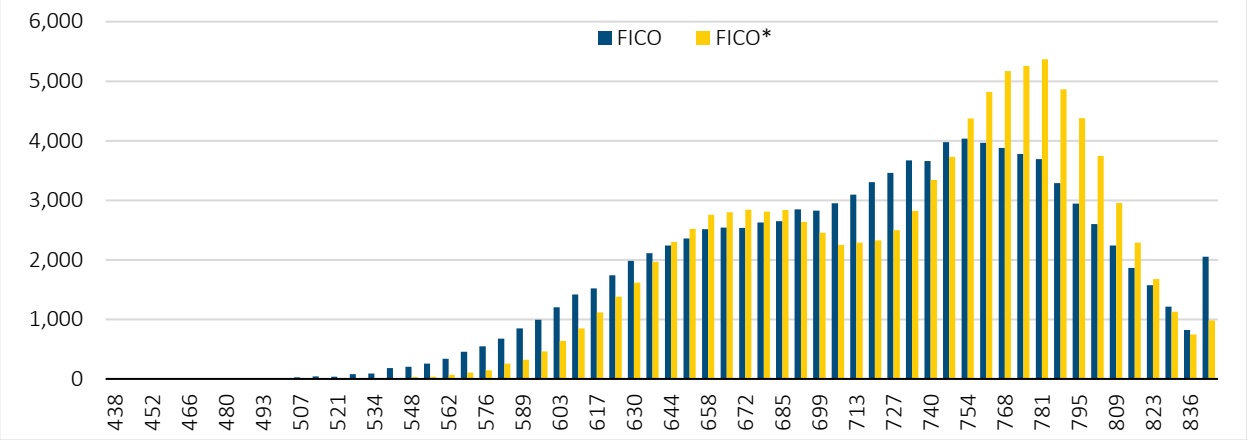


Figure 31 and Figure 32 display the simulated distributions of LTV and DTI. The DTI distribution is right-skewed, and both LTV and DTI are negatively related to unobserved credit strength by construction. Figure 33 contrasts the unobserved, group-neutral score $FICO^*$ with the observed $FICO$ score used by the model. In the simulated sample, mean unobserved $FICO^*$ is essentially the same across groups, whereas mean observed $FICO$ is about 701 for Group a and 736 for Group b because of the embedded measurement shift. The observed score is also bounded within the standard 300–850 range.

3.4.4 MODEL EVALUATION RESULTS

Throughout this subsection, GLM denotes the generalized linear model benchmark, NN denotes the standard neural-network benchmark, and ADV-NN denotes the adversarial neural network. All reported metrics are computed on the held-out evaluation sample. In the frontier figure, higher ROC AUC indicates better ranking performance, while an absolute TPR gap closer to zero indicates more similar true-positive rates across audit groups. The 2% high-risk operating point selects the 2% of cases with the highest predicted risks for a potentially unfavorable action, such as additional review or more restrictive terms. In the 1%, 2%, and 5% figures, a demographic-parity ratio closer to one and an absolute TPR gap closer to zero indicate more similar group outcomes. The highlighted ADV-NN was selected using development-sample validation; the audit label was not used in training, adversarial training, λ selection, or refitting.

Copyright © 2026 Society of Actuaries Research Institute

Figure 34 through Figure 37 compare predicted default probabilities across groups for the GLM baseline and the adversarial model. Relative to the GLM, the adversarial model produces group-specific distributions that are closer together, especially in the upper tail where high-risk flags are most likely to be triggered.

Figure 34
PREDICTED RISK DISTRIBUTIONS FOR THE REFERENCE GROUP

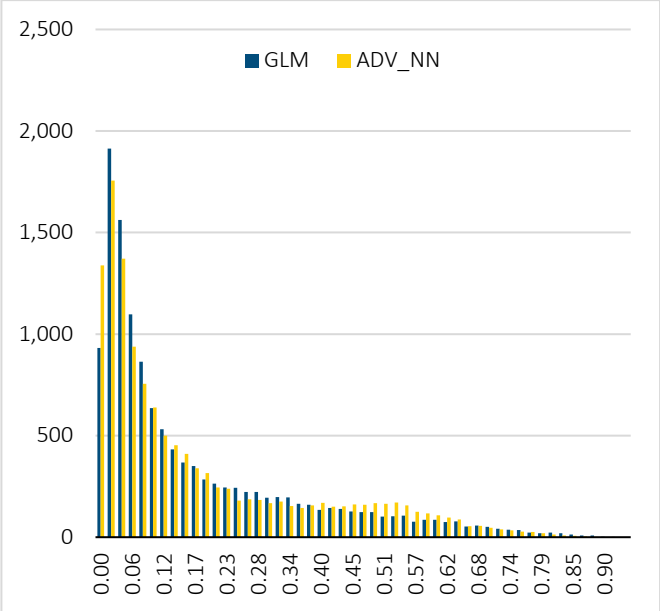


Figure 35
PREDICTED RISK DISTRIBUTIONS FOR THE DISADVANTAGED GROUP

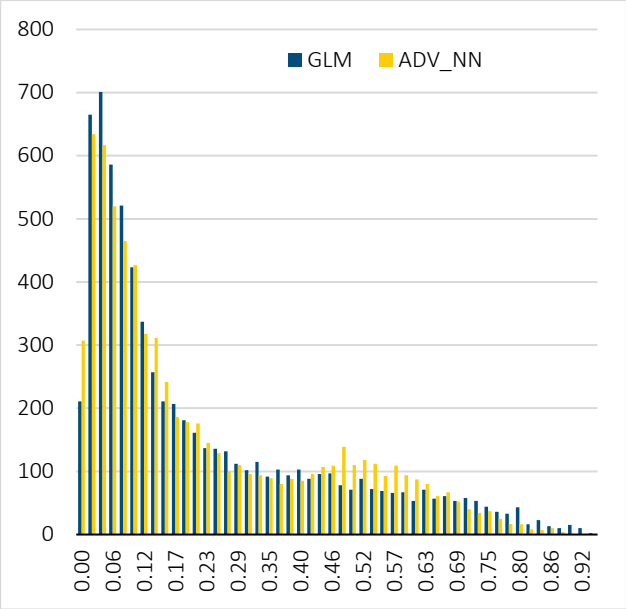


Figure 36
RIGHT TAIL OF PREDICTED RISK DISTRIBUTIONS FOR THE REFERENCE GROUP

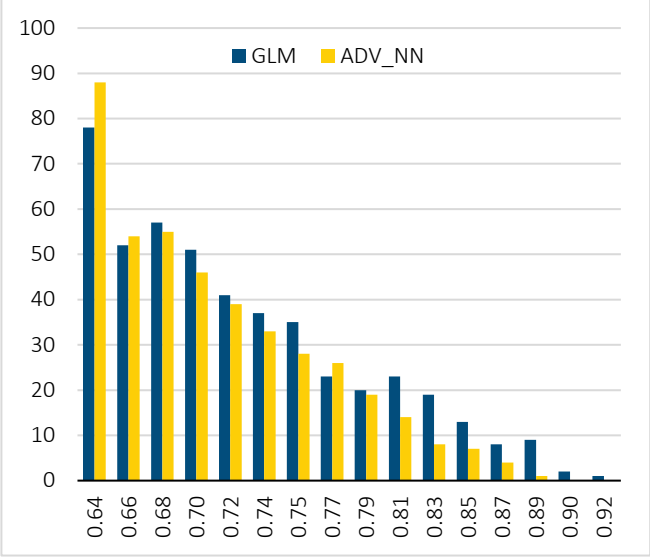


Figure 37
RIGHT TAIL OF PREDICTED RISK DISTRIBUTIONS FOR THE DISADVANTAGED GROUP

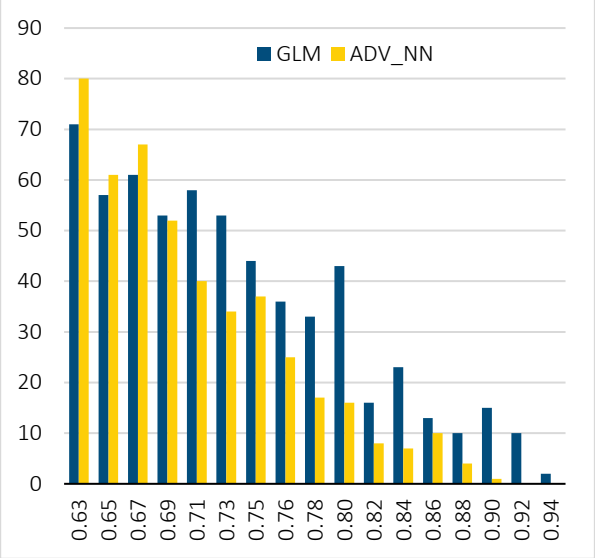


Figure 36 and Figure 37 provide a magnified view of the right tail of the distributions in Figure 34 and Figure 35, focusing on the region that is most relevant for selection into "high-risk" actions. This zoomed view makes the tail behavior easier to interpret: compared with the GLM baseline, the adversarial model reduces the separation

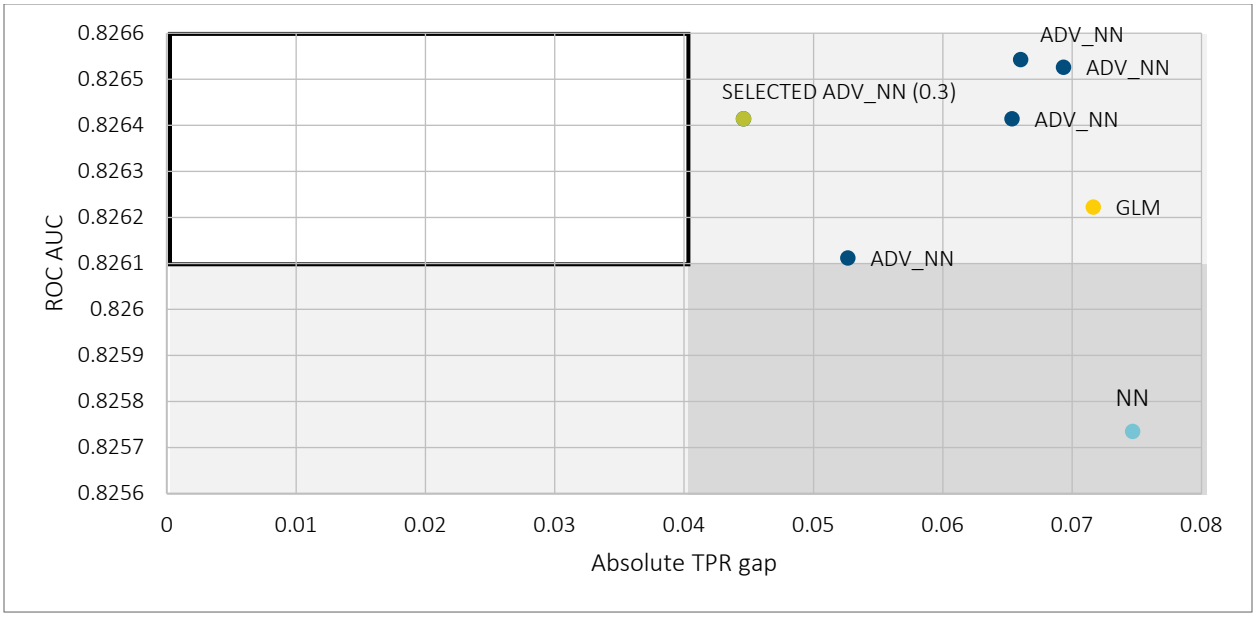
Copyright © 2026 Society of Actuaries Research Institute

between the groups in the highest-risk region, which is consistent with smaller group-audit gaps when the top 1%–5% of predicted risks are flagged for more restrictive treatment.

Figure 38 is the main summary chart for the mortgage case study. Each dark-blue point is one ADV-NN fit at a prespecified value of λ. Higher ROC AUC is better, while a smaller 2% TPR gap is better. The GLM attains a held-out ROC AUC of 0.8262 with a 2% absolute TPR gap of 0.0716. The standard NN attains an ROC AUC of 0.8257 with a gap of 0.0747. The validation-selected ADV-NN attains an ROC AUC of 0.8264 and reduces the gap to 0.0446. The selected adversarial model therefore improves the reported audit diagnostic without a material loss of ranking performance.

Figure 38

FAIRNESS–ACCURACY FRONTIER FROM AN ADVERSARIAL-PENALTY SWEEP AT THE 2% HIGH-RISK OPERATING POINT



Copyright © 2026 Society of Actuaries Research Institute

Figure 39
DEMOGRAPHIC PARITY RATIO OF HIGH-RISK SELECTION RATES ACROSS OPERATING POINTS (1%, 2%, 5%)

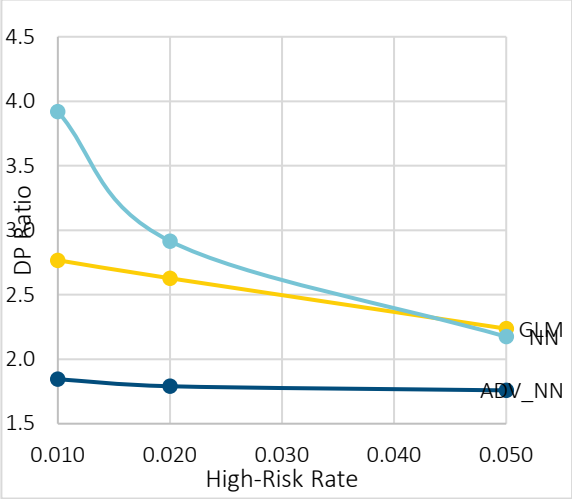


Figure 40
TPR GAP ACROSS OPERATING POINTS (1%, 2%, 5%)

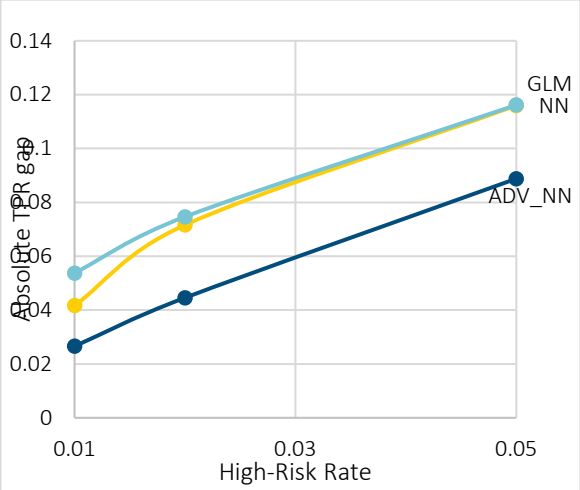


Figure 39 reports demographic-parity ratios across 1%, 2%, and 5% high-risk action rates. At the 2% operating point, the demographic-parity ratio is 1.790 for the selected ADV-NN, compared with 2.628 for the GLM and 2.916 for the NN. Because values closer to one indicate more similar selection rates, the selected ADV-NN attenuates, but does not eliminate, the selection disparity.

Figure 40 shows the same pattern for TPR gaps. The 2% absolute TPR gap is 0.0446 for the selected ADV-NN, compared with 0.0716 for the GLM and 0.0747 for the NN. Comparable reductions appear at the 1% and 5% operating points. Taken together, Figure 38 through Figure 40 suggest that the selected ADV-NN reduces the reported group-audit gaps in a setting containing both an indirectly associated ZIP variable and a biased observed FICO score.

3.5 HEALTH INSURANCE

3.5.1 EMBEDDED STATISTICAL BIAS MECHANISM

The hypothetical health-insurance case study uses health-risk scoring and risk-stratification settings in its illustration. In these processes, insurers or health plans rely on administratively constructed morbidity measures rather than on direct observation of underlying health status. In practice, an observed morbidity score may reflect not only a person's unobserved health risk, but also differences in diagnosis intensity, coding practices, utilization patterns, continuity of care, and the completeness of recorded medical information (Yi et al., 2025). As a result, two individuals with similar unobserved morbidity may nevertheless receive different observed morbidity scores because one is more likely to interact with health-care environments that generate more intensive or more complete documentation.

The outcome variable in this case study is a binary indicator Y that records whether the adverse event of interest occurs over the stated horizon. The predictive models are trained on observed inputs that include age, sex, chronic-condition status, the observed morbidity score M_{obs}, and the Area Deprivation Index, which is operationalized in the simulation through the high-area-deprivation indicator ADI_{high}. The audit label in this case study is $A \in \{a, b\}$, where a denotes the disadvantaged group and b denotes the reference group. This audit label is not used in model training and is retained only for ex post evaluation. The observed indirectly associated variable is the Area Deprivation Index,

Copyright © 2026 Society of Actuaries Research Institute

a community-level measure of neighborhood disadvantage constructed from factors such as income, education, employment, and housing quality, which makes it a natural indirect association when place-based disadvantage is correlated with audit-group membership (Yi et al., 2025; Owen & Teppema, 2023). In addition, the simulation distinguishes between the unobserved morbidity factor M^*, which is risk-relevant and group-neutral by construction, and the observed morbidity score M_{obs}, which is the quantity actually used in model training.

The key distortion mechanism is that the observed morbidity score may be a biased measurement of the unobserved morbidity factor for one group. In particular, individuals in group $a$may have observed values of M_{obs} that are systematically higher than their unobserved values of M^*, even when true health status is the same. The measurement process is represented as

$$M_{\text{obs}} = M^* + \theta 1_{\{A=a\}} + \varepsilon,$$

where $\theta > 0$ is the magnitude of the upward shift for group a and ε is measurement noise. Because higher morbidity implies higher event risk, this shift makes group a appear systematically riskier than an otherwise identical member of group b, even when unobserved health status is the same.

This construction creates two distinct but related channels. First, the Area Deprivation Index is statistically related to the audit group A. Second, the observed morbidity score M_{obs} is a biased measurement of the unobserved, risk-relevant morbidity factor M^*. Together, these channels allow the model to learn subgroup-related distortion even though the audit label itself is excluded from training.

3.5.2 DATA GENERATION PROCESS

The synthetic health-insurance dataset is designed as a stylized underwriting and risk-stratification environment with a binary adverse-event outcome. This setup is motivated by evidence that place-based social determinants and barriers to care can shape both health outcomes and the administrative information available to payers (Barrett & Entzminger, 2022; Owen & Teppema, 2023; Yi et al., 2025). The goal is not to reproduce any one insurer's proprietary portfolio, but to construct a controlled setting in which health risk is group-neutral at the unobserved level while observed underwriting variables can still generate group differences through indirect association and biased measurement.

At a high level, the simulation proceeds as follows:

- generate group membership A, where a is the disadvantaged group and b is the reference group;
- generate the Area Deprivation Index indirectly associated variable, operationalized as ADI_{high}, with high deprivation more prevalent in group a;
- generate age, sex, and an unobserved frailty variable U;
- generate chronic-condition status from a logistic model driven by age, sex, and frailty, but not directly by group membership;
- generate unobserved morbidity M^* as a function of frailty, age, chronic-condition status, and random noise;
- construct the observed morbidity score M_{obs} by shifting M^* upward for group $a$and adding measurement noise;
- standardize unobserved morbidity and generate the binary outcome Y from a logistic model driven by standardized M^*, age, and chronic-condition status; and
- calibrate the intercept so that the overall event rate matches the target level.

Copyright © 2026 Society of Actuaries Research Institute

The interpretation is that health-insurance models often rely on observed morbidity measures because they are scalable, routinely available, and operationally convenient. However, if M_{obs} is a distorted measure of the unobserved signal M^* for one subgroup, then a predictive model trained on M_{obs} may inherit that distortion. The simulation therefore separates the role of the unobserved health signal from the observed morbidity measure so that the effect of measurement bias can be studied directly.

The simulation specifies the underlying risk process before generating observations. Because the true mechanism is known, model predictions can be evaluated against underlying risk rather than observed measurements alone.

An important point is that the outcome equation is part of the simulation design. In actual underwriting or risk-stratification practice, adverse outcomes are observed first, and the analyst estimates a predictive model from data. Here, the process is reversed: the data-generating mechanism is specified first and then sample outcomes from that known structure. Because the data-generating mechanism is specified explicitly, the underlying risk process is known by construction. The fitted models can therefore be evaluated on whether they recover the intended risk signal or respond to distortions in the observed inputs.

Group a comprises 25% of the portfolio, and ADI_{high} is more prevalent in group a than in group b (55% versus 25%). Age is generated from a truncated normal distribution with mean 45 and standard deviation 15 on $[18, 85]$, sex is generated as Female $\sim$ Bernoulli(0.50), and unobserved frailty follows $U \sim N(0,1)$. Chronic-condition status is then generated from a logistic model with coefficients -0.5 on the intercept, 0.45 on scaled age, 0.10 on female, and 0.40 on U. Unobserved morbidity is generated as a function of frailty, age, chronic-condition status, and Gaussian noise, while the observed morbidity score equals unobserved morbidity plus a group-specific upward shift of 2.20 for group a and additive normal noise with standard deviation 0.35. The binary outcome Y is then generated from standardized unobserved morbidity, age, and chronic-condition status, with the intercept calibrated so that the overall adverse-event rate is approximately 12%. Additional details regarding the simulation can be found in Appendix G.

3.5.3 EXPLORATORY DATA PATTERNS IN THE SIMULATION

The following figures illustrate how the simulated data reflect the intended bias mechanism before model fitting begins. Figure 41 and Figure 42 summarize the main features of the simulated health-insurance data. The simulation contains 100,000 members, including 24,910 in Group a and 75,090 in Group b. The high-area-deprivation indicator equals one for 32,593 members and is substantially more common in Group a than in Group b, approximately 55% versus 25%. The sample contains 11,997 adverse events and 88,003 non-events.

Figure 41 displays the age distribution, which is generated from a truncated normal distribution and represents a conventional demographic risk driver. Figure 42 contrasts the observed morbidity score available to the model with the unobserved morbidity factor that drives the true outcome. This contrast makes the embedded bias mechanism visible: the score used operationally by the model is systematically shifted upward for the disadvantaged group even though the unobserved morbidity process is constructed to be group neutral.

Copyright © 2026 Society of Actuaries Research Institute

Figure 41
DISTRIBUTION OF AGE

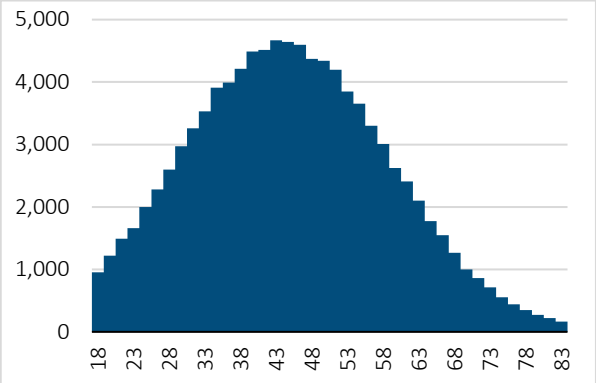
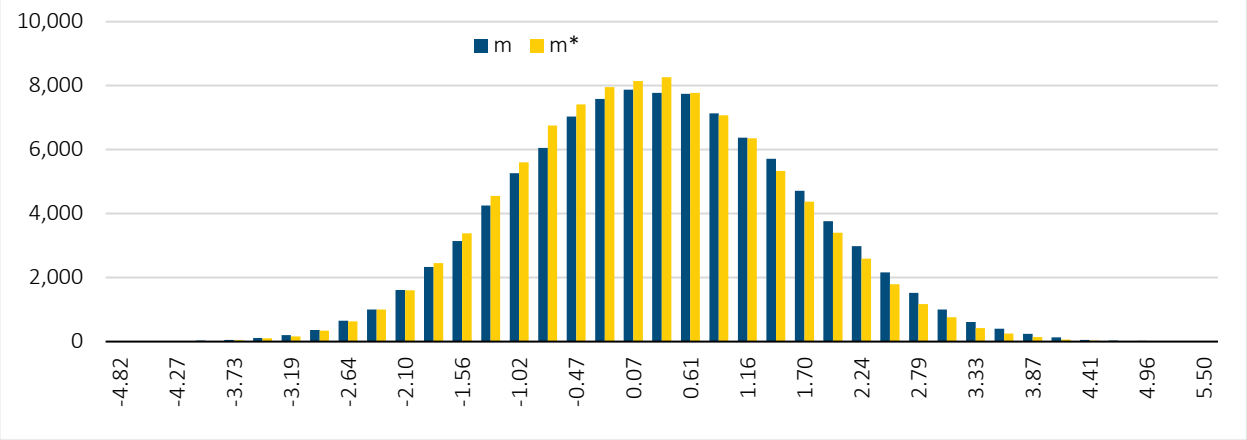


Figure 42
OBSERVED MORBIDITY SCORE M_OBS AND UNOBSERVED MORBIDITY SCORE M*



3.5.4 MODEL EVALUATION RESULTS

Throughout this subsection, GLM denotes the generalized linear model benchmark, NN denotes the standard neural-network benchmark, and ADV-NN denotes the adversarial neural network. All reported metrics are computed on the held-out evaluation sample. In the frontier figure, higher ROC AUC indicates better ranking performance, while an absolute TPR gap closer to zero indicates more similar true-positive rates across audit groups. The 2% high-risk operating point selects the 2% of cases with the highest predicted risks for a triage or intervention decision, which is not necessarily an adverse underwriting action. In the 1%, 2%, and 5% figures, a demographic-parity ratio closer to one and an absolute TPR gap closer to zero indicate more similar group outcomes. The highlighted ADV-NN was selected using development-sample validation; the audit label was not used in training, adversarial training, λ selection, or refitting.

Results in this section are intended to be interpreted as an illustration of how mitigation changes audited disparities when the model observes a biased morbidity measure. Figure 43 through Figure 46 compare the group-specific score distributions and their right tails. Relative to the GLM, the selected ADV-NN produces closer group-specific score distributions and smaller ex post audit gaps in the presence of the biased morbidity and area-deprivation pathways.

Copyright © 2026 Society of Actuaries Research Institute

Figure 43
PREDICTED RISK DISTRIBUTIONS FOR THE REFERENCE GROUP

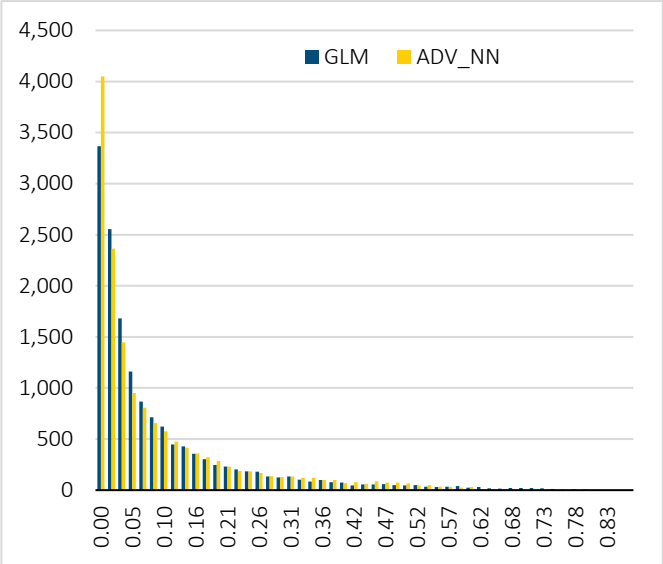


Figure 44
PREDICTED RISK DISTRIBUTIONS FOR THE DISADVANTAGED GROUP

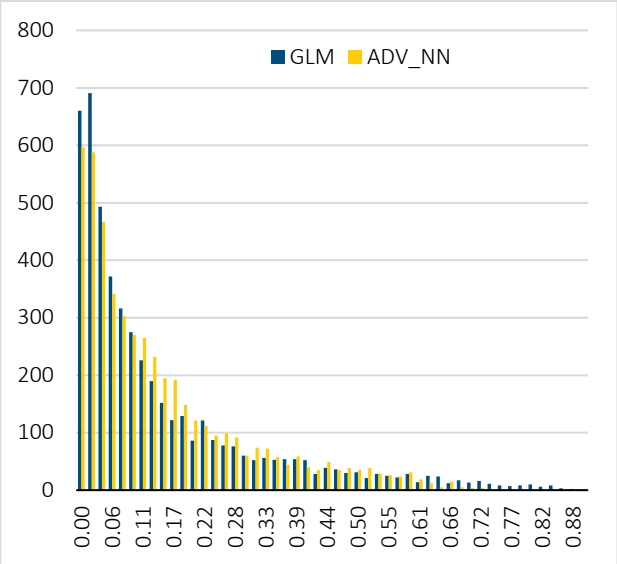


Figure 45
RIGHT TAIL OF PREDICTED RISK DISTRIBUTIONS FOR THE REFERENCE GROUP

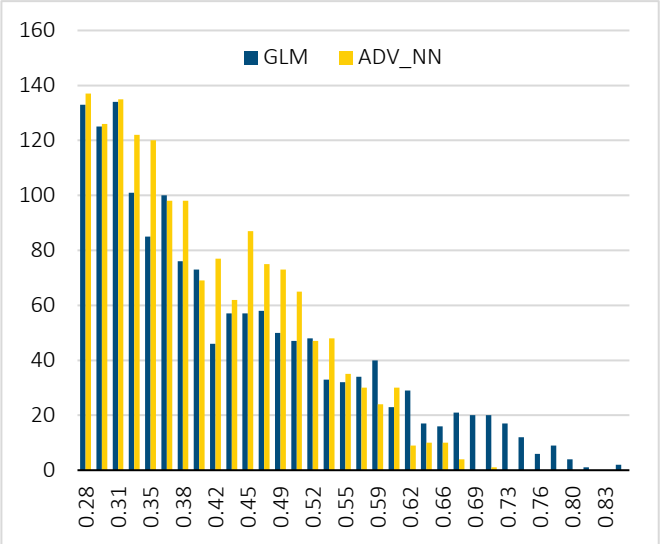


Figure 46
RIGHT TAIL OF PREDICTED RISK DISTRIBUTIONS FOR THE DISADVANTAGED GROUP

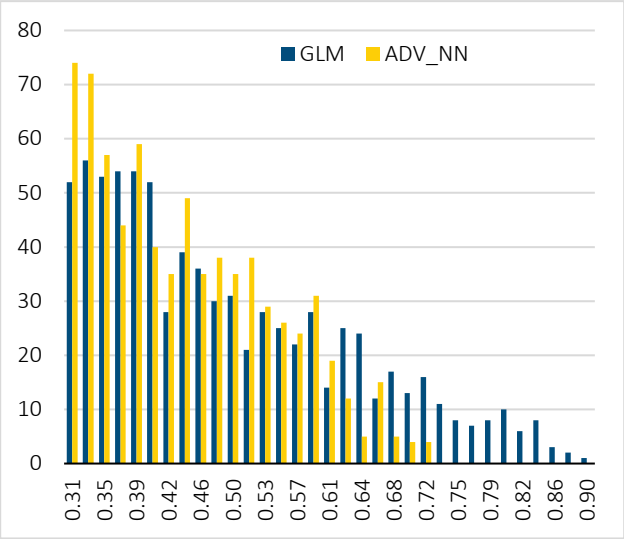


Figure 45 and Figure 46 provide a magnified view of the right tail of the predicted-risk distributions in Figure 43 and Figure 44. This is the region that matters operationally when the insurer or health plan intervenes only on the highest-risk cases. Compared with the GLM baseline, the adversarial model reduces the separation between the group-specific tails, which is consistent with improved fairness metrics at fixed high-risk operating points.

Having established that adversarial training compresses group differences in the highest-risk region, Figure 47 quantifies the fairness-accuracy trade-off across values of the adversarial penalty parameter λ, while Figure 48 and Figure 49 examine fairness metrics across alternative high-risk thresholds.

Copyright © 2026 Society of Actuaries Research Institute

Figure 47

FAIRNESS–ACCURACY FRONTIER FROM AN ADVERSARIAL-PENALTY SWEEP AT THE 2% HIGH-RISK OPERATING POINT

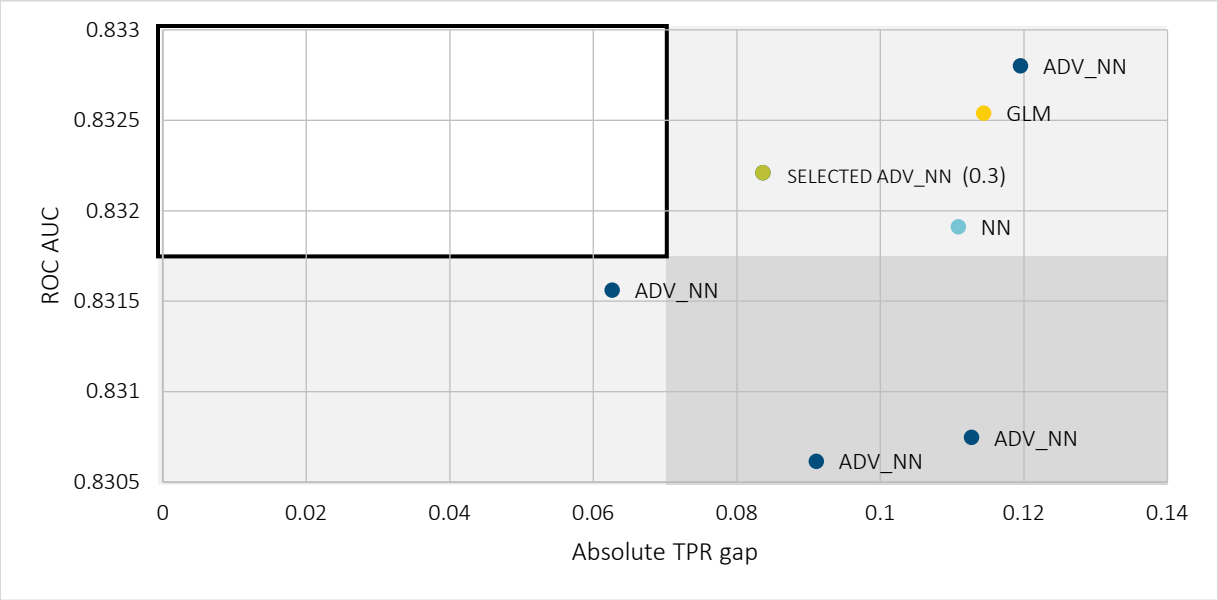


Figure 47 is the main summary chart for the health case study. Each dark-blue point represents one ADV-NN fit at a prespecified value of λ. The vertical axis is ROC AUC, so higher is better; the horizontal axis is the ex post TPR gap at the 2% high-risk operating point, so smaller is better. The figure shows that the GLM attains a held-out ROC AUC of 0.8325 with a 2% TPR gap of 0.1144, while the standard NN attains an ROC AUC of 0.8319 with a gap of 0.1109. The validation-selected ADV-NN attains an ROC AUC of 0.8322 and reduces the gap to 0.0836. The change in predictive ranking is negligible, while the reported TPR-gap diagnostic improves materially.

Figure 48 and Figure 49 extend the comparison across 1%, 2%, and 5% high-risk action rates. At the 2% operating point, the demographic-parity ratio is approximately 2.403 for the selected ADV-NN, compared with 2.740 for the GLM and 2.880 for the NN. The corresponding absolute TPR gap is 0.0836 for the selected ADV-NN, compared with 0.1144 for the GLM and 0.1109 for the NN. Across these operating points, the adversarial neural network yields demographic-parity ratios closer to one than the GLM or the plain neural network, indicating substantially more balanced rates of high-risk flagging across groups.

Copyright © 2026 Society of Actuaries Research Institute

Figure 48

DEMOGRAPHIC PARITY RATIO OF HIGH-RISK SELECTION RATES ACROSS OPERATING POINTS (1%, 2%, 5%)

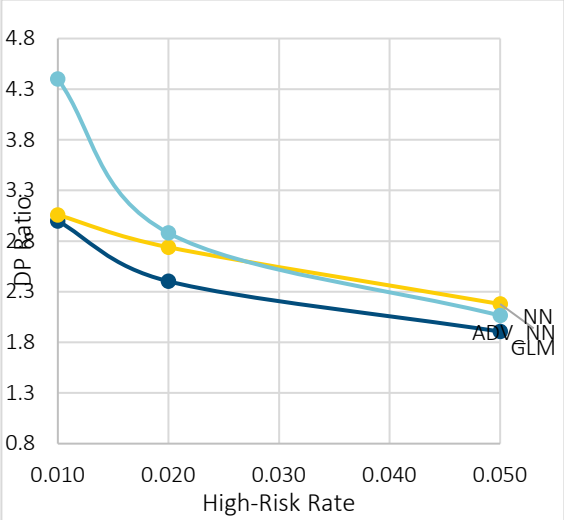
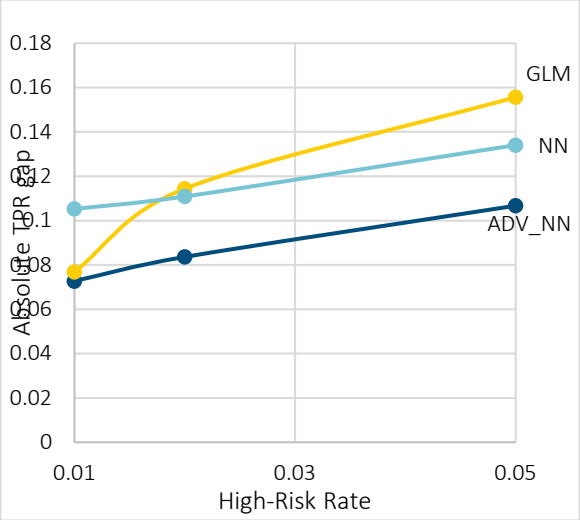


Figure 49

TPR GAP ACROSS OPERATING POINTS (1%, 2%, 5%)



Copyright © 2026 Society of Actuaries Research Institute

Section 4 Indirectly Associated Variables, Measurement Bias, and Group Disparities

4.1 DEFINITIONS AND NOTATION

Let $A \in \{a, b\}$ denote the audit label, where a denotes the disadvantaged group and $b$the reference group. The audit label is deliberately excluded from model development and is retained only for simulation design and ex post audit.

Let Z_{IA} denote the designated observed indirectly associated variable used as the adversarial target and let W denote the remaining observed predictors. The observed feature vector is written as

$$X = (Z_{\mathrm{IA}}, W).$$

In the applications, $Z_{\mathrm{IA}} = T$ for life and auto insurance, $Z_{\mathrm{IA}} = Z$ for credit and mortgage insurance, and $Z_{\mathrm{IA}} = ADI_{\mathrm{high}}$ for health insurance.

Let Y denote the target outcome representing actuarial risk, such as expected loss, claim frequency, claim severity, default, mortality, or another underwriting outcome.

Let the prediction model be written as

$$\hat{Y} = f(X) = f(Z_{\mathrm{IA}}, W).$$

A fitted model can produce systematic group differences if the distribution of $\hat{Y}$ differs across values of A. Formally, this occurs when:

$$\mathcal{L}(\hat{Y} \mid A = a) \neq \mathcal{L}(\hat{Y} \mid A = b),$$

where $\mathcal{L}(\cdot)$ denotes a conditional distribution. Importantly, this can occur even when A is not included directly in X.

4.2 MECHANISMS THAT GENERATE INDIRECTLY ASSOCIATED DISPARITIES

The central statistical point is that explicitly avoiding the use of A does not imply avoiding the use of information about A. If Z_{IA} is correlated with A, then A is partially predictable from the observed data. In that case, the model output

$$\hat{Y} = f(Z_{\mathrm{IA}}, W)$$

can inherit dependence on A, simply because $\hat{Y}$ is a function of variables that carry information about subgroup membership.

A simple way to see this is through conditional expectations. In supervised learning, a well-trained predictor approximates:

$$E[Y \mid X] = E[Y \mid Z_{\mathrm{IA}}, W].$$

If the distribution of (Z_{IA}, W) differs across values of A, then even a model that does not observe A can reproduce differences across groups through the pathway:

Copyright © 2026 Society of Actuaries Research Institute

$$A \rightarrow Z_{\text{IA}} \rightarrow \hat{Y}.$$

This report focuses on two related mechanisms:

1. Indirect association. Indirectly associated variables are observed features that are not themselves the audit label, but that carry enough information about that label to function as an indirect pathway. Common examples include geography, neighborhood characteristics, ZIP-code indicators, territory classifications, area-deprivation measures, and some credit-related or utilization-related variables. Not every weak statistical association is substantively important; the concern arises when the association is strong enough for the fitted model to recover and use group-related information.

2. Measurement bias. Suppose there is an unobserved risk-relevant quantity S^* that the insurer would ideally like to observe, but only an imperfect observed version S is available. A simple representation is:

   $$S = S^* + \theta(A) + \varepsilon,$$

   where $\theta(A)$ is a group-dependent distortion and ε is random noise. If $\theta(A) \neq 0$, then the observed variable S does not measure the unobserved quantity S^* equally across groups. A model trained on S may therefore treat otherwise similar risks differently because it is learning from the distorted observable rather than from the underlying risk signal.

These two mechanisms can also operate jointly. A model may rely on both a biased measurement and an indirectly associated variable that is correlated with A. This is the main mechanism studied in the report: the insurer observes data that are incomplete or systematically distorted for one subgroup, and a predictive model trained on those observables can therefore generate systematically different outputs for risks that are substantively similar at the unobserved level.

4.3 A USEFUL TARGET IN INDIRECT ASSOCIATION AND MEASUREMENT-BIAS SETTINGS

The broad audit objective is to reduce the extent to which the fitted model produces group-differentiated outputs beyond what is justified by underlying risk. At the level of ex post evaluation, one useful conceptual target is:

$$\hat{Y} \perp A \mid Y,$$

or, more generally, approximate conditional independence of predictions and the audit label after conditioning on true outcome or risk strata.

In classification settings, this idea is closely related to equalized-odds-type criteria, which compare error rates across values of A. In regression settings, the analogous idea is that predictions should not differ systematically across audit groups after accounting for the underlying risk the model is intended to measure. In the synthetic simulations, that underlying risk is known by construction from the data-generating process; in operational data, it would generally be unobserved.

However, in the setting studied here, A is deliberately excluded from model training. The practical mitigation target is therefore not to adversarially remove A directly, but to reduce the extent to which the learned representation carries information about the observed indirectly associated variable Z_{IA}. In that sense, the operational target in the adversarial model is better written as

$$h \perp Z_{\text{IA}} \mid Y,$$

Copyright © 2026 Society of Actuaries Research Institute

at least approximately, where h denotes the learned hidden representation. If Z_{IA} carries information about A, then reducing dependence on Z_{IA} can reduce the indirect pathway through which subgroup-related information enters the prediction model.

In this report, group-audit criteria are used as **statistical audit tools** rather than as legal conclusions. Their role is to summarize whether a fitted model appears to rely on distorted or indirect association-laden pathways in ways that produce materially different outputs across values of A. The mitigation mechanism acts on the observed indirectly associated variable Z_{IA}, while the later audit is carried out using A.

Section 5 Predictive Models and Mitigation Approach

5.1 MODELING GOAL

Across insurance lines of business, predictive models are used to translate observed information about an applicant, policyholder, or exposure into a risk score that supports pricing, underwriting, or triage. In this report, the focus is on AI models for insurance pricing and risk assessment, together with a standard actuarial benchmark.

Let Z_{IA} denote the designated observed indirectly associated variable used in the adversarial branch, and let W denote the remaining observed predictors. The observed model input is written as

$$X = (Z_{\mathrm{IA}}, W).$$

In many underwriting settings, Y is binary, for example whether an adverse event occurs within a specified time horizon. In that case, the model produces a predicted probability:

$$\hat{p}(X) = P(Y = 1 \mid X).$$

When the target is continuous, such as expected loss cost or claim severity, the same framework applies with a predicted conditional mean:

$$\hat{\mu}(X) = E(Y \mid X).$$

The audit label A is not included in model training. It is retained only for ex post evaluation. By contrast, the designated observed indirectly associated variable Z_{IA} is part of the ordinary observed data and may also be used in the adversarial regularization term.

5.2 MODEL OVERVIEW

This report compares three model classes trained on the same observed input $X = (Z_{\mathrm{IA}}, W)$. The first is a **generalized linear model (GLM)**, which serves as a transparent actuarial benchmark. The second is a **standard neural network (NN)**, which represents a flexible AI model capable of learning nonlinear effects and interactions automatically. The third is an **adversarial neural network (ADV-NN)**, which augments the neural-network training problem with a penalty designed to make the learned representation less informative about the observed indirectly associated variable Z_{IA}.

Comparing models highlights how design choices affect both accuracy and reliance on indirect association.

These models play different roles, as summarized in Table 1.

Copyright © 2026 Society of Actuaries Research Institute

Table 1
MODEL COMPARISON

Model	Primary Strength	Purpose in Study
GLM	Transparent and interpretable	Actuarial benchmark
NN	Flexible nonlinear learning	AI benchmark
ADV-NN	Reduces reliance on indirect associations	Mitigation approach

5.2.1 GENERALIZED LINEAR MODEL

For a binary target, the GLM constructs a linear score from the observed features and then maps that score into a probability between 0 and 1. A standard logistic specification can be written as

$$P(Y = 1 \mid X) = \frac{\exp(\beta_0 + X^\top \beta)}{1 + \exp(\beta_0 + X^\top \beta)}.$$

Because the structure is relatively simple, GLMs are easier to explain, validate, and audit. In the present context, the GLM is useful not because it is immune to distortion, but because it provides a clear baseline. If the observed inputs are distorted, the GLM can also reproduce those distortions.

5.2.2 NEURAL NETWORK

A standard feed-forward neural network processes the observed input vector X through one or more hidden layers, as illustrated in Figure 50. These hidden layers construct intermediate representations that capture nonlinear transformations and interactions among the original variables. The final layer maps the learned representation into a predicted probability or predicted mean.

For a binary target, this can be written schematically as

$$h = g_\theta(X), \qquad \hat{p}(X) = \sigma(w^\top h + b),$$

where $g_\theta(\cdot)$ denotes the feature-extraction component, h is the learned representation, and $\sigma(\cdot)$ is the logistic link function.

From a practical perspective, a neural network is flexible. It can detect nonlinear patterns and interactions without requiring them to be specified in advance. The same flexibility, however, also means that the network can combine many weak signals into a strong predictive rule. If the observed data contain measurement bias or indirect association structure, a standard NN may learn distorted relationships that are difficult to identify from the fitted model alone.

Figure 50
STANDARD NEURAL NETWORK

5.2.3 ADVERSARIAL NEURAL NETWORK

The adversarial neural network is designed to address that problem during training. Like the standard NN, it learns a representation of the observed predictors and uses that representation to predict the outcome Y. However, it also includes an auxiliary model, called an adversary, that attempts to recover the observed indirectly associated variable Z_{IA} from the same learned representation.

Copyright © 2026 Society of Actuaries Research Institute

The training logic is as follows:

- the predictor is trained to predict Y as accurately as possible;
- the adversary is trained to predict Z_{IA} from the learned representation;
- the predictor is simultaneously trained to make that indirect association-prediction task more difficult.

As a result, as shown in Figure 51, the learned representation is encouraged to retain information that is useful for predicting risk while discarding, as much as possible, information that mainly reveals the observed indirect association.

Formally, let:

$$h = g_\theta(X), \qquad \hat{Y} = f_\theta(X),$$

and let the adversary produce:

$$\widehat{Z_{\mathrm{IA}}} = q_\phi(h).$$

The overall objective combines the predictive loss and the adversarial loss:

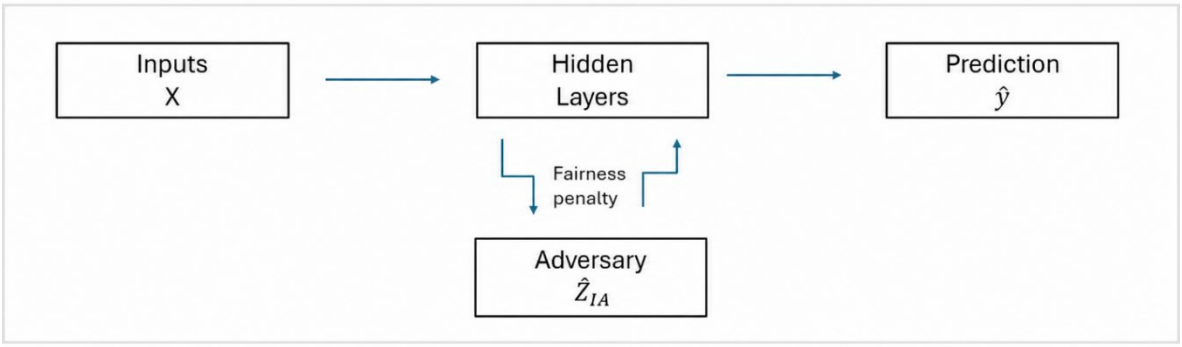
$$L(\theta, \phi) = L_{\text{pred}}(Y, \hat{Y}) + \lambda L_{\text{adv}}(Z_{\mathrm{IA}}, \widehat{Z_{\mathrm{IA}}}).$$

With gradient reversal, the adversarial loss contributes the opposite gradient to the feature extractor: the adversary is trained to minimize its prediction loss, while the feature extractor is trained to make that prediction more difficult. Here, λ is a tuning parameter controlling the strength of the adversarial penalty. When λ is small, training focuses mainly on predictive accuracy. When λ is larger, the model places more weight on suppressing indirectly associated information in the learned representation. In practice, increasing λ can reduce ex post group-difference diagnostics, but may also reduce predictive performance to some extent. The value of λ is selected using the cross-validation procedure summarized in Section 6.5 and detailed in Appendix A.4, rather than by inspecting the held-out group-audit results.

It is therefore important to be careful about interpretation. The ADV-NN does **not** observe or adversarially remove the audit label A. Rather, it acts on the observed indirectly associated variable Z_{IA}, which is part of the insurer's observed data. The logic is indirect: if Z_{IA} is one of the main pathways through which information about A enters the prediction system, then reducing the model's reliance on Z_{IA} can reduce ex post disparities across the audit groups defined by A.

The adversarial approach studied here is closely related to domain-adversarial training and gradient-reversal methods developed in the broader machine-learning literature; see Ganin et al. (2016) and Zhang, Lemoine, and Mitchell (2018).

Figure 51
ADVERSARIAL NEURAL NETWORK



Copyright © 2026 Society of Actuaries Research Institute

Section 6 Evaluation Design and Group-Audit Metrics

6.1 EVALUATION GOALS

Model evaluation in this report has two components. The first is predictive performance, meaning how well the model separates higher-risk from lower-risk observations. The second is ex post group audit, meaning how model-based decisions or selections differ across groups defined by the audit label A.

This distinction is central. The models are trained without using A. The indirectly associated variable Z_{IA} may enter model training because it is observed and belongs to the predictor set, and in the ADV-NN it is also the adversarial target. But the ex-post group-audit diagnostics reported in the evaluation stage are always computed with respect to A, not Z_{IA}.

Evaluating both predictive performance and ex post audit metrics captures different aspects of model quality that may not move together.

6.2 PREDICTIVE PERFORMANCE: ROC CURVE AND AUC

For binary outcomes, predictive performance is commonly evaluated using the **Receiver Operating Characteristic** (ROC) curve and the **Area Under the ROC Curve** (AUC). These tools measure how well a model distinguishes between observations that experience the adverse outcome ($Y = 1$) and those that do not ($Y = 0$).

The ROC curve is constructed by considering every possible threshold that could be used to classify observations as high risk or low risk based on the model's predicted score. For each possible threshold, two quantities are calculated: the true positive rate (the proportion of actual adverse cases correctly identified as high risk) and the false positive rate (the proportion of non-adverse cases incorrectly identified as high risk). The ROC curve plots the true positive rate on the vertical axis against the false positive rate on the horizontal axis as the threshold varies from the strictest to the most permissive classification rule.

The Area Under the ROC Curve (AUC) summarizes the entire ROC curve into a single number between 0 and 1. AUC has a useful probabilistic interpretation: it is the probability that a randomly selected observation with $Y = 1$ receives a higher predicted score than a randomly selected observation with $Y = 0$. In other words, AUC measures how often the model ranks truly higher-risk cases above lower-risk cases. An AUC of 0.5 corresponds to random ranking, meaning the model performs no better than random guessing. An AUC of 1.0 corresponds to perfect prediction, where every adverse case receives a higher predicted score than every non-adverse case. Values between these extremes reflect intermediate levels of the model's ability to correctly separate positive cases from negative cases power.

6.3 FIXED OPERATING POINTS: SELECTING THE TOP $X\%$ AS "HIGH RISK"

Group-audit metrics are evaluated at fixed operating points because operational decisions are typically applied to a small fraction of cases (for example, referral to manual review, additional documentation, a surcharge, coverage restrictions, or declination).

In this report, an operating point is defined by an action rate $x\%$ (for example, 1%, 2%, or 5%). The steps are:

1. The model assigns each case a risk score (often a predicted probability).
2. Cases are ranked from highest score to lowest score.
3. The "high-risk" action is applied to the top $x\%$ of cases.

Copyright © 2026 Society of Actuaries Research Institute

Let $\widehat{D}$ denote the indicator for being selected into the "high-risk" action (so $\widehat{D} = 1$ if selected, and $\widehat{D} = 0$ otherwise). The selection threshold is chosen so that the overall selection rate is approximately $x\%$, i.e.,

$$P(\widehat{D} = 1) \approx x\%.$$

Important clarification: "top $x\%$" refers to the model's ranking, not to a "true top $x\%$ risk group" observed in the data. In real applications (and in simulated data with binary outcomes), one observes only whether Y occurred, not each individual's underlying risk probability. The operating-point approach mimics how risk scores are used in real-world applications: decisions are made based on model rankings.

Also note that in many underwriting contexts, being flagged as "high risk" is an unfavorable action (more friction or a worse offer). In the health-risk application, however, selection may represent a triage or intervention decision rather than an unfavorable underwriting action. This direction matters for interpreting the fairness metrics below.

The particular action rates reported in this document (1%, 2%, and 5%) are chosen for illustration rather than because they are uniquely privileged thresholds. The operating points are varied as a sensitivity analysis. Across the reported applications, the qualitative ordering is broadly similar across operating points: the adversarial specification continues to reduce the reported gaps relative to the GLM and the standard neural network.

6.4 GROUP-AUDIT METRICS REPORTED

At each operating point, the report presents the following group-comparison metrics:

1. Demographic parity ratio. The ratio of the high-risk selection rate for Group $a$to the corresponding rate for Group b. Values closer to one indicate more similar selection rates.
2. Absolute True positive rate (TPR) gap. It compares how often truly adverse cases are correctly selected across groups. Values closer to zero indicate more similar true-positive rates.

Intuitively, demographic parity asks whether the two groups are flagged as high risk at similar rates. The TPR gap asks whether truly adverse cases are captured at similar rates across groups. For the unfavorable high-risk actions studied here, smaller absolute gaps and ratios closer to one indicate more similar operational treatment.

When $\widehat{D} = 1$ represents an unfavorable action, a demographic parity ratio above 1 means that group a is selected more often than group b at the same operating point. The TPR gap helps diagnose whether group differences reflect unequal detection of truly adverse outcomes. Mathematical specifications of these metrics are provided in Appendix B.

6.5 ESTIMATION, TUNING, AND AUDIT WORKFLOW

For each line of business, the simulation produces the audit label A, the observed feature partition $X = (Z_{\mathrm{IA}}, W)$, the designated observed indirectly associated variable Z_{IA}, and the outcome Y. The data are then divided into model-development and held-out evaluation samples.

The GLM and the standard NN are trained on the predictive task:

$$X \to Y.$$

The audit label A is not included in the training inputs.

The ADV-NN is trained on the same predictive task:

$$X \mapsto Y,$$

Copyright © 2026 Society of Actuaries Research Institute

but augments the predictive loss with an adversarial term defined using the observed indirectly associated variable Z_{IA}. Thus, the adversarial branch acts on information already contained in the observed data structure rather than on the audit label used later for ex post evaluation. In implementation terms, the model-development batches contain only the observed predictors, the supervised outcome, and the indirectly associated target for the adversarial branch.

Candidate values of λ are evaluated within the model-development sample using outcome-stratified cross-validation. The procedure first identifies the highest mean validation ROC AUC and retains candidates whose ROC AUC is within 0.005 of that value. Among the retained candidates, it selects the value with the largest validation adversary loss, indicating that the designated indirectly associated variable is hardest to recover from the learned representation while predictive ranking remains close to its best validation level. Ties are resolved using lower log loss, higher average precision, and smaller λ, in that order. The audit label A and the held-out group-audit metrics are not used in selecting λ.

After λ has been selected, the ADV-NN is refitted on the full model-development sample using the chosen value of λ. Only after the models have been fixed are predictive performance and group-audit diagnostics computed on the held-out evaluation sample. At that stage, the audit label A is brought back solely to summarize ex post differences in score distributions, selections, and error rates across groups. These quantities are therefore audit statistics of fitted-model outputs, not ingredients of the fitted score and not inputs to model selection.

The λ-sweep figures reported in the application sections are intended to be interpreted in the same ex post sense. Each point is obtained by refitting the ADV-NN at a prespecified value of λ and then auditing the fitted model afterward. The sweep is included to display the empirical accuracy–audit frontier and to show that the qualitative findings are not driven by one arbitrary value of λ. The held-out audit-label metrics are not used for model selection.

Copyright © 2026 Society of Actuaries Research Institute

Section 7 Conclusion

Across the five synthetic applications, the selected ADV-NN generally reduced the reported group-audit gaps while maintaining ROC AUC close to the benchmark levels. The magnitude of the change varied across applications, and mitigation did not eliminate all disparities.

These simulations are mechanism-based case studies rather than empirical estimates or legal findings. Their practical implication is that when observed insurance data are systematically distorted or indirectly associated with group membership, it is typically beneficial to evaluate model-validity concerns and group-level audit outcomes jointly.

Copyright © 2026 Society of Actuaries Research Institute

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 who generously shared their wisdom, insights, advice, and guidance, and for their diligent work overseeing, reviewing and editing this report for accuracy and relevance.

Project Oversight Group members:

Dorothy L Andrews, ASA, MAAA, PhD
Carlos Arocha, FSA
Joan Barrett, FSA, MAAA
Adrian Clark, FSA< MAAA
Andrew Clark, PhD
Jean-Marc Fix, FSA, MAAA
Robert Gomez, FSA, CERA, MAAA
Sven Haunerland, FSA, CERA, ACIA
Anton Kobelev, ASA, CERA
Sin Rong (Bernice) Lim, FSA, CERA, FCIA
Honora Lo, ASA, MAAA
Jennifer Loftus, FIA, FSIA
Chris Lombardi, FSA, CERA, MAAA
Marilyn McGaffin, ASA, MAAA
Tyson Mohr, FSA, ACAS, MAAA
Norman Niami, FCAS, MAAA
Nicole Persson, ASA, MAAA
Ronald PoonAffat, FSA, MAAA, FIA
Shisheng (Rose) Qian, FSA, CERA
Francis Rahil, FSA, CERA
Bruce B Rosner, FSA, MAAA
Jason Sears, FSA, CERA
Igor Shamshtein, FSA
Ren Zhong, ASA, MAAA

At the Society of Actuaries Research Institute:

Lisa A. Schilling, FSA, EA, FCA, MAAA, Director, Practice Research
Barbara Scott, Sr. Research Administrator

Copyright © 2026 Society of Actuaries Research Institute

Appendix A: Mathematical Specification of Models

A.1 GENERALIZED LINEAR MODEL

Let $X \in \mathbb{R}^p$ be the observed feature vector and let $Y \in \{0,1\}$ be the binary outcome. The predicted probability is denoted by $\hat{p}(X) = \hat{P}(Y = 1 \mid X)$.

The GLM uses a logistic link with linear predictor

$$\eta(X) = \beta_0 + X^\top \beta.$$

The fitted probability is

$$\hat{p}(X) = \sigma\{\eta(X)\} = \frac{1}{1 + \exp[-\eta(X)]}.$$

Here, $\eta(X)$ is the linear score and $\sigma(\cdot)$ maps that score to a probability in $(0,1)$.

A.2 STANDARD NEURAL NETWORK (EXAMPLE WITH TWO HIDDEN LAYERS)

A feed-forward neural network maps the observed feature vector X into a learned representation h. For example, with two hidden layers,

$$h_1 = \phi(W_1 X + b_1),$$
$$h_2 = \phi(W_2 h_1 + b_2),$$
$$h = h_2,$$

where $\phi(\cdot)$ is a nonlinear activation function. The prediction layer maps h into a fitted probability,

$$\hat{p}(X) = \sigma(W_o h + b_o).$$

Regularization methods, such as weight decay or dropout, may be used during training. They are omitted from the notation to keep the model description compact.

A.3 ADVERSARIAL NEURAL NETWORK (ADV-NN)

A.3.1 PREDICTOR AND OBSERVED INDIRECTLY ASSOCIATED ADVERSARY VARIABLE

The adversarial neural network uses the same observed feature vector X and the same binary outcome Y as the standard neural network. Let Z_{IA} denote the observed indirectly associated variable used in the adversarial branch. The audit label A is not used in ADV-NN training, model selection, or refitting. It is retained only for ex post evaluation after the fitted model has been fixed.

The predictor first learns a representation

$$h = f_{\theta_f}(X),$$

and then maps this representation into an outcome prediction

$$\hat{p}(X) = \sigma\{g_{\theta_y}(h)\}.$$

Copyright © 2026 Society of Actuaries Research Institute

The adversarial branch attempts to recover the observed indirectly associated variable Z_{IA} from the same representation h. The adversarial branch is outcome-conditional: one adversary is used for observations with $Y = 0$ and another adversary is used for observations with $Y = 1$. Thus,

$$\hat{q}_0(Z_{\mathrm{IA}} \mid h) = a_{\theta_{z0}}(h), \qquad \text{for observations with } Y = 0,$$

and

$$\hat{q}_1(Z_{\mathrm{IA}} \mid h) = a_{\theta_{z1}}(h), \qquad \text{for observations with } Y = 1.$$

When Z_{IA} is binary, each adversary is a two-class classifier. The use of outcome-conditional adversaries is intended to reduce recoverable information about Z_{IA} within outcome strata rather than only in the pooled sample.

A.3.2 LOSS FUNCTIONS AND TRAINING OBJECTIVE

Let $\mathcal{L}_Y$ denote the predictive loss for Y, and let $\mathcal{L}_Z$ denote the adversarial loss for predicting the observed indirectly associated Z_{IA}. For binary Y, the predictive loss is binary cross-entropy:

$$\mathcal{L}_Y = -\frac{1}{n}\sum_{i=1}^{n}[Y_i \log \hat{p}_i + (1 - Y_i)\log(1 - \hat{p}_i)].$$

For the adversarial branch, let $\ell_Z(\hat{q}, Z_{\mathrm{IA}})$ denote cross-entropy for predicting Z_{IA}. The outcome-conditional adversarial loss is

$$\mathcal{L}_Z = \frac{1}{n}\sum_{i:Y_i=0} \ell_Z\{\hat{q}_{0i}, Z_{\mathrm{IA}_i}\} + \frac{1}{n}\sum_{i:Y_i=1} \ell_Z\{\hat{q}_{1i}, Z_{\mathrm{IA}_i}\}.$$

The training objective can be written as the min–max problem

$$\min_{\theta_f, \theta_y \theta_{z0}, \theta_{z1}} \max \left[\mathcal{L}_Y(\theta_f, \theta_y) - \lambda \mathcal{L}_Z(\theta_f, \theta_{z0}, \theta_{z1})\right].$$

Equivalently, the adversaries are trained to predict Z_{IA} from h, while the feature extractor is trained to make that indirect association-prediction task more difficult while still preserving information useful for predicting Y. In implementation, this is achieved through a gradient-reversal layer. The parameter $\lambda \geq 0$ controls the strength of the adversarial penalty.

The audit label A does not appear in this objective. It is not included in X, is not used as an adversarial target, and is not used to select λ. After the GLM, NN, and ADV-NN have been fitted, A is brought back only to compute ex post audit diagnostics such as selection-rate ratios and true-positive-rate gaps.

A.4 IMPLEMENTATION DETAILS

The simulations use $100{,}000$ observations in each line of business. Each simulated dataset contains the binary outcome Y, the observed features X, the observed indirectly associated target Z_{IA}, and the audit label A. The audit label A is excluded from all model-fitting inputs and is used only after fitting.

For each line of business, the development sample is split into training and validation folds using outcome-stratified cross-validation. The same observed input structure is used for the GLM, the standard NN, and the ADV-NN. Continuous predictors are standardized within the fitting sample, and the same scaling transformation is applied to the validation or held-out sample.

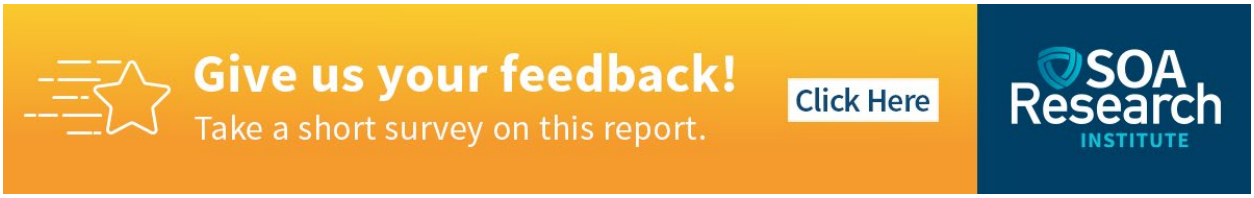
The neural-network models are implemented in PyTorch and trained with the Adam optimizer. A common training configuration is used across the five lines of business: the batch size is 1,024, the learning rate is 0.001, and the

Copyright © 2026 Society of Actuaries Research Institute

weight-decay parameter is 0.0001. The ADV-NN is trained for 10 epochs. The first five epochs constitute a prediction-only warm-up period, during which the predictor is trained on Y without activating the adversarial penalty. The adversarial objective is activated for the remaining five epochs. Thus, the warm-up is defined in epochs rather than as a fixed number of gradient steps. The elements that vary by line of business are the observed predictor set, the designated adversarial target Z_{IA}, and the candidate grid of λ values. Continuous predictors are standardized using parameters estimated from the relevant fitting sample and the same transformation is applied to the corresponding validation or held-out sample.

Candidate adversarial strengths are selected from a prespecified grid of λ values. For each candidate λ, the ADV-NN is trained on outcome-stratified fitting folds and evaluated on the corresponding validation fold. The validation summaries include log loss, ROC AUC, average precision, and the outcome-conditional validation adversary loss. The adversary is trained to predict Z_{IA} from the learned representation using cross-entropy loss, separately within the $Y = 0$ and $Y = 1$ strata. Under the common adversary architecture and training procedure, a larger validation adversary loss therefore means that Z_{IA} is more difficult to recover from the learned representation. The selection rule first retains candidate values whose cross-validated ROC AUC is within 0.005 of the best candidates and then, among those candidates, selects the value with the largest validation adversary loss. Ties are broken by lower log loss, higher average precision, and smaller λ, in that order. The selected value is then used to refit the ADV-NN on the full development sample before final held-out evaluation.

The ex post audit step is separate from model fitting. Once the fitted models are fixed, predicted scores are evaluated on held-out data. The audit label A is then used to compute group-comparison metrics at fixed operating points, such as the top 1%, 2%, and 5% of predicted risks. These audit metrics summarize fitted-model behavior and are not used as training targets.



Copyright © 2026 Society of Actuaries Research Institute

Appendix B: Mathematical Specification of Evaluation Metrics

All metrics in Appendix B are computed after model estimation using the audit label $\mathbf{A}$. They are ex post audit summaries. They are not training inputs, and they are not defined with respect to the adversarial indirectly associated variable Z_{IA}.

B.1 FIXED OPERATING POINT AND DECISION INDICATOR

Let $s(X_i)$ denote the model score used for ranking (for example, $s(X_i) = \hat{p}(X_i)$). For an operating point $r\%$, define a threshold τ_r such that selecting the highest-scoring cases yields an overall selection rate of approximately $r\%$. The decision indicator is:

$$\widehat{D} = \mathbb{1}\{s \geq \tau_r\}, \qquad P\big(\widehat{D} = 1\big) \approx r\%.$$

The indicator function $\mathbb{1}\{\cdot\}$ equals 1 if the condition inside braces is true and equals 0 otherwise.

B.2 AUC (RANKING PERFORMANCE)

Let S_1 denote the score for a randomly selected observation with $Y = 1$ and let S_0 denote the score for a randomly selected observation with $Y = 0$. Then:

$$\mathrm{AUC} = P(S_1 > S_0) + \frac{1}{2}P(S_1 = S_0).$$

Interpretation: AUC is the probability that a positive case receives a higher score than a negative case (with ties counted as half).

B.3 DEMOGRAPHIC PARITY (SELECTION RATE) METRIC

Demographic parity ratio:

$$R_{\mathrm{DP}} = \frac{P\big(\widehat{D} = 1 \mid A = \mathrm{a}\big)}{P\big(\widehat{D} = 1 \mid A = \mathrm{b}\big)}.$$

R_{DP} measures the gap on a multiplicative scale. Values of R_{DP} closer to 1 indicate more similar selection rates.

B.4 ABSOLUTE TRUE POSITIVE RATE GAP

True positive rate (TPR) within group a:

$$\mathrm{TPR}(a) = P\big(\widehat{D} = 1 \mid Y = 1, A = a\big).$$

TPR gap:

$$\Delta_{\mathrm{TPR}} = |\mathrm{TPR(a)} - \mathrm{TPR(b)}|.$$

Interpretation: Δ_{TPR} near 0 means the model selects truly adverse cases at similar rates across groups.

Copyright © 2026 Society of Actuaries Research Institute

Appendix C: Data-Generating Process of Life Insurance

C.1 NOTATION AND BASIC FUNCTIONS

Let $A \in \{a, b\}$ denote group membership, with Group a representing the disadvantaged group and Group b the reference group. Let T denote the binary indirectly associated variable, Z^* the unobserved credit-related underwriting score, Z the observed score, Age the applicant age, BMI the body-mass-index variable, S the smoker indicator, C the chronic-condition indicator, R the risky-occupation indicator, and Y the binary adverse-event indicator.

The logistic link is

$$\sigma(x) = \frac{1}{1 + \exp(-x)}.$$

C.2 GROUP AND INDIRECTLY ASSOCIATED VARIABLE GENERATION

Group membership is generated exogenously as

$$P(\mathrm{Group} = a) = 0.35.$$

Conditional on group membership, the indirectly associated variable T is generated as

$$P(T = 1 \mid \mathrm{Group} = a) = 0.60,$$

and

$$P(T = 1 \mid \mathrm{Group} = b) = 0.20.$$

This creates an indirect association because T is materially associated with Group even though Group itself is excluded from model fitting.

C.3 UNOBSERVED SCORE AND BIASED MEASUREMENT

The unobserved score is generated independently of group membership:

$$Z^* \sim N(700, 50^2).$$

The observed score is generated by applying a group-dependent downward shift and adding measurement noise:

$$Z = Z^* - \theta \mathbf{1}\{\mathrm{Group} = a\} + \varepsilon_Z, \qquad \theta = 40, \qquad \varepsilon_Z \sim N(0, 25^2).$$

The observed score is then clipped to the range $[300{,}850]$:

$$Z \leftarrow \min\{850, \max(300, Z)\}.$$

Because larger values of Z^* correspond to lower risk in the outcome model, subtracting 40 points for Group a makes that group appear riskier on the observed score scale, even though Z^* itself is group neutral.

C.4 COVARIATES AND ADDITIONAL OBSERVED RISK FACTORS

Age is generated from a truncated normal distribution:

$$\mathrm{Age} \sim N(45, 15^2) \quad \text{truncated to } [20{,}75].$$

The standardized age variable is

Copyright © 2026 Society of Actuaries Research Institute

$$\text{Age}_g = \frac{\text{Age} - 45}{10}.$$

BMI category probabilities differ slightly by group. For Group A, the normal, overweight, and obese category probabilities are 0.35, 0.375, and 0.275. For Group B, the corresponding probabilities are 0.40, 0.35, and 0.25.

$$P(BMI\ category = normal \mid Group\ a) = 0.35,\ P(BMI\ category = normal \mid Group\ b) = 0.40,$$
$$P(BMI\ category = overweight \mid Group\ a) = 0.375,\ P(BMI\ category = overweight \mid Group\ b) = 0.35,$$

and

$$P(BMI\ category = obese \mid Group\ a) = 0.275,\ P(BMI\ category = obese \mid Group\ b) = 0.25.$$

Conditional on the category, BMI is generated from a truncated normal distribution:

$$BMI \mid \text{normal} \sim N(22, 2.5^2) \quad \text{truncated to } [16,25],$$
$$BMI \mid \text{overweight} \sim N(27.5, 2.5^2) \quad \text{truncated to } [25,30],$$

and

$$BMI \mid \text{obese} \sim N(32, 3^2) \quad \text{truncated to } [30,45].$$

The standardized BMI variable is

$$BMI_g = \frac{BMI - 27}{5}.$$

Smoking status is generated as

$$S \sim \text{Bernoulli}\{\sigma(-1.10 + 0.10\text{Age}_g)\}.$$

Chronic-condition status is generated as

$$C \sim \text{Bernoulli}\{\sigma(-1.75 + 0.40BMI_g + 0.35\text{Age}_g)\}.$$

Risky occupation is generated from a territory-linked probability:

$$P(R = 1 \mid T) = 0.10 + 0.10T.$$

Thus, in the simulation, BMI has a mild group association, and risky occupation is linked to the indirectly associated territory variable T. The central embedded-bias mechanism remains the combination of the T indirect-association channel and the measurement-bias channel through Z, with Y generated from unobserved risk drivers rather than from Group directly.

C.5 OUTCOME MODEL AND INTERCEPT CALIBRATION

Outcome generation uses a logistic simulation model. The coefficient values are β_{Age} = 0.10, β_{BMI} = 0.10, β_{S} = 0.40, β_{C} = 0.80, β_{R} = 0.40, γ_{Z} = -1.50, β_{Age^2} = 0.05, β_{BMI^2} = 0.06, $\beta_{\text{Age,BMI}}$ = 0.08, $\beta_{S,\text{Age}}$ = 0.15, $\beta_{C,BMI}$ = 0.20, and γ_{Z^2} = 0.12. The linear predictor excluding the intercept is

$$\eta = \beta_{\text{Age}}\text{Age}_g + \beta_{BMI}BMI_g + \beta_S S + \beta_C C + \beta_R R + \gamma_Z \frac{Z^* - 700}{50} + \beta_{\text{Age}^2}\text{Age}_g^2 + \beta_{BMI^2}BMI_g^2 + \beta_{\text{Age},BMI}\text{Age}_g BMI_g$$
$$+ \beta_{S,\text{Age}} S\text{Age}_g + \beta_{C,BMI} CBMI_g + \gamma_{Z^2}\left(\frac{Z^* - 700}{50}\right)^2.$$

The intercept β_0 is chosen numerically so that the simulated event rate matches the target $\tau = 0.15$:

Copyright © 2026 Society of Actuaries Research Institute

$$\frac{1}{n}\sum_{i=1}^{n}\sigma\left(\beta_0+\eta_i\right)\approx 0.15.$$

The event probability is

$$P(Y=1\mid Z^*,\text{Age},BMI,S,C,R)=\sigma(\beta_0+\eta),$$

and the binary outcome is sampled as

$$Y\sim\text{Bernoulli}\{\sigma(\beta_0+\eta)\}.$$

The outcome equation does not include Group or T as direct predictors. It is driven by the unobserved score Z* and observed risk factors Age, BMI, S, C, and R. Because R is generated with a higher probability when T = 1, this simulation includes a territory-linked observed risk factor in addition to the Z measurement-bias pathway.

C.6 WHERE GROUP DISPARITIES ENTER

Group disparities can arise through the indirectly associated variable T, the biased observed score Z, and the territory-linked risky-occupation feature R.

The fitted model observes Z rather than Z*. It may therefore treat the downward shift in Z for Group A as a genuine risk difference. This remains the central embedded-bias mechanism in the life-insurance simulation, while the study allows BMI composition and risky occupation to vary with group or territory as described above.

Copyright © 2026 Society of Actuaries Research Institute

Appendix D: Data-Generating Process of Auto Insurance

D.1 NOTATION AND BASIC FUNCTIONS

Let A denote the audit group, with Group a representing the disadvantaged group and Group b the reference group. Let T denote territory, Age driver age, Young the youthful-driver indicator, M annual mileage, $M_h = log(1 + M / 12{,}000)$ the transformed mileage measure, V^* the unobserved violation count, V the recorded violation count, P_prior the prior at-fault claim flag, and Y the next-period claim indicator. Let $\sigma(x) = 1 / (1 + exp(-x))$ denote the logistic link.

D.2 GROUP AND TERRITORY GENERATION

A is generated exogenously with P(Group a) = 0.35. Conditional on A, territory is generated with P(T = 1 | Group a) = 0.65 and P(T = 1 | Group b) = 0.15. Territory therefore functions as an observed indirectly associated variable because it is correlated with the audit group even though A itself is excluded from model fitting.

D.3 AGE, MILEAGE, AND TRANSFORMED EXPOSURE

Age is generated from a truncated normal distribution centered at 40 with standard deviation 12 on the interval $[18, 85]$. The youthful-driver indicator equals $1_{\{Age < 25\}}$. Annual mileage is generated from a gamma distribution calibrated so that mileage is centered near 12,000 miles. The model uses the transformed exposure measure $M_h = log(1 + M / 12{,}000)$, which allows the effect of mileage to increase with exposure but at a diminishing rate.

D.4 UNOBSERVED VIOLATIONS, RECORDED VIOLATIONS, AND PRIOR CLAIMS

The unobserved violation count is generated from a Poisson model with log-rate $\log \lambda^* = c_0 + c_M M_h + c_Y Young$. Territory does not enter this unobserved violation equation. Recorded violations are then constructed as V = V* + Δ, where Δ ~ Poisson(2.00 T). The surveillance uplift therefore affects the observed violation count only in T = 1 territories. The prior at-fault claim flag is generated from a logistic model driven by log(1 + V*) and Young, again without a direct territory term.

D.5 OUTCOME MODEL AND INTERCEPT CALIBRATION

The next-period claim indicator is generated from a logistic model with linear predictor $\eta = \beta_M M_h + \beta_V log(1 + V^*) + \beta_Y Young + \beta_P P_{prior} + \varphi M_h log(1 + V^*)$. The intercept is chosen numerically so that the average claim rate is approximately 15%. Conditional on η, the claim probability is $P(Y = 1) = \sigma(\beta_0 + \eta)$, and Y is sampled from a Bernoulli distribution with that probability.

D.6 WHERE GROUP DISPARITIES ENTER

The outcome equation for Y does not include A explicitly, and it also omits a direct territory term. Group disparities can nevertheless emerge through two channels. First, the indirect association channel: A influences territory T, and T is part of the observed predictor set. Second, the measurement channel: the fitted model observes recorded violations V rather than unobserved violations V^*, and the difference $V - V^*$ is larger in the higher-surveillance territory. This structure isolates the fairness question to territory-linked recording distortion rather than true territory differences in unobserved claim risk.

Copyright © 2026 Society of Actuaries Research Institute

Appendix E: Data-Generating Process of Credit Insurance

E.1 NOTATION AND BASIC FUNCTIONS

Let $A_i \in \{a, b\}$ denote group membership, let Z_i denote the ZIP-code indirectly associated variable, let $S_i^\star$ denote unobserved credit strength, let S_i denote the observed score used by the model, let D_i denote debt-to-income, let L_i denote the recent delinquency count, and let $Y_i \in \{0,1\}$ denote one-year default. Define

$$\sigma(x) = \frac{1}{1 + \exp(-x)}.$$

Standardize the unobserved score:

$$\tilde{S}_i^\star = \frac{S_i^\star - \mathbb{E}[S^\star]}{\mathrm{sd}(S^\star)}.$$

E.2 GROUP AND INDIRECTLY ASSOCIATED VARIABLE GENERATION

The parameterization used for the reported figures sets $n = 100{,}000, P(A = a) = 0.35, P(Z = 1 \mid A = a) = 0.65$, and $P(Z = 1 \mid A = b) = 0.25$. Let $G_i = 1_{\{A_i = a\}}, G_i \sim Bernoulli(0.35)$, in symbolic form,

$$Z_i \mid G_i \sim \mathrm{Bernoulli}(0.65 G_i + 0.25(1 - G_i)).$$

In the simulation, the disadvantaged share is roughly one-third of the portfolio and the indirect association is substantially more common in the disadvantaged group.

E.3 UNOBSERVED CREDIT STRENGTH AND BIASED OBSERVATION

Unobserved credit strength is generated as

$$S_i^\star \sim \mathcal{N}(\mu_S, \sigma_S^2),$$

independently of group. Measurement noise is drawn as

$$\varepsilon_i^S \sim \mathcal{N}(0, \tau_S^2),$$

and the observed score used by the model is

$$S_i = S_i^\star - \theta\, \mathrm{G_i} + \varepsilon_i^S.$$

Here $\theta > 0$ represents the group-specific downward shift in observed credit quality. For the reported simulation, μ_S= 700, σ_S = 60, τ_S = 25, and θ = 45.

E.4 OTHER COVARIATES

Debt-to-income is generated from a lognormal distribution and clipped to a plausible range:

$$\log D_i \sim \mathcal{N}(\mu_D, \sigma_D^2),$$

$$D_i = \min\{1.50, \max\{0.05, \exp(\log D_i)\}\}.$$

Recent delinquency count is generated as

$$L_i \sim \mathrm{Poisson}(\lambda_L).$$

Copyright © 2026 Society of Actuaries Research Institute

These covariates are generated independently of group so that the primary source of disparity is the biased observed score together with the correlated indirectly associated variable.

E.5 OUTCOME MODEL AND INTERCEPT CALIBRATION

The one-year default outcome is generated from a logistic model driven by the standardized unobserved score, debt-to-income, and delinquency history. For the reported simulation, α_S = -1.20, α_D = 1.25, and α_L = 0.60:

$$\eta_i^{(0)} = \alpha_S \tilde{S}_i^{\star} + \alpha_D D_i + \alpha_L L_i.$$

The intercept α_0 is chosen numerically so that the average default probability matches the target default rate τ_{target} specified in the simulation configuration:

$$\frac{1}{n}\sum_{i=1}^{n} \sigma\left(\alpha_0 + \eta_i^{(0)}\right) \approx \tau_{\text{target}}.$$

Conditional on α_0,

$$p_i = \sigma\left(\alpha_0 + \eta_i^{(0)}\right),$$

$$Y_i \sim \text{Bernoulli}(p_i).$$

In the simulation, the target default rate is 12%, creating a realistic but moderately imbalanced classification problem.

E.6 WHERE GROUP DISPARITIES ENTER

The outcome model for Y_i does not include A_i directly. Nevertheless, group disparities can emerge through two channels. First, the indirect association channel: the ZIP variable Z_i is correlated with A_i and can therefore leak group information into the model representation. Second, the measurement channel: the fitted model observes S_i rather than $S_i^{\star}$, and S_i is shifted downward for the disadvantaged group. This design mirrors a credit-setting concern in which the model learns from a biased observed score even though true default risk is generated from a group-neutral unobserved credit factor.

Copyright © 2026 Society of Actuaries Research Institute

Appendix F: Data-Generating Process of Mortgage Insurance

F.1 NOTATION AND BASIC FUNCTIONS

Let A denote the audit group, with Group a representing the disadvantaged group and Group b the reference group. Let Z denote the ZIP / neighborhood indirectly associated variable, $FICO^*$ the unobserved group-neutral credit score, FICO the observed score used by the model, LTV loan-to-value, DTI debt-to-income, and Y the binary default outcome. Let $\sigma(x) = 1 / (1 + exp(-x))$ denote the logistic link.

F.2 GROUP AND INDIRECTLY ASSOCIATED VARIABLE GENERATION

A is generated exogenously with $P(Group\ a) = 0.35$. Conditional on A, the ZIP indirectly associated variable is generated with $P(Z = 1 \mid Group\ a) = 0.60$ and $P(Z = 1 \mid Group\ b) = 0.20$. This creates a neighborhood-based indirect association through which group status can be partially inferred from observable features even though A itself is excluded from the fitted model.

F.3 UNOBSERVED CREDIT SCORE AND BIASED MEASUREMENT

The unobserved score $FICO^*$ is generated independently of A from a bimodal mixture intended to mimic the common empirical pattern of a lower-score band and a higher-score band. The observed score is then generated as $FICO = clip(FICO^* - 35 \cdot 1_{\{A=a\}} + \varepsilon, 300, 850)$, where ε is additive normal noise with standard deviation 25. Because lower observed FICO implies higher predicted default risk, this measurement equation makes Group a appear systematically riskier than an otherwise identical borrower with the same unobserved credit quality.

F.4 COVARIATES (LTV AND DTI)

LTV and DTI are generated as additional mortgage-risk drivers that depend on unobserved credit strength but not directly on A. LTV is generated so that stronger unobserved credit is associated with lower leverage on average. DTI is generated from a right-skewed lognormal construction so that stronger unobserved credit is associated with lower repayment burden on average.

F.5 OUTCOME MODEL AND INTERCEPT CALIBRATION

The default outcome is generated from a logistic model driven by unobserved credit quality and the two financial ratios. Using centered covariates, the linear predictor is $\eta = -1.15 \cdot std(FICO^*) + 3.0 \cdot (LTV - 0.80) + 2.3 \cdot (DTI - 0.35) + 1.0 \cdot (LTV - 0.80)(DTI - 0.35)$. The intercept is chosen numerically so that the average default rate is approximately 20%. Conditional on η, the default probability is $P(Y = 1) = \sigma(\alpha_0 + \eta)$, and Y is sampled from a Bernoulli distribution with that probability.

F.6 WHERE GROUP DISPARITIES ENTER

The outcome model for Y does not include A explicitly. Group disparities can nevertheless emerge through two channels. First, the indirect association: A influences Z, and Z is included among the observed predictors. Second, the measurement channel: the fitted model observes $FICO$ rather than $FICO^*$, and $FICO$ is systematically shifted downward for Group a. The design therefore mirrors a mortgage-underwriting concern in which audit-group attributes are unavailable for pricing, yet correlated indirectly associated variables and biased measurements can still create disparities in fitted-model outputs.

Copyright © 2026 Society of Actuaries Research Institute

Appendix G: Data-Generating Process of Health Insurance

G.1 NOTATION AND BASIC FUNCTIONS

Let $A_i \in \{0,1\}$ denote group membership, where $A_i = 1$ corresponds to Group a and $A_i = 0$ to Group b. Let ADI_i denote the high-area-deprivation indirectly associated variable, let Age_i denote age, let Female_i denote sex, let U_i denote unobserved frailty, let C_i denote chronic-condition status, let $M_i^\star$ denote unobserved morbidity, let $M_{i,\mathrm{obs}}$ denote observed morbidity, and let $Y_i \in \{0,1\}$ denote the binary adverse event. Define

$$\sigma(x) = \frac{1}{1+\exp(-x)}.$$

Scaled age is defined as

$$\mathrm{AgeScaled}_i = \frac{\mathrm{Age}_i - 45}{10}.$$

G.2 GROUP, INDIRECTLY ASSOCIATED VARIABLE, AGE, SEX, AND FRAILTY GENERATION

Group membership is generated as

$$A_i \sim \mathrm{Bernoulli}(0.25).$$

Conditional on group, the area-deprivation indirectly associated variable is generated as

$$\mathrm{ADI}_i \mid A_i \sim \mathrm{Bernoulli}(0.55\,1\{A_i = 1\} + 0.25\,1\{A_i = 0\}).$$

Age is generated from a truncated normal distribution,

$$\mathrm{Age}_i \sim \mathrm{TruncNorm}(45, 15^2; 18,85),$$

sex is generated as

$$\mathrm{Female}_i \sim \mathrm{Bernoulli}(0.50),$$

and unobserved frailty is generated as

$$U_i \sim \mathcal{N}(0,1).$$

G.3 CHRONIC CONDITION AND UNOBSERVED MORBIDITY

Chronic-condition status is generated from a logistic model that depends on age, sex, and frailty but not directly on group:

$$p_{C,i} = \sigma(-0.5 + 0.45\,\mathrm{AgeScaled}_i + 0.10(\mathrm{Female}_i - 0.5) + 0.40\,U_i),$$

$$C_i \sim \mathrm{Bernoulli}(p_{C,i}).$$

Unobserved morbidity is then generated as

$$\varepsilon_i^M \sim \mathcal{N}(0, 0.45^2),$$

$$M_i^\star = 0 + 0.90\,U_i + 0.35\,\mathrm{AgeScaled}_i + 0.80\,C_i + \varepsilon_i^M.$$

Copyright © 2026 Society of Actuaries Research Institute

Thus, unobserved morbidity is group-neutral conditional on the simulated covariates and frailty.

G.4 OBSERVED MORBIDITY AND MEASUREMENT BIAS

The observed morbidity score used by the fitted model is generated as

$$\varepsilon_i^{\text{obs}} \sim \mathcal{N}(0, 0.35^2),$$

$$M_{i,\text{obs}} = M_i^\star + 2.20\, A_i + \varepsilon_i^{\text{obs}}.$$

Because the observed morbidity score is shifted upward for Group a, the disadvantaged group appears systematically less healthy on the observable scale even when unobserved morbidity is the same.

G.5 OUTCOME MODEL AND INTERCEPT CALIBRATION

Unobserved morbidity is standardized before the binary outcome is generated and then generates the binary adverse event from a logistic model:

$$\widetilde{M}_i^\star = \frac{M_i^\star - \mathbb{E}[M^\star]}{\text{sd}(M^\star)},$$

$$\eta_i^{(0)} = 0.8\, \widetilde{M}_i^\star + 0.70\, \text{AgeScaled}_i + 0.25\, C_i.$$

The intercept β_0 is chosen numerically so that the average event probability matches the target rate $\tau = 0.12$:

$$\frac{1}{n}\sum_{i=1}^{n} \sigma\left(\beta_0 + \eta_i^{(0)}\right) \approx 0.12.$$

Conditional on β_0,

$$p_i = \sigma\left(\beta_0 + \eta_i^{(0)}\right),$$

$$Y_i \sim \text{Bernoulli}(p_i).$$

G.6 WHERE GROUP DISPARITIES ENTER

The outcome model for Y_i does not include group membership directly. Group disparities can nevertheless emerge because the fitted model is trained on observed predictors that include both $M_{i,\text{obs}}$ and ADI_i. The indirect association comes from the correlation between ADI_i and group membership, while the measurement channel comes from the upward shift in $M_{i,\text{obs}}$ for Group a. These channels mirror realistic settings in which biased observed health-status measures and neighborhood variables can create disparate outcomes even when the true event mechanism is group-neutral.

Copyright © 2026 Society of Actuaries Research Institute

References

Barrett, J. C., & Entzminger, S. D. (2022). How does where you live impact your health? Society of Actuaries Research Institute.

Eling, M., Nuessle, D., & Staubli, J. (2022). The impact of artificial intelligence along the insurance value chain and on the insurability of risks. The Geneva Papers on Risk and Insurance—Issues and Practice, 47(2), 205–241.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59), 1–35.

Krafcheck, E., Balnozan, I., & Huang, F. (2026). Fairness metrics for life insurance. Society of Actuaries Research Institute.

Owen, R., & Teppema, S. (2023). Social, physical, and cultural determinants of health: Their incorporation into health care data and actuarial workstreams. Society of Actuaries Research Institute.

Smith, L. T., Pirchalski, E., & Golbin, I. (2022). Avoiding unfair bias in insurance applications of AI models. Society of Actuaries Research Institute.

Yi, R., Milton-Hall, J., Gersch, N. R., Merkey, B., & Nicholson, M. (2025). Integration of social determinants of health into Medicaid managed care risk adjustment: Considerations and financial impacts. Society of Actuaries Research Institute.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 335–340. https://doi.org/10.1145/3278721.3278779.

Copyright © 2026 Society of Actuaries Research Institute

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

Copyright © 2026 Society of Actuaries Research Institute