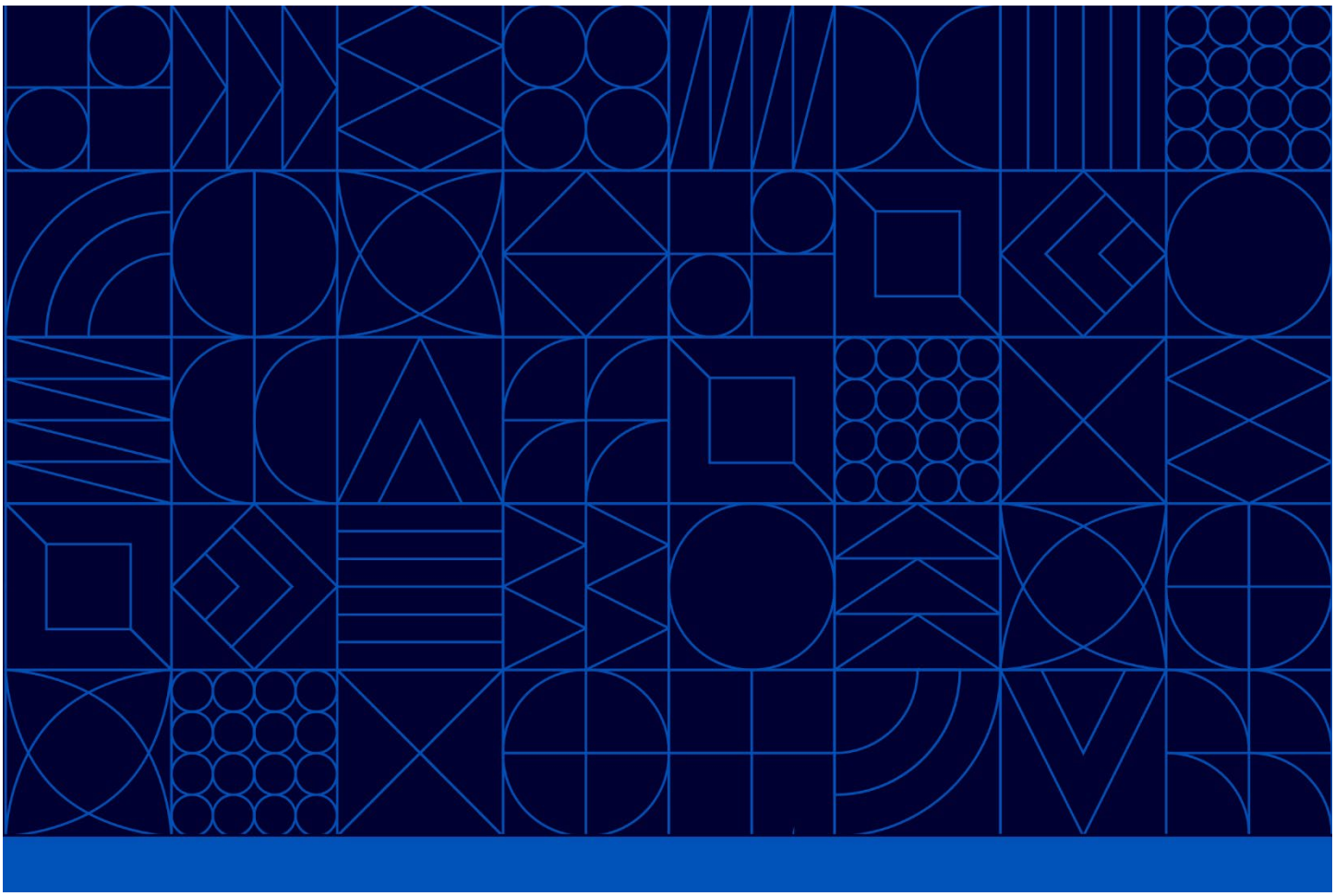
August | 2026

Evaluating Representativeness in Healthcare Claims Data

A Framework for Representativeness-Based Fairness in Actuarial Applications of AI



SOA Research INSTITUTE
Societal Purpose

SOA Research INSTITUTE
Societal Purpose

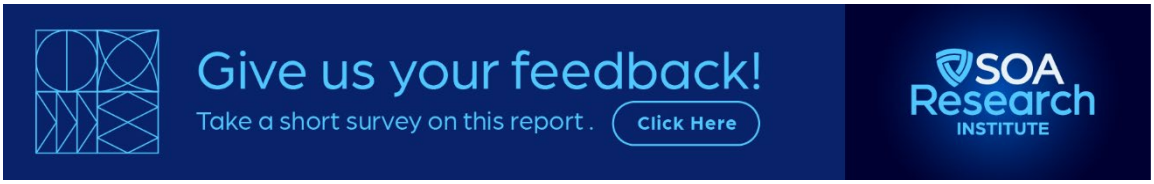
Evaluating Representativeness in Healthcare Claims Data

A Framework for Representativeness-Based Fairness in Actuarial Applications of AI

Authors Robert D. Lieberthal, PhD

Stacy Chen, MPH

Jawand Singh, BS



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

Contents **3**

Executive Summary **4**

Section 1 Introduction: Background and Literature Review **6**
1.1 Introduction and Background 6
1.2 Bias Across AI/ML Model Development 7
1.3 Data Characteristics in Actuarial Practice 8
1.4 Methodological Approaches to Addressing Non-representative Data 9
1.5 Recommendations Based on Key Literature 11

Section 2 Representativeness Assessment **13**
2.1 Conceptual Framework 13
2.2 Three-Stage Bias Assessment Method 14
2.3 Data Requirements for Representativeness Assessment 17

Section 3 Actuarial Application **19**
3.1 Actuarial Practice in the AI Era 19
3.2 Pricing Applications 20
3.3 Risk Adjustment 20
3.4 Utilization Management and Operational Applications 21
3.5 Provider Payment and Episode-Based Models 21

Section 4 Data Generation **22**
4.1 Synthetic Data for Research Application 22
4.2 Scenarios Modeled 25

Section 5 Representativeness Results **27**
5.1 Overall Results 27
5.2 Scenario 1 27
5.3 Scenario 2 29
5.4 Scenario 3 31

Section 6 Model Card and Decision Assistance Tool **36**
6.1 Introduction to Model Cards 36
6.2 Model Card Generator and Decision Assist Tool for Actuarial Applications 37
6.3 Decision Assistance Tool 42
6.4 Actuarial Practice Extensions 43

Section 7 Implications and Conclusions **45**
7.1 Limitations 45
7.2 Tips for Implementation 45
7.3 Implementation Guidance 47
7.4 Implementation with R and Python 47
7.5 Extension to Other Practice Areas 48

Section 8 AI Note **49**

Section 9 Acknowledgments **50**

Appendix: Open-Source Packages for Implementation **51**

References **52**

About The Society of Actuaries Research Institute **56**

Copyright © 2026 Society of Actuaries Research Institute

Executive Summary

This research demonstrates that representativeness in healthcare claims data and potential bias are central determinants of whether artificial intelligence (AI) and machine learning (ML) models are fit for actuarial use. Representation is the extent to which the data used to train a model is reflective of the true population and disease burden the model is intended to serve. Healthcare administrative claims reflect who accessed care, which may differ from those who needed it. Populations facing barriers to care, such as rural residents, those with limited income, and those systematically underdiagnosed in clinical settings, may appear in claims data at rates that do not reflect their true disease burden. When predictive models are trained on non-representative data, models may learn and amplify these gaps, producing outputs that may be accurate on average while failing specific subgroups.

This report presents a framework for evaluating and addressing representativeness in healthcare claims data used for AI and machine learning models in actuarial applications. The framework organizes bias assessment across three stages during which bias can potentially be introduced. In this study, unless specifically noted otherwise, "bias" refers to a systematic difference in how a model represents, predicts for, or performs across patient populations (Hasanzadeh et al., 2025). Other named forms of bias discussed in this report (for example, confirmation bias) are identified explicitly where they appear. The three stages where bias can enter models described in this report are: *input bias* arising from non-representative training data, *model bias* arising from algorithmic choices, and *application bias* arising from deployment context. A biased model can appear acceptably accurate when evaluated overall while failing materially for underrepresented subgroups, creating actuarial risk when used for pricing, screening, or care management applications. Each type of bias, as well as potential mitigation techniques are demonstrated through three synthetic data scenarios.

Across these scenarios, this study demonstrates concrete findings:

- Propensity score matching methods partially recover accuracy but leave residual bias in urban-rural disparities;
- Missingness-aware imputation substantially outperforms standard deletion approaches; and
- Representation mitigation results from explicit tuning to specific definitions rather than acceptance of the default application without future analysis.

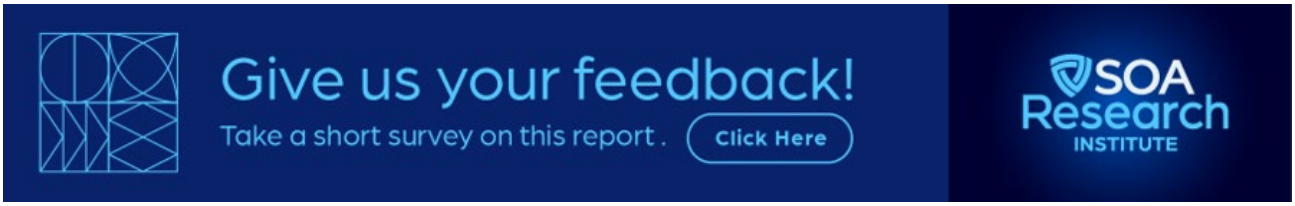
These methodological insights build on the representativeness framework and fairness assessment principles established in the literature review, which provide the theoretical foundation linking data quality to model performance across populations. Together, these scenarios demonstrate that standard performance metrics may systematically conceal subgroup disparities; that the appropriate mitigation strategy depends on which form of impact is prioritized; and that intersectional assessment is necessary when multiple demographic factors jointly determine access and utilization. For each scenario, this research assesses and illustrates specific bias mitigation tools; a list of these tools appears in the appendix.

This project also provides a Model Card and Decision Assist Tool for the diagnosis, mitigation, and documentation of bias as part of the modeling process. "Model cards" are a common

Copyright © 2026 Society of Actuaries Research Institute

method in the technology industry for demonstrating the methods used to develop a model or software, as well as actions taken to ensure high-quality output. The model card generator is part of an Excel workbook that guides practitioners through documenting model variants across structured sections. That workbook also includes a decision tool function that compares up to three model variants and suggests both which performs the best and how much improvement may be needed to get high quality model output.

Effective representation bias mitigation depends on the type of representativeness failure, the actuarial context, and the impact being prioritized. The Model Card and Decision Assist Tool provided here will help actuaries exercise professional judgment, comply with modeling requirements, and deploy AI systems that work for a plan or company as a whole and are reliable in the populations and subpopulations they are intended to serve.



Copyright © 2026 Society of Actuaries Research Institute

Section 1 Introduction: Background and Literature Review

This literature review examines data representativeness challenges in AI/ML models used for actuarial healthcare applications. Administrative claims data often under-represents or mis-represents populations facing barriers to care. This is a data quality issue that introduces bias and can distort risk adjustment, insurance pricing, and care management algorithms. A literature review of prior research identifies ways in which bias may propagate through three stages across the AI/ML development lifecycle. Methodological solutions for mitigation include propensity score methods, entropy balancing, and metrics aligned with equal performance principles. Representativeness metrics, missing data treatment, and validation approaches can support actuaries in evaluating the reliability of AI/ML models, as well as the application of methods for bias assessment, mitigation, and methodology documentation, as described in Sections 3–5.

1.1 Introduction and Background

1.1.1 Project Motivation and Rationale

The Society of Actuaries Research Institute (SOA) commissioned this research to fill a gap: as AI and ML tools are adopted rapidly across health insurance and actuarial practice, there is limited practical guidance for actuaries on how to evaluate whether the data used to train these models adequately represents the populations they will serve. Administrative healthcare claims data are generated through healthcare utilization rather than population-level sampling, meaning they may systematically exclude or undercount populations that face access barriers. When predictive models are trained on such data without accounting for these gaps, they can perform well on average while having a detrimental impact on specific subgroups, creating both actuarial risk and potential for unintended harm (Society of Actuaries, 2025).

In response, this research develops and demonstrates a practical, three-stage framework for identifying and addressing representativeness failures in healthcare claims data used for AI/ML model development. The framework is designed to be actionable and provides actuaries with quantitative tools for diagnosing bias at the data input, model development, and deployment stages. This report illustrates these tools through controlled synthetic data scenarios built on realistic clinical pathways and introduces a structured model card and decision assistance tool that actuaries can adopt to document and manage representativeness risk. The approach taken by the research team prioritizes methods that translate directly to operational actuarial practice, including open-source implementations in R and Python, a practical Excel-based tool, and examples drawn from health insurance applications (note that many, if not most, of these packages are available in proprietary statistical and actuarial tools).

1.1.2 Prior Work on AI, ML, and Bias

A recent SOA panel discussion characterized AI as a family of techniques in which machines perform tasks that would otherwise require human reasoning, with ML representing the subset that learns from historical data to generate predictions (Poon Affat, 2025). ML-based tools have been applied across a range of operational contexts in insurance and rely on historical data to generate predictions. These approaches include traditional methods, such as regression-based and tree-based models, as well as deep learning. Interest and use of AI/ML have increased

Copyright © 2026 Society of Actuaries Research Institute

exponentially, with a 2024 survey indicating 84% of health insurers reporting use in some capacity (National Association of Insurance Commissioners (NAIC), n.d.). These algorithms are increasingly applied to a range of applications, such as disease management programs, prior authorization, and fraud claim detection.

The growth of AI/ML in healthcare has outpaced the development of policies for representation and reliability, among others. When predictive models are trained on biased data, they can aggravate disparities in healthcare resource allocation, treatment recommendations, and actuarial decision-making. In particular, the SOA AI Working Group emphasized the importance of "ensuring AI-driven models do not introduce unintended biases that could lead to unfair discrimination in healthcare outcomes" (Poon Affat, 2025).

Healthcare administrative claims data reflect healthcare utilization, which may differ from population health needs, thus creating a fundamental challenge. Since claims data reflect only those who successfully accessed and navigated the healthcare system, they exclude those facing structural barriers to care, such as rural populations or uninsured individuals (Davis et al., 2022). Additionally, clinical bias can be perpetuated due to systematically underdiagnosed or misdiagnosed conditions among certain groups or disease states. In general, coding disparities, such as inconsistent coding quality, administrative burden, and upstream clinical documentation problems (for example, incomplete notes, ambiguous terminology, or colloquial descriptions), also reduce the validity of data (Lucyk et al., 2017). As a result, AI/ML models using claims data can amplify pre-existing disparities.

1.2 Bias Across AI/ML Model Development

In this study, "bias" refers to a systematic difference in how a model represents, predicts for, or performs across patient populations, and particularly differences that disadvantage groups already facing barriers to care (Hasanzadeh et al., 2025). Bias across model development can consist of input bias (data issues), model bias, or application bias, as discussed in the following section.

1.2.1 Input Bias Due to Non-Representative Training Data

Researchers have repeatedly found that healthcare datasets used in AI/ML development are predominantly derived from nonrepresentative populations. For example, across 6,000+ genome-wide studies, 90% of the participants were of European descent (Cheung et al., 2025). Bozkurt et al. (2020) found that nearly two-thirds of reviewed studies omitted race and ethnicity from reporting, demonstrating how commonly these variables are absent from clinical AI literature. Further, "...studies mentioned these often did not report if they were included as model inputs." Additionally, measurement bias due to inconsistent diagnostic practices, scanner/protocol variation across healthcare organizations, and variable coding inconsistencies across healthcare organizations can further introduce unintended errors into training data (Lucyk et al., 2017; Hasanzadeh et al., 2025).

1.2.2 Model Bias

Bias can also occur through the modeling choices and assumptions made by analysts. While all statistical models are based on a set of assumptions, many ML and "deep learning models...operate as "black boxes," where the outputs are presented as a fait accompli without

Copyright © 2026 Society of Actuaries Research Institute

the ability to scrutinize the models' underlying decision-making processes" (Cheung et al., 2025). Outputs may be presented without the understanding of the model's underlying decision-making processes (Hasanzadeh et al., 2025). One well-known example is a commercial algorithm that underestimated the health needs of a cohort of patients because healthcare spending was used as a proxy for disease burden, which did not consider the barriers to healthcare access that the cohort encountered (Obermeyer et al., 2019).

Another major analytical issue for analysts is the treatment of missing data. Missing covariates are more common among patients with less access to care. For example, patients may not have annual checkups if they are not insured. The literature also finds evidence of confirmation bias among practitioners: missing lab results tend to correspond to healthy results as doctors don't collect or record data if they are not deemed relevant (Smith, 2021). For all these reasons, the standard assumption that data are Missing Completely at Random (MCAR) would lead to bias. This causes patients to be disproportionately excluded when analysts drop missing cases or use assumptions that are not based on data or experience when implementing imputation methods (Jeanselme et al., 2022). Evidence shows that no single imputation strategy, including mean, median, nearest neighbors, or multiple imputation approaches, reliably reduces bias, and exclusion disproportionately removes minoritized populations from the learning data (Jeanselme et al., 2022).

1.2.3 Application Bias: Deployment in Clinical and Insurance Settings

Predictive outputs may misinform clinical decisions due to the problem of "application bias." Application bias refers to the misapplication of results, even when they are accurate, or the failure to address inaccuracies before model deployment. For example, automation bias, in which users over-trust models, has been demonstrated in clinical settings: radiologists of all experience levels showed decreased diagnostic accuracy when exposed to incorrect imaging suggestions from AI (Hasanzadeh et al., 2025). This demonstrates that even highly experienced professionals are vulnerable to over-reliance on model recommendations. Similarly, concerns have emerged in insurance settings. Auditors report that staff frequently accepted model-recommended denials without reviewing documentation (Ross, 2024).

Dismissal bias, in which users ignore high false-positive alert systems, can occur in clinical and payer settings. For example, a study of intensive care units (ICUs) finds that clinicians ignored the majority of the 1.1 million arrhythmia alarms due to high false-positive rates (Hasanzadeh et al., 2025). Similar patterns emerge in fraud detection systems where claims analysts disregard automated alerts due to high false-positive rates. Finally, feedback loop bias can occur when a user's flawed actions that were informed by model recommendations that are not validated, and are then incorporated back into training data, which reinforces distortions (Hasanzadeh et al., 2025).

1.3 Data Characteristics in Actuarial Practice

Actuarial work in healthcare relies on administrative claims data, risk scoring, and predictive models, with increasing use of AI/ML tools. Working with these models raises familiar questions actuaries address in traditional practice: Is the data appropriate for the intended purpose? How reliable are model outputs? What limitations affect interpretation? The framework presented here addresses these questions for AI/ML applications using claims data.

Copyright © 2026 Society of Actuaries Research Institute

- **Data Assessment.** Assessing claims data requires evaluation of whether available information is sufficient for the analytical purpose, whether data reflect populations to which results will be applied, and whether systematic limitations introduce bias. When models are trained on datasets that underrepresent populations facing access barriers, coding variations, or enrollment gaps, documentation of these deficiencies helps practitioners assess whether limitations are material for specific applications. The bias assessment framework in Section 2 provides quantitative approaches for characterizing representation gaps.
- **Modeling Considerations.** Model development requires attention to input selection, validation approaches, and performance assessment. For AI/ML models using claims data, this includes evaluating whether learned relationships remain stable across populations and whether models produce materially different results for similar risks. Effective validation would assess performance across relevant subgroups, not just overall metrics. The model bias mitigation approaches in Section 2.2.2 address these concerns.
- **Communication and Documentation.** Clear documentation supports appropriate model use through transparent description of data sources, explicit statement of assumptions, acknowledgment of limitations, and specification of intended applications. Model cards (Section 6) provide structured templates for organizing this information to enable practitioners to assess whether a model is fit for their specific purpose.
- **Application Context.** Different actuarial applications impose different requirements. Pricing emphasizes calibration accuracy, risk adjustment requires complete diagnostic capture, and utilization management needs consistent performance across access patterns. Section 2.3 discusses how these context-specific requirements affect what validation evidence is needed and whether documented limitations are acceptable for intended use.

1.4 Methodological Approaches to Addressing Non-Representative Data

1.4.1 Representativeness Metrics

To address non-representative data, analysts can specify the target population to which inferences are intended to generalize (Rudolph et al., 2023). Common approaches include representative sampling (sample composition matches population), representative estimates (sample produces accurate population-level estimates), and representative estimates of heterogeneous effects (model performs well across a variety of subpopulations). Methods such as propensity score weighting, inverse probability weighting, and standardization can improve sample representativeness.

Representative results can also be operationalized and measured and are more applicable to model construction using administrative data, as simple random sampling from an external population using observational data is often challenging. Demonstrating generalizability and representativeness in the target population increases the likelihood that results will hold for out-of-sample applications. Actuarial science and health services research offer tools to improve the reliability of analyses based on non-representative datasets. Matching, stratification, inverse probability of treatment weighting (IPTW), covariate adjustment, and entropy balancing can reduce confounding from observable variables and can improve representativeness of analytic samples (Harvey et al., 2017). In addition, Debray (2026) specifically points out that,

Copyright © 2026 Society of Actuaries Research Institute

"Techniques such as propensity matching, inverse probability weighting, instrumental variable analysis can help reduce confounding, though their effectiveness depends on the availability of high-quality data and the appropriate specification of adjustment models."

1.4.2 Defining and Assessing Fairness

To assess fairness, one must have a working definition of algorithmic fairness. For this study, the definition introduced by Rajkomar et al., known as equal performance, is used. This approach evaluates whether a model is equally accurate for protected and non-protected groups (Rajkomar et al., 2018). This approach emphasizes parity in model behavior across groups, and that omitting sensitive model features (for example, protected class variables) does not prevent bias with respect to those features, as models can infer these characteristics from correlated patterns in data. Equal performance can be operationalized through several metrics:

- **Equal Sensitivity**, which requires the model to identify true cases at the same rate across groups. This is especially important for false negatives that disproportionately impact certain cohorts, such as those that have been historically underrepresented (Obermeyer et al., 2019).
- **Equal Specificity,** which requires similar false-positive rates across groups. This is important because when false positives occur, they can result in costly and unnecessary interventions.
- **Equalized Odds**, which requires both equal sensitivity and equal specificity across groups. Although conceptually appealing, achieving equalized odds can require lowering accuracy for one subgroup, which raises ethical concerns regarding the principle of beneficence (a core principle that mandates researchers protect human subjects from harm and maximize potential benefits while minimizing risk from the 1979 Belmont report) (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979).

These fairness metrics align with the principles of broader ethical traditions that govern human subjects' research. The Belmont Report emphasizes the principle of beneficence as "maximizing benefits while minimizing the potential harm" (The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979).

1.4.3 Intersectionality and Fairness Assessment

Standard fairness metrics evaluate model performance along single demographic dimensions, comparing outcomes—for example, between people living in different geographic areas, or between income levels—without considering how these dimensions interact. Intersectionality holds that individuals occupying multiple disadvantaged positions face distinct forms of disadvantage that cannot be understood by examining each dimension in isolation (Flores et al., 2023; Choi et al., 2026). In claims data, this means a model appearing fair along health status and income dimensions separately may still disadvantage low-income patients of one specific racial, ethnic, or religious group in ways that marginal fairness assessment cannot detect.

Recent ML fairness research has formalized the distinction between marginal and intersectional debiasing. Lett et al. (2025), working with emergency admission prediction models, demonstrate an approach that applies fairness constraints at the level of intersecting demographic groups rather than optimizing along each demographic dimension separately, the key methodological

Copyright © 2026 Society of Actuaries Research Institute

contrast being that marginal debiasing can satisfy group-level constraints while leaving intersectional subgroups poorly served (Lett et al., 2025). For claims-based actuarial models, this implies that representativeness assessment should evaluate performance across combinations of relevant factors, not only factor by factor, and that bias mitigation methods should be selected and evaluated against intersectional performance targets explicitly. Scenario 3 demonstrates these principles by examining how age and income jointly affect colorectal cancer screening rates.

1.5 Recommendations Based on Key Literature

1.5.1 Organizational Support

With AI/ML systems more frequently used for risk prediction, utilization management, claims adjudication, and fraud detection processes, addressing potential sources of bias is increasingly important (Gervasi et al., 2022). Organizations currently vary widely in their monitoring and evaluation process for AI/ML outputs. Engaging a team with diverse backgrounds and perspectives to lead efforts to design algorithms or evaluate the potential purchase of algorithms and provide feedback can help (Cheung et al., 2025). Next, in the implementation phase, tools are needed for the end users of the AI/ML product. This can be an opportunity for the use of model cards to increase interpretability and provide potential limitations (Sendak et al., 2020).

1.5.2 Data Transparency

To ensure interpretability and reproducibility, model developers can report all model features, including demographic characteristics and their distribution. Additionally, when feasible, developers can promote open-source code by uploading code to an online repository (Bozkurt et al., 2020). Finally, developers can articulate which priority fairness metrics are appropriate to the context in addition to using them for parameter adjustment. It can also be helpful to prioritize among these example fairness criteria depending on the type of predictive error that would potentially pose the greatest risk for creating disparities in its clinical application (e.g., false negatives, false positives, reduced uptake).

1.5.3 Methodological Adjustment

Analysts will likely want to carefully consider how to handle missing data to reduce or, if possible, eliminate model bias. If the rate of missing covariates is inversely correlated with access to care, mitigation strategies or dropping observations can increase disparities. There is no single best imputation strategy, but evidence suggests that dropping observations with missing covariates can lead to biased results (Jeanselme et al., 2022). Methods to make samples more representative of the target population include statistical techniques such as inverse probability weighting, entropy balancing, standardization, stratification, and other propensity score methods (Harvey et al., 2017). Impact analyses, such as counterfactual analysis, can be used as well.

One method can be to test the model's potential biases by adding counterfactual datapoints through modifying key features. Another method, Counterfactual Fairness with Regularization (CFR), "removes the direct effect of sensitive attributes on the predicted outcome while preserving...accuracy..." (Jui and Rivas, 2024). This approach uses regularization to encourage

Copyright © 2026 Society of Actuaries Research Institute

the model to make similar predictions for similar individuals but with different sensitive attributes. Other potential methods to consider include resampling and re-labeling, adversarial techniques, thresholding, balancing instance distribution, disparate impact remover, and embedding data into a lower dimension.

1.5.4 Validation and Accuracy

Performance evaluation should cover standard accuracy metrics alongside the fairness metrics the model was explicitly optimized for, since reporting only the former can obscure disparate performance across subgroups (Rajkomar et al., 2018). These metrics include:

Table 1
Validation, Accuracy, and Statistical Metrics

Metric	Acronym	Use in the report
Sensitivity	N/A	Sections 2.2.2, 4.1.4, 5.4
Positive predictive value	PPV	General
Area under the curve	AUC	Section 5
Receiver operating characteristic	ROC	Sections 2.2.3, 6.1
Area under the receiver operating characteristic curve	AUC-ROC, AUROC	Section 5.4
Random forest	RF	Section 5.4
Average treatment effect	ATE	Section 2.2.2
Average treatment effect on the treated	ATT	Section 2.2.2
Standard deviation	SD	Section 5.4
False positive rate	FPR	Section 5.3

External validation on data that reflects the intended clinical context is essential for the purposes of generalizability and application in practice. Performance estimates are much more likely to transfer to the populations a model is meant to serve when the context is reflective of the intended use of a model (Rudolph et al., 2023). Central to that transfer is a clearly specified target population. Generalizability claims distinguish between at least two distinct bases: statistical generalizability, which rests on sampling design, weighting, or stratification, and scientific generalizability, which rests on subject-matter reasoning about mechanism and context. Conflating them risks misapplication of findings. Explicitly stating which form was the goal of a given model would strengthen both transparency and reproducibility.

1.5.5 Documentation Approaches

Following these concerns, states have begun to implement regulatory safeguards with a focus on fairness. New York, for example, now requires insurers to demonstrate that their AI and other data systems do not proxy for protected classes and result in unfair discrimination (New York State Department of Financial Services, n.d.). To comply, insurers will want to consider implementing policy safeguards alongside methodological solutions to address fairness and representativeness. Sendak et al. (2020) proposed a standardized disclosure format for clinical AI tools, analogous to a drug facts label, covering intended use, performance characteristics, and training data provenance.

Copyright © 2026 Society of Actuaries Research Institute

Section 2 Representativeness Assessment

2.1 Conceptual Framework

In a claims data context, representativeness is a measurable relationship between an analytic claims dataset and a clearly defined target population. The focus is on assessing whether analytical results and model predictions would be expected to hold in the target population rather than attempting to achieve representative sampling. This distinction matters because claims data are generated through healthcare utilization patterns rather than population-based sampling designs, creating potential gaps. Any such gaps may need to be characterized and addressed.

The evaluation framework presented in this report structures the assessment process into three sequential stages.

- First, explicitly defining the target population for each application, including all relevant eligibility criteria, geographic boundaries, plan characteristics, and enrollment requirements. This specification establishes the reference standard against which to evaluate the claims data.
- Second, quantifying the magnitude and nature of distributional differences between the dataset and target population across demographic, geographic, and clinical dimensions.
- Third, assessing whether potential observed representation gaps create material risks for model reliability, accuracy, or compliance.

This staged approach allows actuaries to systematically evaluate data quality concerns that arise specifically from representation issues rather than other data quality challenges.

The framework recognizes that claims data serve dual purposes. They function as administrative records designed for payment processing and reimbursement, and are used for predictive modeling, risk classification, and financial projections. This analytic use introduces potential representation challenges because the claims generation process may systematically exclude or underrepresent individuals who face barriers to healthcare access.

Achieving or improving representativeness in claims data requires careful specification of the population the data are intended to represent and distinguishing between the enrolled population captured in claims data and the broader target population of interest for actuarial analysis. The enrolled population includes only individuals with active coverage who generated claims during the measurement period, while the target population may include potential enrollees, market segments, or broader demographic groups relevant to pricing, risk adjustment, or strategic planning decisions.

Claims data as administrative records reflect successful navigation of multiple access points in the healthcare system. To have a claims record, an individual must have insurance coverage, encounter a healthcare provider, receive appropriate services, and have those services coded and submitted for payment. Each step introduces potential selection effects. Rural populations may have limited provider access. Low-income individuals may delay care due to cost-sharing requirements. Certain conditions may be systematically undercoded due to clinical documentation practices or reimbursement incentives. The methodology requires explicit

Copyright © 2026 Society of Actuaries Research Institute

documentation of these known limitations and their relevance to the specific actuarial application.

The distinction between representative sampling and representative results has practical implications for how to validate models and communicate limitations. Representative sampling would require that the claims dataset mirrors the target population's demographic composition, which is rarely achievable with administrative data. Representative results require that the relationships between model inputs and outcomes remain stable across population subgroups, even when the claims dataset's composition differs from the target population. This weaker assumption allows for statistical adjustment techniques while acknowledging residual uncertainty about model performance in underrepresented segments.

2.2 Three-Stage Bias Assessment Method

The three-stage bias assessment framework evaluates potential bias across the AI/ML lifecycle, including input bias, model bias, and application bias as discussed in the literature review.

Figure 1
Three-Stage Bias Assessment Framework

Source: authors

2.2.1 Stage 1: Input Bias Assessment

Stage 1 evaluates potential bias caused by model inputs. This stage focuses on the evaluation of the magnitude, statistical significance, and clinical/actuarial relevance of observed differences to assess whether input data are appropriate for the intended modeling purpose.

Magnitude measures can be used to quantify the absolute and relative differences in key characteristics such as demographic, clinical, and utilization across subgroups. These measures help identify representation and distributional imbalances, such as whether groups are underrepresented relative to the target population or whether the covariate distribution differs meaningfully across groups. To examine statistical significance of how the sample differs from the population of interest, statistical testing such as confidence intervals and hypothesis testing can be used. However, to assess clinical/actuarial significance, it is important to interpret statistical significance along with magnitude measures. This is particularly valuable in large samples where small differences reach statistical significance without real-life relevance. Additionally, the common use of a uniform criteria for rejecting a null hypothesis (e.g., P-value less than .05) may be considered arbitrary, especially when balance can be improved through statistical techniques. For this reason, the authors suggest using caution when using hypothesis tests such as t-tests for assessing covariate balance.

Copyright © 2026 Society of Actuaries Research Institute

Other ways to assess balance across subgroups include standardized mean differences (SMD) and variance ratios. “The standardized mean difference (SMD) is the difference in the means of each covariate between groups standardized by a standardization factor so that it is on the same scale for all covariates (Griefer, Noah, 2023)” (New York City Center for Innovation through Data Intelligence, 2025; Greifer, 2025). SMDs provide a scale-free measure of imbalance: while values closer to zero indicate better balance, recommended thresholds for important covariates are between 0.1 and 0.5. The variance ratio is simply calculated as the ratio of the variance of a covariate in one group to that of the other. A variance ratio of one indicates similar variances of the subgroups and larger deviations suggest heterogeneity. The demonstration of the use of these measures is achieved with the R package MatchIt (Greifer, 2025).

2.2.2 Stage 2: Model Bias Mitigation

Stage 2 focuses on mitigating the bias caused by input bias and mitigating bias created during model development by addressing imbalances or missingness in covariate distributions across subgroups. When balance is achieved, model estimates are less sensitive to specification choices and can more closely approximate underlying relationships of interest. No single method provides a definitive method to address imbalanced or underrepresented groups in data. However, there are multiple complementary approaches to mitigate model bias. Choosing the best approach depends on the model objective, data structure, and sample size considerations.

While there are a variety of methods for propensity scoring, all “...models can be conceptualized as methods to create observation weights that adjust covariate distributions to reduce selection bias” (Bishop et al., 2018). Post-adjustment balance is typically assessed using standardized mean differences and diagnostics of common support to help ensure adequate overlap of propensity score distributions across groups. For these analyses, inverse probability weighting (IPW) using logistic regression can be used to estimate propensity scores, which are then used to construct IPW weights that rebalance the comparison groups (e.g., individuals seen by primary care providers vs. specialists, the scenario described below in Section 4.2.2). Statistical packages such as MatchIt or TriMatch in R or pysmatch in Python can implement these methods (HanCheng, 2025; Bryer, 2025; Greifer, 2025).

Stabilized weights can be constructed by multiplying the raw IPW by the marginal probability of treatment, which compresses the weight distribution and reduces the influence of extreme values. These weights are then incorporated into the model to produce population-level estimates. The estimand is determined by how the weights are specified: ATE weights rescale the sample to represent the full population, while ATT weights rebalance the comparison group to match the characteristics of the treated group. The choice between them follows from the research question—whether the goal is a population-average effect or an effect specific to those who received the intervention. Like propensity score modeling, IPW is inherently iterative; analysts adjust the model specification and recheck covariate balance until adequate balance is achieved. IPW can be implemented using the ipw package in R (Tran et al., 2025).

Entropy balance (EB) is an optimization approach that calibrates sample weights by imposing moment-matching constraints on the covariate distribution of the unexposed target. By directly enforcing balance on selected covariates and their specified moments, EB achieves covariate balance a priori, without relying on iterative propensity score estimation. Researchers identify the covariates, interactions, the moments to be balanced, and acceptable tolerance levels. EB,

Copyright © 2026 Society of Actuaries Research Institute

then, identifies the optimal set of weights that satisfies these constraints while minimizing deviation from the original weight distribution through an entropy-minimization criterion. Compared with propensity score-based methods, EB can achieve superior covariate balance with greater efficiency, especially in settings where the treated group constitutes a large share of the sample, as balance is enforced through calibration rather than reweighting based on estimated treatment probabilities. This method can be implemented in R using the ebal package (Hainmueller, 2022).

Stratification and calibration weighting methods provide an additional set of tools for mitigating model bias when external population benchmarks are available. Post-stratification adjusts sample weights so that marginal distributions of key variables (e.g., age, sex, or geographic region) match known population totals. This approach is most effective when strata are well populated, and population margins are reliably measured. When balance is required across multiple dimensions simultaneously, raking (iterative proportional fitting) extends post-stratification by iteratively adjusting weights to align with several marginal distributions at once, improving representativeness while preserving sample size.

Calibration weighting generalizes these approaches by adjusting weights to satisfy a set of linear constraints that align weighted covariate moments with population targets. Unlike traditional post-stratification, calibration weighting allows flexibility in the choice of constraints and can incorporate continuous variables and interaction terms. These methods are particularly useful in observational and survey-based settings where selection probabilities vary across subgroups and where aligning the analytic sample with a target population is essential for reducing potential bias in downstream model estimation. These methods can be implemented in R using the survey package (Lumley et al., 2025).

The handling of missing data is an important source of potential bias in modeling, especially when missingness is correlated with demographic or other sensitive attributes. It is important to note that higher imputation accuracy does not necessarily mean higher fairness, and current literature points to biases resulting from all imputation methods, but that it is context dependent. Pattern-mixture models can be used to assess whether bias from the violation of MCAR severely impacts model output by providing bounds on estimates under alternative missing not at random (MNAR) scenarios. Pattern-mixture models can be implemented with the R package missingHE, and multiple imputations can be implemented with the R package mice (Gabrio, 2025; Buuren et al., 2025).

To assess whether findings and fairness properties hold under different assumptions, results can be stress-tested against alternative analytic choices. Examples include choosing alternative covariate definitions, functional forms, algorithms for regression, tree-based, and other methods and alternative weighting and balancing approaches. Additional analyses can evaluate the robustness of the model to missing data assumptions and alternative outcome definitions. Other potential sensitivity analyses to consider include resampling and re-labeling, adversarial techniques, thresholding, and embedding data into a lower dimension (Jui and Rivas, 2024). Sensitivity analyses may provide transparency into the robustness of a model and may shed light on potential limitations.

Copyright © 2026 Society of Actuaries Research Institute

2.2.3 Stage 3: Application Bias Mitigation

Stage 3 focuses on bias that can arise in the application phase, when a model is applied to a population after development. Ongoing monitoring is required to evaluate whether models perform consistently across subgroups to ensure fairness. Reporting performance both overall and stratified by key demographics/subgroups is typical (Collins et al., 2024, Mitchell, 2019). Subgroup performance is commonly assessed using mean squared error (MSE) and mean absolute error (MAE). Demographic or other subgroup characteristics are typically used to assess and balance the population. For binary outcomes, similar subgroup summaries using brier scores and classification error rates can be used.

Calibration monitoring can evaluate whether predicted risks correspond to observed actual rates within subgroups using calibration plots (observed vs. predicted risk). Discrimination metrics can be assessed overall and by subgroups using ROC curves/AUC and precision-recall curves. Finally, ensuring equal performance across subgroups, would include monitoring fairness metrics. Fairness metrics evaluate distribution of a model's predictions and downstream decisions across subgroups. A list of typical fairness metrics is shown in Table 2. These metrics differ from the common prediction metrics described above. Several types of fairness metrics are commonly used and reported depending on the model context (Rountree et al., 2025).

Table 2
Fairness Metrics

Metric	Equation	Meaning
Equalized Odds	$P(\hat{Y} = 1 \mid S = a, Y = y) = P(\hat{Y} = 1 \mid S = b, Y = y)$	Both groups have equal true positive and false positive rates
Equalized Opportunity	$P(\hat{Y} = 1 \mid S = a, Y = 1) = P(\hat{Y} = 1 \mid S = b, Y = 1)$	Both groups have equal true positive rates (a relaxed version of equalized odds)
Predictive Equality	$P(\hat{Y} = 1 \mid Y = 0, S = a) = P(\hat{Y} = 1 \mid Y = 0, S = b)$	The rate of false positive is independent of the sensitive attribute
False Negative Rate Parity	$P(\hat{Y} = 0 \mid Y = 1, S = a) = P(\hat{Y} = 0 \mid Y = 1, S = b)$	The rate of false negative is independent of the sensitive attribute
Predictive Parity	$P(\hat{Y} = 1 \mid Y = 1, S = a) = P(\hat{Y} = 1 \mid Y = 1, S = b)$	Model precision is the same for both groups
Demographic Parity	$P(\hat{Y} = 1 \mid S = a) = P(\hat{Y} = 1 \mid S = b)$	The prediction should be independent of the sensitive attribute

Equalized Odds and Predictive Parity equations can be found in Yusha'u and Abdullahi, 2026.

2.3 Data Requirements for Representativeness Assessment

Applying the framework (Section 2.2) to claims data requires datasets with sufficient demographic, geographic, and clinical information to support the distributional comparisons and quantitative metrics specified in Section 2. At minimum, evaluation demands individual-level records with member demographics (e.g., age, sex, geographic location), enrollment information (coverage type, effective dates, plan characteristics), utilization summaries (allowed amounts, service categories, provider specialties), and clinical indicators (diagnosis codes, procedure codes, chronic condition flags). However, the specific dimensions requiring detailed assessment depend on the actuarial application context and any known systematic gaps in administrative claims data.

Copyright © 2026 Society of Actuaries Research Institute

Clinical condition representation, especially for systematically under-documented conditions in mental health, chronic pain, and social determinants of health, creates model bias in risk adjustment algorithms, care management targeting protocols, and utilization management decision rules that rely on diagnosis code patterns as proxies for true disease burden. By designing synthetic data scenarios that introduce controlled representation gaps along these three dimensions, the quantitative impact of non-representative data on actuarial model outputs is demonstrated while providing practitioners with concrete examples of how representation assessment methods detect and measure these systematic biases in claims datasets used for AI/ML model development.

The synthetic data scenarios serve a specific methodological purpose: demonstrating that the representativeness metrics and bias assessment methods developed in Section 2 can reliably detect known, intentionally introduced representation gaps. The primary objective is internal validity, showing that systematically introducing urban-rural diagnosis differentials, clinical documentation variation, or intersectional screening gaps allows for the quantitative measures to successfully identify these biases and quantify their magnitude. The “ground truth” version of the data set is used to demonstrate the value of the decision assistance tool included as part of the workbook described in Section 6.

Copyright © 2026 Society of Actuaries Research Institute

Section 3 Actuarial Application

Representativeness requirements and assessment priorities vary substantially across actuarial applications due to differences in prediction targets, financial materiality thresholds, regulatory constraints, and the consequences of model error. Practitioners can adapt the three-stage bias assessment framework based on their specific application context, emphasizing different representativeness dimensions and applying different materiality thresholds depending on how models are deployed and what decisions they support.

An important distinction affects how representation gaps are best addressed across different applications. Statistical methods may be methodologically sound (unbiased under stated assumptions) yet actuarially inappropriate for specific contexts. For example, multiple imputation methods that assume Missing at Random (MAR) may produce unbiased parameter estimates in a statistical sense, but imputing utilization for individuals who face access barriers could overstate expected costs. Conversely, imputation or weighting methods may be needed to estimate true clinical burden despite incomplete documentation, because payment mechanisms assume complete capture of diagnostic information. This distinction between methodologically valid statistical procedures and actuarially appropriate modeling choices suggests the need for practitioners to consider whether representativeness gaps reflect temporary data collection limitations that can be mitigated through statistical adjustment, or whether they represent persistent differences between observed populations and target applications that require explicit modeling through segmentation, credibility weighting, or disclosed limitations.

The framework's three stages apply universally, but emphasis differs by context.

- Pricing prioritizes input bias and demographic balance.
- Risk adjustment prioritizes clinical documentation and coding intensity.
- Utilization management prioritizes application monitoring and decision-level fairness.

Practitioners determine materiality thresholds based on the financial magnitude of decisions, regulatory requirements, and operational consequences of model errors. Section 4 demonstrates these principles through scenarios chosen to reflect challenges across multiple application types, enabling practitioners to identify which scenario most closely resembles their data quality concerns. Actuarial applications with detailed implications are discussed in the following subsections.

3.1 Actuarial Practice in the AI Era

Actuaries are increasingly asked to evaluate, validate, and sign off on AI and ML models that inform consequential decisions such as risk scores that drive plan revenue and the impact of prior authorization algorithms that affect member care. This shift represents both an opportunity and a responsibility. The methods described in this report are grounded in statistical and technical literature that directly applies to healthcare claims data. This research is intended to show how to detect bias quantitatively, which mitigation strategies are appropriate in which contexts, and how to document findings in a way that supports professional judgment and downstream use.

For actuaries working in areas where clinical burden translates directly into revenue, representativeness is directly linked to financial risk. For example, Hierarchical Condition

Copyright © 2026 Society of Actuaries Research Institute

Category (HCC) risk scores used to calibrate capitation payments depend on complete diagnostic capture. When systematic documentation gaps cause certain conditions that are more common in rural areas, in populations seen primarily by primary care physicians, or in lower-income communities, risk scores underestimate actual disease burden. The Stage 1 and Stage 2 bias assessment methods in this report provide a structured way to quantify and address these potential gaps before they propagate into financial projections. Similar considerations apply in ACA marketplace risk adjustment, Medicaid managed care, and self-funded employer plan actuarial work, where population heterogeneity and access patterns vary substantially.

The practical implication is that avoiding introducing or amplifying bias in AI-driven work can be enabled using tools like the ones described in this report. If a model systematically underperforms for rural members, low-income members, or those primarily seen by primary care providers, that disparity is likely to surface as an actuarial accuracy problem. The Model Card and Decision Assist Tool developed as part of this project gives practitioners a structured starting point for documenting these concerns and communicating them to model users and decision-makers.

3.2 Pricing Applications

Insurance pricing emphasizes Stage 1, input bias assessment, focusing on geographic rating factors and demographic mix, since rate filings depend on utilization assumptions that must hold across the marketed population. Systematic gaps in rural or small market segments can distort rating structures when applied to populations with fundamentally different provider networks and access patterns. Thus, for pricing applications, practitioners would likely prioritize demographic representativeness (age, sex, geography) over clinical detail, since pricing models typically use broader risk categories than risk adjustment algorithms.

Stage 3 application, potential bias monitoring, focuses on calibration accuracy and rate stability over time, assessing whether premium rates remain adequate as the enrolled population evolves. Materiality thresholds align with statutory rate filing requirements and competitive market constraints, where even small systematic errors compound across large membership bases.

3.3 Risk Adjustment

Medicare Advantage, Affordable Care Act (ACA) marketplace, and self-funded employer risk adjustment applications all would benefit from Stage 1 assessment of clinical condition representation, because payment mechanisms depend on complete diagnostic capture. Systematic underrepresentation of populations with access barriers directly translates to revenue forecast errors and quality bonus penalties. For risk adjustment applications, practitioners would likely emphasize clinical documentation completeness and coding intensity variation (Stage 2 model bias) more than demographic balance, since prospective risk scores derived from incomplete utilization data introduce systematic bias into financial projections.

Stage 3 monitoring tracks diagnosis and treatment capture rates, risk score drift, and Star rating impacts across population segments. Provider specialty and geographic variation in coding practices may create material bias even when demographic distributions appear balanced. Detecting such potential bias requires ongoing validation that observed diagnosis patterns reflect true clinical burden rather than documentation differences.

Copyright © 2026 Society of Actuaries Research Institute

3.4 Utilization Management and Operational Applications

Prior authorization algorithms, care management targeting, and fraud detection systems face distinct representativeness challenges because decision thresholds calibrated to observed utilization patterns may inappropriately restrict access for populations whose legitimate care needs differ from reference populations. For these applications, practitioners would likely prioritize Stage 3 application bias monitoring over input data balance, since errors in individual-level decisions accumulate through operational deployment rather than aggregate financial projections. False positive rates (unnecessary denials, inappropriate targeting) and false negative rates (missed opportunities, undetected fraud) require separate assessment across demographic and access-related subgroups. Representativeness concerns focus less on whether training data match target populations and more on whether deployed algorithms maintain consistent accuracy when applied to members with different baseline characteristics, network access, or care-seeking patterns than those well-represented in development data.

3.5 Provider Payment and Episode-Based Models

Episode groupers, bundled payment algorithms, and readmission prediction models depend on representative training data capturing complete care episodes across diverse care-seeking patterns. Potential representation gaps may introduce systematic errors when algorithms trained on populations with established primary care relationships are applied to members with fragmented care access. For these applications, practitioners would likely assess completeness of episode capture (Stage 1) and validate that cost, or utilization predictions remain stable across subgroups with different care coordination patterns (Stage 3). Network configuration and referral patterns affect which services appear in claims data for episode attribution, which may create representation challenges distinct from demographic imbalance.

Copyright © 2026 Society of Actuaries Research Institute

Section 4 Data Generation

4.1 Synthetic Data for Research Application

For this research, data is generated to represent three representation dimensions that may generate bias across multiple actuarial functions: geographic distribution (urban vs. rural populations), clinical condition prevalence (focusing on systematically under-documented conditions), and intersectional bias across multiple dimensions (in this case, age and income). These three dimensions directly affect actuarial work while reflecting well-documented potential systematic gaps in claims data quality. Geographic representation, particularly rural population underrepresentation, may introduce bias into commercial insurance geographic rating factors and provider network adequacy assessments. Clinical condition identification is key for risk adjustment mechanisms such as Medicare Advantage. Age and income distribution gaps affect actuarial work across multiple contexts, such as employer-provided health plans with diverse workforces.

4.1.1 Scenario Design Principles—Overarching Design Philosophy

Synthetic data scenarios for this research follow an overarching design philosophy that prioritizes testing specific, measurable potential representation problems relevant to actuarial decision-making contexts. Each scenario targets a distinct potential bias mechanism. This focused approach helps ensure scenarios remain testable using the quantitative metrics established in Section 2. Each scenario generates a sufficient sample size and includes both a “biased” dataset reflecting systematic representation gaps and a “reference” or “ground truth” population dataset representing the target population against which representativeness is assessed.

The data structure mimics actual administrative claims data architecture, incorporating the elements detailed in Table 3 across multiple scenarios: patient demographics enabling distributional comparisons, insurance type indicators, clinical data supporting risk adjustment validation, and diagnosis utilization measures quantifying access pattern differences. This standardized data structure allows consistent application of the bias assessment framework across different representation scenarios while maintaining alignment with data elements actuaries encounter in operational claims datasets used for AI/ML model development.

Copyright © 2026 Society of Actuaries Research Institute

Table 3
Data Elements Across the Scenarios

Category	Rationale	Examples
Patient Demographics	Core dimensions for assessing representativeness	Age, sex, geographic location
Insurance Type	Different payer types have distinct population characteristics and selection effects	Commercial, Medicare, Medicaid
Clinical Data	Enables testing of clinical representation bias; required for hierarchical condition category (HCC) and other risk adjustment model validation	Diagnosis codes, Procedure codes, Service dates
Utilization	Measures healthcare access and engagement patterns; identifies underutilization in underrepresented groups	Number of visits, emergency department (ED) visits, Hospitalizations, Pharmacy claims
Costs	Enables financial impact assessment of representation issues; required for pricing, rating, and reserving applications	Total, patient responsibility
Provider Characteristics	Identifies access barriers and network adequacy issues; enables assessment of care quality variation	Specialty, location, practice type

4.1.2 Scenario Design Principles—Scenario Selection Rationale

The three scenarios were selected to demonstrate how representation gaps affect actuarial applications across different bias mechanisms while remaining technically feasible and providing novel methodological guidance. Each scenario addresses operational challenges actuaries face. Geographic disparities affect commercial rating factors, clinical documentation variation impacts risk adjustment payments, and intersectional bias affects care management algorithms. The scenarios map to the three-stage bias framework, with Scenario 1 demonstrating input bias assessment, Scenario 2 illustrating model bias from coding variation, and Scenario 3 showing application bias monitoring. Synthea™'s disease modules provided validated clinical pathways for the selected conditions, with diagnostic workflows and screening protocols that could be systematically varied by access parameters. The scenarios focus on potential representation challenges with limited actuarial methodology rather than extensively studied disparities.

4.1.3 Synthea Implementation

Synthea, an open-source synthetic patient generator, provides the technical foundation for creating claims datasets with controlled representation characteristics for this research (Walonoski et al., 2018). The platform generates realistic synthetic healthcare data by simulating patient lifecycles through clinical pathways based on published disease progression models, treatment guidelines, and epidemiological evidence. For this project, Synthea configuration requires specification of demographic distribution parameters (age structure, sex ratios, geographic distribution by census tract, or county Federal Information Processing Series (FIPS) codes) that allow creation of both reference populations matching target actuarial populations and intentionally biased datasets reflecting systematic underrepresentation patterns.

Clinical pathway modeling within Synthea determines utilization patterns by simulating care-seeking behavior, provider encounters, diagnostic workflows, and treatment protocols, which

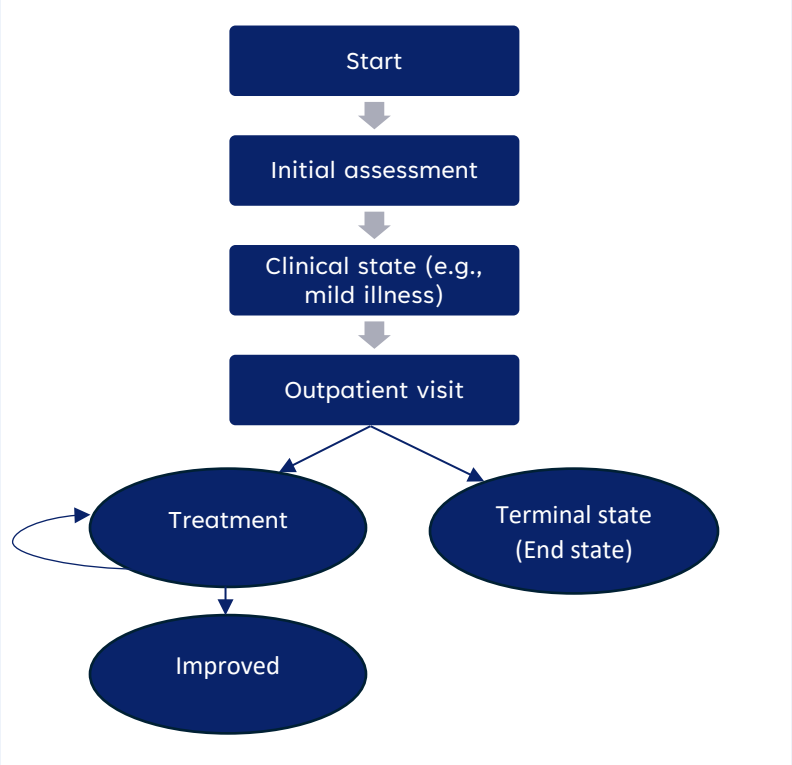
Copyright © 2026 Society of Actuaries Research Institute

produces clinical data using Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) terminology and International Classification of Diseases, version 10 (ICD-10) diagnosis codes (Centers for Medicare & Medicaid Services (CMS), 2026; National Library of Medicine (NLM), 2019). The Markov state transition workflow programmatically controls population generation parameters, executes batch runs for multiple representation scenarios, and exports results. This implementation approach allows reproducible generation of claims datasets where representation gaps are introduced systematically through controlled parameter specifications rather than post-hoc sampling or deletion, ensuring that synthetic data scenarios reflect plausible claims generation processes that mirror real-world systematic biases that can arise from differential healthcare access, provider documentation practices, and insurance coverage patterns. Figure 2 shows an example of a generic Synthea module.

4.1.4 Data Characterization Requirements

Confidence in synthetic data accuracy derives from complementary validation strategies. First, internal validity testing confirms that the data generation process successfully implements intended representation mechanisms, where introduced biases produce statistically detectable differences in expected directions (e.g., rural underrepresentation creates measurable geographic imbalance). Second, external validity assessment compares synthetic data characteristics to established methods of data analysis and assessment. This dual approach validates that synthetic scenarios reflect plausible claims data characteristics while containing controlled representation gaps of known magnitude. Practitioners applying the framework to operational data benefit from this validation because it demonstrates method sensitivity and specificity in controlled conditions before deployment to real-world assessment problems.

Figure 2
Generic Synthea Data Generation



Source: authors

Copyright © 2026 Society of Actuaries Research Institute

Data characterization establishes baseline distributional properties and quantifies the magnitude of representation gaps between biased datasets and reference populations. Distribution comparisons across key dimensions (e.g., demographics, geography, clinical conditions, utilization patterns) provide the foundational evidence that representation problems exist and require methodological intervention to achieve representativeness. Balance tables complement visual diagnostics by presenting quantitative summaries of pre-adjustment imbalances. Continuous variables are reported as means and standard deviations, and categorical variables as counts and proportions. These characterization outputs document the specific representation challenges present in each synthetic scenario, establish the baseline against which bias mitigation methods will be evaluated in Stage 2, and provide the reference distributions needed for calculating representation metrics.

4.2 Scenarios Modeled

4.2.1 Urban/Rural Disparities in Sleep Apnea Diagnosis

Scenario 1 examines geographic representation bias by comparing sleep apnea diagnosis rates across urban and rural populations. Sleep apnea serves as a clinical condition because diagnosis requires specialized testing infrastructure (sleep studies) and specialists who concentrate in metropolitan areas, creating natural access-dependent variation. The reference population specification uses Rural-Urban Commuting Area (RUCA) codes, "...a classification scheme allowing for flexible, census tract delineation of rural and urban areas throughout the United States and its territories" (Economic Research Service, n.d.). The biased dataset introduces systematic rural underrepresentation by reducing diagnostic testing utilization in RUCA codes 4–10 while maintaining metropolitan diagnosis patterns, creating geographic coding intensity variation where true prevalence remains constant but diagnostic capture declines with rurality.

4.2.2 Clinical Documentation Bias in Type 2 Diabetes

Scenario 2 examines clinical documentation variation by comparing Type 2 diabetes coding patterns across provider settings. Type 2 diabetes is prone to documentation bias because testing coding varies substantially across provider specialties despite equivalent pharmaceutical evidence. The reference population specification uses published population health survey prevalence rates, with target distributions reflecting clinical burden documented through validated assessments rather than administrative claims coding patterns. The scenario's biased dataset introduces systematic undercoding and missingness through provider-level documentation variation. Primary care settings exhibit lower capture and documentation of hypertriglyceridemia compared to specialists (see, for example, Leinung et al., 2000 for evidence that this happens in practice and Cleveland Clinic, 2022), and these cases are considered missing data. This masking is applied as a modification to the existing synthetic data. Key variables include diagnosis code prevalence by provider specialty, clinical presentation, and socioeconomic variables.

4.2.3 Intersectional Bias—Age and Income Impacts on Colorectal Cancer Screening Rates

Scenario 3 examines intersectional bias by assessing how age and income jointly affect colorectal cancer screening rates. Colorectal cancer (CRC) screening provides a test case for intersectional bias because clinical guidelines establish age-based screening recommendations,

Copyright © 2026 Society of Actuaries Research Institute

while income affects access to screening modalities through time, transportation, and follow-up coordination barriers. The reference population specification uses census-based age distributions combined with census-tract income data, adjusted for insurance coverage patterns. The biased dataset introduces compounding representation gaps where screening rates decline with both younger age and decreasing income, creating an interaction effect where low-income adults at younger ages, and therefore not yet eligible for Medicare and newly eligibility for CRC screening show the largest screening deficits. This scenario explicitly tests whether single-dimension fairness metrics can mask compounding disadvantages when multiple access barriers combine.

Copyright © 2026 Society of Actuaries Research Institute

Section 5 Representativeness Results

5.1 Overall Results

Across the three synthetic scenarios, the results show that conventional overall performance statistics can obscure material subgroup disparities in healthcare claims-based AI/ML applications. In each scenario, the representativeness problem emerged through a different mechanism: geographic underdiagnosis in sleep apnea, documentation-related missingness in Type 2 diabetes, and intersectional access disparities in colorectal cancer screening. The results also show that no single mitigation strategy dominates across all fairness objectives examined. Instead, the appropriate mitigation approach depends on the application context, the type of representation failure, and the specific impact definition being prioritized, such as equal opportunity, demographic parity, or minimax subgroup performance. The existence of the "true" model also demonstrates that, in general, mitigation mechanisms based on the data available to actuaries and analysts is likely to mitigate bias, but not completely or comprehensively eliminate it.

5.2 Scenario 1

Scenario 1 evaluated the effects of geographically patterned underdiagnosis on a sleep apnea prediction task. The synthetic population comprised 22,903 patients (81% urban, 19% rural) with 2,160 sleep apnea cases overall (9.4% prevalence). Bias was introduced by "masking" rural diagnoses, where sleep apnea cases simulated by Synthea were systematically "hidden" from the model to reflect underdiagnosis due to access issues (Cowan, 2026). The masking reduced observed prevalence from 9.43% to 8.92% overall and removed 117 rural cases from the observed dataset. Before label masking, sleep apnea prevalence was comparable across geographic groups (urban 9.50%; rural 9.13%). After applying a 30% masking rate to diagnosed rural cases (117 of 390 rural cases), rural prevalence fell to 6.39% (a decline of 2.74 percentage points), while urban prevalence was unchanged at 9.50%. The characteristics of the population are shown in Table 4 and Table 5.

Copyright © 2026 Society of Actuaries Research Institute

Table 4
Stage 1—Baseline Population Comparison (Unbiased)

Characteristic	Overall (N = 22,903)	No Sleep Apnea (N = 20,743)	Sleep Apnea (N = 2,160)
Geography			
Rural	4,271 (18.6%)	3,881 (18.7%)	390 (18.1%)
Urban	18,632 (81.4%)	16,862 (81.3%)	1,770 (81.9%)
Sex			
Female	10,476 (45.7%)	9,870 (47.6%)	606 (28.1%)
Male	12,427 (54.3%)	10,873 (52.4%)	1,554 (71.9%)
Age Group (years)			
60–69	1 (0.0%)	1 (0.0%)	0 (0.0%)
70–79	3,794 (16.6%)	3,454 (16.7%)	340 (15.7%)
80–89	4,332 (18.9%)	3,941 (19.0%)	391 (18.1%)
90+	14,776 (64.5%)	13,347 (64.3%)	1,429 (66.2%)
Clinical Characteristics			
Hypertension	10,863 (47.4%)	9,292 (44.8%)	1,571 (72.7%)
Congestive Heart Failure (CHF)	2,368 (10.3%)	2,139 (10.3%)	229 (10.6%)
Alcohol Use	499 (2.2%)	438 (2.1%)	61 (2.8%)
Obesity (BMI ≥ 30)	4,891 (21.4%)	4,440 (21.4%)	451 (20.9%)
Current Smoker	11 (0.0%)	11 (0.1%)	0 (0.0%)

Note: Values may not sum to 100% due to rounding.

Table 5
Stage 1—Baseline Population Comparison (Biased)

Characteristic	Overall (N = 22,903)	No Sleep Apnea (N = 20,743)	Sleep Apnea (N = 2,160)
Geography			
Rural	4,271 (18.6%)	3,998 (19.2%)	273 (13.4%)
Urban	18,632 (81.4%)	16,862 (80.8%)	1,770 (86.6%)
Sex			
Female	10,476 (45.7%)	9,906 (47.5%)	570 (27.9%)
Male	12,427 (54.3%)	10,954 (52.5%)	1,473 (72.1%)
Age Group (years)			
60–69	1 (0.0%)	1 (0.0%)	0 (0.0%)
70–79	3,794 (16.6%)	3,472 (16.6%)	322 (15.8%)
80–89	4,332 (18.9%)	3,957 (19.0%)	375 (18.4%)
90+	14,776 (64.5%)	13,430 (64.4%)	1,346 (65.9%)
Clinical Characteristics			
Hypertension	10,863 (47.4%)	9,382 (45.0%)	1,481 (72.5%)
Congestive Heart Failure (CHF)	2,368 (10.3%)	2,147 (10.3%)	221 (10.8%)
Alcohol Use	499 (2.2%)	442 (2.1%)	57 (2.8%)
Obesity (BMI ≥ 30)	4,891 (21.4%)	4,462 (21.4%)	429 (21.0%)
Current Smoker	11 (0.0%)	11 (0.1%)	0 (0.0%)

Note: Values may not sum to 100% due to rounding.

Three XGBoost classification models were then trained on 85% of the synthetic population of 22,903 patients (81% urban, 19% rural) and evaluated on a 15% hold-out test set. The results of

Copyright © 2026 Society of Actuaries Research Institute

this model were quite favorable. Overall AUC was nearly identical across the true (0.687), biased (0.680), and post-mitigation (0.674) models, suggesting comparable predictive performance under standard evaluation. However, the subgroup results tell a different story. Rural recall dropped from 0.587 in the true model to 0.111 in the biased model, an 81% reduction, after 30% of rural sleep apnea diagnoses were masked to simulate underdiagnosis. Urban recall declined only modestly from 0.597 to 0.570. The mitigation method used, propensity score matching (PSM), mitigated this bias, and partially recovered rural recall to 0.206 but did not achieve parity with the urban group (urban recall: 0.454 in the PSM model).

These results confirm that standard aggregate metrics are insufficient for bias detection, and that propensity score adjustment alone does not fully address systematic underrepresentation in the training data. Demographic distributions were identical across biased and reference datasets by design, as the bias mechanism operated exclusively through diagnostic code masking rather than enrollment selection. These findings also illustrate the core representativeness problem motivating the report. A model can appear acceptably accurate when evaluated overall while failing materially for underrepresented or access-constrained subgroups, creating actuarial risk when used for pricing, screening, or care management applications. The model card technique described in Section 6 of this report further illustrates the identification and documentation of both the bias and the mitigation approach.

5.3 Scenario 2

Scenario 2 examined how provider-linked documentation bias can distort a Type 2 diabetes prediction model when missingness is handled naively. This is designed to show that the treatment of missing data when data is systematically missing can impact model performance and subgroup fairness. Statistics for the mask rate application are shown in Table 6.

Table 6
Stage 2—Mask Rate for Hypertriglyceridemia (Documentation Bias)

Metric	Value
Patients with hypertriglyceridemia (true)	5,282
Patients masked (under-documented)	3,458
Actual mask rate	65.5%

The bias can also be seen in the number of patients whose diagnosis was missing, which was nearly 3,500 patients out of 33,200, as shown in Table 7 and Table 8.

Table 7
Stage 2—Mask Rate for Hypertriglyceridemia (Documentation Bias)

Provider Type	N (Patients)	N (True Diagnosis)	N (Masked)	Mask Rate
Primary Care Physician (PCP)	24,908	3,908	3,321	84.98
Specialist	8,292	1,374	137	9.97

Table 8
Stage 2—Diagnosis of Hypertriglyceridemia

Characteristic	Observed (biased documentation) N = 33,200	Characteristic N = 33,200
Hypertriglyceridemia	1,824 (5.5%)	5,282 (16%)

Copyright © 2026 Society of Actuaries Research Institute

The data was then used to train four models against the same diabetes outcome. The models were:

1. Baseline model with true hypertriglyceridemia,
2. Biased model with observed hypertriglyceridemia,
3. Imputed model with a filled-in hypertriglyceridemia feature,
4. Missingness-aware model with observed hypertriglyceridemia plus a missingness indicator.

It was evaluated on a 15% hold-out test set. Missingness was higher among those who were seen by a primary care provider compared to a specialist, reflecting real-world biases. Overall AUC was nearly identical across the baseline (0.991–0.994), biased (0.987–0.995), imputed (0.990–0.995), and missingness-aware (0.991–0.994) models, suggesting comparable predictive performance at a high level.

Subgroup and fairness metrics told a different story. When missing data was simply dropped (the biased model), the gap between PCP and Specialist group recall and F1 increased from 0.063 to 0.094 and 0.025 to 0.047, respectively. This omission also distorted the model's underlying logic, notably shifting feature importance: reliance on hypertriglyceridemia dropped significantly (-0.039 difference) while reliance on A1c increased (+0.035 difference). Most importantly, this biased treatment of missingness exacerbated fairness disparities, driving both Equal Opportunity Difference and Equalized Odds Difference up from 0.063 in the true model to 0.095. These results are shown by provider type in Table 9.

Table 9
Stage 2—Diagnosis of Hypertriglyceridemia

Model	Provider	Sample Size (n)	AUC	Precision	Recall	F1
Baseline	PCP	3713	0.991	0.918	0.843	0.879
Baseline	Specialist	1266	0.994	0.902	0.906	0.904
Biased (Dropped missing)	PCP	3713	0.987	0.898	0.825	0.860
Biased (Dropped missing)	Specialist	1266	0.995	0.895	0.919	0.907
Imputed	PCP	3713	0.990	0.927	0.847	0.885
Imputed	Specialist	1266	0.995	0.935	0.906	0.920
Missingness-aware	PCP	3713	0.991	0.912	0.860	0.885
Missingness-aware	Specialist	1266	0.994	0.894	0.910	0.902

Mitigation strategies successfully recovered performance, and their effectiveness depended on the targeted fairness metric as shown in Table 10. Imputation optimized demographic representation, improving the Demographic Parity Difference to 0.026 and the Demographic Parity Ratio to 0.846. However, the Missingness-Aware approach was superior for restoring error-rate parity, successfully reducing both Equal Opportunity Difference and Equalized Odds Difference to 0.050. These results confirm that simply dropping systematically missing data can introduce bias, and that different mitigation strategies optimize for different fairness definitions.

Imputation improved demographic parity, while a missingness-aware approach better restored equal opportunity. This reinforces the broader point that fairness optimization is not one-dimensional and is more successful when approaches are selected based on intended use.

Copyright © 2026 Society of Actuaries Research Institute

Table 10
Stage 2—Diagnosis of Hypertriglyceridemia

Model	Demographic Parity Diff[1]	Demographic Parity Ratio[2]	Equal Opportunity Diff[3]	Equalized Odds Diff[4]
Baseline (true ht)	0.0318	0.8204	0.0626	0.0626
Biased (observed ht)	0.0357	0.8025	0.0948	0.0948
Imputed (filled-in ht)	0.0263	0.8461	0.0592	0.0592
Missingness-aware (ht+missing)	0.0301	0.8321	0.0500	0.0500

1 Difference in Selection Rates. Ideal value = 0
2 Ratio of Selection Rates. Ideal value = 1.0
3 Difference in True Positive Rates (Recall). Ideal value = 0
4 Maximum of TPR difference and FPR difference. Ideal value = 0

5.4 Scenario 3

Scenario 3 focused on intersectional bias in colorectal cancer screening, where age and income jointly shape access and observed screening behavior. This scenario demonstrated that intersectional bias can be difficult to spot based on aggregate assessments and examined how to mitigate biases by assessing multiple dimensions jointly. Colorectal cancer screening provides a test case for intersectional bias because clinical guidelines establish age-based screening recommendations and recently changed its recommendation to include younger cohorts, while income affects access to screening modalities through time, transportation, and follow-up coordination barriers. Four income bands (<30K, 30–59K, 60–99K, and ≥100K) were used along with six age bands (50–54, 55–59, 60–64, 65–69, 70–74, and 75–80). This results in 24 income -by-age cells (e.g., 50–54 with <30K, 50–54 with 30–59K, etc.).

This analysis replicated the intersectional fairness framework described in Lett et al. (2025) using a synthetic colorectal cancer (CRC) screening cohort. Rather than emergency admissions, the outcome of interest was unmet screening need: an individual is classified as needing screening (outcome = 1) if they have not been screened in the last five years and are eligible based on age and risk criteria. The primary impact in this context is a false negative or failing to identify an unscreened individual who needs intervention.

In this model, no biased labels were used, and all models were trained on screening during the past five years. Instead, a bias augmentation approach was applied to 1,807 train / 1,697 test observations in two younger, lower income brackets (age {50–54, 55–59} x income {<30k, 30k–59k}). Ambulatory visits over the last two years were adjusted downward by one standard deviation, while comorbidities were increased by 0.5 standard deviations. Table 11 compares the disadvantaged subgroup (age 50–59, income <$60k) to the non-disadvantaged subgroup after synthetic bias augmentation. Disadvantaged individuals are expected to show fewer ambulatory visits (downward shift of −1.0 SD, reflecting access barriers) and higher comorbidity counts (upward shift of +0.5 SD, reflecting unmanaged chronic disease burden). Both features also carry additional measurement noise. Screening status reflects the true (unbiased) outcome label. Note that the screening rates in the disadvantaged group are higher than in the advantaged group at baseline because Synthea generates higher screening rates for younger adults or those with more recent screening. This also demonstrates the value of comparing baseline statistics to expected statistics rather than simply relying on the raw comparison of metrics when assessing potential bias.

Copyright © 2026 Society of Actuaries Research Institute

Table 11
Scenario 3—Cohort Characteristics After Noise Augmentation

Characteristic	Overall (N = 20,277[1])	Characteristic Age 50–59, Income < $60k) (N = 3,504[1])	Non-disadvantaged (N = 16,773[1])
Age (years)			
Mean (SD)	65.53 (8.66)	54.61 (2.85)	67.82 (7.66)
Median (Q1, Q3)	66.00 (58.00, 73.00)	55.00 (52.00, 57.00)	69.00 (62.00, 74.00)
Annual Income (USD)			
Mean (SD)	68,595.33 (98,899.31)	30,126.90 (15,769.26)	76,631.67 (106,765.37)
Median (Q1, Q3)	47,544.00 (24,972.00, 80,914.00)	29,959.50 (16,856.00, 43,173.50)	56,662.00 (27,841.00, 89,337.00)
BMI			
Mean (SD)	28.59 (2.42)	28.82 (2.33)	28.54 (2.44)
Median (Q1, Q3)	28.10 (27.60, 30.00)	28.40 (27.70, 30.10)	28.10 (27.60, 30.00)
Sex			
Female	9,708 (48%)	1,743 (50%)	7,965 (47%)
Male	10,569 (52%)	1,761 (50%)	8,808 (53%)
Obesity			
No	4,292 (21%)	715 (20%)	3,577 (21%)
Yes	15,985 (79%)	2,789 (80%)	13,196 (79%)
Hypertension			
No	11,624 (57%)	2,272 (65%)	9,352 (56%)
Yes	8,653 (43%)	1,232 (35%)	7,421 (44%)
Ambulatory Visits (Last 2 Years)[2]			
Mean (SD)	10.77 (26.91)	6.09 (20.63)	11.75 (27.95)
Median (Q1, Q3)	5.00 (0.00, 9.00)	0.00 (0.00, 0.00)	5.00 (3.00, 9.00)
Comorbidity Count[2]			
Mean (SD)	1.68 (0.97)	1.88 (1.15)	1.64 (0.93)
Median (Q1, Q3)	2.00 (1.00, 2.00)	1.84 (1.03, 2.66)	2.00 (1.00, 2.00)
Colorectal Cancer Screening Status			
Needs Screening	9,363 (46%)	1,298 (37%)	8,065 (48%)
Screened	10,914 (54%)	2,206 (63%)	8,708 (52%)

1 n / N (%)
2 Directional bias augmentation applied to the disadvantaged subgroup: ambulatory visits shifted −1.0 SD (access barrier) + noise (1.0 SD); comorbidity count shifted +0.5 SD (unmanaged burden) + noise (1.0 SD). Values floored at 0.

Gradient boosting classification models were trained, validated, and tested on a 50/50 data split of a cohort of 20,277 individuals aged 45–75 to predict colorectal cancer screening need, specifically targeting patients who had unmet screening needs (if they had not been screened in the last five years). Age and income were categorized into 24 intersectional bands.

Three statistical models were trained and evaluated:

- a baseline random forest model with no fairness constraints;
- a model with a post-hoc calibration applied to base model predictions to iteratively target “calibration fairness,” in other words, to ensure that the model’s scores are equally trustworthy across income-age intersection groups; and

Copyright © 2026 Society of Actuaries Research Institute

- a model with reweighting during the training process to reduce false negative rates (FNR).

These different fairness adjustments to the same base machine learning model demonstrate that the fairness objective choice depends on the exact definition.

These results confirm that meaningful fairness techniques must be explicitly tuned against the precise definition rather than applied by default. Table 12 demonstrates that fairness adjustments did not significantly reduce overall performance of the model.

- The Baseline RF achieves an AUROC of 0.698 and AUPRC of 0.739, with an overall FNR of 47.1%, meaning nearly half of all truly unscreened individuals are missed.
- Multicalibration produces nearly identical recall (52.5%) and slightly lower AUROC (0.674) and AUPRC (0.657), indicating that the local probability adjustments reduced overall disparities slightly without improving screening sensitivity.
- The FNR-reweighted RF maintains the same AUROC (0.698) and AUPRC (0.740) as the baseline but shifts its operating point: recall drops from 52.9% to 47.1% overall, and the positive rate falls from 37.3% to 29.8% (threshold rises from 0.440 to 0.510). This reflects the model's reweighting strategy which concentrates weight on hard-to-detect positives in disadvantaged subgroups, it becomes more selective overall while improving coverage in the worst-off groups specifically.

Table 12
Scenario 3—Model Performance

Model	AUC	AUPRC	Recall	F1	Positive Rate
Baseline	0.698	0.739	0.529	0.585	0.373
Multicalibration RF	0.674	0.657	0.525	0.583	0.370
FNR-reweighted RF	0.698	0.740	0.471	0.573	0.298

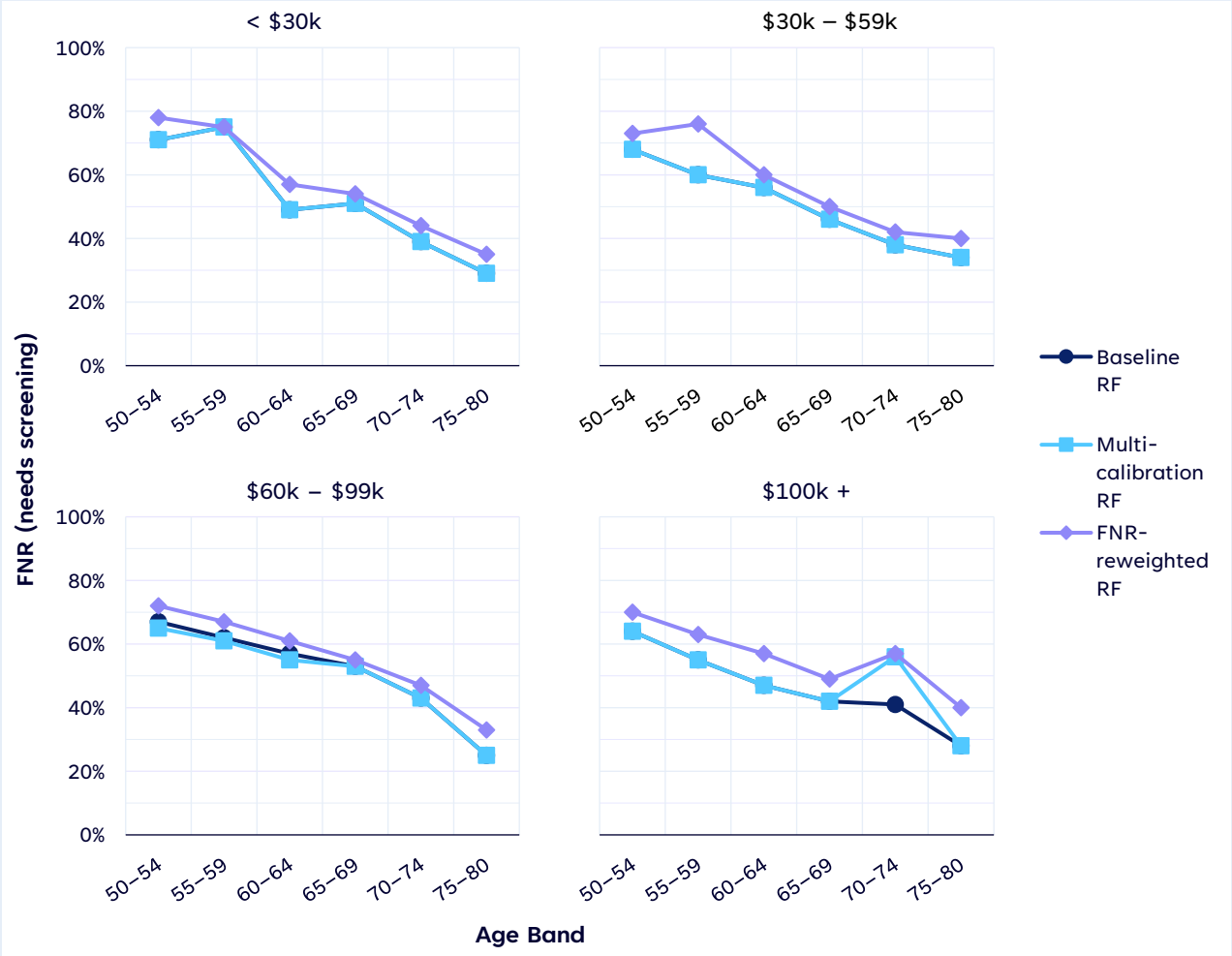
The same approach is applied to Intersectional FNR as shown in Table 13. The main finding is that multicalibration produced no change in intersectional FNR. The gap remained at 50.2% with the identical worst-case group. This is expected as multicalibration optimizes for calibration within subgroups (agreement between predicted probability and observed outcome rate), not for FNR parity. When predicted probabilities shift to improve calibration, the rank ordering and, therefore, the classification decisions at a fixed threshold may be largely unchanged. FNR-reweighted RF reduced the FNR gap by 5.4 percentage points (50.2% → 44.8%). The worst group shifted from 55–59 / <$30k to 50–54 / <$30k, indicating the reweighting improved coverage for the 55–59 cell at the cost of slightly increased FNR in the 50–54 cell. The maximum FNR rose slightly (75.5% → 77.6%), illustrating the classic minimax fairness tradeoff: the algorithm redistributes errors more evenly but cannot eliminate them. The minimum recall across groups also decreased from 24.5% to 22.4%, consistent with trading off some best-case performance to improve worst-case performance. These results are also shown by subgroup in Figure 3.

Copyright © 2026 Society of Actuaries Research Institute

Table 13
Scenario 3—Intersectional FNR

Model	Min FNR	Max FNR	FNR Gap	Worst Group
Baseline	0.254	0.755	0.502	Age 55–59; <$30k
Calibration Fair	0.254	0.755	0.502	Age 55–59; <$30k
FNR-reweighted	0.327	0.776	0.448	Age 55–59; <$30k

Figure 3
False Negative Rate (FNR) by Intersectional Group (Age X Income)



Source: authors. Lower FNR = fewer missed unscreened people; flatter lines = fairer across groups.

The results illustrate a foundational principle of algorithmic fairness: the choice of a fairness algorithm must match the operational impact. In a screening context, where the primary negative impact is failing to identify unscreened individuals, FNR parity (equal opportunity) is the appropriate objective. Multicalibration produced no change in the intersectional FNR gap (50.2% in both Baseline and Multicalibration RF), because it targets calibration rather than classification equity. The FNR-reweighted approach, which directly penalizes false negatives in high-FNR subgroups during training, reduced the gap from 50.2% to 44.8%, though the worst-case FNR in the most disadvantaged cell (50–54 / <$30k) remained high at 77.6%.

Copyright © 2026 Society of Actuaries Research Institute

The results across all three scenarios and overall demonstrate the need for careful assessment of how a model is developed and the need for potential mitigation measures. The report next discusses the model card approach to documentation of model building, along with a decision assistance method built into the project tool for determining whether to apply mitigation techniques.

Copyright © 2026 Society of Actuaries Research Institute

Section 6 Model Card and Decision Assistance Tool

6.1 Introduction to Model Cards

6.1.1 Model Cards as "Nutrition Labels"

Model cards provide standardized, at-a-glance documentation of AI/ML systems, analogous to nutrition labels on food products. Just as consumers use nutrition labels to quickly assess whether a product meets their dietary needs and constraints, actuaries can use model cards to rapidly evaluate whether an AI/ML model is fit for their intended purpose. This "nutrition label" approach transforms complex technical documentation into an accessible format that highlights the most decision-relevant information: what the model does (Technical Specifications), what it should and should not be used for (Intended Use), which population characteristics may affect performance (Factors), and how well it performs (Metrics) (Ezi et al., 2022). The nutrition label metaphor emphasizes the desire for model documentation to be immediately interpretable by practitioners who need to make fit-for-purpose determinations and the technical developers who built the model.

In this project, model cards serve as the primary interface between the framework (Section 2.1) and practical implementation decisions. They synthesize results from representativeness assessments, statistical adjustment methods, and validation testing, along with model information, into a format that supports actuarial judgment about whether a model's known limitations are acceptable for a given use case. By standardizing how representativeness information is documented and presented, model cards enable actuaries to systematically evaluate whether claims data used in AI/ML development reflects the populations to which the model will be applied, making visible the data quality considerations essential for responsible actuarial practice.

6.1.2 HuggingFace Standard

Several model card frameworks exist for healthcare AI applications. This includes specialized templates for clinical end users and regulatory compliance (Sendak et al., 2020; Coalition for Health AI, 2024). This project adopts the HuggingFace model card structure as its foundational template for several methodological reasons. The HuggingFace framework, substantially developed as an open standard through the HuggingFace Model Card Guidebook, emphasizes rigorous quantitative analysis with explicit requirements for disaggregated performance reporting across demographic, geographic, and clinical subgroups (Ezi et al., 2022). This aligns directly with actuarial practice, where understanding model performance variation across risk classification or pricing factors is fundamental to assessing whether a model is fit for purpose. The framework requires both unitary results (performance by individual factors) and intersectional results (performance across combinations of factors), enabling actuaries to identify representation gaps that might be masked in aggregate statistics. This granular analytical approach maps naturally onto the bias assessment framework (input bias, model bias, application bias) by providing structured documentation of where disparities emerge in the model development and deployment pipeline.

From a methodological standpoint, HuggingFace model cards place greater emphasis on comprehensive data documentation compared to compliance-oriented alternatives. The

Copyright © 2026 Society of Actuaries Research Institute

framework's separate sections for Evaluation Data and Training Data, each requiring details on datasets, data splits, preprocessing steps, and representativeness assessment, support rigorous data quality standards. This separation enables independent assessment of whether training data representativeness issues have been adequately addressed through validation on more representative test datasets, or whether systematic representation gaps persist across the development lifecycle.

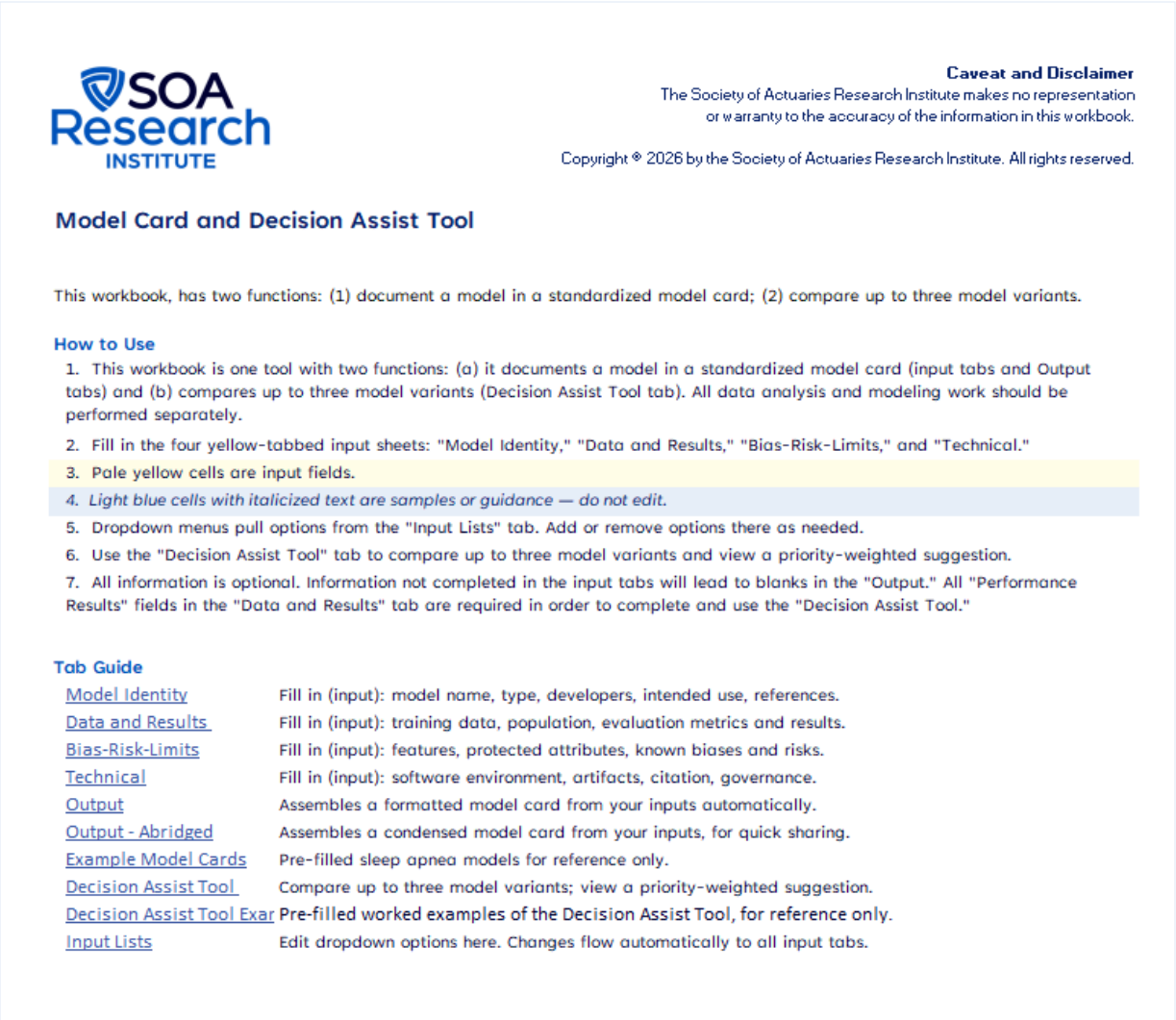
Additionally, the HuggingFace framework's widespread adoption, open-source implementation, and machine-readable format provide practical advantages for developing reproducible actuarial tools. Since the original proposal in 2018, HuggingFace has become a major platform for hosting model cards, with all HuggingFace-shared models on the Hub accompanied by model cards implemented as README files in model repositories (Ezi et al., 2022). This extensive implementation experience has resulted in well-documented best practices, automated generation tools, and a large community of practice. The framework is designed to integrate with automated model documentation systems, enabling actuaries to develop standardized packages and reporting workflows that generate model cards as part of model validation pipelines. This technical flexibility supports consistency across analyses and reduces manual documentation effort while maintaining comprehensive coverage of representativeness assessment requirements.

6.2 Model Card Generator and Decision Assist Tool for Actuarial Applications

The framework described above is implemented as the SOA Research Institute Model Card and Decision Assist Tool in an Excel workbook developed as a companion deliverable. The workbook guides practitioners through documenting a model for bias and representativeness. It provides four structured input tabs organized by topic: "Model Identity," "Data and Results," "Bias-Risk-Limits," and "Technical." Pale yellow cells are input fields; white cells with royal blue text are labels; italicized text in light blue cells is guidance on what to include. Dropdown menus for common values (model type, geographic scope, programming language) pull from a configurable "Input Lists" tab. The "Output" tab generates a formatted model card once the inputs are complete. The cover sheet of the workbook is shown below in Figure 4.

Copyright © 2026 Society of Actuaries Research Institute

Figure 4
Model Card Generator and Decision Assist Tool Structure (Screenshot)



SOA Research INSTITUTE

Caveat and Disclaimer
The Society of Actuaries Research Institute makes no representation or warranty to the accuracy of the information in this workbook.

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

Model Card and Decision Assist Tool

This workbook, has two functions: (1) document a model in a standardized model card; (2) compare up to three model variants.

How to Use

1. This workbook is one tool with two functions: (a) it documents a model in a standardized model card (input tabs and Output tabs) and (b) compares up to three model variants (Decision Assist Tool tab). All data analysis and modeling work should be performed separately.
2. Fill in the four yellow-tabbed input sheets: "Model Identity," "Data and Results," "Bias-Risk-Limits," and "Technical."
3. Pale yellow cells are input fields.
4. *Light blue cells with italicized text are samples or guidance — do not edit.*
5. Dropdown menus pull options from the "Input Lists" tab. Add or remove options there as needed.
6. Use the "Decision Assist Tool" tab to compare up to three model variants and view a priority-weighted suggestion.
7. All information is optional. Information not completed in the input tabs will lead to blanks in the "Output." All "Performance Results" fields in the "Data and Results" tab are required in order to complete and use the "Decision Assist Tool."

Tab Guide

Model Identity	Fill in (input): model name, type, developers, intended use, references.
Data and Results	Fill in (input): training data, population, evaluation metrics and results.
Bias-Risk-Limits	Fill in (input): features, protected attributes, known biases and risks.
Technical	Fill in (input): software environment, artifacts, citation, governance.
Output	Assembles a formatted model card from your inputs automatically.
Output - Abridged	Assembles a condensed model card from your inputs, for quick sharing.
Example Model Cards	Pre-filled sleep apnea models for reference only.
Decision Assist Tool	Compare up to three model variants; view a priority-weighted suggestion.
Decision Assist Tool Exar	Pre-filled worked examples of the Decision Assist Tool, for reference only.
Input Lists	Edit dropdown options here. Changes flow automatically to all input tabs.

Source: authors.

Two other important features are the naming of model card elements in Column A of each input tab and the “Input Lists” tab. The model card elements in Column A have specific names that connect to the corresponding Output element. For example, “Model Name / Title” in the “Model Identity” tab is also in the Output and Output—Abridged tabs. Users who wish to change the names of fields or add fields are free to do so. However, this may cause the underlying functions in the model to “break” or to fail to function without additional corresponding changes to the “Output” tabs by the user.

6.2.1 Inputs

The workbook is a Microsoft Excel-based tool that is designed for an actuary, actuarial team, or other user who has already begun to develop models. Once results are available, the user can use the model card generator within the workbook to document the work that has been done, assess the need for additional modeling, and identify potential for bias through the decision

Copyright © 2026 Society of Actuaries Research Institute

assist portion of the workbook. This can also be performed in an iterative fashion, with initial model documentation, development, and refinement. There are no required elements of the model card, so teams can use the generator at any point in the modeling process.

Each of the four input tabs captures different elements of a model:

- **"Model Identity"** captures who built the model, what it does, and how it should and should not be used. Within this tab, Model Details provides the technical foundation for assessing whether a model's development approach aligns with the intended actuarial application.
- **"Data and Results"** documents the training dataset, population, sampling design, and evaluation metrics, including subgroup recall.
- **"Bias-Risk-Limits"** records features used, protected attributes, and known sources of bias by type, such as selection, measurement, and label bias. Inputs also include whether a formal disparate impact analysis was performed.
- **"Technical"** covers software environment, model artifacts, citation, and governance, including review cadence and sign-off.

Pre-filled example cards for the three sleep apnea variants are available in the "Example Model Cards" tabs (one Abridged and one Full) as a reference.

The Excel workbook is designed to be a free-standing and self-explanatory tool including both the model card and decision assistance function that accompanies this report. Additional details for key sections are discussed below.

6.2.1.1 Model Identity

The Model Identity tab documents the model's provenance, including who developed it, when, under what constraints, and using which methodological approaches. For actuaries evaluating vendor-supplied models or assessing internally developed tools, these specifications enable initial screening for appropriateness before conducting detailed validation work.

The Intended Use section explicitly delineates the actuarial contexts for which the model was designed and validated, including identification of some situations where the model's training assumptions or data characteristics may not hold. Model summary is intended for the user to specify the exact actuarial functions the model supports (e.g., prospective risk adjustment, commercial small group pricing for fully insured products, retrospective utilization review). For each application, documentation is intended to identify the decision context, typical magnitude of financial impact, and regulatory environment. This specificity enables actuaries to assess whether their use case falls within the model's validated scope.

Out-of-Scope Use identifies specific applications for which the model is not intended to be deployed without additional validation or modification. Common exclusions include application to populations with substantially different demographic composition than the training data, use for individual-level predictions when trained only on aggregate data, deployment in environments with different variable restrictions than those assumed during development, or application to products with benefit designs that differ materially from the training data.

6.2.1.2 Data and Results

Population and Sampling entries identify the demographic, geographic, clinical, and administrative characteristics that may influence model performance and data

Copyright © 2026 Society of Actuaries Research Institute

representativeness. Unlike clinical risk adjustment models, where factors directly predict outcomes, model cards use factors to organize the assessment of representation gaps and performance variation. Core demographic factors include age bands, sex, and geographic region at appropriate granularity (e.g., three-digit ZIP, county, state). Clinical factors relevant to claims data applications typically include chronic condition indicators, utilization intensity measures, and pharmacy utilization patterns. Administrative factors address plan design characteristics, network participation, and claims coding completeness that may vary systematically across populations.

Performance measures focus on whether the model performs consistently across population segments. For classification models, measures include sensitivity, specificity, area under the ROC curve, and subgroup-specific recall, reported separately for relevant demographic and geographic groups. To enable the detection of representation problems that aggregate metrics would mask, all performance measures must be disaggregated.

Training Data documentation addresses the core data quality metrics, including what data were used, how representative they are of the target population, and what limitations or deficiencies exist. The training data section specifies data sources (e.g., commercial claims database, Medicare fee-for-service sample, integrated delivery system records), coverage period, and population characteristics. Documenting training data decisions is important because these decisions directly affect representativeness. For example, excluding members with short enrollment periods, limiting to specific geographic regions, or filtering based on data completeness criteria all create potential representation gaps to be evaluated against the intended deployment population.

Detailed documentation accompanies data preprocessing steps because seemingly technical decisions can have substantial actuarial implications. Handling of missing data (exclusion, imputation, flagging), coding of categorical variables (granularity of age bands, geographic aggregation level), and treatment of outliers all affect which population segments are well-represented in the final model. The documentation characterizes data quality metrics, including completeness (percentage of records with key variables populated), consistency (coherence of longitudinal records), and accuracy (validation against external benchmarks where available).

Evaluation Data documentation follows a parallel structure but emphasizes how the validation dataset addresses potential representativeness concerns in training data. If training data come from a single geographic region or plan type, evaluation data from different regions or products provide evidence about model portability. Temporal validation using hold-out periods assesses whether the model remains stable as clinical coding practices and utilization patterns evolve. The section is intended to explicitly compare the demographic, geographic, and clinical composition of training versus evaluation datasets, documenting any systematic differences and their implications for model generalizability. For vendor models, this section enables evaluation of whether the model has been validated on data similar to their own book of business, or whether additional validation work is required before deployment.

6.2.1.3 Bias-Risk-Limits

The Bias-Risks-Limits section of the workbook documents the features used, the protected attributes involved, and the known sources of bias, risk, and limitation for this model. This sheet captures why a model may perform unevenly across groups, which complements the metrics on

Copyright © 2026 Society of Actuaries Research Institute

the Data and Results sheet, which capture whether it does. These entries become the "Bias, Risks, and Limitations" section of the model card. Documenting limitations, even where they are uncertain, may help future users of the model to understand certain design features or modeling choices. The bias types in the workbook map to this report's three-stage view of bias: selection and measurement bias are forms of input bias; label bias can be entered at input or model stages; distribution shift is primarily an application-stage risk.

6.2.1.4 Technical

Technical specifications include the model architecture (e.g., gradient boosted trees, neural networks, generalized linear models), training algorithms, and hyperparameter settings. Developer information identifies the organization responsible for model development and maintenance. This section is also intended to specify the original population for which the model was developed, including geographic scope, plan types, and time period, establishing the baseline against which representativeness will be assessed.

6.2.2 Outputs

The tool produces a structured model card organized into four sections. The first three establish context: Model Identity records the model name, type, version, developers, and licensing; Intended Use specifies direct and downstream applications, the populations to which the model is expected to generalize, and explicit out-of-scope uses; and Training Data and Population documents the dataset, its source, preprocessing steps, sampling design, and key demographic and geographic characteristics of the underlying population.

The remaining three sections address performance and accountability. Evaluation and Performance reports aggregate metrics alongside subgroup-level recall, making disparities across population groups visible in a standardized format. Bias, Risks, and Limitations documents the features and protected attributes used, classifies known bias by type (selection, measurement, and label), and records whether a formal disparate impact analysis was conducted. Technical Specifications covers the software environment, model artifacts, citation, and governance, including review cadence and sign-off authority.

The Output – Abridged tab presents a version of the model card with key data elements. These have been selected by the research team and can be edited by the user. In addition, this tab has been formatted for printing on a single double-sided piece of paper. Users who want to add or remove elements are free to do so; however, this may alter the formatting designed into the default version of the model.

6.2.3 Examples

To illustrate the adapted framework, three model cards were developed for the Scenario 1 sleep apnea models: a card for the true (unbiased) model, one for the biased model trained on masked rural diagnoses, and one for the propensity score matched model. The tool generates one model card at a time from the inputs on the four input tabs. The three cards shown in the workbook were produced by running the tool once per model variant and were placed side by side to make the comparison visible; the workbook does not generate multiple cards simultaneously. Each card follows the model card approach within the tool, documenting model details, intended use, bias and limitations, training data and procedure, and evaluation results, including subgroup recall by urban/rural geography. Together, the three cards make visible how

Copyright © 2026 Society of Actuaries Research Institute

training data decisions propagate into measurable performance disparities, and how those disparities are partially, but not fully, addressed by propensity score matching. **Pre-filled versions of all three example model cards are available in the accompanying Excel file that is published along with this report.**

6.3 Decision Assistance Tool

6.3.1 Inputs from Model Card

The SOA Model Card and Decision Assist Tool is a companion Excel-based decision assistance tool designed to support model selection among the three model variants. The workbook shows as an example the three models from Scenario 1: the true model, the biased model, and the propensity score matched model. Those using the tool can input their own values based on the models they are assessing. Rather than requiring practitioners to manually reconcile competing performance metrics, the tool accepts user-specified priority weights and returns a structured recommendation. In practice, the use of multiple models allows for assessment and comparison for model selection.

The decision assistance tool is based on the model card inputs that the users provide. These include the statistics about the model, including overall performance and performance in subgroup analysis. The decision assistance tool then uses these to calculate an equity ratio and a recall gap. The decision assistance tool will populate the input items from the model card inputs for the first model, and users can enter up to two additional models as they would like. Users enter weights reflecting their application priorities across seven metrics: overall accuracy, recall, precision, F1 score, equity score, recall, and precision.

6.3.2 Outputs and Decision Assistance Tool

Each model's performance on each metric is scored using a red/yellow/green threshold system, where green indicates acceptable performance, yellow indicates marginal performance, and red indicates values below an acceptable threshold. The tool computes a weighted composite score across models and displays the suggested model prominently. A "Suggested Model" section at the bottom identifies specific metrics where the model still performs poorly, giving practitioners a starting point for additional adjustment or targeted documentation in the model card.

Figure 5 shows the tool interface. With equal weighting across metrics, the model named "True" is reflecting its balance of aggregate accuracy and subgroup performance. Practitioners whose application prioritizes rural equity over overall accuracy may shift weights accordingly and observe how the recommendation responds. **The tool is available alongside the model cards in the accompanying Excel file that is published along with this report.** In this example, the biased model scores low on representation, while the model named "True" scores highly. The propensity score matched version sits in the middle, showing the improvement that is generated in this example by adjusting the original model outputs through a propensity score matched version.

Copyright © 2026 Society of Actuaries Research Institute

Figure 5
SOA Model Card Generator Decision Assistance Tool (Screenshot)

Model Selection Decision Assist Tool Worked Examples				Caveat and Disclaimer The Society of Actuaries Research Institute makes no representation or warranty to the accuracy o the information in this workbook
This sheet is a worked example with pre-filled values. Model 1 ('Biased') was trained on misrepresentative data; Models 2 and 3 are alternatives. Change the Priority Weight to see how the suggestion responds. Subgroup Recall Ratio = Subgroup B Recall ÷ Subgroup A Recall (closer to 1.0 = smaller disparity).				
Priority Weight One setting applies to all three models.	Balanced			*Accuracy First = 70% AUC / 30% ratio; Balanced = 50/50; Parity First = 30% AUC / 70% ratio. Weights are shown on the Input Lists tab.*
Metric	**Model 1**	**Model 2**	**Model 3**	**Notes**
Model Name / Variant	**Biased**	**True**	**Propensity-Corrected**	*e.g., Biased, True, Propensity-Corrected*
AUC	0.6803	0.82	0.79	*Area under ROC curve (0 to 1)*
Accuracy	0.7416	0.78	0.765	*Overall classification accuracy*
Precision	0.176	0.22	0.2	*Positive predictive value*
Overall Recall	0.4798	0.55	0.52	*Sensitivity across all groups*
Subgroup A Recall	0.5698	0.56	0.54	*Urban Recall (Subgroup A)*
Subgroup B Recall	0.1111	0.54	0.48	*Rural Recall (Subgroup B)*
F1 Score	0.2549	0.3277	0.297	*Harmonic mean of precision and recall*
Calculated Scores (auto-updates as you enter data above)				
Subgroup Recall Ratio (B	0.195	0.964	0.889	*Subgroup B Recall ÷ Subgroup A Recall. Closer to 1.0 = smaller disparity.*
Recall Gap (A minus B)	0.459	0.020	0.020	*Absolute recall gap in percentage points. Smaller gap = smaller disparity.*
Composite Score	**0.438**	**0.892**	**0.839**	*Weighted score based on priority. Higher = stronger candidate.*
Suggestion				
Suggested Model	**True**			*Model with the highest composite score given the selected Priority Weight*
Bias Diagnosis				
Flags fire automatically from the Subgroup Recall Ratio above. Thresholds: ratio < 0.80 = High Concern (orange); 0.80–0.90 = Monitor (yellow); ≥ 0.90 = Acceptable (green). A recall gap > 0.20 may indicate a large clinical/actuarial impact.				
Disparity Signal	**High Concern**	**Acceptable**	**Monitor**	*Orange = ratio < 0.80; Yellow = 0.80–0.90; Green = ≥ 0.90*
Recall Gap Flag	**Large gap (>20pp) — likely significant impact**	**Small gap (<10pp)**	**Small gap (<10pp)**	*Absolute percentage-point gap in recall between Subgroup A and B*
Likely Cause	Severe underdiagnosis or systematic exclusion of subgroup in training data — check label/diagnosis bias	No significant disparity detected	Mild disparity — check proxy variables and sampling balance	*Based on the Subgroup Recall Ratio. Review the Bias-Risk-Limits tab for confirmation.*
Suggested Mitigation(s)	(1) Propensity score reweighting of training sample (2) Re-stratified sampling to balance subgroups (3) Adversarial debiasing during model training (4) Separate models per subgroup if populations are clinically	No mitigation indicated — document the subgroup recall ratio in the model card and set a monitoring threshold	(1) Audit proxy variables (e.g. zip code, provider type) (2) Subgroup calibration post-hoc (3) Add fairness constraint to training objective	*Mitigation methods ranked by invasiveness. Lower numbers = try first.*

Source: authors.

6.4 Actuarial Practice Extensions

Production actuarial model cards would augment the core sections with claims-data-specific elements, such as data lineage, coding-intensity variation, plan-design effects, and application-specific validation requirements. These extensions are out of scope for the educational model cards developed for this research, which use synthetic data and a single actuarial scenario. Practitioners applying this framework to operational models may want to consider them (see, e.g., Luckyk et al., 2017 or Gidwani et al., 2025 for additional examples and specific applications). In addition, the need to monitor and, potentially, to periodically update the model card is built into the design of this tool. The question of review cadence is also explicitly built into the Technical tab of the spreadsheet.

Copyright © 2026 Society of Actuaries Research Institute

As an Excel workbook, dynamically revising the tool is relatively simple. The decision assist function can be used to select a model, whether the suggested version or another version, and that can become the baseline for comparison using the decision assist feature when the model results are re-evaluated in the future. However, there is no explicit requirement to revise the model cards in the future, and monitoring and revision of the model is outside of the scope of the model card generator portion of the tool designed here.

Copyright © 2026 Society of Actuaries Research Institute

Section 7 Implications and Conclusions

7.1 Limitations

7.1.1 Synthetic Data

This research has several limitations to consider when applying the framework. First, synthetic data generated by Synthea, while externally benchmarked, cannot fully replicate the complex dependencies and administrative idiosyncrasies of real claims data. Controlled simulation isolates specific bias mechanisms but may understate how multiple sources of bias compound in practice. Second, propensity score matching did not fully restore parity in Scenario 1, and the degree to which any statistical mitigation can recover from severe underrepresentation remains an open question that depends heavily on the missingness mechanism and available covariates.

7.1.2 Additional Areas of Practice

The framework was developed and tested on generic, broadly insured synthetic populations. Calibration to commercial-type plans, Medicare Advantage, Medicaid, or self-insured employer populations would require additional validation. In addition, fairness metrics presented in Table 2 are not simultaneously achievable in most real settings; practitioners typically must choose among competing definitions based on the specific impact(s) they most want to avoid. Finally, the model card structure proposed here does not replace actuarial judgment about fitness-for-purpose, and completion of a "checklist" or other similar document does not constitute formal validation.

7.2 Tips for Implementation

7.2.1 Bias Reduction in Practice

Based on the framework and results from all three scenarios, the following general tips for actuaries working with AI/ML models on claims data arise from this research:

- Disaggregating model performance by subgroup evaluating subgroups jointly as well as separately can help identify when a model may appear acceptably accurate when evaluated overall while failing materially for underrepresented or access-constrained subgroups. Scenario 1 showed that the overall AUC was stable (0.674–0.687), while rural recall collapsed by 81%, a disparity that was invisible in aggregate metrics. Scenario 3 extended this further: single-dimension assessments of age or income alone missed the compounding disadvantage faced by young, low-income patients, whose screening deficit was largest precisely at the intersection of those two characteristics.
- Reporting subgroups by geography, age group, clinical cohort, and any other relevant variables, as well as their intersections can help avoid overlooking potential bias that is hidden in the aggregate.

7.2.2 Iterative Process

Applying the three-stage bias assessment framework iteratively, rather than as a one-time check, can be very helpful. Any representation problems identified in Stage 1 may propagate through modeling choices and persist into deployment. To avoid or limit such unintended

Copyright © 2026 Society of Actuaries Research Institute

propagation, revisit the assessment whenever data sources change, the model is applied to a new population, or monitoring detects performance drift in any subgroup.

It may be helpful to expect that different mitigation methods optimize for different things and explore as many of them as is feasible. In Scenario 1, propensity score matching partially improved rural recall but did not close the gap. In Scenario 2, imputation improved demographic parity while a missingness-aware model was superior for error-rate parity; neither dominated on all metrics. Combining complementary approaches and using sensitivity analyses to understand how robust results are to those choices can help practitioners optimize results.

Model cards are a living document; they are most useful when updated at each validation checkpoint, when data sources change, and when deployment context shifts. For vendor-supplied models, receiving model card documentation as part of procurement can be helpful to validate that the training and evaluation populations are sufficiently similar to a company's own book of business.

7.2.3 Metric Selection and Disclosures

Selecting fairness metrics before modeling, not after, and tuning mitigations directly to those metrics can help avoid post-hoc metric selection ("metric shopping"). In Scenario 3, mitigation strategies that were not optimized for intersectional recall widened the minimax recall gap compared to the biased model, despite improving aggregate performance. The choice of fairness criterion shapes which mitigation strategies are appropriate and whether they conflict with overall accuracy. The authors suggest explicitly choosing fairness criteria grounded in the metrics that matter most for the actuarial application and documenting choices with rationale in the model card.

Disclosing known representation gaps explicitly may be helpful, even when they cannot be fully addressed. Transparent documentation of where a model is likely to underperform enables downstream users to apply appropriate judgment. Disclosure can indicate responsible deployment and support practice in model documentation and communication.

7.2.4 Ethical Considerations and Limitations

Ethical considerations for actuarial applications include the context in which models are applied. For healthcare claims data, key considerations may include handling of protected health information and appropriate use of potentially sensitive variables like geography or utilization patterns that may correlate with protected or other sensitive characteristics. Given that the data used in this project was synthetic, these specific considerations do not apply but may need to be considered when applying the results and techniques.

Limitations directly affect the fitness-for-purpose assessment, so comprehensive documentation can help model users ascertain whether a model is appropriate for the task at hand. Potential limitations include, but are not limited to:

- Known representation gaps identified through the bias assessment. Clearly describing them along with their potential impact on model performance can help model users spot potential concerns about fitness for purpose.
- Data quality issues that could not be fully resolved, such as coding intensity variation across provider systems, incomplete capture of out-of-network utilization, or limited

Copyright © 2026 Society of Actuaries Research Institute

geographic diversity in training data. Documenting and communicating the issues to model users can help them make informed decisions about fitness for purpose.

- Methodological limitations, including simplified model architecture for operational efficiency, constraints on variable selection due to regulatory requirements, or validation challenges due to limited sample sizes in key subpopulations. All of these may affect a model's fitness for purpose.

Comprehensive documentation emphasizes to users the need to assess whether the documented limitations are acceptable given their specific application and the magnitude of financial or operational decisions being supported by the model.

7.3 Implementation Guidance

Model card implementation integrates documentation requirements into existing actuarial validation workflows rather than creating separate compliance processes. Initial model card drafts created during a project's planning phase document intended use, target population, and anticipated data sources, providing a framework for validation testing. As validation proceeds, performance metrics populate incrementally, with sections reviewed during validation checkpoints to ensure documentation remains current. This iterative approach prevents documentation from lagging behind development, creating incomplete or inaccurate model cards that provide limited value for deployment decisions.

Maintenance protocols establish triggers for model card updates based on material changes in data sources, model architecture, deployment context, or observed performance. Annual refreshes verify that documentation remains current even when no substantive model changes have occurred. Material changes triggering immediate updates include shifts in data sources affecting representativeness assessments, algorithm modifications changing model architecture, regulatory changes affecting intended use specifications, or performance monitoring detecting degradation in subgroup-specific metrics.

Scenario-specific application demonstrates how the three synthetic data scenarios inform model card completion for models trained on data with known representation characteristics. For example, models trained on data resembling:

- Urban/Rural disparities in sleep apnea diagnosis, geographic limitation disclosures are important, noting that rural population underrepresentation may affect model performance in frontier regions.
- Clinical Documentation Bias in Type 2 diabetes implications, documentation that provider specialty affects diagnosis capture rates is key, with limitations noting that systematic clinical documentation gaps may underestimate disease burden when applied to populations with different coding intensity patterns.
- Intersectional bias in age and income impacts on colorectal cancer screening rates implications, age distribution documentation, and restrictions on application to age ranges adequately represented in training data are important.

7.4 Implementation with R and Python

A key practical message of this research is that implementing the bias assessment framework is substantially easier than it may appear. All analyses in this report were conducted in R and

Copyright © 2026 Society of Actuaries Research Institute

Python using freely available, well-documented open-source packages. Propensity score methods, entropy balancing, fairness metric calculation, and model card generation are achievable with open-source model package calls. Appendix A lists the specific packages used for each component of the analysis. Actuaries with moderate coding experience in either R or Python can replicate or adapt these methods for their own data without requiring deep machine learning expertise.

For actuaries new to these tools, the three synthetic scenarios in this report function as worked examples: the code and data are available as online appendices, and the scenarios are designed to be adapted to different conditions and populations. The Excel Model Card and Decision Assist Tool requires no coding at all and can be used immediately alongside any modeling workflow.

7.5 Extension to Other Practice Areas

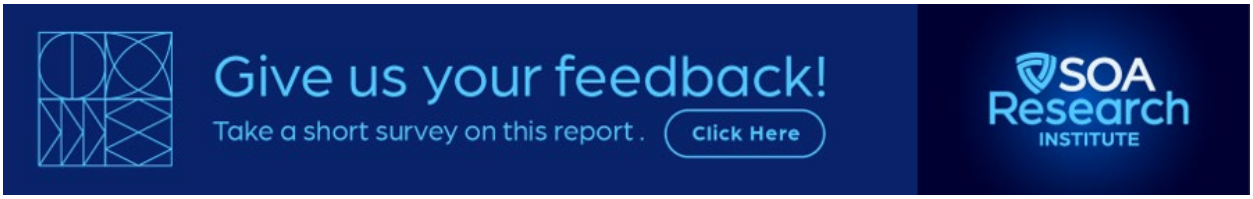
This research was developed in a healthcare claims context, and the representativeness framework and methods may be applicable across a range of actuarial applications. In general, predictive models are trained on historical data that may not reflect the full population of interest; they can be reviewed and may be improved using the techniques highlighted here. Potential geographic and demographic representation gaps in claims experience data used for pricing models, for example, are a widespread challenge. Similarly, underwriting models that reflect historical applicant pools may differ from future markets. Data used for training purposes in other lines with healthcare data, such as disability and long-term care insurance, may underrepresent populations with different access to diagnostic services. This study's three-stage bias assessment framework applies in each of these contexts, although the specific metrics and mitigation methods differ. Practitioners in health and non-health practice areas interested in applying these concepts can look for external benchmarks appropriate to their data, which may include census data for geographic representativeness or industry experience tables for demographic comparisons, as reference populations against which to assess their training datasets.

Copyright © 2026 Society of Actuaries Research Institute

Section 8 AI Note

Claude.AI, Grammarly, Turnitin, and Microsoft Word Editor were used in the preparation of this material. These tools were used for the synthesis and review of the material created by the authors.

Claude Code, OpenAI Codex, OpenAI ChatGPT, GitHub Copilot, and Gemini (CLI and Antigravity) were used in the development of the extended Synthea data, as well as model card development. All code was reviewed, revised, and integrated into the code base by the authors.



Copyright © 2026 Society of Actuaries Research Institute

Section 9 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 for their diligent work overseeing the research framework and deliverables, providing substantive feedback on methodology and findings, and reviewing and editing this report for accuracy and relevance.

Thanks to Sara Teppema, FSA, MAAA, FCA, who was an Advisor for the project team and a reviewer for the report and project deliverables.

Project Oversight Group members

Dorothy Andrews, ASA, MAAA, PhD

Andrew Dillworth, FSA, MAAA

Robert Gomez, FSA, MAAA, CERA

Matthew Kelsey, ASA

Eileen Luxton, FSA, FCIA

Marilyn McGaffin, ASA, MAAA, FLMI

Tri Pham, FSA, MAAA

Tony Pistilli, FSA, MAAA, CERA, FCA

Rose Qian, FSA, CERA

Becky Sheppard, FSA, MAAA

Justus Tulowiecki, ASA

At the Society of Actuaries Research Institute

Rob Montgomery ASA, MAAA, FLMI

Lisa Schilling, FSA, EA, FCA, MAAA

Barbara Scott, Senior Research Administrator

Copyright © 2026 Society of Actuaries Research Institute

Appendix: Open-Source Packages for Implementation

Table 14
Open-Source Packages

Package	Language	Use	Link
ebal	R	Entropy matching for balanced samples	https://cran.r-project.org/web/packages/ebal/index.html
Fairlearn	Python	"Fairlearn is an open-source, community-driven project to help data scientists improve fairness of AI systems."	https://fairlearn.org/
Hugging Face	YAML	"Model cards are files that accompany the models and provide handy information."	https://huggingface.co/docs/hub/model-cards
ipw	R	Inverse probability weighting	https://cran.r-project.org/web/packages/ipw/index.html
MatchIt	R	"Selects matched samples of the original treated and control groups"	https://cran.r-project.org/web/packages/MatchIt/index.html
mice	R	"Multiple imputation using Fully Conditional Specification (FCS)"	https://cran.r-project.org/web/packages/mice/index.html
missingHE	R	Evaluation with missing outcome data	https://cran.r-project.org/web/packages/missingHE/index.html
pysmatch	Python	Propensity score matching	https://miaohancheng.com/pysmatch/
randomForest	R	"Classification and regression based on a forest of trees using random inputs, based on Breiman (2001)"	https://cran.r-project.org/web/packages/randomForest/index.html
survey	R	"Analysis of complex survey samples"	https://cran.r-project.org/web/packages/survey/index.html
Synthea	Java	"SyntheaTM is a Synthetic Patient Population Simulator. The goal is to output synthetic, realistic (but not real), patient data and associated health records in a variety of formats."	https://github.com/synthetichealth/synthea
TriMatch	R	Propensity score matching technique for categorical valued (three or more values) treatments."	https://cran.r-project.org/web/packages/TriMatch/index.html
XGBoost	R	"Extreme Gradient Boosting, which is an efficient implementation of the gradient boosting framework from Chen & Guestrin (2016)"	https://cran.r-project.org/web/packages/xgboost/index.html

Source for quotations: Package documentation.

Copyright © 2026 Society of Actuaries Research Institute

References

Bishop, Crystal D., Walter L. Leite, and Patricia A. Snyder. 2018. “Using Propensity Score Weighting to Reduce Selection Bias in Large-Scale Data Sets.” *Journal of Early Intervention* 40 (4): 347–62. https://doi.org/10.1177/1053815118793430.

Bozkurt, Selen, Eli M. Cahan, Martin G. Seneviratne, et al. 2020. “Reporting of Demographic Data and Representativeness in Machine Learning Models Using Electronic Health Records.” *Journal of the American Medical Informatics Association* 27 (12): 1878–84. https://doi.org/10.1093/jamia/ocaa164.

Bryer, Jason. 2025. *TriMatch: Propensity Score Matching of Non-Binary Treatments*. Version 1.0.1. Released December 8. https://cloud.r-project.org/web/packages/TriMatch/index.html.

Buuren, Stef van, Karin Groothuis-Oudshoorn, Gerko Vink, et al. 2025. *Mice: Multivariate Imputation by Chained Equations*. Version 3.19.0. Released December 10. https://cran.r-project.org/web/packages/mice/index.html.

Centers for Medicare & Medicaid Services (CMS). 2026. “ICD-10.” March 10. https://www.cms.gov/medicare/coding-billing/icd-10-codes.

Cheung, C. Ronny, Mark Butler, Carolyn Cooper, and Harpreet Dhanoa. 2025. “Using Artificial Intelligence and Machine Learning to Promote Child Health Equity.” *Pediatrics* 156 (Supplement 1): e2025070739L. https://doi.org/10.1542/peds.2025-070739L.

Choi, Sugy, Ninez A. Ponce, and Sandro Galea. 2026. “Advancing the Science and Scholarship of Health Equity.” *JAMA Health Forum* 7 (1): e256576. https://doi.org/10.1001/jamahealthforum.2025.6576.

Cleveland Clinic. 2022. “Hypertriglyceridemia: Causes, Risk Factors & Treatment.” Snapshots. Accessed April 17, 2026. https://my.clevelandclinic.org/health/diseases/23942-hypertriglyceridemia.

Coalition for Health AI. 2024. *The CHAI Applied Model Card*. Version 0.1. November. https://mc.chai.org/v0.1/documentation.pdf.

Collins, G. S., K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J. B. Reitsma, M. van Smeden, A. L. Boulesteix, J. C. Camaradou, L. A. Celi, S. Denaxas, A. K. Denniston, B. Glocker, R. M. Golub, H. Harvey, G. Heinze, M. M. Hoffman, et al. 2024. “TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods.” BMJ (Clinical research ed.), 385, e078378. https://doi.org/10.1136/bmj-2023-078378

Cowan, Don. 2026. “Masking.” The Science of Machine Learning & AI. https://www.ml-science.com/masking.

Davis, Andy, Asif Dhar, Neal Batra, Jay Bhatt, Wendy Gerhardt, and Brian Rush. 2022. “Economic Cost of Health Disparities: US Healthcare Can’t Afford Health Inequities.” Deloitte Insights, June 22. https://www.deloitte.com/us/en/insights/industry/health-care/economic-cost-of-health-disparities.html.

Copyright © 2026 Society of Actuaries Research Institute

Debray, Thomas P. A., Tri-Long Nguyen, and Robert W. Platt. 2026. *Comparative Effectiveness and Personalized Medicine Research Using Real-World Data*. 1st ed. Boca Raton, FL: Chapman and Hall/CRC. https://doi.org/10.1201/9781003300809.

Economic Research Service. n.d. "Rural-Urban Commuting Area Codes." U.S. Department of Agriculture. Last modified December 9, 2025. Accessed April 24, 2026. https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes.

Ezi, Ozoani, Marissa Gerchick, and Mitchell Margaret. 2022. "Model Card Guidebook." Hugging Face. https://huggingface.co/docs/hub/en/model-card-guidebook.

Flores, Efrén J., Ruth C. Carlos, and Lucy B. Spalluto. 2023. "Health Equity Is Intersectional." *Journal of the American College of Radiology : JACR* 20 (7): 621–22. https://doi.org/10.1016/j.jacr.2023.05.001.

Gabrio, Andrea. 2025. *missingHE: Missing Outcome Data in Health Economic Evaluation*. Version 1.5.1. Released July 2. https://cran.r-project.org/web/packages/missingHE/index.html.

Gervasi, Stephanie S., Irene Y. Chen, Aaron Smith-McLallen, et al. 2022. "The Potential for Bias in Machine Learning and Opportunities for Health Insurers to Address It." *Health Affairs* 41 (2): 212–18. https://doi.org/10.1377/hlthaff.2021.01287.

Gidwani, R., V. Yank, S. M. Asch, L. Burgette, A. Kofner, A. Peltz, and Z. Wagner. 2025. High-deductible health plans and receipt of guideline-concordant care for adults with chronic illness. JAMA Network Open 8(4): e258045.

Greifer, Noah. 2025. *MatchIt*. Released May 29. https://cran.r-project.org/web/packages/MatchIt/vignettes/MatchIt.html.

Hainmueller, Jens. 2022. *Ebal: Entropy Reweighting to Create Balanced Samples*. Version 0.1-8. Released June 9. https://cran.r-project.org/web/packages/ebal/index.html.

HanCheng, Miao. 2025. *Pysmatch Propensity Score Matching (PSM)*. Jupyter Notebook. Released May 29. https://github.com/miaohancheng/pysmatch.

Harvey, Raymond A., Jennifer D. Hayden, Pravin S. Kamble, Jonathan R. Bouchard, and Joanna C. Huang. 2017. "A Comparison of Entropy Balance and Probability Weighting Methods to Generalize Observational Cohorts to a Population: A Simulation and Empirical Example." *Pharmacoepidemiology and Drug Safety* 26 (4): 368–77. https://doi.org/10.1002/pds.4121.

Hasanzadeh, Fereshteh, Colin B. Josephson, Gabriella Waters, Demilade Adedinsewo, Zahra Azizi, and James A. White. 2025. "Bias Recognition and Mitigation Strategies in Artificial Intelligence Healthcare Applications." *NPJ Digital Medicine* 8 (1): 154. https://doi.org/10.1038/s41746-025-01503-7.

Jeanselme, Vincent, Maria De-Arteaga, Zhe Zhang, Jessica Barrett, and Brian Tom. 2022. "Imputation Strategies Under Clinical Presence: Impact on Algorithmic Fairness." *Proceedings of Machine Learning Research* 193: 12–34.

Jui, Tonni Das, and Pablo Rivas. 2024. "Fairness Issues, Current Approaches, and Challenges in Machine Learning Models." *International Journal of Machine Learning and Cybernetics* 15 (8): 3095–125. https://doi.org/10.1007/s13042-023-02083-2.

Copyright © 2026 Society of Actuaries Research Institute

Leinung, Matthew C., Andrew G. Gianoukakis, Daniel W. Lee, Stacey L. Jeronis, and James Desemone. 2000. “Comparison of Diabetes Care Provided by an Endocrinology Clinic and A Primary-Care Clinic.” *Endocrine Practice* 6 (5): 361–66. https://doi.org/10.4158/EP.6.5.361.

Lett, Elle, Shakiba Shahbandegan, Yuval Barak-Corren, Andrew M. Fine, and William G. La Cava. 2025. “Intersectional and Marginal Debiasing in Prediction Models for Emergency Admissions.” *JAMA Network Open* 8 (5): e2512947. https://doi.org/10.1001/jamanetworkopen.2025.12947.

Lucyk, Kelsey, Karen Tang, and Hude Quan. 2017. “Barriers to Data Quality Resulting from the Process of Coding Health Information to Administrative Data: A Qualitative Study.” *BMC Health Services Research* 17 (1): 766. https://doi.org/10.1186/s12913-017-2697-y.

Lumley, Thomas, Peter Gao, Ben Schneider, and Thomas Lumley. 2025. *Survey: Analysis of Complex Survey Samples*. Version 4.4-8. Released August 28. https://cran.r-project.org/web/packages/survey/index.html.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, et al. 2019. “Model Cards for Model Reporting.” In Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (pp. 220-229).

National Association of Insurance Commissioners (NAIC). n.d. “NAIC Survey Reveals Majority of Health Insurers Embrace AI.” NAIC. Accessed November 28, 2025. https://content.naic.org/article/naic-survey-reveals-majority-health-insurers-embrace-ai.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Washington, DC: U.S. Department of Health and Human Services. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html.

National Library of Medicine (NLM). 2019. “SNOMED CT.” NIH.Gov, September 18. https://www.nlm.nih.gov/healthit/snomedct/index.html.

New York City Center for Innovation through Data Intelligence. 2025. *YouthNPower Direct Cash Transfer Pilot: Year 1 Report*. New York, NY. https://www.nyc.gov/assets/cidi/downloads/pdfs/YouthNPower-Report-CIDI-2025.pdf.

New York State Department of Financial Services. n.d. “Insurance Circular Letter No. 7 (2024): Use of Artificial Intelligence Systems and External Consumer Data and Information Sources in Insurance Underwriting and Pricing.” Accessed November 28, 2025. https://www.dfs.ny.gov/industry-guidance/circular-letters/cl2024-07.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” *Science* 366 (6464): 447–53.

Poon Affat, Ronald L. 2025. “AI Risk Management Frameworks: An Expert Panel Discussion.” Society of Actuaries, April. https://www.soa.org/resources/research-reports/2025/ai-risk-management-frameworks/.

Copyright © 2026 Society of Actuaries Research Institute

Rajkomar, Alvin, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. 2018. "Ensuring Fairness in Machine Learning to Advance Health Equity." *Annals of Internal Medicine* 169 (12): 866–72. https://doi.org/10.7326/M18-1990.

Ross, Casey. 2024. "AI versus AI: The Emerging Arms Race over Health Insurance Denials." *STAT*, December 12. https://www.statnews.com/2024/12/12/artificial-intelligence-appealing-health-insurance-denials/.

Rountree, Lillian, Yi-Ting Lin, Chuyu Liu, et al. 2025. "Reporting of Fairness Metrics in Clinical Risk Prediction Models Used for Precision Health: Scoping Review." *Online Journal of Public Health Informatics* 17 (March): e66598. https://doi.org/10.2196/66598.

Rudolph, Jacqueline E., Yongqi Zhong, Priya Duggal, Shruti H. Mehta, and Bryan Lau. 2023. "Defining Representativeness of Study Samples in Medical and Population Health Research." *BMJ Medicine* 2 (1): e000399. https://doi.org/10.1136/bmjmed-2022-000399.

Sendak, Mark P., Michael Gao, Nathan Brajer, and Suresh Balu. 2020. "Presenting Machine Learning Model Information to Clinical End Users with Model Facts Labels." *NPJ Digital Medicine* 3 (1): 41. https://doi.org/10.1038/s41746-020-0253-3.

Smith, Timothy M. 2021. "4 Widespread Cognitive Biases and How Doctors Can Overcome Them." American Medical Association, February 4. https://www.ama-assn.org/about/ethics/4-widespread-cognitive-biases-and-how-doctors-can-overcome-them.

Society of Actuaries, 2025. "Evaluating Representation in Health Claims Data." https://www.soa.org/research/opportunities/2025/evaluating-representation-in-health-claims-data/.

Tran, Hung Thai, Willem M. van der Wal, and Ronald B. Geskus (maintainer 2011-2022). 2025. *Ipw: Estimate Inverse Probability Weights*. Version 1.2.2. Released December 8. https://cran.r-project.org/web/packages/ipw/index.html.

Walonoski, Jason, Mark Kramer, Joseph Nichols, et al. 2018. "Synthea: An Approach, Method, and Software Mechanism for Generating Synthetic Patients and the Synthetic Electronic Health Care Record." *Journal of the American Medical Informatics Association* 25 (3): 230–38. https://doi.org/10.1093/jamia/ocx079.

Yusha'u, Abubakar Sadiq, and Aminu Aliyu Abdullahi. 2026. "Six Approaches to Measuring Algorithmic Bias: An Empirical Evaluation of Fairness Metrics in Machine Learning." *International Journal of Scientific Research* 12 (1).

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 Bryn Mawr Ave, Suite 1000
Chicago, IL 60631
www.SOA.org

Copyright © 2026 Society of Actuaries Research Institute