

FAIRNESS ASSESSMENT IN CLINICAL PREDICTION MODELS

Subgroup performance, parity diagnostics and
threshold-dependent trade-offs

Technical report

Planimeter Kft.

Version 1.0.0

24 September 2026

Methodological demonstration only. Parity metrics are descriptive model-audit diagnostics, not proof of discrimination or clinical fairness. The models are not validated clinical decision tools.

Executive summary

Objective. To demonstrate why acceptable aggregate model performance is not sufficient for clinical model evaluation, and how subgroup metrics, parity diagnostics and threshold sensitivity can reveal materially different model behaviour.

Data. The released analysis contains 11,173 complete observations from the Framingham teaching dataset, split deterministically into 7,821 training and 3,352 test observations. Sex is used as the illustrative subgroup variable.

Key result. Overall test performance was reasonable for both models (AUC 0.725 for the random forest and 0.749 for logistic regression), but subgroup metrics differed substantially. In the test set, male and female event rates were 33.6% and 18.3%, respectively, and sensitivity, specificity, predictive values and AUC showed different subgroup patterns.

Interpretation. Fairness is not a single property of a model. Different parity definitions answer different questions, and changing the classification cutoff can materially alter the parity profile even when overall accuracy changes little.

1. Data and analytical framework

1.1 Source data

The analysis uses the Framingham cardiovascular teaching dataset distributed with the R package riskCommunicator. The objective is methodological: to illustrate subgroup model auditing using realistic clinical variables, not to develop or validate a deployable cardiovascular risk score.

1.2 Analysis population and split

The analysis retains the binary cardiovascular outcome and baseline predictors used in the Insight: sex, age group, age, systolic and diastolic blood pressure, BMI, total cholesterol, current smoking and diabetes. Rows with missing selected values are removed. A fixed seed (20260610) creates the 70/30 train-test split.

Characteristic	Value
Prepared analysis dataset	11,173 observations
Training set	7,821
Held-out test set	3,352
Test-set men	1,466
Test-set women	1,886

1.3 Models

Two models are fitted on identical training data: a conventional logistic regression and a probability random forest using ranger with 500 trees. Predictions are evaluated on the same held-out test set. The primary binary classification cutoff is 0.50.

1.4 Subgroup performance and fairness diagnostics

Traditional evaluation is performed separately by sex using AUC, accuracy, sensitivity, specificity, PPV and NPV. The fairmodels analysis then expresses several familiar error and classification quantities as Female/Male parity ratios, with Male specified as the illustrative privileged group solely to reproduce the analytical demonstration.

- Accuracy Equality: female accuracy divided by male accuracy.
- Equal Opportunity: female sensitivity (true-positive rate) divided by male sensitivity.
- Predictive Equality: female false-positive rate divided by male false-positive rate.
- Predictive Parity: female PPV divided by male PPV.
- Statistical Parity: female positive-classification rate divided by male positive-classification rate.

A ratio of 1 represents parity for that particular definition. Deviation from 1 is descriptive; it is not, by itself, evidence of discrimination or proof that a model is clinically inappropriate.

1.5 Aggregate Parity Deviation (APD)

To summarise the five selected parity ratios, this case study introduces Aggregate Parity Deviation (APD):

$$APD = \sum |\log(\text{parity ratio})|$$

Lower APD means that the selected ratios are collectively closer to 1. APD is not a fairmodels statistic and should not replace inspection of the component metrics. Different configurations can produce similar APD values while having different clinical consequences.

1.6 Cutoff sensitivity

For the random forest, the classification threshold is varied from 0.10 to 0.90 in steps of 0.05. At each cutoff, overall accuracy and the five parity ratios are recalculated, followed by APD when all ratios are defined. This demonstrates that fairness diagnostics are threshold-dependent rather than fixed properties of a fitted probability model.

1.7 Reproducibility

The public package contains the compact analysis dataset, fixed split, development-run predictions and metrics, two cleaned R scripts, a data dictionary, validation outputs and the technical report. All public script paths are relative.

2. Results

2.1 Population differences come first

The outcome is not equally distributed across the two test-set groups. Men have 493 cardiovascular events among 1,466 observations (33.6%), whereas women have 346 events among 1,886 observations (18.3%). This prevalence difference is important context because several performance and parity measures depend on the underlying outcome distribution.

Test subgroup	Observed event rate
Male	33.6% (493/1,466)
Female	18.3% (346/1,886)

2.2 Overall performance can conceal subgroup behaviour

Model	Overall test performance
Random forest	AUC 0.725; accuracy 75.0% at cutoff 0.50
Logistic regression	AUC 0.749; accuracy 76.4% at cutoff 0.50

The aggregate results alone would not reveal how differently errors are distributed across groups. Subgroup AUC is 0.656 in men versus 0.733 in women for the random forest, and 0.709 versus 0.743 for logistic regression.

2.3 Multi-metric subgroup profile

Random forest metric	Male Female
Accuracy	65.9% 82.0%
Sensitivity	20.1% 11.6%
Specificity	89.1% 97.9%
PPV	48.3% 54.8%
NPV	68.8% 83.1%

For the random forest, women have higher accuracy and specificity but lower sensitivity. This is exactly the kind of pattern that an overall accuracy number cannot describe.

2.4 Parity metrics answer different questions

Parity metric	Logistic regression Random forest
Accuracy Equality	1.189 1.245
Equal Opportunity	0.398 0.576
Predictive Equality	0.202 0.197
Predictive Parity	0.946 1.135
Statistical Parity	0.229 0.277

At cutoff 0.50, the five selected ratios produce APD = 4.222 for logistic regression and APD = 3.808 for the random forest. The individual ratios show why the single summary should not be interpreted alone: some dimensions are near parity while others are far from 1.

2.5 Fairness changes when the threshold changes

Random-forest cutoff	Accuracy APD
0.45	75.4% 3.53
0.50	75.0% 3.81
0.70	75.0% 1.95

The 0.50 and 0.70 cutoffs have almost identical overall accuracy (approximately 75%), yet their APD values differ markedly (3.81 versus 1.95). This does not establish 0.70 as a better clinical threshold. It shows that a model can retain similar aggregate accuracy while redistributing positive classifications and errors across subgroups.

At the highest evaluated cutoffs (0.75 and above), some parity components become undefined because one or more subgroup classification quantities collapse to zero. That is itself a useful governance signal: threshold selection can make apparently simple fairness summaries unstable or non-identifiable.

3. Interpretation and governance implications

This analysis supports three practical conclusions. First, aggregate performance should be treated as the start of model evaluation, not the end. Second, subgroup fairness diagnostics are inherently multi-dimensional: accuracy parity, opportunity parity, predictive parity and error-rate parity need not move together. Third, operational choices such as a classification threshold are part of the fairness assessment because they change the realised error profile.

The results should therefore be used diagnostically. A model audit can identify where subgroup behaviour differs, quantify those differences and show how they respond to design choices. Deciding whether a difference is clinically acceptable requires information outside the metric itself: intended use, harms of false positives and false negatives, treatment pathways, prevalence, calibration, stakeholder priorities and regulatory context.

4. Limitations

- The Framingham dataset is used as a teaching dataset; these models are not intended for clinical deployment.
- The analysis uses a binary sex variable from the source dataset and therefore does not represent the full complexity of sex or gender.
- A single deterministic train-test split is used for the demonstration; repeated validation or external validation would be required for a production model.
- Subgroup event prevalence differs substantially, so observed metric differences should not automatically be labelled model bias.
- Male is designated as the privileged group only as an analytical convention for this example, not as a normative claim.
- APD is a case-study summary invented for this project. It is not part of fairmodels and can conceal clinically important differences among its component ratios.
- Threshold analysis is descriptive. Selecting a clinical cutoff requires explicit utility, harm and decision-context considerations.

5. Reproducibility and validation

The released dataset contains 11,173 rows and 11,173 unique source-row identifiers, with no duplicate identifiers and no missing values in the released fields. The fixed split reproduces 7,821 training and 3,352 test observations.

During package preparation, the supplied development outputs were independently recalculated from the released CSV files. All 165 of 165 numerical checks passed within numerical tolerance, including subgroup metrics, parity ratios, APD and the full cutoff table. The packaging environment did not include an R runtime, so the cleaned R scripts were structurally reviewed but should receive one final local execution before publication.

6. AI use statement

Generative AI tools were used as a support tool during drafting, code development, review and validation. The analytical design, interpretation, final materials and responsibility for the released version remain with Planimeter Kft.

References

riskCommunicator: Framingham teaching dataset distributed with the R package riskCommunicator.

fairmodels: R package for fairness diagnostics built on the DALEX model-explanation framework.

DALEX: model-agnostic explainability and model-auditing framework used by fairmodels. Planimeter Kft. Fairness Assessment in Clinical Prediction Models. Biostat Insight Series.