# D-BIAS Analytical Report

*Dataset Bias Detection and Fairness Evaluation*

dataset
2025-11-09
Generated by D-BIAS System

Fairness Health Score

**42.00**

## Executive Summary

• Of course. Here is a clear and structured analysis of the biases detected in the `heart_disease_cleaned.csv` dataset, designed for both non-technical and technical audiences.

**Analysis of Detected Data Biases**

This report breaks down the potential biases in your heart disease dataset. Understanding these issues is crucial for building a fair, accurate, and reliable predictive model.

**1. Categorical Imbalance:** `sex`

• * **Severity: High** (This is a critical issue that will almost certainly lead to a biased model.)

• * **Meaning:** The dataset is heavily skewed towards one gender. **78.9% of the patients are male**, meaning there are nearly four male patients for every one female patient.

• * **Harm:** A model trained on this data will become an "expert" on heart disease in males but will lack sufficient data to learn the unique patterns, symptoms, and risk factors for females. It

may learn to associate male characteristics with heart disease simply because they are more common in the data.

• * **Impact:** A diagnostic tool built from this data could be significantly less accurate for women. It might **fail to detect heart disease in female patients** or underestimate their risk, leading to delayed diagnosis and poorer health outcomes. For example, since classic heart attack symptoms can differ between sexes, the model may not learn to recognize female-specific symptoms.

• * **Fix:**

• * **Data Collection:** If possible, gather more data on female patients to balance the dataset.

• * **Sampling Techniques:** Use methods like **oversampling** (duplicating records of female patients) or **undersampling** (removing records of male patients) to create a more balanced training set.

• * **Algorithmic Approach:** Assign a higher "class weight" to female instances during model training, forcing the model to pay more attention to the underrepresented group.

**2. Categorical Imbalance:** `fbs` **(Fasting Blood Sugar > 120 mg/dl)**

• * **Severity: High** (A critical imbalance that can cause the model to ignore a key risk factor.)

• * **Meaning:** The vast majority of patients (**85.0%**) in the dataset have normal fasting blood sugar levels (recorded as 'False'). The model has very few examples of patients with high blood sugar, a key indicator of diabetes.

• * **Harm:** The model might incorrectly conclude that high fasting blood sugar is not an important predictor of heart disease, simply because it sees so few cases. It may learn to ignore this feature altogether.

• * **Impact:** The model could **underestimate the heart disease risk for pre-diabetic or diabetic individuals**. This is dangerous, as diabetes is a major comorbidity for cardiovascular disease. A patient's high blood sugar might be overlooked as a critical warning sign.

• * **Fix:**

• * Similar to the `sex` feature, use **oversampling** to increase the number of patients with high blood sugar in the training data or apply **class weights** to make these instances more influential.

**3. Outlier Bias:** `chol` **(Cholesterol)**

• * **Severity: High** (This points to a critical data quality problem that can corrupt the model's logic.)

• * **Meaning: 20.0% of the** `chol` **values are outliers**. The data is "left-skewed," which suggests an unusual number of very low values. In many versions of this dataset, this is caused by cholesterol being recorded as `0`, which is biologically impossible and likely represents missing data.

• * **Harm:** These erroneous '0' values will drastically skew statistical calculations like the average and standard deviation. A model will learn a completely distorted relationship between cholesterol and heart disease.

• * **Impact:** The model might learn that a cholesterol level of 180 is "high risk" because the average is artificially pulled down by the zeros. This could lead to **false positives (unnecessary alarms for healthy patients)** or, worse, misclassifying a genuinely high-risk patient as normal because their cholesterol level doesn't seem high compared to the skewed data.

• * **Fix:**

• * **Data Cleaning:** This is a data quality issue first and foremost. Treat the `0` values as missing data.

• * **Imputation:** Replace the missing values with a reasonable estimate, such as the **median or mean cholesterol** calculated from the valid (non-zero) entries.

• * **Removal:** If the number of affected rows is small, they could be removed, but with 20% affected, imputation is a better strategy.

**4. Numeric Correlation Bias:** `ca` ↔ `num` **(Number of Blocked Vessels ↔ Heart Disease)**

• * **Severity: Moderate** (A strong, clinically relevant relationship that can make a model less robust.)

• * **Meaning:** There is a strong positive correlation (**r=0.574**) between `ca` (the number of major vessels blocked) and `num` (the heart disease diagnosis). In simple terms, patients with more blocked vessels are very likely to have heart disease.

• * **Harm:** While this correlation is expected and medically sound, it can cause a model to become a "one-trick pony." The model might rely almost exclusively on this one powerful feature, ignoring other subtle but important predictors like blood pressure, age, or lifestyle factors.

• * **Impact:** The model may become brittle. It might **fail to diagnose heart disease in its early stages**, where major vessels are not yet significantly blocked but other risk factors are present. If the `ca` measurement is unavailable or inaccurate for a new patient, the model's prediction could be completely wrong.

• * **Fix:**

• * **Regularization:** Use modeling techniques (like Lasso or Ridge regression) that penalize overly complex models and prevent any single feature from having a disproportionate influence.

• * **Feature Importance Analysis:** After training, check if the model is overly reliant on `ca` and ensure other features contribute to the final prediction.

**5. & 6. Other Moderate Numeric Correlations (`oldpeak` ↔ `num` and `age` ↔ `ca`)**

• * **Severity: Moderate** (These are notable relationships that require monitoring.)

• * **Meaning:**

• * `oldpeak` ↔ `num` (**r=0.446**): The degree of ST depression during exercise (`oldpeak`) is strongly linked to a heart disease diagnosis.

• * `age` ↔ `ca` (**r=0.417**): Older patients in this dataset tend to have more blocked vessels.

• * **Harm & Impact:** These correlations present similar risks to the `ca` ↔ `num` bias. A model could over-rely on `oldpeak` and become less sensitive to other factors. The `age` ↔ `ca` correlation (known as multicollinearity) can make it difficult to determine the independent impact of age versus blocked vessels, making the model harder to interpret.

• * **Fix:** The same fixes apply: use **regularization**, monitor feature importance, and for multicollinearity (`age` ↔ `ca`), consider if both features are needed or if one can be removed without losing significant information.

**Overall Assessment and Recommendations**

**Overall Reliability Assessment**

• The dataset contains clinically relevant and powerful predictors for heart disease. However, its reliability in its raw state is **low to moderate**. Critical data quality issues (cholesterol outliers) and severe representation biases (gender and fasting blood sugar) must be addressed before this data can be used to train a trustworthy model.

### Fairness & Ethical Implications

• The most significant ethical concern is the **severe underrepresentation of women (78.9% male)**. A model trained on this data poses a high risk of performing inequitably, potentially leading to poorer diagnostic accuracy and health outcomes for female patients. This could perpetuate existing gender disparities in cardiovascular medicine. Any model built from this data must be rigorously tested for gender-based performance differences.

### Concluding Summary: Dataset Fairness Health Score

• Qualitatively, this dataset has a **"Poor" fairness health score** in its current state. While it holds valuable medical information, the combination of severe gender imbalance and critical data errors means that using it without significant mitigation would likely result in an unfair, unreliable, and potentially harmful predictive tool.

### Actionable Recommendations

• To improve this dataset and prepare it for fair and reliable modeling, follow these steps:

1. **Clean the Data First:**

• * **Address** `chol` **outliers:** Investigate the `chol` values of `0`. Treat them as missing data and use a sound imputation method (e.g., replace with the median of non-zero values) to correct this feature.

2. **Mitigate Representation Bias:**

• * **Balance the** `sex` **and** `fbs` **features:** Use a combination of techniques like **SMOTE** (an advanced oversampling method), random oversampling, or assigning class weights during model training to ensure underrepresented groups are not ignored.

3. **Build a Robust Model:**

• * **Use Regularization:** Employ models with built-in regularization (e.g., Logistic Regression with L1/L2 penalty, XGBoost) to prevent over-reliance on dominant features like `ca` and `oldpeak`.

• * **Validate for Fairness:** After training, do not just look at overall accuracy. **Disaggregate performance metrics by gender**. Ensure the model's accuracy, precision, and recall are acceptable for both male and female subgroups. If not, revisit the mitigation steps.

• Top detected bias types: Categorical Imbalance (3), Numeric Correlation Bias (3), Outlier Bias (1)

• Reliability: {'fairness_score': 42, 'message': 'Moderate dataset size — reliability good but not fully stable. High bias risk — dataset requires correction before modeling.', 'n_columns': 16, 'n_rows': 919, 'reliability_level': 'moderate'}

## Mapped Bias Analysis

### Categorical Imbalance

Feature: **sex** ● High

'Male' dominates 78.9% of 'sex' values (entropy=0.74).

Feature: **fbs** ● High

'False' dominates 85.0% of 'fbs' values (entropy=0.61).

Feature: **exang** ● Moderate

'False' dominates 58.0% of 'exang' values (entropy=0.98).

### Numeric Correlation Bias

Feature: **age** ↔ **ca**

Moderate

Strong correlation r=0.417.

Feature: **oldpeak** ↔ **num**

Moderate

Strong correlation r=0.446.

Feature: **ca** ↔ **num**

Moderate

Strong correlation r=0.574.

### Outlier Bias

Feature: **chol**

High

20.0% of 'chol' values are outliers (left-skewed).

## Visualizations

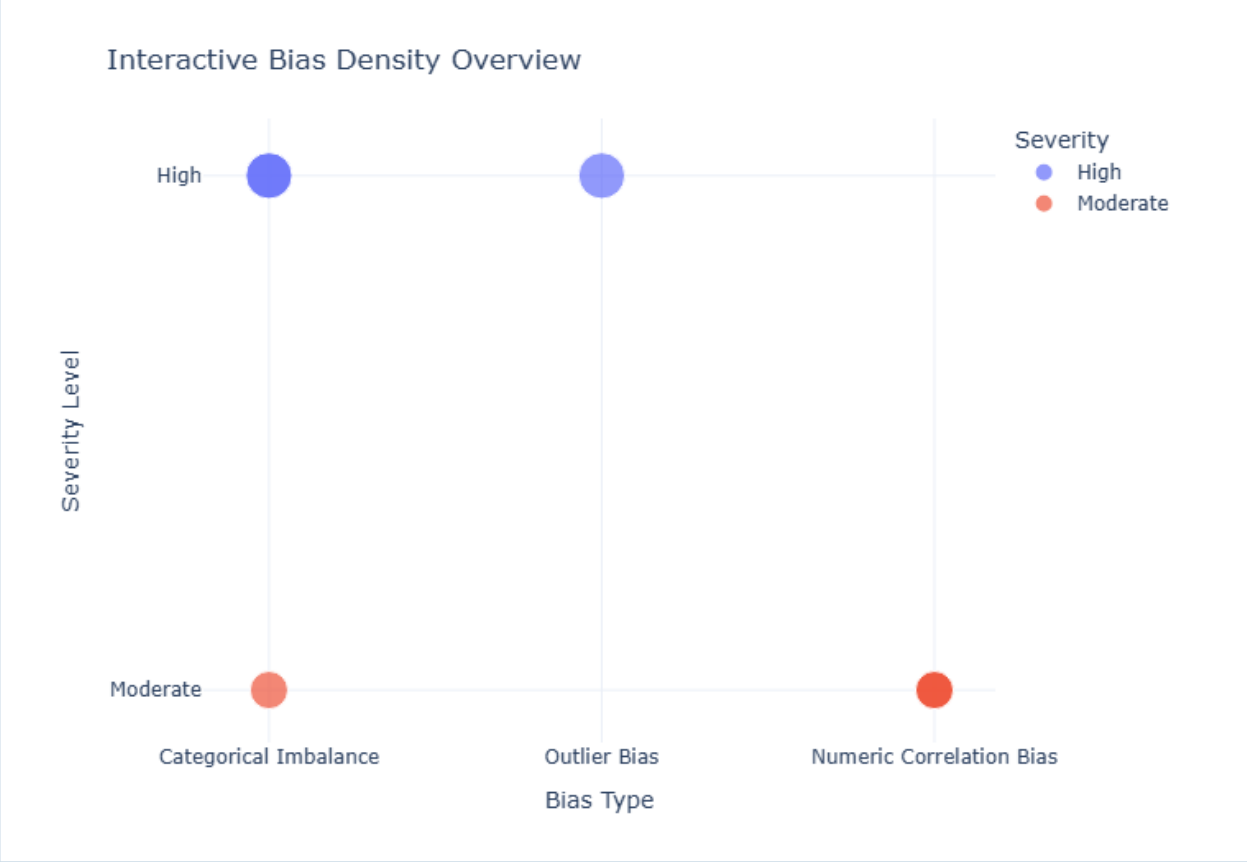


*Figure 1 — Bias Distribution by Type*

Density and distribution of detected bias types and their severity levels.

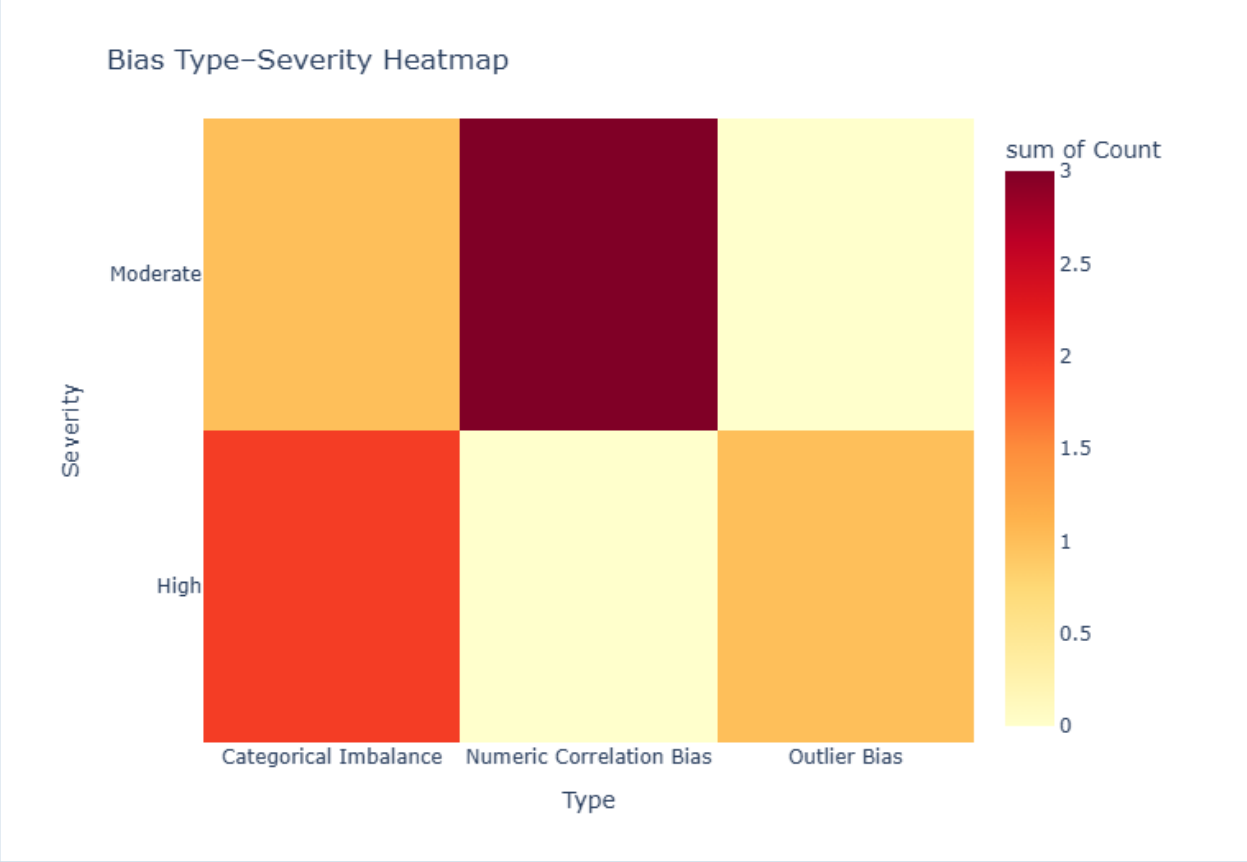


*Figure 2 — Fairness Trend Over Time*

Temporal trend or aggregated pattern indicating how fairness metrics evolve.

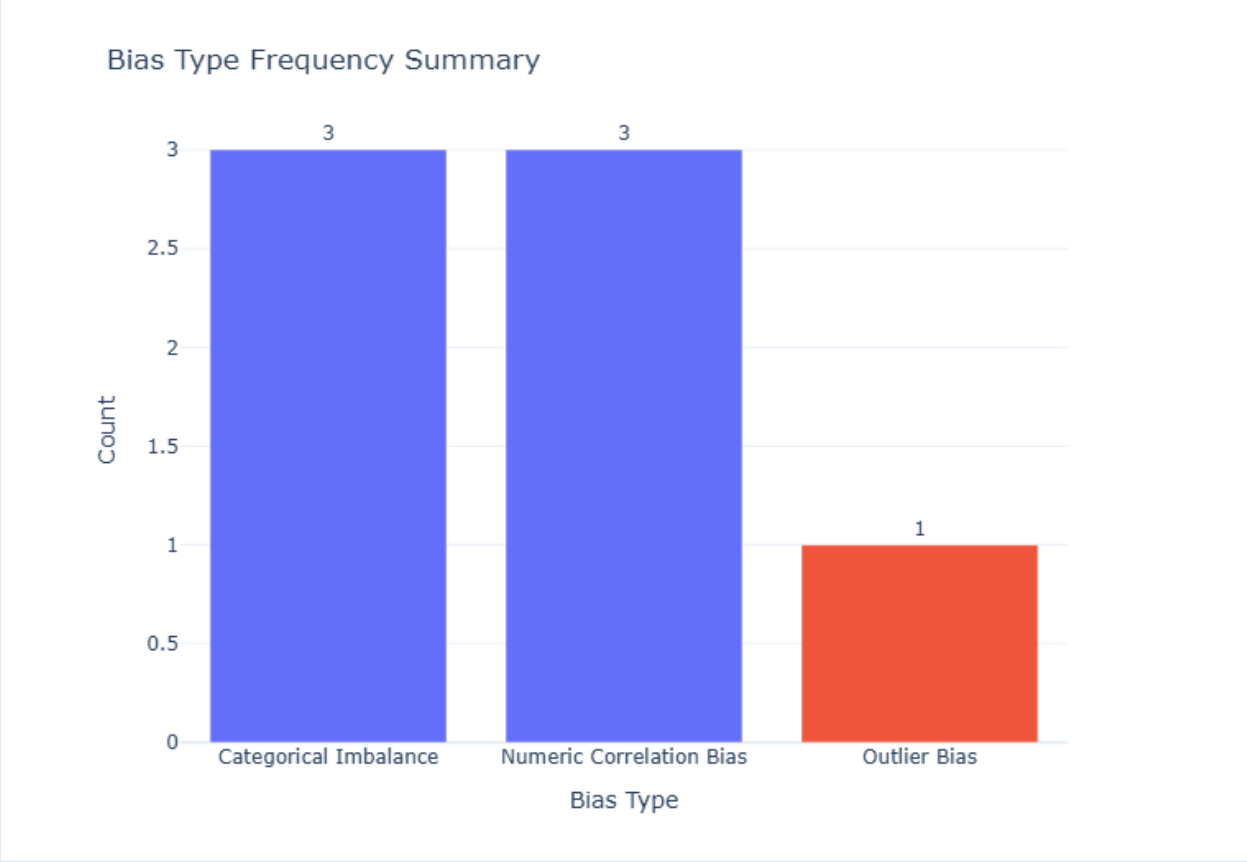


*Figure 3 — Bias Severity Breakdown*

Counts and proportions of bias severities across detected types.

## Recommendations & Actions

## Other Recommendations

1. To improve this dataset and prepare it for fair and reliable modeling, follow these steps:

2. **Clean the Data First:**

3. **Address** `chol` **outliers:** Investigate the `chol` values of `0`. Treat them as missing data and use a sound imputation method (e.g., replace with the median of non-zero values) to correct this feature.

4. **Mitigate Representation Bias:**

5. **Balance the** `sex` **and** `fbs` **features:** Use a combination of techniques like **SMOTE** (an advanced oversampling method), random oversampling, or assigning class weights during model training to ensure underrepresented groups are not ignored.

6. **Build a Robust Model:**