HIGH SCHOOL EDITION
Journal of Student Research

Volume 12 Issue 3 (2023)

Impact of Gender Bias in Training Data for Machine Learning Models predicting Myocardial Infarction

Victoria Harding Bradley

Menlo School

ABSTRACT

The use of biomarkers reference ranges derived from clinical trials to detect and diagnose the presence of cardiovascular disease (CVD) and the occurrence of myocardial infarction (MI) is well established. However, the predominance of older male participants in these trials has been shown to contribute to increased rate of misdiagnoses among women. The application of Machine Learning (ML) to medical diagnosis promises the potential to improve accuracy. However, ML also has the potential to perpetuate this problem of gender bias in trial data leading to worse outcomes for female patients. This research found that models trained using data containing only male patient data were less accurate at predicting MI among female patients than those trained using data sets with greater representation of females. For the model trained only with male patient data, this increased false negative rate corresponds to 2% of female patients with MI not being correctly diagnosed. Addressing this issue will require (a) the collection of more female patient data to support the construction of training data sets which accurately reflect the patient population (b) consistent reporting of gender mix and other demographic information such as age when ML model performance is reported.

Background

Medical background

A biomarker is a characteristic that can be measured and evaluated as an indicator of a normal biological process, or where there has been a change to that process, such as in the event of myocardial infarction (MI). There have been hundreds of biomarkers proposed for the detection and diagnosis of cardiovascular disease (CVD) and MI. Examples of biomarkers which are useful in detecting and diagnosing CVD include proteins released into the bloodstream such as Troponin (I and T), Cholesterol readings (hdl/ldl) and blood pressure measurement.

For each biomarker, there is a reference range that is considered normal. In the event of CVD, and more specifically MI the deviation of the biomarker from the normal range can lead to a diagnosis of a clinical problem. The most frequently used biomarkers to diagnose MI (Hachey et al, 2017) are: Cardiac Troponin, Creatinine Kinase (CK), CK-MB and Myoglobin.

Morrow et al. (Morrow 949-952) identified three considerations in the use of biomarkers: biomarker measurement and handling; the data that contributes to the biomarker and how the biomarker affects management of illness. The effectiveness depends on the use of the correct reference ranges for the specific biomarkers. These reference ranges are derived from repeated clinical trials and can differ by sex and by age category.

Historically, clinical trials have been dominated by male participants, typically in the middle to older age brackets. If the reference ranges of so many biomarkers are different for male and female patients, and for different age categories within the male and female groups, the use of ranges from potentially biased trials may result in an incorrect diagnosis or a missed diagnosis. Given the predominance of older male participants in the trials, there is a risk that MI and CVD might not be correctly identified in females, and in younger patients leading to worse medical outcomes.

Machine Learning background

Machine learning (ML) is a subset of Artificial Intelligence which uses data to build models to predict outcomes. Classification is a subclass within ML where the objective is to predict whether a data point is a member of a class or not. Two of the most common classification algorithms are Random Forest and Gradient Boosting. (Urbanowicz, Browne, 2017).

In Machine Learning, models are trained using a data set (the training data set) consisting of a set of features (candidate predictors of the outcome) and an outcome value (known as the Class). Model accuracy is assessed by comparing output values predicted by the model with the actual value in the test data set. It is assumed that the training and test data sets are representative of the same population from which they have been drawn at random. Deviation from this assumption in the training data (bias) will impact the accuracy of the model.

The impact of bias in training data sets has been extensively explored in connection with facial recognition models. Research has highlighted that lack of ethnic representation in facial training data leads to models with lower accuracy for members of the under-represented ethnicities (Zewe, 2022).

Literature Review

Medical literature explores the effectiveness of biomarkers for diagnosing CVD and MI as well as the gender bias potentially inherent in those biomarkers. Puntmann defines a biomarker as a “characteristic that can be objectively measured and evaluated as an indicator of normal biological processes” (Puntmann 2009). According to Dhingra (Dhingra et al, 2017) advances in biomarker research and developments related to CVD over the past 30 years have led to more sensitive screening methods, a greater emphasis on its early detection and diagnosis, and improved treatments resulting in more favorable clinical outcomes in the community.

Evidence of the difference in reference ranges between men and women was set out by Apple (Apple et al. 2003) where the analysis of seven CK-MB mass assays in 696 clinical trial subjects showed that the 99th centile upper reference range was up to three times higher in men than women. All assays showed both a 1.2- to 2.6 times higher 99th percentile upper range for male subjects when compared with female subjects, as well as mean concentrations significantly higher for the male subjects.

The high specificity of cardiac Troponin T and Troponin I has resulted in these biomarkers becoming increasingly integral to the diagnosis of myocardial infarction MI. Similar to CK-MB, there have been notable gender differences for Troponin where the 99th centile reference limits in males are up to two times the levels of those of females. Specifically, cardiac Troponin T was measured at the level of 20 ng/L for the male subjects compared with 13 ng/L for female subjects. Cardiac Troponin I was measured at 36 ng/L men in contrast to 15 ng/L for women (Shah 2017).

Therefore, the use of a uniform threshold for cardiac troponin does not provide equivalent prediction in men and women, with lower thresholds needed for women to provide comparable risk satisfaction. The most recent scientific statement from the American Heart Association highlights the underrepresentation of women in clinical trials. As Kim (Kim, 2022) communicates in his article, these sex differences manifest in the diagnosis of heart disease, as women are 21% less likely than men to be told they are at risk of CAD and are less likely to self-identify as being at risk. There has also been extensive research into the use of Machine Learning as an additional and potentially more accurate diagnostic tool for diagnosing a wide range of medical conditions including MI. In the area of cardiac medicine, research has been published on which algorithm leads to the most accurate MI diagnosis (for example Chakraborty,2021) as well as on the specific challenges of diagnosing women with MI (for example Mansoor, 2017).

Methods

The analysis was conducted using google colab, a cloud-based python execution product from Google Research and machine learning algorithms from the sklearn python library. The analysis focused on a publicly available data set containing 1,319 patient records each with gender, age, glucose levels, CK-MB and Troponin level and the presence of MI (Sozan, Maghdid et al, 2022). The data set used was constructed to "collect characteristics of Heart Attack or factors that contribute to it." and the following features are included for each patient: Age, gender, heart rate (impulse), systolic BP (pressurehigh), diastolic BP (pressurelow), blood sugar(glucose), CK-MB, Test-Troponin (troponin) as well as the MI diagnosis. For the purpose of this research, CK-MB and Test-Troponin were selected as the biomarker features for model training, because they are commonly used to detect MI and CVD.

The data set was split by gender and sequence of 6 training data sets created each with an increasing proportion of female subjects included (0% female, 20% female, 35% female, 60% female, 65% female and 80% female). Models were trained using the sequence of training data sets using each of two ML algorithms: Random Forest Classifier and Gradient Boosting Classifier generating a total of 12 models. The quality of each model, in each series, was assessed against two test data sets drawn at random from the full data set, one with 50% female/50% male and one with 100% females.

Data Preparation

Each of the biomarkers (Troponin and CK-MB) were transformed to ensure similar range and variance. Gender was numerically encoded (female=0; male = 1) and the presence of a MI diagnosis was numerically encoded (no MI present = 0; MI present = 1).

The different training and test data sets were created by taking a random sample of patients in required proportion by gender for the required mix. The order of the data in training and test data sets were randomized to avoid any order bias. For example, the following python code was used to achieve this step (in this case a training set of 400 males, and a test set of 400 females):

```
grouped_heart_data = heart_data.groupby(['gender'])
male_patients = grouped_heart_data.get_group(1)
female_patients = grouped_heart_data.get_group(0)
male_patients  = male_patients.sample(frac = 1)
female_patients  = female_patients.sample(frac = 1)
male100pc = male_patients[0:400]
female100pc = female_patients[0:400]
male100pc  = male100pc.sample(frac = 1)
female100pc  = female100pc.sample(frac = 1)
train_x = male100pc[male100pc.columns[6:8]]
train_y = male100pc ["Class"]
test_x = female100pc[female100pc.columns[6:8]]
test_y = female100pc ["Class"]
```

Training and evaluation of individual models

Using each training data set, a model was constructed using each of the two selected algorithms. Each model was then evaluated using the 100% female and 50% female: 50% male test data sets. The following python code was used to build the Random Forest Classifier and Gradient Boosting Classifier:

```
// Random Forest Classifier
model = RandomForestClassifier()
model.fit(train_x, train_y)
predicted_y = model.predict(test_x)
model.score(test_x, test_y)
// Gradient Boosting Classifier
gb_clf = GradientBoostingClassifier(n_estimators=20, learning_rate=0.5, max_features=2, max_depth=2, random_state=0)
gb_clf.fit(train_x, train_y)
gbpredicted_y100 = gb_clf.predict(test_x100)
gb_clf.score(test_x100, test_y100)
```

The evaluation stage consisted of calculating the accuracy and a confusion matrix. The following python code was used to achieve this step for each model:

```
cm = confusion_matrix(test_y100, predicted_y100)
disp = ConfusionMatrixDisplay (confusion_matrix=cm,
                              display_labels = None)
disp.plot()
```

Results

Data analysis

The data contains 449 female records and 870 male records and is similar to the data used for setting biomarkers benchmarks, with more males overall and female patients tending to be older with fewer diagnoses of MI than the male patients in the data set.

Table 1: Gender distribution among data set overall and by MI diagnosis

Gender	% of population	% of patients with MI diagnosis	% of patients with no MI diagnosis
Male	66%	70%	60%
Female	34%	30%	40%

In the data, males were overrepresented compared to the general population (66% of the patients were male) and more likely to have MI diagnosis compared to women (70% of MI diagnosed patients were male).

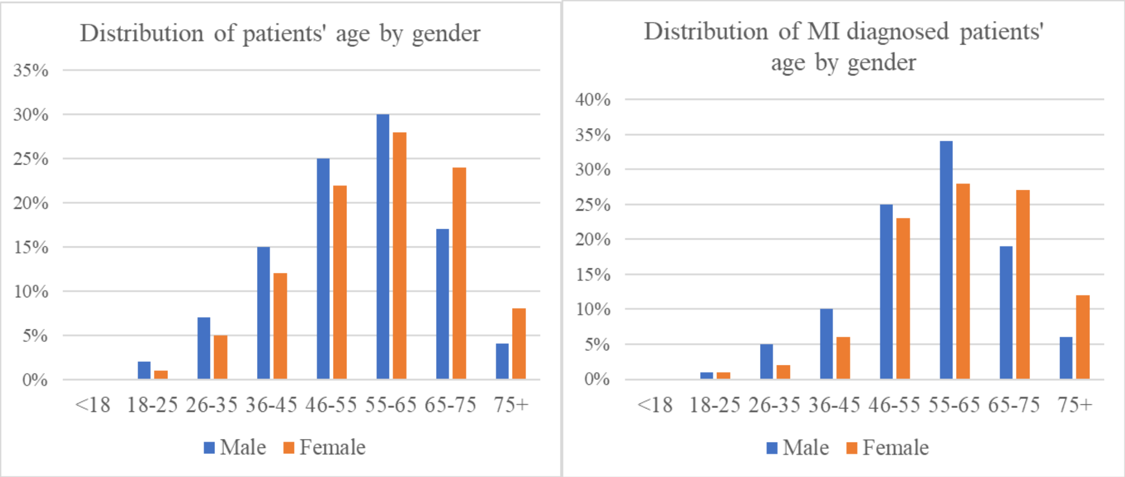

Figure 1: Bar graph of distribution of patient age categories (overall and among those with MI diagnosis)

Overall, patients included in the data set tended to be middle aged or older. Female patients in the data set tended to be even older with a higher proportion over 65 (24%) compared to among the male patients (17%). The gender bias was greater among those with MI diagnosis with 25% of male MI diagnoses 65+ compared to 39% of female MI diagnoses. This confirms that this data set is similar to those used to set reference ranges as reported in the research.

Table 2: Average biomarker score by gender among patients with MI and without MI diagnosis

	No MI diagnosed		**MI Diagnosed**	
	CK-MB	**Troponin**	**CK-MB**	**Troponin**
Male	2.65756	0.040482	23.05517	0.619984
Female	2.399995	0.00648	23.74931	0.458684

Table 2 shows the average biomarker score for CK-MB and Troponin among patients with or without a MI diagnosis for the male and female patients. The average Troponin scores among both those female patients with no MI diagnosis and those with MI diagnosis is much lower than among the Male patients with the same diagnosis status. The CK-MB scores in this data set do not exhibit a similar degree of gender difference.

Table 3: Model results

	Random Forest Classifier						Gradient Boosting Classifier					
(%male)	Test data 50%			Test data 0%			Test data 50%			Test data 0%		
Training set:	Ac	#fp	#fn	Ac	#fp	#fn	Ac	#fp	#fn	Ac	#fp	#fn
100%	0.98	0	4	0.96	1	9	0.98	1	4	0.97	1	10
80%	0.99	1	1	0.97	2	7	0.99	1	1	0.98	2	6
65%	0.99	0	4	0.98	0	5	0.98	0	8	0.97	1	8
50%	1.0	0	0	1.0	0	0	1.0	0	0	1.0	0	0
35%	0.99	0	1	0.99	1	2	0.99	0	1	0.99	0	1
20%	0.98	1	6	0.99	1	1	0.97	5	7	0.98	1	3

Ac: accuracy; #fp: number of false positives; #fn: number of false negatives

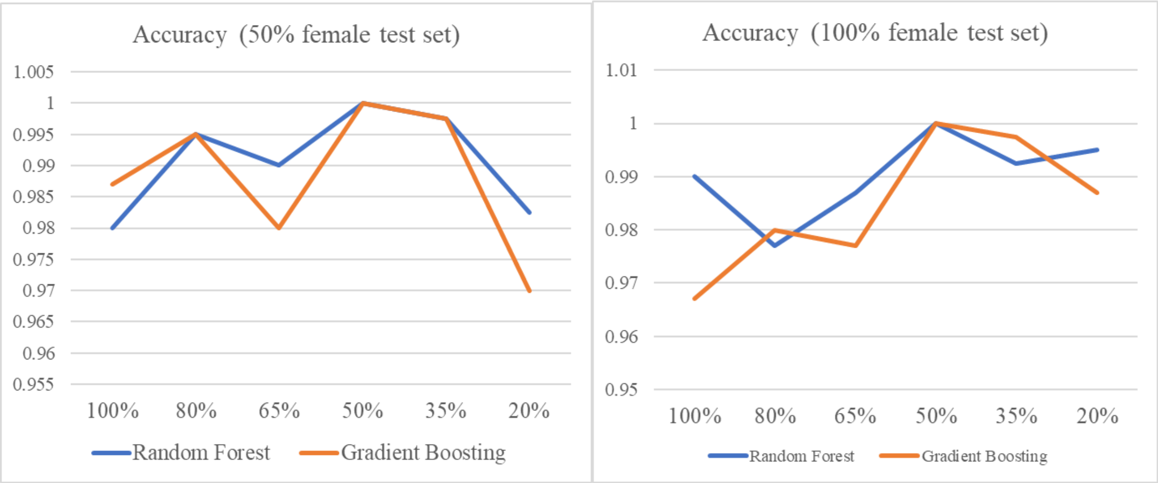

Figure 2: Comparing accuracy of Random Forest and Gradient Boosting Classifier models trained using different proportions of male patients when predicting MI using a test set made up of 50% and 100% female

The performance of the series of Random Forest and Gradient Boosting models trained with the training data sets was compared for both the 100% and 50% female test data. Figure 2 shows that the accuracy of models built using Random Forest and Gradient Boosting is similar with no clear pattern of difference in accuracy between the two algorithms.

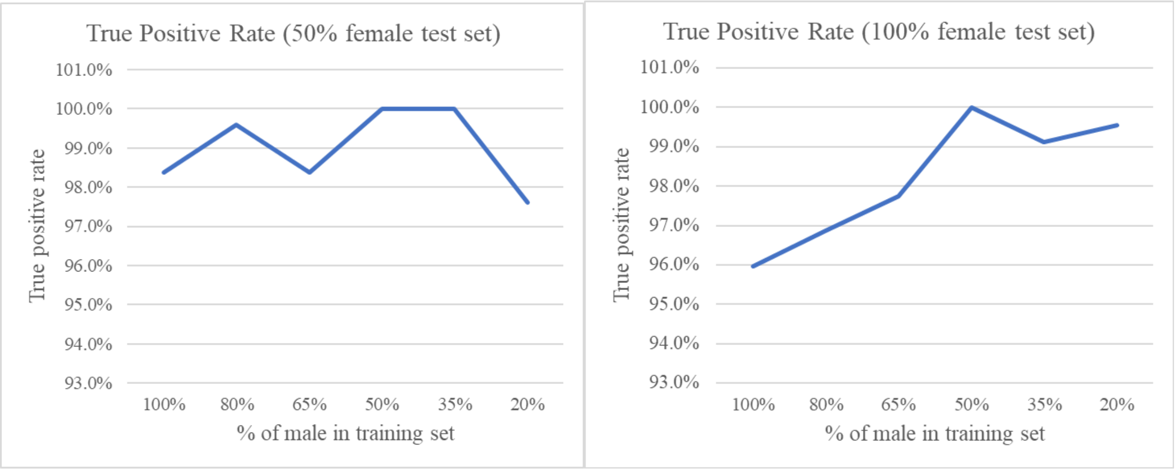

Figure 3: True positive rate of Random Forest models trained using different proportions of male patients when predicting MI using a test population made up of 50% and 100% female

The true positive rate (= true positive /(true positives + false negatives)) was calculated for the models trained using Random Forest. For the 100% female test data set, the graphs show that the model trained using 100% male patient data performed worst and the performance improves as the proportion of female patients in the training set increases with the best performance achieved for the 50%:50% female:male training data. Beyond this point, the performance is similar at greater proportions of female patients. For the 50%:50% female:male test data (reflecting the true population mix), the True Positive rate is higher until the training set reaches the 50% female :50% male proportions.

Discussion

It should be noted that the model tuning to increase the accuracy of the models was outside of the scope of this research and was not performed. While model tuning is a standard part of developing ML models, the objective of this research was to determine the performance of models not tuned for accuracy with female patients given that the lack of specific tuning for female patients reflects the usual research practices.

The Random Forest Classifier and the Gradient Boosting classifier were selected because they are widely used classification ML algorithms. Reviewing Table 3, it is apparent that the models generated from either method produce very similar levels of accuracy.

In evaluating ML model performance intended for medical diagnosis, the standard of accuracy required is very high because each false negative represents a missed MI diagnosis with the consequential impact on patient outcomes. Similarly, each false positive represents misdirected treatment with the potential patient impact of further unnecessary investigations and treatment as well as costs.

The 100%:0% male:female training set leads to the least accurate model when tested on the 100% female test data set. For the Random Forest Classifier, the accuracy was 0.97, with 1 false positive and 9 false negatives. For the Gradient Boosting Classifier, there was an accuracy of, again, 0.97 and the model diagnosed 1 false positive, and 10 false negatives out of the 400 patients in the test data set. These are a significantly lower accuracy level than for any other model and this relationship is also visible from the graph of true positive rates (Figure 3).

It is clear that the most accurate model for the 50:50 test data (close to the actual population mix), is that trained using the 50:50 training data. For the Random Forest Classifier model, the accuracy was 0.99 and the model did not diagnose a false positive, but it did diagnose four false negatives. For the Gradient Boosting Classifier, the accuracy was also 0.99 and the model diagnosed one false negative, and one false positive.

Conclusion

The research literature on gender bias in biomarker reference ranges notes that under-representation of female patients in trial data contributes to the under-diagnosis of MI among younger female patients. In the languages of ML, this corresponds to a higher rate of false negatives. This research has shown that under-representation of female patients in ML training data sets is likely to lead to similarly higher levels of false negatives among female patients if ML models are used for diagnostic purposes.

The impact of under-representation by gender or ethnicity in training data sets is generally acknowledged as leading to models which perform less well for those under-represented populations. When applied to diagnosis of potentially life-threatening or life-altering medical conditions (including MI), the impact of this is likely to lead more directly to real-world patient harm and is therefore of greater concern.

Addressing this issue will require the collection of more female patient data to enable the construction of training data sets which accurately reflect the patient population. To allow researchers and medical practitioners to interpret published research correctly, it is also recommended that research publications require consistent reporting of gender mix and other demographic information such as age when ML model performance is being reported.

References

"A Comparative Study of Myocardial Infarction Detection from ECG Data Using Machine Learning", Aritra Chakraborty, Santanu Chatterjee, Koushik Majumder, Rabindra Nath Shaw & Ankush Ghosh, Lecture Notes in Networks and Systems book series (LNNS,volume 218), 2021

"Benchmarks for the assessment of novel cardiovascular biomarkers.", Morrow, D. A., and J. A. de Lemos. *Circulation*, vol. 115, no. 8, 2007, pp. 949-952

"Can machine-learning models overcome biased datasets?", Adam Zewe, February 2022, https://news.mit.edu/2022/machine-learning-biased-data-0221

"Introduction to Learning Classifier Systems", Ryan J. Urbanowicz , Will N. Browne, Springer, 2017

"Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: A machine learning approach", Hend Mansoor, PharmD, Islam Y. Elgendy, Richard Segal, Anthony A. Bavry, Jiang Bian, Care of patients with cardiovascular disorders, vol 46, issue 6, p405-411, 2017

"Scikit-learn: Machine Learning in Python" https://scikit-learn.org/stable/

An Extensive Dataset for the Heart Disease Classification System - Sozan S. Maghdid, Tarik A. Rashid, Published: 17 February 2022 https://data.mendeley.com/datasets/65gxgy2nmg

Trends in Use of Biomarker Protocols for the Evaluation of Possible Myocardial Infarction - Brian J. Hachey, Michael C. Kontos, L. Kristin Newby, Robert H. Christenson, W. Frank Peacock, Katherine C. Brewer and James McCord Originally published 22 Sep 2017 https://doi.org/10.1161/JAHA.117.005852 Journal of the American Heart Association. 2017;6:e005852