Responsible AI

Question 1 - Problem statement

1. We have an AI tool that screens patients with Acute Coronary Syndrome to make decisions about their treatment. In particular, the algorithm decides if they should receive percutaneous coronary intervention (PCI) or the more invasive coronary artery bypass grafting (CABG). The AI algorithm is proprietary, and we do not know what risk factors are taken into account. We decide to investigate its performance in Caucasian vs. Asian patients by comparing the algorithm decisions with a ‘gold standard’ generated by consensus from experts based on relevant clinical information which does not include race. We find the following numbers:

For Caucasian patients:

True Positives = 100; True Negatives = 2600; False Positives = 200; False Negatives = 100;

For Asian patients: True Positives = 300; True Negatives = 1900; False Positives = 500; False Negatives = 300;

Based on this information answer if the algorithm is racially bias and why. Identify specifically, which measure(s) of discrimination performance is problematic and how this could negatively affect one of the racial groups.

|  |  |  |  |
| --- | --- | --- | --- |
| **Measure** | **Derivations** | Caucasian | Asian |
| Sensitivity or True Positive Rate (TPR) | TPR = TP / (TP + FN) | 0.5000 | 0.5000 |
| Specificity (SPC) or True Negative Rate (TNR) | SPC = TN / (FP + TN) | 0.9286 | 0.7917 |
| Precision or Positive Predictive Value (PPV) | PPV = TP / (TP + FP) | 0.3333 | 0.3750 |
| Negative Predictive Value (NPV) | NPV = TN / (TN + FN) | 0.9630 | 0.8636 |
| Fall-out or False Positive Rate (FPR) | FPR = FP / (FP + TN) | 0.0714 | 0.2083 |
| False Discovery Rate (FDR) | FDR = FP / (FP + TP) | 0.6667 | 0.6250 |
| Miss Rate or False Negative Rate (FNR) | FNR = FN / (FN + TP) | 0.5000 | 0.5000 |
| Accuracy (ACC) | ACC = (TP + TN) / (P + N) | 0.9000 | 0.7333 |
| F1 Score (F1) | F1 = 2TP / (2TP + FP + FN) | 0.4000 | 0.4286 |
| Balanced Accuracy (BA) | (TPR + TNR) / 2 | 0.7143 | 0.6458 |

Comparing the calculations:

Sensitivity (TPR): Both groups have the same sensitivity (0.5), indicating no bias in detecting positive cases.

Specificity (TNR): Specificity is higher for Caucasian patients (0.929) compared to Asian patients (0.792). This indicates the algorithm is better at identifying negative cases in Caucasian patients.

Positive Predictive Value (PPV): PPV is slightly higher for Asian patients (0.375) than for Caucasian patients (0.333), indicating when the algorithm predicts a positive case, it is more accurate for Asian patients.

Negative Predictive Value (NPV): NPV is higher for Caucasian patients (0.963) compared to Asian patients (0.864), meaning the algorithm is more reliable in predicting negative cases for Caucasian patients.

False Positive Rate (FPR): FPR is significantly lower for Caucasian patients (0.071) compared to Asian patients (0.208), meaning Asian patients are more likely to receive a false positive result.

False Negative Rate (FNR): Both groups have the same false negative rate (0.5), indicating no bias in missing positive cases.

Accuracy: Accuracy is higher for Caucasian patients (0.9) compared to Asian patients (0.733).

Balanced Accuracy (BA): BA is higher for Caucasian patients (0.715) compared to Asian patients (0.646).

F1 Score: F1 score is slightly higher for Asian patients (0.429) compared to Caucasian patients (0.4).

Conclusion

The algorithm shows racial bias primarily in specificity (TNR) and false positive rate (FPR). These discrepancies suggest that the algorithm is less accurate at identifying true negative cases and has a higher rate of false positives for Asian patients compared to Caucasian patients. This bias can negatively affect Asian patients by leading to more unnecessary invasive procedures, increased medical risks, and higher healthcare costs due to higher false positive rates. Addressing these discrepancies requires re-evaluating and potentially retraining the AI model to ensure equitable performance across different racial groups.

Question 2

2.1. Which of the following is an example of deployment bias?

a. An algorithm to estimate risk of secondary cardiovascular disease is applied to a population receiving a new statin therapy not accounted for at the time of model training.

b. An algorithm estimates risk of primary cardiovascular disease on the basis of grey vs not grey hair colour.

c. An algorithm embedded in a finger pulse oximeter is trained to measure oxygen saturation using people with white skin colour and it is used for people also with white skin colour.

d. An algorithm to detect cancer from breast screening mammograms is trained using radiologists opinions as outcome labels.

Answer: a

2.2. While testing the usability and safety of an AI system used to personalise medication dosing for babies you find that critical errors happen occasionally when a clinician inputs the weight for the patient in the wrong units. Which of the following solutions to this problem follow the principle of ‘safety by design’?

a. The user is not allowed to enter weight manually, instead, weight is uploaded from a scale via Bluetooth

b. The system alerts the clinician when it is likely that a wrong weight has been entered

c. The manual includes more precise information about how units are to be entered

d. The users are trained specifically on how to enter weight and units into the system

Answer: a,b

2.3. Why do we care about the uncertainty in the prediction of risk for an individual?

a. We want to avoid estimations of low risk with high uncertainty

b. We want to avoid estimations of 50% risk with high uncertainty

c. We want to know if the average risk in the training data is a good representation of the average risk in the target population

d. We want to know the relationship between the known probability distribution and the unknown outcome

Answer: c

Question 3 – Problem statement

You have been asked to compare the performance of three diagnostic algorithms (A, B and C) based on their predictions on a test set of 10 patients. The three algorithms provide the same diagnostic probability for each patient, but different prediction intervals (PI) as follows:

A table with numbers and symbols

Description automatically generated

3.a. Ignoring the small test sample size, is algorithm A better or worse than B? Explain your answer.

3.b. Ignoring the small test sample size, is algorithm A better or worse than C? Explain your answer.

Prediction intervals (PIs) provide a range within which the true value is expected to lie with a certain probability. A narrower PI indicates higher confidence in the prediction, while a wider PI suggests more uncertainty. The range of the PIs for each algorithm were calculated and show in table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Patient# | Disease | Prediction | PI Range (A) | PI Range (B) | PI Range (C) |
| 1 | 0 | 0.01 | 0.06 | 0.03 | 0.06 |
| 2 | 0 | 0.11 | 0.1 | 0.06 | 0.22 |
| 3 | 0 | 0.43 | 0.22 | 0.11 | 0.22 |
| 4 | 0 | 0.36 | 0.36 | 0.18 | 0.36 |
| 5 | 0 | 0.58 | 0.4 | 0.2 | 0.08 |
| 6 | 1 | 0.43 | 0.5 | 0.25 | 0.06 |
| 7 | 1 | 0.61 | 0.5 | 0.25 | 0.24 |
| 8 | 1 | 0.88 | 0.2 | 0.1 | 0.32 |
| 9 | 1 | 0.75 | 0.1 | 0.05 | 0.2 |
| 10 | 1 | 0.94 | 0.1 | 0.04 | 0.16 |
| Average |  |  | 0.254 | 0.127 | 0.192 |

3.a Comparing algorithm A and B across patients, algorithm B consistently provides narrower intervals, suggesting algorithm B has higher confidence in its predictions. Thus, algorithm B is better.

3.b. For Patients 1, 3, and 4, the PIs of Algorithms A and C are same. For patients 2, 8, 9, 10, algorithm A provides narrower PIs. For patients 4, 5, 6, algorithm C provides narrower PIs. The average PI range for Algorithm C (0.192) is narrower than the average PI range for Algorithm A (0.254), indicating algorithm C has generally higher confidence in its predictions. Thus, algorithm C is better.