Case Studies

Scenario 4: Fairness in AI-Assisted Screening for Diabetic Retinopathy

Background: An AI tool is being developed to assist ophthalmologists in screening for diabetic retinopathy from images of the retina acquired using a retinal scanner. Diabetic retinopathy is a condition that can lead to blindness if not detected and treated early. The condition is more prevalent in certain populations, such as Hispanic and African American communities. Current screening relies on experts reviewing the retinal images, but this process misses about 15% of positive cases, leading to preventable vision loss.

Development and Training: The AI tool will be trained using retinal images as input data and expert diagnoses as output labels. Training and evaluation will be based on data acquired from a network of 10 hospitals across the USA. The goal is to make the tool available globally to improve screening for diabetic retinopathy.

Questions to consider:

- What dangers/risks of the use of AI for this problem can you identify at this stage?
- How would you go about addressing these?
- What fairness metric(s) do you think might be appropriate when assessing the AI tool for potential bias?

REMINDER - DEFINITIONS OF FAIRNESS

- *False Negative Rate (FNR): the rate at which positive cases are missed by the classifier*
- *Equalised odds - equal true positive rate (TPR) & false positive rate (FPR) for each protected group*
- *Equal opportunity - only equalise either FPR or FNR, not both*

Suggested answers/discussion points:

- What dangers/risks of the use of AI for this problem can you identify at this stage?
    - **Demographic Information Recording:** If demographic information (such as race, sex, and age) is not recorded, it will be challenging to assess and ensure the fairness of the tool.
    - **Bias in Training Data:** It is known that there is melanin (the chemical that causes skin pigmentation) in the retina, so different races' retinas will have different pigmentation. The training data might predominantly consist of patients from certain ethnic groups (e.g., Caucasians), which could limit the tool's effectiveness in detecting the condition in other populations, such as Hispanic and African American communities where the prevalence is higher.
    - **Generalization to Different Retinal Imaging Devices:** Older or different models of retinal imaging devices might produce lower-quality images, potentially affecting the AI tool's accuracy. This could be a significant issue in parts of the world with older medical infrastructure.
- How would you go about addressing these?
    - **Broader Data Acquisition:** Acquire training data from a diverse range of sources, especially from regions with higher prevalence of diabetic retinopathy, such as Hispanic and African American communities.
    - **Recording Demographic Information:** Ensure that demographic information (race, sex, age) is recorded and utilized during both the training and evaluation phases. This data is crucial for assessing and improving the fairness of the tool.
    - **Thorough Testing Across Demographics:** Test the AI tool's performance rigorously on different demographic groups. This involves using held-out test sets from various populations and environments to ensure the tool is not biased towards any particular group.
- What fairness metric(s) do you think might be appropriate when assessing the AI tool for potential bias?
    - **False Negative Rate (FNR):** Given the severe consequences of missed diagnoses (false negatives), controlling the FNR is crucial.
    - **Equal Opportunity:** Ensure that the true positive rate is similar across different demographic groups. This helps in mitigating bias where one group might have a lower chance of a correct diagnosis.
    - **Calibration:** Check if the predictions are reliable and consistent across different groups. This means that for any predicted probability, the actual outcomes should match, regardless of the demographic group.

Conclusion: This scenario highlights the importance of fairness in AI-powered medical screening tools. While the AI significantly improves outcomes for high-prevalence groups, it underscores the need for inclusive and diverse training datasets to ensure equitable healthcare for all populations. Balancing sensitivity and specificity across different demographics is crucial for the global success of AI in medical applications.