Case Studies

<u>Scenario 5:</u> AI-Assisted Detection of Pediatric Bone Fractures

Background: An AI tool is being developed to assist radiologists in detecting bone fractures in pediatric patients using X-ray images. Pediatric bone fractures are relatively rare compared to the vast number of normal X-ray images taken for other reasons, such as routine check-ups or minor injuries that do not result in fractures. Additionally, certain racial groups, such as Asian children, may have smaller bone structures, which can affect the visibility and detection of fractures compared to Caucasian children.

Development and Training: The dataset for training the AI tool contains a large number of normal X-ray images and a relatively small number of fracture images. The proportion of racial minorities such as Asian children in the dataset is small.

<u>Questions to consider:</u>

- 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*
- *Demographic parity - equal chance of being classified positive for each protected group*
- *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?
  - **Data Imbalance:** The dataset is highly imbalanced, with far more normal X-ray images than images showing fractures. Additionally, it has insufficient representation of racial minorities, such as Asian children, whose smaller bone structures may be harder to detect fractures in.
  - **Overfitting to Non-Fracture Data:** The AI model might overfit to the abundant non-fracture data, resulting in high accuracy on normal cases but poor performance in detecting fractures, especially in underrepresented groups.
  - **Generalization Issues: The** tool may not generalize well to different racial groups if the fracture patterns in these groups are not adequately represented in the training data.
- How would you go about addressing these?
  - **Resampling Techniques:** Apply resampling techniques such as oversampling the minority class (fracture images) and underrepresented racial groups, or undersampling the majority class (normal images) to balance the dataset.
  - **Synthetic Data Generation:** Use techniques like SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic fracture images and augment the minority class, including for racial minorities.
  - **Data Augmentation:** Implement data augmentation strategies to artificially increase the number of fracture images and images from underrepresented racial groups by applying transformations such as rotations, flips, and brightness adjustments.
  - **Cost-Sensitive Learning:** Incorporate cost-sensitive learning approaches where misclassifying a fracture (false negative) is penalized more heavily than misclassifying a non-fracture (false positive).
  - **Anomaly Detection Algorithms:** Use anomaly detection algorithms that are particularly suited for identifying rare events, improving the model's ability to detect fractures across all racial groups.
  - **Regular Retraining and Validation:** Continuously retrain and validate the AI model with new data, including data from different hospitals and patient populations, to ensure it remains effective and generalizes well.
- What fairness metric(s) do you think might be appropriate when assessing the AI tool for potential bias?
  - **False Negative Rate (FNR):** Minimizing the FNR is crucial to ensure that the AI tool does not miss actual fracture cases, which can lead to serious health consequences.
  - **Precision-Recall Curve:** Evaluate the model's performance using the precision-recall curve, which is more informative than the ROC curve in the context of imbalanced datasets.
  - **F1 Score:** Use the F1 score, the harmonic mean of precision and recall, to balance the trade-off between false positives and false negatives.

- **Equal Opportunity:** Ensure that the true positive rate (recall) is similar across different racial groups to mitigate bias and ensure fairness in fracture detection.
- **Calibration:** Verify that the predicted probabilities are reliable and consistent across different racial groups, meaning that for any predicted probability, the actual outcomes should match across all groups.

Conclusion: This scenario underscores the challenges of data imbalance and the importance of addressing potential racial biases in medical imaging AI tools. By implementing strategies to balance the dataset and using appropriate fairness metrics, the AI tool can improve fracture detection in pediatric patients from diverse backgrounds, leading to better health outcomes and reduced disparities in medical care.