Case Studies

Scenario 3: AI Tool for Automated Diagnosis of Tuberculosis (TB) from Chest X-Rays

A team of researchers is developing an AI tool for automated diagnosis of tuberculosis (TB) from chest X-rays. The intention is to use the tool to automate interpretation of X-rays in parts of the world where there is a shortage of skilled radiologists to perform this task. The AI tool will be trained using two datasets – one of confirmed TB cases from a hospital in India, and one of healthy controls from a hospital in the USA. A random subset of the data from both sites will be held out for testing.

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*
- *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?
  - There are known differences between sexes in X-ray image quality due to the presence of breast tissue, and this has been shown to lead to bias in trained AI models.
  - Not specifically related to bias/fairness – but presumably the X-ray images between the USA and India sites would be acquired using different equipment. There is a danger of "shortcut learning", i.e. differences in image characteristics between sites might be used as proxy features for TB.
  - There may also be other differences in the patient cohorts between USA and India (e.g. size, body fat levels) which could lead to changes in image appearance and potential shortcut learning.
  - It is not mentioned whether demographic information will be recorded for bias assessment. It is important to assess performance for protected groups such as different races and sexes.
- How would you go about addressing these?
  - Try to acquire data from a broader range of sites, and mix TB/control cases within each site.
  - Make sure demographic information is recorded, e.g. race, sex, age.
  - Thoroughly test performance of the tool on different demographic groups.
  - Investigate whether the fairness metrics employed generalize to other domains, e.g. data from other held-out test hospitals
- What fairness metric(s) do you think might be appropriate when assessing the AI tool for potential bias?
  - TB is known to have a higher prevalence in men than women (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2860414/) so demographic parity might not be appropriate.
  - Perhaps Equalised Odds might be preferable?

Conclusion: This case study highlights the importance of addressing data imbalance and ensuring fairness in AI systems for medical diagnostics. By diversifying the training dataset, collaborating with global hospitals, applying bias mitigation techniques, and using appropriate fairness metrics such as demographic parity, equal opportunity, and calibration, the AI tool can be developed to provide accurate and equitable TB diagnosis from chest X-rays. This approach will help improve health outcomes, promote equitable healthcare, and reduce disparities in medical care, particularly in regions with limited access to skilled radiologists.