**Title**: Analyzing the equalized odds fairness of mortality classification for infection patients

**Background**:

Research has shown a strong association between sensitive attributes (including poor socioeconomic status and ethnic minorities) and significant health inequalities leading to individuals having an increased risk of infection and poor outcomes. The impact of which has been emphasised during the COVID-19 pandemic. As such when developing artificial intelligence (AI) solutions, it is important to ensure they are un-biased by using representative datasets and fairness metrics. Equalized odds (EO) can be considered the most relevant measure of fairness in this scenario given we want to acknowledge and ideally minimize false positives (i.e., predicting survival for patients who die) as well as obtain equal performance across the sensitive attribute classes.

**Methods**: A many-to-many long short-term memory recurrent neural network (LSTM-RNN) was developed that uses patient features from MIMIV-IV including lab test results and clinical parameters to predict the outcome of patients receiving antibiotics within the intensive care unit. The model was used to generate predictions for all individuals within an unseen test set. Results were then broken down by the sensitive attribute classes gender, socioeconomic status (i.e., insurance type), and ethnicity.

**Results**: To attain EO the true positive rate (TPR) and false positive rate (FPR) must be equal or at least similar across relevant sensitive attribute groups. EO was achieved across gender with a TPR of 0.82 and 0.79, and a FPR of 0.42 and 0.44, for males and females respectively. Reasonable results were also obtained for socioeconomic status (Medicaid TPR 0.82, FPR 0.34; Medicare TPR 0.76, FPR 0.43; Other TPR 0.85, FPR 0.45). However, performance across ethnic groups was not very consistent (White TPR 0.85, FPR 0.48; Black TPR 0.78, FPR 0.41; Hispanic TPR 0.78, FPR 0.45; Asian TPR 0.79, FPR 0.30; Native TPR 0.57, FPR Nan; Other TPR 0.71, FPR 0.30; Unknown TPR 0.63, FPR 0.28). With model outputs particularly differing for American natives who are infrequently present in the dataset.

**Conclusions**: The model demonstrated some EO fairness across genders, but ethnicity biases were present. Future research will investigate how to ensure datasets are representative and best mitigate these biases within AI models, to obtain consistent performance across the intended patient population.