A) Ethics & Bias (10 points)

How might biased training data affect patient outcomes in the case study?

- In the hospital readmission prediction model, biased training data can lead to serious and systemic consequences:

a) Over prediction for Marginalized Groups: Historical data may reflect systemic inequities, such as increased readmission rates for low-income or minority patients. These disparities are often the result of social and economic barriers rather than clinical need. If an AI model learns these patterns without correction, it may disproportionately flag these patients as high-risk, leading to unnecessary stress, medical overreach, or denial of appropriate services.

b) Under prediction for Underrepresented Groups: Certain populations—such as those in rural regions, indigenous communities, or undocumented individuals—may not be adequately represented in the training data. As a result, the model may overlook important risk factors or misclassify these patients, potentially leading to preventable adverse outcomes due to lack of timely intervention.

c) Erosion of Trust and Institutional Risk: If patients and clinicians observe that predictions are systematically biased; it may erode trust in the healthcare system's fairness. Clinicians may refuse to adopt AI recommendations, and hospitals may face reputational and legal risks. In healthcare, biased decisions are not just technical failures—they are ethical and legal liabilities.

Strategy to Mitigate Bias:

- We propose using fairness-aware learning via reweighing during the preprocessing phase. This technique adjusts the importance of data points during training by assigning higher weights to instances from underrepresented or disadvantaged groups. It ensures that the model learns to recognize patterns equitably without discarding any useful information.

  [Code snippet available in the accompanying notebook file]

- After reweighing, the model is retrained on the transformed dataset. Post-training, fairness can be quantitatively assessed using metrics such as:

a) Disparate Impact Ratio – compares the likelihood of favourable outcomes across groups
b) Equal Opportunity Difference – measures difference in true positive rates between groups
c) Demographic Parity – evaluates prediction balance across protected attributes
d) These metrics support transparent reporting and compliance with fairness standards.

B) Trade-offs (10 points)

Trade-off Between Interpretability and Accuracy in Healthcare:

In healthcare, the priority is not just predictive performance—it is actionable insight. Even if deep learning models achieve higher accuracy, their black-box nature makes them difficult to justify in clinical contexts. A highly accurate model that cannot explain its decisions risks being rejected by practitioners and regulators.

On the other hand, interpretable models like Logistic Regression, Decision Trees, or XGBoost with SHAP offer sufficient accuracy while maintaining explainability. These models enable:

a) Clinical validation of predictions
b) Documentation for compliance (HIPAA, GDPR)
c) Human-in-the-loop decision support

[Code snippet available in the accompanying notebook file]

For example, SHAP visualizations can show how patient features (e.g., length of stay, comorbidities, previous readmissions) contribute to the prediction. This aids clinical reasoning and supports transparency in care delivery.

Impact of Limited Computational Resources on Model Choice:

Resource constraints are common in hospital IT environments—particularly in rural or underfunded settings. In these cases:

a) Avoiding resource-intensive models like deep RNNs or large transformers is crucial
b) Choosing lighter models like LightGBM, Random Forest, or Logistic Regression helps ensure reliable operation on CPU-based systems
c) Optimization techniques like quantization, model pruning, or exporting to ONNX format enable real-time inference on edge devices or minimal hardware setups

Such trade-offs allow for responsible and sustainable deployment without sacrificing equity, accuracy, or ethical integrity. Models must be designed to fit both the technical environment and the lives they impact.