Part 3: Critical Thinking (20 points)

1. Ethics & Bias (10 points)

How biased training data might affect patient outcomes

Biased training data can lead to unequal or harmful decisions for certain patient groups. Some examples:

a) Underrepresented groups receive inaccurate risk scores.
   If minority patients, rural populations, or low-income patients appear less frequently in the dataset, the model may systematically underpredict their readmission risk, leading to fewer follow-up resources, poorer monitoring, and worse outcomes.

b) Historical healthcare inequalities become encoded into the model.
   If past medical records reflect unequal access to treatments, then the model may "learn" that these patients are lower priority, reinforcing discrimination.

c) False positives burden certain groups.
   Overpredicting readmission risk for some populations could cause unnecessary interventions, extra testing, or anxiety while wasting hospital resources.

d) Reduced trust in AI.
   If some groups consistently receive incorrect or unfair predictions, this damages trust in the system and may discourage patients from seeking care.

e) False negatives: High-risk patients not flagged for follow-up, increasing chances of readmission or complications.

One strategy to mitigate this bias

Stratified Sampling & Fairness Audits: Ensure balanced representation across age, gender, ethnicity, and socioeconomic status during data collection. Regularly audit model performance across subgroups to detect and correct disparities.

2. Trade-offs (10 points)

Trade-off between interpretability and accuracy in healthcare

In healthcare:

a) Interpretable models (Logistic Regression, Decision Trees)
   Pros: Easy for clinicians to understand, easier to justify to regulators, high transparency.

Cons: May be less accurate for complex medical patterns.

b) Highly accurate but less interpretable models (Deep Neural Networks)

Pros: Capture complex relationships and often produce higher predictive performance.

Cons: Harder to explain why a prediction was made, which can reduce clinician trust and complicate ethical/legal accountability.

Trade-off:

Healthcare often prioritizes interpretability due to patient safety and legal risks, but must balance it with the need for high accuracy to avoid missed diagnoses or incorrect predictions. Many hospitals choose a combination:

- use the more accurate model for predictions,
- provide explanations using tools like SHAP,
- And keep a simple baseline model for transparency.

Impact of limited computational resources on model choice

If the hospital has limited computational resources (older servers, restricted cloud use, no GPUs):

Lightweight models are preferred: Logistic Regression, Naive Bayes, Small Decision Trees

These models train quickly, require minimal memory, and are easy to deploy.

Heavy models may be impractical eg Deep learning models.

Such models need more RAM, CPU time, and may increase latency during prediction.

Result:

The hospital might choose a simpler, more efficient model to ensure: fast prediction times, lower maintenance cost, no strain on hospital IT infrastructure.