Part 3: Critical Thinking (20 points)

Ethics & Bias (10 points)

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

Biased training data can significantly undermine the reliability, fairness, and safety of predictive models in healthcare. In this case study, where the goal is to predict hospital readmission within 30 days, biased data may lead to the following consequences:

1. **Disparities in Care**: If the dataset underrepresents specific populations — such as low-income patients, ethnic minorities, or rural populations — the model may perform poorly for these groups. This can result in inaccurate risk scores and overlooked readmission risks, perpetuating existing disparities in healthcare access and outcomes.
2. **False Confidence or Alarm**: Patients who resemble the majority of the training data might be predicted more accurately, while predictions for minority or outlier patients could be erroneous — either underestimating true risk (leading to missed follow-up) or overestimating it (causing unnecessary interventions).
3. **Reinforcement of Systemic Inequities**: Historical healthcare data often reflect unequal treatment, insurance barriers, and bias in clinical decision-making. Using such data without critical evaluation could bake those inequalities into the model, leading to automation of past injustices.
4. **Erosion of Trust**: Patients and clinicians may lose trust in AI tools if outcomes appear unfair or opaque, especially in life-impacting domains like hospital discharge planning and readmission prevention.

Suggested Strategy to Mitigate Bias:

A robust bias mitigation plan should include **three layers**:

- **Data-level intervention**: Ensure balanced representation by collecting and curating a diverse dataset. Use *stratified sampling* across variables like ethnicity, gender, age, and socioeconomic status to prevent overfitting to the majority group.
- **Model-level intervention**: Employ *fairness-aware algorithms* that include regularization techniques to penalize disparities in model outcomes across sensitive groups. Introduce fairness constraints or use adversarial debiasing techniques to reduce group-specific prediction error.
- **Post-hoc analysis**: Regularly audit the model for *fairness metrics* (e.g., equal opportunity, disparate impact) and visualize performance differences across subgroups. Involve clinical experts and ethicists in reviewing flagged disparities.

Trade-offs (10 points)

Discuss the trade-off between model interpretability and accuracy in healthcare.

In healthcare, the **trade-off between interpretability and accuracy is critical** because decisions directly impact patient safety, clinician trust, and regulatory compliance. Here's how the trade-off plays out:

1. **Interpretability**:
    - Simple models like **logistic regression, decision trees, and rule-based systems** are transparent and easy to understand.
    - Clinicians can trace how each input (e.g., length of stay, lab results) influences the output (e.g., readmission risk).
    - Interpretability is essential for gaining trust, explaining decisions to patients, and meeting legal obligations (e.g., under HIPAA or GDPR).
2. **Accuracy**:
    - Complex models like **XGBoost, Random Forests, or deep neural networks** often achieve higher accuracy, especially with non-linear relationships and high-dimensional data.
    - However, these models function more like "black boxes," making it difficult to explain individual predictions.
3. **The Trade-off**:
    - **High interpretability → Lower complexity → Potential loss in accuracy**
    - **High accuracy → More complexity → Reduced transparency**

Healthcare demands a balance. For high-risk decisions (e.g., discharging a patient), transparency might outweigh small gains in accuracy.

How limited computational resources affect model choice:

Hospitals with constrained computing environments — especially in low-resource settings — must prioritize:

- **Efficiency**: Use models that train and infer quickly, like **logistic regression, shallow decision trees**, or **LightGBM with limited depth**.
- **Low hardware dependency**: Avoid GPU-intensive models (like deep learning) that require advanced infrastructure.
- **Model size and latency**: Deploy smaller models to ensure real-time predictions in integrated systems like EHRs.

In these settings, it's often better to trade off marginal accuracy for faster, interpretable models that are easier to deploy and maintain. Lightweight models can still perform well with good preprocessing, feature engineering, and cross-validation.