Critical Thinking (20 points)

Ethics & Bias (10 points)

Impact of Biased Training Data on Patient Outcomes

Biased training data in the readmission prediction system could have severe, life-altering consequences for patients:

1. Disparate Care Quality: If the model is trained primarily on data from affluent, insured patients, it may systematically underpredict readmission risk for marginalized groups (low-income, minority, uninsured patients). These patients would then not receive the enhanced follow-up care and resources they need, leading to worse health outcomes and potentially preventable complications.
2. Reinforcement of Historical Inequities: If historical data reflects existing healthcare disparities (where certain groups received poorer quality care leading to higher readmissions), the model could learn to associate these demographic characteristics with higher risk without addressing the root causes. This creates a vicious cycle where biased predictions lead to biased resource allocation, further exacerbating health disparities.
3. Misallocation of Critical Resources: High-risk patients from underrepresented groups might be misclassified as low-risk, denying them access to crucial post-discharge support like nursing follow-ups, medication management, or transportation assistance.
4. Loss of Trust in Healthcare System: Repeated algorithmic discrimination could erode trust in the healthcare institution among vulnerable populations, causing them to delay seeking care or disengage from preventive services.

Strategy to Mitigate Bias

Implement Fairness-Aware Model Development with Disparity Testing:

1. Stratified Performance Analysis: Regularly evaluate model performance across different demographic subgroups (race, gender, socioeconomic status, insurance type) rather than just overall metrics.
2. Bias Detection Metrics: Calculate and monitor fairness metrics such as:

    - Equalized Odds: Ensure similar false positive/negative rates across groups
    - Demographic Parity: Similar prediction rates across protected attributes
    - Predictive Rate Parity: Similar precision across groups
3. Preprocessing Intervention: Apply reweighting techniques to training data to balance representation of underrepresented groups, or use adversarial debiasing where the model simultaneously learns to predict readmission while becoming incapable of predicting protected attributes.

Trade-offs (10 points)

Model Interpretability vs. Accuracy in Healthcare

The tension between interpretability and accuracy represents a fundamental challenge in healthcare AI:

Interpretability-First Approach (e.g., Logistic Regression, Decision Trees):

- Advantages: Clinicians can understand the reasoning behind predictions, verify medical plausibility, and maintain clinical oversight. This builds trust and facilitates integration into clinical workflows.
- Disadvantages: May sacrifice predictive performance for complex, non-linear relationships in medical data, potentially missing important risk factors.

Accuracy-First Approach (e.g., Deep Learning, Ensemble Methods):

- Advantages: Can capture complex interactions between hundreds of clinical variables, potentially identifying novel risk patterns and achieving higher predictive accuracy.
- Disadvantages: Black-box nature makes it difficult to explain predictions to patients or clinicians, raising ethical concerns and potentially leading to blind reliance on algorithmic outputs.

Recommended Balance: In healthcare, where decisions have profound consequences, a hybrid approach is preferable. Use inherently interpretable models when possible, or employ sophisticated explainability techniques (SHAP, LIME) with complex models. The

optimal balance depends on the specific clinical context—higher-stakes decisions (like life-threatening conditions) warrant more interpretability, even at some accuracy cost.

Impact of Limited Computational Resources on Model Choice

Limited computational resources would significantly constrain model selection and deployment:

1. Shift Toward Simpler Models: Would favor traditional machine learning models (Logistic Regression, Random Forests) over deep learning architectures, as they require less computational power for both training and inference.
2. Feature Engineering Priority: Would emphasize careful feature selection and engineering to reduce dimensionality, rather than relying on models that automatically learn feature interactions from high-dimensional data.
3. Compressed Deployment Requirements: Would necessitate:
    - Model quantization to reduce memory footprint
    - Batch processing instead of real-time predictions
    - Simplified ensemble methods (fewer trees in Random Forests)
    - Possibly rule-based systems for initial screening
4. Trade-off Acceptance: Would require accepting slightly lower predictive performance in exchange for operational feasibility, focusing on "good enough" models that can run reliably on existing hospital infrastructure without requiring expensive hardware upgrades.
5. Cloud Consideration: Might explore hybrid approaches where intensive training occurs periodically in cloud environments, with lighter inference models deployed locally in the hospital system.