How biased training data might affect patient outcomes

Biased training data can lead to **systematic misclassification** of certain patient groups. Examples:

- **Underprediction for vulnerable groups** (e.g., low-income, elderly, non-English speakers): These patients may be incorrectly labeled as "low risk," reducing access to follow-up interventions and increasing their real readmission risks.
- **Overprediction for certain demographics** (e.g., patients with chronic conditions): This may lead to unnecessary interventions, resource strain, and reduced trust in the system.
- **Reinforcement of existing healthcare disparities:** If historical care patterns were unequal, the model may learn and perpetuate those inequities.

Overall, **biased outputs affect clinical decisions**, quality of care, and can worsen health equity.

2. Strategy to mitigate this bias

Fairness-aware evaluation and rebalancing:

- **Audit model performance** separately across demographic groups (age, sex, socioeconomic proxies, language).
- If disparities are found, apply a mitigation such as:
  Reweighting or resampling (e.g., upsampling underrepresented groups or applying fairness-aware loss weighting) so the model learns balanced patterns.

(Other valid strategies include adversarial debiasing, threshold adjustments per subgroup, and removing problematic features—but the one above satisfies the "1 strategy" requirement.)

Trade-offs (10 points)

1. Interpretability vs. accuracy in healthcare

- **High interpretability (e.g., logistic regression, decision trees):**
  - Pros: Easy for clinicians to understand and trust; supports regulatory requirements and informed decision-making.
  - Cons: May not capture complex relationships in clinical data, leading to lower predictive performance.
- **High accuracy models (e.g., gradient boosting, neural networks):**
  - Pros: Typically achieve better predictive performance.
  - Cons: Harder for clinicians to interpret; may reduce trust, hinder explainability, and complicate clinical validation.

Trade-off:
Healthcare often prioritizes interpretability *unless* the accuracy gain is large enough and accompanied by strong explainability tools (e.g., SHAP). Models must support safe, justified clinical decisions.

2. Impact of limited computational resources on model choice

- The hospital may need a model that is **lightweight, fast to train, and fast to run** on existing on-premise systems.
- **Constraints lead to choosing simpler models,** such as:
    - Logistic regression
    - Small decision trees
    - Random forests with limited depth
    - LightGBM with small tree sizes
- **Complex models** (deep neural networks, NLP-heavy architectures) might be impractical due to:
    - Long training time
    - High memory/GPU requirements
    - Slow inference, which is problematic for real-time clinical workflows

Therefore, limited computing resources push the choice toward **interpretable, efficient algorithms** that still deliver acceptable accuracy.