Ethics & Bias (10 points)

Impact of Biased Training Data

- Biased training data — such as underrepresentation of certain ethnic groups or missing social determinants — can lead to **systematic underestimation or overestimation of risk**.
- This might cause **high-risk patients to be overlooked**, or unnecessary interventions for low-risk patients, reducing care quality and potentially reinforcing existing health disparities.

Mitigation Strategy

- **Bias Auditing with Fairness Toolkits** (e.g., IBM AI Fairness 360):
    - Evaluate performance across subgroups (race, gender, income).
    - Implement **reweighting techniques** or **adversarial debiasing** to balance predictions.
    - Engage clinicians to interpret fairness metrics and co-develop ethical thresholds.

Trade-Offs (10 points)

Interpretability vs Accuracy

- **Highly accurate models** like deep neural nets may find complex patterns but are often opaque — making it hard for clinicians to trust or act on predictions.
- **Interpretable models** (e.g., decision trees, logistic regression) offer clear reasoning, which is critical in healthcare where accountability and clinical judgment matter.
- Trade-off: Prioritize interpretability when decisions impact health, regulation, and liability — even if it costs a bit in predictive power.

Model Choice with Limited Resources

- **Lightweight Models Preferred**:
    - Logistic regression, Naive Bayes, or shallow decision trees are computationally efficient.
    - They require fewer resources for training and real-time inference.
- **Alternative**:
    - Use pre-trained models or cloud-based APIs to offload computation while maintaining scalability and accuracy.