Part 3: Critical Thinking

---

1. Ethics & Bias

How might biased training data affect patient outcomes?
Biased training data can cause the model to systematically underperform for underrepresented groups such as ethnic minorities or low-income patients. For instance:

- The model may underpredict readmission risk for these groups due to lack of examples.
- Patients in need of care may be overlooked, resulting in worse health outcomes.
- These effects reinforce and widen existing healthcare disparities.

Unchecked bias in clinical AI can lead to substandard decisions and inequitable care delivery.

One strategy to mitigate this bias:
To address bias, the model development process should ensure data representativeness and apply fairness-aware techniques. This includes:

- Gathering a more diverse dataset or oversampling underrepresented groups.
- Using debiasing methods such as sample re-weighting or adding fairness constraints.
- Continuously monitoring performance across demographic subgroups using fairness metrics.
- Including clinicians in the review process to assess outputs for real-world validity.

---

2. Trade-offs

What is the trade-off between model interpretability and accuracy in healthcare?
Complex models like neural networks and ensemble methods can detect subtle patterns and offer high accuracy, but they are typically opaque ("black boxes"). Simpler models such as logistic regression or decision trees are easier to interpret and explain to clinicians but may be less accurate.

In healthcare, interpretability is often prioritized because:

- Clinicians must understand and justify model decisions.
- Explainable predictions are critical for building trust and accountability.

A slightly less accurate but transparent model may be preferred in clinical environments.

How would limited computational resources affect model choice?
In hospitals with limited processing capabilities (e.g., no GPUs or limited IT infrastructure), computationally efficient models are preferred. This may include:

- Using simpler models like logistic regression or shallow decision trees.
- Applying compression techniques like pruning or quantization for deployment.
- Prioritizing reliability and speed over marginal gains in predictive performance.

In resource-constrained settings, stability and efficiency often outweigh marginal accuracy gains.

Part 4: Reflection & Workflow Diagram

---

1. Reflection

What was the most challenging part of the workflow, and why?
The biggest challenge was managing competing priorities: achieving fairness, interpretability, and accuracy while respecting strict data privacy regulations. For instance:

- Optimizing model performance while preventing bias.
- Ensuring the model is clinically interpretable without sacrificing predictive power.
- Adhering to HIPAA while trying to leverage detailed patient data.

These trade-offs required careful decision-making at every stage of the project.

How would you improve your approach with more time or resources?
Given more time and resources, the approach could be strengthened by:

- Expanding the dataset to include more diverse patient demographics.
- Iterating on model design with clinical experts to refine features.
- Conducting real-world testing (pilot study) before deployment.
- Using advanced techniques such as explainable ensemble models or privacy-preserving methods like federated learning.

These steps would enhance both model performance and ethical robustness.

---

2. AI Development Workflow Diagram

Below is a simplified AI workflow to predict patient readmission risk. This flowchart illustrates each phase of the machine learning pipeline from problem identification to monitoring: