Part 3: Critical Thinking

---

1. Ethics & Bias

How biased training data might affect patient outcomes

Biased training data can seriously distort an AI model's predictions and, in healthcare, this directly impacts patient safety and fairness.
For example, if the hospital's historical data over-represents certain demographic groups (e.g., older adults or patients from urban areas) and under-represents others (e.g., rural or low-income patients), the model may learn patterns that are **not generalizable**.
Consequently, patients from under-represented groups might be **incorrectly assessed as low-risk** and not receive the necessary follow-up care—leading to higher readmission rates, delayed treatment, or even life-threatening situations.

In addition, if the model learns from data reflecting **systemic healthcare disparities** (such as differences in access to post-discharge care or medication adherence), it might perpetuate or even amplify these inequalities rather than correct them.
Ultimately, biased predictions reduce trust in the AI system and could expose the hospital to ethical, reputational, and legal risks.

Strategy to mitigate bias

A key mitigation strategy is **bias-aware data auditing and balanced sampling**:

1. **Conduct bias audits** before training — check for representation gaps across gender, age, ethnicity, and socioeconomic groups.
2. **Balance the dataset** through oversampling under-represented classes (e.g., using SMOTE) or re-weighting observations so all patient groups contribute equally to model learning.
3. **Evaluate fairness metrics** (e.g., equal opportunity difference, demographic parity) alongside traditional accuracy metrics.
   This ensures the model performs equitably across diverse patient populations and supports more ethical decision-making.

---

2. Trade-offs

Interpretability vs. Accuracy

In healthcare, there is an inherent **trade-off between model interpretability and accuracy**. Highly accurate models like deep neural networks or ensemble methods (e.g., Gradient Boosted Trees) often behave as "black boxes," making it difficult for clinicians to understand *why* certain patients are flagged as high risk.
In contrast, simpler models such as **logistic regression** or **decision trees** are easier to interpret and explain to clinicians, regulators, and patients—but may have lower predictive performance.

In a hospital setting, **interpretability is often prioritized** because clinical decisions must be explainable and accountable.
However, a practical compromise can be achieved by:

- Using **explainable AI tools** such as **SHAP** or **LIME** to interpret complex models, or
- Combining interpretable baseline models for deployment with more complex models for research and internal validation.

This balance ensures clinicians can trust and act on model outputs while maintaining strong performance.

Impact of limited computational resources on model choice

If the hospital has limited computational capacity (e.g., older servers or no GPU infrastructure), it may not be feasible to train or serve large deep learning models.
In such cases:

- The team might choose **lightweight, resource-efficient algorithms** like **logistic regression, random forest, or gradient boosting with limited depth**, which provide solid performance on structured EHR data.
- **Batch inference** rather than real-time scoring can reduce resource load.
- Alternatively, the hospital could adopt a **cloud-based deployment model** with proper security controls if on-premise hardware cannot support model operations.

In short, computational constraints encourage the selection of simpler, more efficient models, even if that slightly reduces predictive accuracy, in favor of reliability, speed, and cost-effectiveness.