Part 3: Critical Thinking (20 Points)

A. Ethics & Bias (10 Points)

Question:
How might biased training data affect patient outcomes in the case study?
Suggest 1 strategy to mitigate this bias.

Answer:
Biased training data in healthcare AI systems can perpetuate and even amplify existing inequalities. In the case of a model predicting patient readmission risk, if the data is skewed—perhaps containing fewer records from rural hospitals, lower-income communities, or marginalized ethnic groups—the model may generalize poorly to those populations. This leads to **systemic underestimation** or **overestimation** of readmission risks, causing patients to either receive insufficient care or be subjected to unnecessary interventions (Obermeyer et al., 2019).

One well-documented example of bias in healthcare AI was found in an algorithm used across the U.S. that underestimated the health needs of Black patients because it used historical healthcare spending as a proxy for need—ignoring structural barriers to access that kept spending artificially low in those populations. This led to millions of Black patients receiving poorer-quality care (Obermeyer et al., 2019).

To mitigate this bias, organizations must prioritize **inclusive data practices**. This includes not only stratified sampling but also **data augmentation**, where underrepresented groups are synthetically enhanced, and **bias detection tools** like IBM's AI Fairness 360 Toolkit and Google's What-If Tool are integrated during model development and validation. Furthermore, assembling **diverse interdisciplinary teams**—including ethicists, clinicians, and community reps—can guide ethical decisions during dataset creation and model evaluation (Bellamy et al., 2018).

Finally, ongoing **post-deployment monitoring** is vital. Bias can creep in after deployment due to real-world changes (e.g., new policies or treatments). Therefore, models should be audited regularly, with outcomes disaggregated by demographics to catch emerging disparities early.

References:
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). *Dissecting racial bias in an algorithm used to manage the health of populations*. Science, 366(6464), 447–453.
- Bellamy, R.K.E., et al. (2018). *AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias*. IBM Research.

B. Trade-offs (10 Points)

Question:
Discuss the trade-off between model interpretability and accuracy in healthcare.
If the hospital has limited computational resources, how might this impact model choice?

Answer:
In the healthcare sector, model interpretability is not just a technical preference—it's a **clinical and ethical necessity**. Doctors and hospital administrators need to understand and trust the decisions

made by AI systems. Complex models like deep neural networks may offer higher predictive accuracy by identifying subtle nonlinear patterns in Electronic Health Records (EHRs), but their **opaque logic** makes it difficult to explain their reasoning. This can lead to **reduced trust**, **hesitation to act on AI advice**, and challenges in clinical accountability (Doshi-Velez & Kim, 2017).

For example, a deep learning model may flag a patient as high risk for readmission, but if the clinician cannot explain *why*, they may hesitate to act on that prediction. This reduces trust in the model and may lead to it being underused, regardless of its accuracy. On the other hand, models like **logistic regression or decision trees** are inherently transparent. Clinicians can clearly see that, say, a high heart rate and low medication adherence led to a high readmission score—empowering them to take targeted actions.

This trade-off also extends to **regulatory compliance**. Frameworks like GDPR in Europe and HIPAA in the U.S. may require explainability for automated decisions that affect individuals. Interpretable models reduce the legal and ethical risks associated with black-box algorithms in healthcare (Doshi-Velez & Kim, 2017).

When computational resources are constrained—common in many hospitals in developing countries or rural areas—complex models requiring GPUs and high memory may not be feasible. In these cases, lightweight models not only save costs but also allow for **real-time deployment**, such as integrating directly into EHR systems or mobile apps used by health workers. **Model distillation**, where a complex model's knowledge is transferred to a simpler one, can be another practical solution in such contexts.

In conclusion, the most effective approach may involve a **hybrid strategy**: use complex models during initial research to gain insights and then develop interpretable surrogate models for deployment—striking a balance between accuracy, interpretability, and feasibility.

References:

- Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv preprint arXiv:1702.08608.
- Caruana, R., et al. (2015). *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission*. Proceedings of the 21st ACM SIGKDD.