# Paper: A tutorial on fairness in machine learning in healthcare

Link: <https://arxiv.org/pdf/2406.09307>

**What is the problem being solved?**

* Fairness in ML Models for Healthcare
* Fair = does not discriminate against an individual or group

**Why is it important?**

* Can perpetuate or exacerbate existing inequalities if not designed to account for fairness.
* Particularly in disease diagnosis and risk prediction, tend to underperform for minority groups, such as certain ethnicities or genders.

**Previous work:**

* Although fairness is well-researched in ML, its application in healthcare is still developing, with only a small percentage of studies evaluating fairness in ML models.
* Reviews of electronic health record (EHR) models often do not explore potential biases.

**Why ML Models Become Unfair**:

* **Bias in Data**: healthcare data most often measures and categorizes people -> societal bias, sampling method, time of collection, or data quality issues -> fail to represent the population of interest
* **Bias in Models**: Selection of inappropriate outcomes or evaluation metrics during training
* **Bias in Deployment**: Over-reliance on or ignoring model outputs

**Fairness criteria**:

* **Group Fairness**: predict similarly across different protected groups
  + Independence: *its predictions do not depend on the protected attribute*
  + Separation: *the prediction does not depend on the protected attribute within the positive and negative classes*
  + Sufficiency: *the label does not depend on the protected attribute given the prediction*
* **Individual Fairness**: similar predictions to similar individuals based on user-defined similarity metrics
* **Causal Fairness**:link observed disparities in model performance to their underlying cause (unfairness cause)
* Hard to satisfy all fairness criteria (i.e., group, individual, and causal fairness), as they may conflict with each other.