🎯 Bias and Fairness in Medical AI

1 Dataset Bias

💡 Key Insight

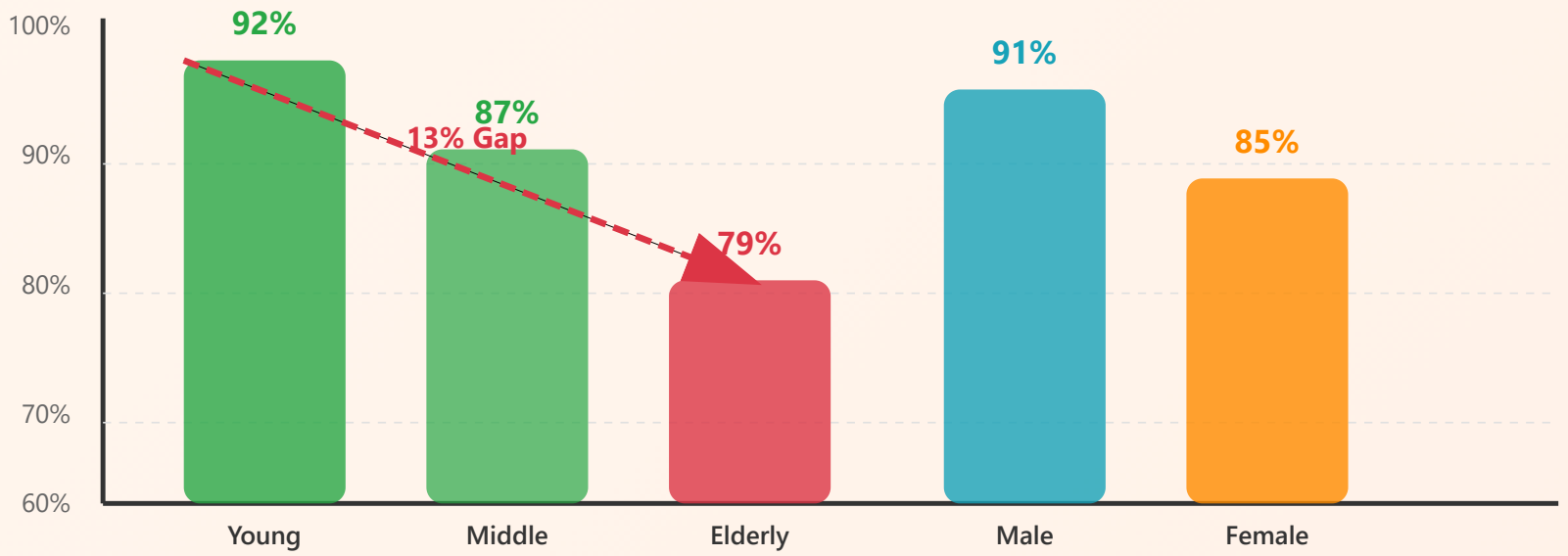
Biased training data leads to models that perform poorly on underrepresented groups, potentially causing life-threatening diagnostic errors in vulnerable populations.

67%

Accuracy drop on darker skin tones in some dermatology AI systems

2 Demographic Disparities

Performance by Demographic Group



Key Insight

Performance gaps across demographics can result from biological differences, healthcare access patterns, and data quality variations—not just algorithmic bias.

3x

Higher rate of undetected hypoxemia in Black patients vs White patients with pulse oximeters

3 Fairness Metrics

Multiple Definitions of Fairness

Choose based on clinical context

Demographic Parity

$$P(\hat{Y}=1|A=0) = P(\hat{Y}=1|A=1)$$

Equal positive prediction rates across groups

Equalized Odds

Equal TPR & FPR

True/False positive rates equal across groups

Equal Opportunity

Equal TPR

Equal sensitivity for disease detection

Calibration

$$P(Y=1|\hat{Y}=p, A)$$

Predicted probabilities match true outcomes

Individual Fairness

Similar → Similar

Similar individuals get similar predictions

Counterfactual

Causal Fairness

Prediction unchanged if sensitive attribute changes

⚠ Impossibility Theorem

You cannot simultaneously satisfy all fairness definitions! Choose metrics based on your clinical application and ethical priorities.

Screening Tasks

Risk Prediction

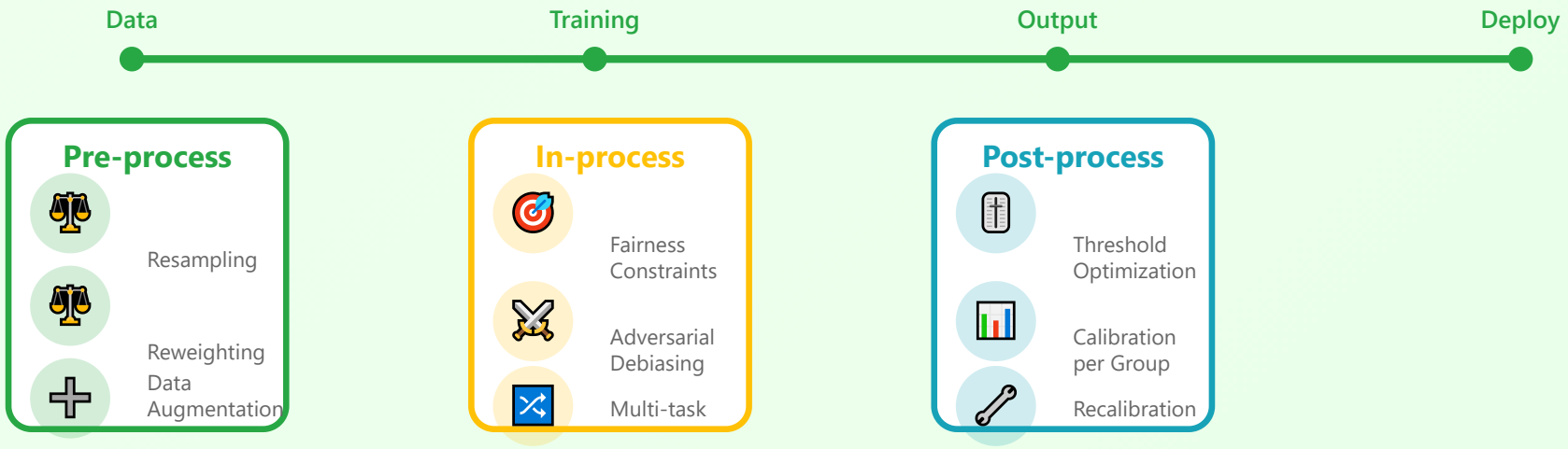
Use **Equal Opportunity**
Ensure all groups have equal chance of disease detection

Use **Calibration**
Predicted probabilities should be reliable across groups

Resource Allocation

Consider **Demographic Parity**
Resources distributed proportionally

4 Mitigation Strategies



Key Insight

Best results often come from combining multiple approaches across the ML pipeline. No single technique solves all bias problems.

12%

Performance improvement on underrepresented hospitals using adversarial debiasing

5 Continuous Monitoring

Key Insight

Model performance can degrade over time due to data drift, population changes, or clinical practice shifts. Continuous monitoring enables early detection and rapid response.

Alert Threshold

>5%

Performance drop in any subgroup triggers investigation

Disparity Threshold

>10%

Performance gap between groups requires mitigation

Review Frequency

Weekly

Regular audits to catch emerging issues early

🎯 Key Takeaways

1. No Perfect Solution

Multiple fairness definitions exist, and satisfying all simultaneously is mathematically impossible. Choose

2. Proactive Approach

Address bias throughout the entire ML lifecycle—from data collection to continuous monitoring.

wisely.

3. Measure Everything

Always report performance metrics disaggregated by demographic subgroups, not just overall accuracy.

4. Combine Strategies

Best results come from combining pre-processing, in-processing, and post-processing mitigation techniques.

5. Never Stop Monitoring

Continuous monitoring is essential—models can degrade over time even after careful development.

6. Stakeholder Engagement

Involve clinicians, patients, and ethicists in fairness decisions—it's not just a technical problem.

The Ultimate Goal

Build AI systems that extend high-quality, equitable healthcare to **ALL patients**, regardless of age, sex, race, ethnicity, or socioeconomic status.