

Algorithmic Bias and Equity

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 23 of "Healthcare Analytics and AI: Building Systems That Actually Work"

The essentials: how bias enters, how to measure it, and why race-blind is not fair.

One-sentence thesis. A model that faithfully learns an unjust world will faithfully reproduce it, so fairness is never a property you can assume, it is an output you must measure across groups before you deploy.

PART 1: Where Bias Enters

The Four Sources of Bias

Algorithmic bias almost never comes from malicious code; it comes from data, and it enters through four doors. **Historical bias** means the training data faithfully records real, ongoing discrimination, the way the eGFR race adjustment gave Black patients falsely better kidney scores. **Label bias** means the outcome variable is a biased proxy for what you actually care about, as when spending stands in for sickness. **Selection bias** means the training population is not representative, as with dermatology AI trained on Northern European cohorts. **Measurement bias** means the instrument itself fails for some groups: pulse oximeters miss hypoxemia on dark skin **about three times as often**. Almost all of it is introduced at the data-

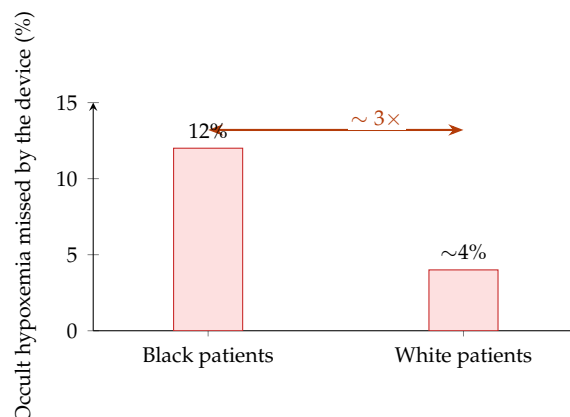
generation stage, long before anyone writes a line of model code.

ANALOGY

The model is a mirror with good math. If the room it reflects is distorted, so is the image, and polishing the glass does not fix the room.

The pulse oximeter is the visceral case: a hardware calibration built for light skin sent Black patients home with untreated low oxygen. Bias here decides who gets oxygen and who does not.

Watch out. You cannot audit any of this away by inspecting the algorithm. You have to interrogate where the data came from and what the label really measures.



Measurement bias you cannot audit out of the code. Pulse oximeters are calibrated on light skin, so roughly one in eight Black patients had dangerously low oxygen the device did not register, about three times the rate in White patients (Sjoding et al., 2020). No model touched this data; the error entered at the sensor.

Bias type	What goes wrong	Healthcare example	Detection cue
Historical	Data records past discrimination	Optum cost-as-need; eGFR race factor	Outcomes track a protected attribute
Label	Outcome is a biased proxy	Spending is not sickness	Proxy-to-construct gap
Selection	Training population unrepresentative	DeepDerm light-skin cohorts	Underrepresented subgroups
Measurement	Instrument fails for a group	Pulse oximetry on dark skin	Per-group error gaps

Demographic Parity

The first fairness question is the bluntest one: does the model flag each group at the same rate? Demographic parity demands equal selection rates across groups, so if the model flags 20 percent of white patients as high-risk, it should flag roughly 20 percent of every other group. This is the fairness of equal output volume, like insisting every neighborhood gets the same number of ambulances regardless of how many calls each one makes.

Formally, the selection rate must not depend on the protected attribute A :

$$P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b)$$

But when one group genuinely has higher prevalence, say higher readmission from decades of underinvestment in post-discharge support, forcing equal flag rates underpredicts risk for the sicker group. In healthcare, where the point is to find patients who need more care, parity can paradoxically harm the people it is meant to help.

Equalized Odds

A better question is whether the model's mistakes fall equally on everyone. Equalized odds asks for equal true-positive and equal false-positive rates across groups: different base rates are allowed, but the error burden must be shared. Different neighborhoods can have different call volumes, the reasoning goes, yet no neighborhood should wait longer for the ambulance to actually arrive.

Formally, the prediction must be independent of the group once you condition on the true outcome y :

$$P(\hat{Y} = 1 \mid Y = y, A = a) = P(\hat{Y} = 1 \mid Y = y, A = b) \quad \forall y$$

This is usually the more defensible criterion because it preserves real prevalence differences while ensuring

that false negatives (missed diagnoses, delayed treatment) do not concentrate on one group; in screening where a miss is the worst outcome (missed cancer, missed sepsis), equal sensitivity should take precedence. The catch is that equalized odds needs ground-truth labels that are themselves unbiased, which Part 1 shows is often not the case.

Calibration Within Groups and the Impossibility Result

The third question asks whether a risk score means the same thing for everyone: when the model says 70 percent, does 70 percent of that group actually have the outcome? Calibration within groups demands exactly that, a predicted 70 percent should mean 70 percent actual risk inside each group.

Formally, the observed rate must match the predicted probability within every group a :

$$P(Y = 1 \mid \hat{p} = p, A = a) = p \quad \text{for every group } a$$

The trouble is that you cannot have everything at once: the Chouldechova and Kleinberg impossibility results proved that demographic parity, equalized odds, and calibration cannot all hold simultaneously except in trivial cases (equal base rates or perfect prediction).

ANALOGY

A short blanket. Pull it up to cover one fairness definition and you expose another; you cannot cover all three at once.

Any fairness audit must therefore make an explicit, documented choice about which criterion to prioritize, and that choice is normative, not technical. It has clinical consequences, so it belongs to the institution, not to a library default. A vendor claiming a model is simply “fair” is either naive or hiding the tradeoff, so ask which metric they optimized and what they sacrificed.

Metric	Requires	Best when	Conflicts with
Demographic parity	Equal selection rate	Equal-access mandates	Calibration
Equalized odds	Equal TPR and FPR	Missed cases are costly	Demographic parity
Calibration by group	Score equals actual risk	Risk stratification	Equalized odds
Predictive parity	Equal PPV	Communicating risk	Equalized odds

Why Race-Blind Design Fails

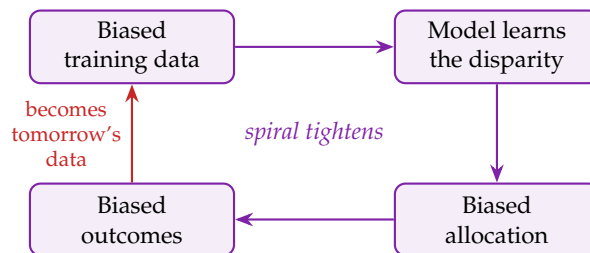
Deleting race from the features does not make a model race-blind; it makes the model hunt for race elsewhere. Zip code predicts race with high accuracy because of residential segregation, so remove it and the model reaches for pharmacy density, insurance type, or utilization history instead. It is like removing the label from a folder while leaving all the contents inside: the information is still there. In a world where race correlates with income, insurance, and utilization, there is no truly race-neutral feature set, and the only way to detect proxy discrimination is to test outputs across groups, not to inspect inputs. A model with zero protected attributes can still produce racially disparate care.

SDOH: The Double-Edged Sword

Social determinants of health drive an estimated **30 to 55 percent** of outcomes. Excluding them does not remove their effect; the model picks up housing instability through surrogates (frequent ED visits,

missed appointments) and penalizes patients for the consequences of their circumstances. Including them improves accuracy (fine-tuned LLMs find adverse SDOH in notes with 93.8 percent accuracy versus 2 percent from diagnostic codes) but risks redlining-by-algorithm. SDOH are a thermometer that reads the room's temperature: useful for turning on the heat, dangerous if you use it to decide who deserves a blanket. The fix is to separate the predictive question (who is at risk) from the allocation question (who gets more resources). The same "high-risk" flag can justify avoiding a patient (payer logic) or investing in them (population-health logic); the algorithm does not choose, the institution does.

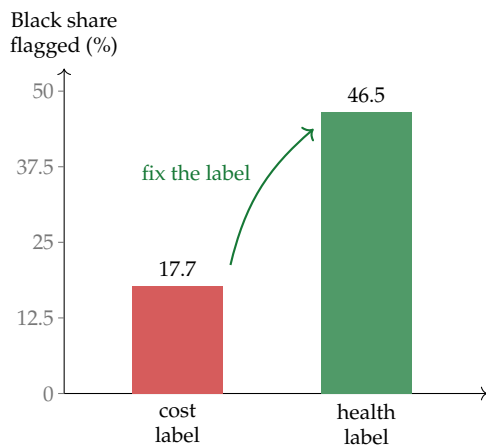
Watch out. The feedback loop is the deadliest dynamic. Biased predictions cause biased allocation, which produces biased outcomes, which become tomorrow's training data. Each turn tightens the spiral, exactly how the Optum algorithm ran for years undetected.



The bias feedback loop. Biased predictions drive biased allocation, which produces biased outcomes that re-enter training, exactly how the Optum algorithm ran undetected for years.

The Optum Case: Every Failure at Once

Obermeyer and colleagues (2019) dissected a commercial risk algorithm serving roughly 200 million people that used future cost as a proxy for health need. Because Black patients historically receive less care, they generated lower costs, so the algorithm scored them lower. At the same risk score, Black patients were **26.3 percent** sicker. Correcting the label (predicting health outcomes instead of cost) would have raised the Black share of patients flagged for extra care from 17.7 percent to 46.5 percent and reduced bias by **84 percent**. It was, in effect, grading students on how much their parents spent on tutoring, then concluding the under-resourced ones did not need help. No one intended to discriminate: engineers chose a reasonable-sounding label, built a well-performing model, and shipped it. The bias was invisible until an outside team audited outputs by race, which is why fairness auditing cannot be optional, and why it must examine outputs stratified by group, not just inputs.



The Optum correction. Predicting health outcomes instead of cost raised the Black share flagged for extra care from 17.7 to 46.5 percent, cutting bias by 84 percent, with no change to the algorithm.

Debiasing: Pre, In, and Post-Processing

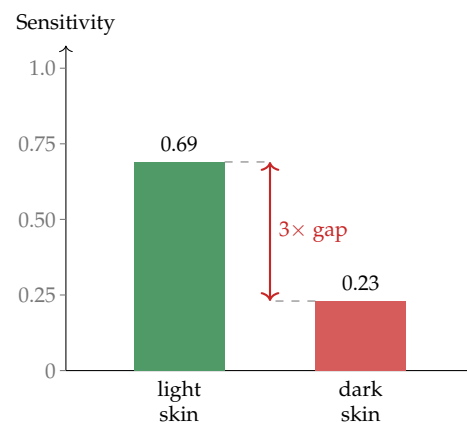
There are three points of intervention. **Pre-processing** fixes the data (reweight or resample so groups are balanced, or choose a better label). **In-processing** constrains training (Fairlearn's ExponentiatedGradient enforces a fairness constraint during fitting). **Post-processing** adjusts outputs (Fairlearn's ThresholdOptimizer sets per-group thresholds to equalize odds after the model is trained). It is the choice of fixing a leak at the reservoir, at the pump, or at the tap: each

works, each costs something different. The Optum fix was pre-processing (change the label), and it was the cheapest and most durable. Post-processing is fast but needs the protected attribute at inference time and can trade away accuracy. Tools exist (Fairlearn, Aequitas, SHAP); the unresolved variable is institutional willingness.

Watch out. Every mitigation moves the accuracy-fairness frontier. A drop from 0.81 to 0.78 AUROC that equalizes false negatives across races is a values decision, not a bug, and it should be made openly with legal exposure in view (ACA Section 1557, Colorado's AI Act with penalties up to 200,000 dollars per violation, Texas SB 1188).

Skin Tone and Gender: Bias Hiding in Plain Sight

Stanford's DeepDerm scored 0.69 sensitivity on light skin but only 0.23 on dark skin, a threefold gap traceable to light-skin-dominated training data. A review of 30 liver-disease algorithms missed 44 percent of cases in women versus 23 percent in men, and the best-performing models had the largest gap: a stethoscope tuned to one chest that goes quiet on every other. These gaps widen survival disparities, and none of the 30 liver studies even stratified results by sex. The bias was not discovered because no one looked. Sex-stratified and skin-tone-stratified evaluation must be mandatory, not optional.



Stanford's DeepDerm: melanoma sensitivity of 0.69 on light skin collapses to 0.23 on dark skin, a threefold gap inherited from a light-skin-dominated training set.

Should You Build This?

Build it when auditing outputs by demographic group is a mandatory gate in the deployment pipeline and the label measures what you actually care about; the Optum repair was a label change, not an algorithm change, and fairer models were never impossible, no one had checked. **Think twice when** the outcome variable is a proxy shaped by unequal access (spending standing in for need), or when the training

data underrepresents the populations the tool will serve. Those two conditions sit behind every case in this chapter, from dermatology classifiers to liver-disease screening. **Who bears the risk:** patients who never learn an algorithm scored them and simply do not get the call, and, increasingly, the deployer, because with the ACA Section 1557 rule and state statutes in force, “we did not know the algorithm was biased” is no longer a defense.

Quick Self-Test

1. Name the four sources of bias and give the detection cue for each.
2. Why can demographic parity harm the higher-prevalence group it is supposed to protect?
3. State the impossibility result in one sentence and explain why it forces an institutional choice.
4. Removing race from the features did not make a model fair. Why not, and how would you actually detect the remaining bias?
5. In the Optum case, what was the biased label, and why did fixing it (not the algorithm) reduce bias by 84 percent?
6. A mitigation drops AUROC from 0.81 to 0.78 but equalizes false-negative rates by race. Is that a bug? Who decides?

If You Remember One Thing

Fairness is measured at the output, not assumed at the input. Demographic parity, equalized odds, and calibration cannot all hold at once, so every deployment makes a normative choice about which one matters, and that choice belongs to the institution. The Optum algorithm was not malicious, illegal, or badly coded. It optimized a reasonable-sounding proxy and harmed millions, and no one caught it until someone audited the outputs by race.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552