

Fairness and Equity Metrics

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Concept guide, "Healthcare Analytics and AI: Building Systems That Actually Work"
The essentials: what to know, why it matters at the bedside, and how models fail.

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 is introduced at the data-generation stage, long before anyone writes model code. There are four ways in. **Historical bias** means the training data faithfully records real, ongoing discrimination, the way the eGFR race adjustment gave Black patients falsely better kidney scores and delayed transplant referral. **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, like dermatology AI trained almost entirely 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, a hardware calibration built for light skin that sent Black patients home with untreated low oxygen.

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.

Watch out. You cannot audit these away by inspecting the algorithm. You have to interrogate where the data came from and what the label really measures, because the bias here decides who gets oxygen and who does not.

Bias type	What goes wrong	Healthcare example	Detection cue
Historical	Data records past discrimination	Optum cost-as-needed; 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 simplest thing “fair” could mean is equal output volume: if the model flags 20 percent of white patients as high-risk, it should flag roughly 20 percent of every other group. That is demographic parity, equal selection rates across groups.

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

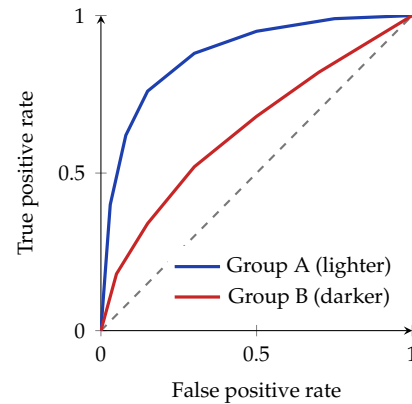
It is the fairness of the equal-ambulance rule: every neighborhood gets the same number of ambulances regardless of how many calls it makes. **Watch out.** When one group genuinely has higher prevalence (higher readmission from decades of underinvestment in post-discharge support), forcing equal flag rates underpredicts risk for the sicker group, so parity can paradoxically harm the people it is meant to protect.

Equalized Odds

A better question than “equal volume” is “equal error.” Equalized odds demands an equal true-positive rate and an equal false-positive rate across groups: different base rates are allowed, but the burden of mistakes must be shared. Neighborhoods can have different call volumes, but no neighborhood should wait longer for the ambulance to actually arrive.

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

This is usually the more defensible metric: it preserves real prevalence differences while keeping false negatives (missed diagnoses, delayed treatment) from concentrating on one group. Where a miss is the worst outcome (missed cancer, missed sepsis), equal sensitivity should take precedence. **Watch out.** Equalized odds needs ground-truth labels that are themselves unbiased, which Part 1 shows is often not the case.



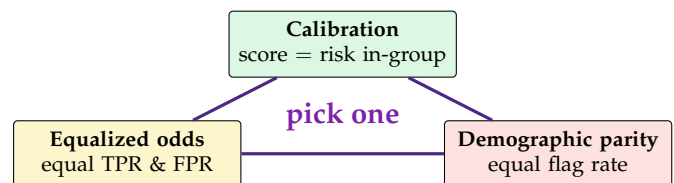
Discrimination is not shared equally. One model ranks well for group A and near chance for group B, an equalized-odds violation that a single pooled AUROC hides completely.

Calibration Within Groups and the Impossibility Result

A third notion of fairness asks that the number mean the same thing in every group: a predicted 70 percent should correspond to 70 percent actual risk whether the patient is in group a or group b .

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

Here is the trap. The Chouldechova and Kleinberg impossibility results proved that demographic parity, equalized odds, and calibration cannot all hold at once except in trivial cases (equal base rates or perfect prediction). It is a short blanket: pull it up to cover one definition and you expose another. Any fairness audit must therefore make an explicit, documented choice about which criterion to prioritize, and that choice is normative, not technical, so it belongs to the institution, not a library default. **Watch out.** A vendor claiming a model is simply “fair” is either naive or hiding the tradeoff; ask which metric they optimized and what they sacrificed.



The fairness impossibility result (Chouldechova; Kleinberg): when group base rates differ, these three cannot all hold at once, so every deployment picks which to satisfy.

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

■ The Optum Case: Every Failure at Once

The clearest real-world proof that a well-built model can still be unjust is the Optum case. Obermeyer and colleagues (2019, Science) 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**.

🔗 ANALOGY

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, and the bias was invisible until an outside team audited outputs by race. This is why fairness auditing cannot be optional, and why it must examine outputs stratified by group, not just inputs.

■ Why Race-Blind Design Fails

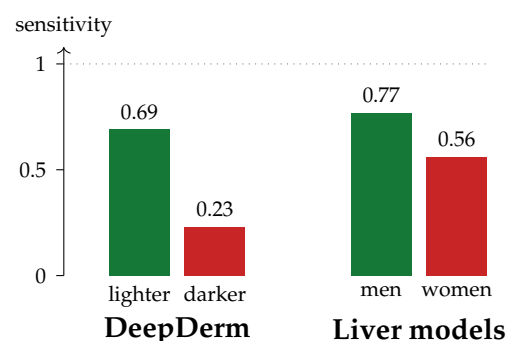
The intuitive fix, delete race from the features, does not make a model race-blind; it makes the model hunt for race elsewhere. Zip code predicts race with high accuracy because of residential segregation; remove it and the model reaches for pharmacy density, insurance type, or utilization history. It is like removing the label from a folder while leaving all the contents inside; the information is still there. Race also lives directly inside clinical algorithms: the eGFR race coefficient and older spirometry “correction factors” hard-coded worse-fitting biology into the math itself. In a world where race correlates with income, insurance, and utilization, there is no truly race-neutral feature set, so 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.

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

Skin Tone and Sex Gaps

Headline accuracy is a marketing number; subgroup-disaggregated accuracy is a clinical one, because the average is a photograph taken from far enough away that the disparities blur out. Stanford's DeepDerm scored 0.69 sensitivity on light skin but only **0.23** on dark skin, a threefold gap traced to light-skin-dominated training data (one widely used dataset held a single Fitzpatrick type VI image, 0.07 percent). A review of 30 liver-disease algorithms missed **44 percent** of cases in women versus 23 percent in men, and none of the 30 studies had even stratified results by sex. These gaps widen existing survival disparities: because melanoma already kills Black patients at higher rates through later-stage diagnosis, a model that confidently calls a dark-skin lesion benign amplifies an existing inequity under the cover of technological authority.

Watch out. “Works well on average” is not acceptable when the average hides a threefold gap that maps onto real health inequities. Evaluate every model stratified by skin tone, age, and sex before deployment; if performance drops for a subgroup, fix the model or explicitly restrict its intended-use population.



Headline accuracy hides subgroup gaps. DeepDerm caught 69% of cancers on lighter skin but 23% on darker; across 30 liver-disease models sensitivity fell from 77% in men to 56% in women.

SDOH: The Double-Edged Sword

Social determinants drive an estimated 30 to 55 percent of outcomes, which makes them a double-edged sword. Excluding them does not remove their effect; the model picks up housing instability through surrogates (frequent ED visits, missed appointments) and quietly penalizes patients for 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, 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 under payer logic or investing in them under population-health logic, and the algorithm does not choose, the institution does.

Debiasing: Pre, In, and Post-Processing

Once you have measured a disparity, you can intervene at three points, like fixing a leak at the reservoir, at the pump, or at the tap. **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. The Optum fix was pre-processing, changing the label, which is the cheapest and most durable;

post-processing is fast but needs the protected attribute at inference time and can trade away accuracy. The 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).

Stage	Where it acts	Example tool	Trade-off
Pre-processing	The data / label	Reweighting; relabel target	Durable, needs new data
In-processing	The training loop	Fairlearn ExponentiatedGradient	Ties fairness to one model
Post-processing	The outputs	Fairlearn ThresholdOptimizer	Needs protected attr. at inference

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, not a technical, 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, and at which processing stage would you intervene?

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