# Bias and Ethics in Medical Artificial Intelligence

**Abstract:**
Artificial Intelligence (AI) is revolutionizing healthcare at a fast pace, with huge leaps in diagnostics, predictive modelling, and clinical decision-making. While these improvements have been valuable, the incorporation of AI has also raised serious issues of bias and fairness. Bias can arise from unrepresentative or skewed datasets, historical inequities present within medical records, or algorithmic design decisions that inadvertently disadvantage certain groups. These biases have the potential to widen present health disparities, unfairly impacting patients along lines of race, gender, socioeconomic status, or geography.

This study delves into the aspects of bias in healthcare AI, classifying them as data-driven bias, measurement bias, and algorithmic bias. The research also assesses fairness metrics such as demographic parity, equal opportunity, and equalized odds to draw frameworks for auditing and enhancing equity in AI systems. Ethical and regulative issues like transparency, informed consent, and regulation in line with GDPR and HIPAA are deeply analysed.

To meet these challenges, the research suggests a fairness-aware system that combines varied data collection, algorithmic debiasing methods, and explainable AI (XAI) to maximize trust in clinicians. Case studies highlight both the dangers of biased AI and the value of fairness-aware interventions. Finally, fairness in AI healthcare systems is not just a technical requirement but a moral obligation in order to produce equitable and trustworthy medical results.

## Introduction:

Machine learning for healthcare (MLHC) is a rapidly developing field that requires a balance between machine learning and healthcare perspectives to maximize benefits and minimize risks. Progress has been made in developing MLHC models, designing and reporting clinical trials, and establishing regulatory evaluation protocols. However, issues of fairness, bias, and unintentional disparate impact remain unaddressed. These concerns have been highlighted in recent research, including racial imbalances in facial recognition and gender classification algorithms, natural language processing output, and racial bias in bail and criminal sentence algorithms. This paper aims to address these issues by examining the fairness in recent guidelines on reporting MLHC models, clinical trials, and regulatory approval. The paper also emphasizes the importance of justice in MLHC and how this can be operationalized in the MLHC setting.

## Background:

Artificial Intelligence (AI) has come to be recognized as a revolutionary technology in medicine, with unprecedented potential for diagnostics, disease prediction, treatment planning, and targeted medicine. AI models trained on massive amounts of medical data have achieved performance matching, and in some instances surpassing, that of human experts for applications in radiology, pathology, and risk assessment. The potential of AI is that it could enhance efficiency, decrease costs, and augment the precision of clinical decision-making.

Despite these possibilities, there are difficulties that cannot be ignored. The most urgent of these is the problem of bias and unfairness in AI systems. Healthcare choices have direct implications for human lives, and unjust or discriminatory results can result in serious harm, most of all to vulnerable groups. Bias can stem from biased data sets, measurement errors, or algorithmic design decisions, leading to disparate treatment recommendations across groups based on race, gender, socioeconomic status, or geography.

**Problem Statement:**

Even with increased use of AI in medicine, numerous research studies confirm that discriminatory algorithms tend to worsen existing health inequities instead of alleviating them.

For instance, a generally accepted healthcare risk-prediction model created by Optum in the United States was shown to predict systematically lower health requirements for Black patients than white patients. The algorithm employed previous healthcare expenditures as a health needs proxy because in the past, less money was allocated to Black patients with the same medical conditions; therefore, the system determined—erroneously— that the Black patients needed less care. Consequently, fewer Black patients were identified for additional medical attention, demonstrating how discriminatory design decisions may directly disadvantage marginalized groups.

Such instances prove that medical AI bias is not a hypothetical possibility but an actual and pressing issue. Biased algorithms undermine the credibility, reliability, and ethics-permitting acceptability of AI-based healthcare systems. Making medical AI fair is not just a technical problem but also a moral and regulatory imperative.

**Objectives:**

Artificial intelligence (AI) systems hold the ability to transform clinical practice, such as enhancing diagnosis precision and surgical decision-making, with decreased costs and manpower. These systems can, however, reinforce social inequities or show bias, either in terms of race or gender. These biases can take place pre-processing, during or post-processing of developing AI models, and hence it is imperative to recognize and counter possible biases to facilitate the proper and trustworthy use of AI models in clinical applications.

To reduce bias issues at the time of model development, we reviewed recent works on various debiasing strategies in biomedical natural language processing (NLP) or computer vision (CV). Then we talked about the techniques, like data perturbation and adversarial learning, that have been used in the biomedical field to resolve bias.

**Scope:**

The research is concerned with AI healthcare applications and bias and fairness, especially in diagnostic software, predictive modelling, and clinical decision support systems. Though the scope does not cover all uses of AI in medicine (e.g., robotic surgery), algorithmic fairness and fair outcomes remain the focus.

**Contributions:**

By putting together existing understanding of the sources of bias in healthcare AI, differing approaches to fairness assessment, consideration of ethical and legal implications, and proposing a paradigm that is fairness-conscious and supported by real case studies, this work contributes. All these efforts aim to guarantee trust in the next generation of healthcare technology by guiding the development of AI systems that are fair and accurate.

**Methodology:**

This research focuses on a qualitative analytical approach combined with a focused technical review to comprehend and assess some of the main challenges associated with bias and fairness in Artificial Intelligence applications within healthcare. Most of the literature presented focuses on works published between 2018 and 2025, from well-renowned academic databases such as IEEE Xplore, PubMed, SpringerLink, ScienceDirect, and arXiv. These sources have been selected for their demonstrated credibility in the fields of computer science, biomedical engineering, and healthcare informatics.

The first step of the methodology was to identify and categorize the kinds of biases that are common in healthcare AI systems. Three broad categories include data-driven bias, emanating from unrepresentative or imbalanced datasets; measurement bias, due to inconsistent or incomplete clinical data; and algorithmic bias, resulting from model design or optimization objectives that favour certain outcomes over others. Analysis of case studies and real-world incidents, such as the Optum health risk prediction algorithm, highlighted how these biases can produce disparate patient outcomes for each category.

In the second phase, fairness metrics used to quantify equity in AI decision-making were evaluated. Metrics of demographic parity, equal opportunity, equalized odds, and calibration error were considered in terms of their relevance and applicability in a clinical setting. These metrics were analysed not only in theoretical terms but also in respect to their computational implementation within common machine learning frameworks like scikit-learn and TensorFlow.

The third phase comprised the review of various debiasing strategies used within AI systems, including reweighting, data augmentation, adversarial learning, and fairness-constrained optimization. These techniques were compared across various AI model architectures — including Logistic Regression, Decision Trees, Random Forests, CNNs, and Transformers — regarding their appropriateness in healthcare scenarios.

Ethical and regulatory perspectives were integrated by reviewing international guidelines and policies, with a particular focus on GDPR, HIPAA, and the World Health Organization's Ethical Guidelines for AI in Health (2021). Such a multi-dimensional methodology ensures that both the technical and ethical dimensions of bias are analysed in-depth, and a balanced framework for understanding fairness in AI-driven healthcare systems is offered.

**Fairness Metrics:**

Fairness in health care AI systems must be quantified by measures that capture disparity in model performance across subgroups, either demographic or clinical.

These metrics allow us to answer whether an algorithm produces equitable outcomes and also where its biases may be. The most widely adopted fairness metrics in machine learning are listed below, especially relevant for medical applications like diagnosis, triage, and risk prediction.

**1. Demographic Parity (Statistical Parity):**

This metric requires that the predictions of an algorithm are independent of sensitive attributes such as race, gender, or socioeconomic background. Demographic parity, formally, holds when:

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

where *Y* is the predicted outcome and *A* represents a protected attribute

(for example: gender, ethnicity)

In health care, this translates to all groups having an equal chance of being flagged as high-risk or recommended for certain treatment, irrespective of their group membership.

Example:

If an AI model predicts the risk of diabetes, then, assuming identical clinical indicators, it should identify male and female patients as high-risk at similar rates.

**2. Equal Opportunity**

Equal opportunity emphasizes that the true positive rate across groups should be equal. It requires that patients who truly have a disease should have an equal chance of being correctly identified regardless of their demographic attributes. Mathematically:

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

This is particularly important in the case of medical screening or diagnostic applications, since missing a positive case in one group can lead to delayed or inadequate treatment.

Example:

In a model for diagnosing breast cancer, a woman of any ethnicity should have the same probability of correctly being identified as having the disease if she actually does.

**3. Equalized Odds:**

Equalized odds extend the notion of equal opportunity by considering FPRs in addition to TPRs. Here, the **true positive rate** and the **false positive rate** must be equal for both groups:

$$P(Y\hat{} = 1 \mid Y = y, A = a) = P(Y\hat{} = 1 \mid Y = y, A = b) \;\; for\, y \in \{0,1\}$$

This makes sure that any AI system would not systematically over or underpredict the outcomes of a particular group. In health, this would keep the diagnostic balance: it avoids under and overdiagnosis.

Example:

A cardiac risk assessment model should not yield disproportionately higher false alarms for one ethnic group versus another, lest it lead to unnecessary procedures or anxiety.

**4. Calibration and Predictive Parity:**

Calibration checks whether predicted probabilities match actual outcomes within each subgroup. A model is said to be calibrated if:

$$P( Y = 1 \mid Y^{=}p, A = a\, ) = p$$

for all probability scores *p*

Another related concept is predictive parity, which ensures that the PPV, or positive predictive value, is identical across groups; i.e., when a model predicts a positive outcome, the likelihood of it being correct is identical across demographics.

Example: In the case of a hospital's AI predicting a 70 percent likelihood that a patient will develop sepsis, about 70 percent of the patients flagged should actually develop sepsis.

**Interpretation in Healthcare Context:**

Each fairness metric provides a different lens to analyse equity in AI systems. Demographic parity focuses on population-level balance and may not consider medical relevance. One of the key points of equal opportunity is to correctly identify sick patients, which is quite relevant in a clinical setting. Equalized odds balances sensitivity and specificity across groups, aligning with diagnostic fairness. Calibration provides trust in the predicted probabilities for clinical decision-making. In practice, perfection in all the metrics is hardly attainable because of trade-offs between accuracy, fairness, and clinical utility. Thus, developers and health professionals have to choose which fairness metrics best fit ethical priorities and real-world consequences of model errors in patient care.

## Literature Review:

### Generalised Biases of AI in Healthcare:

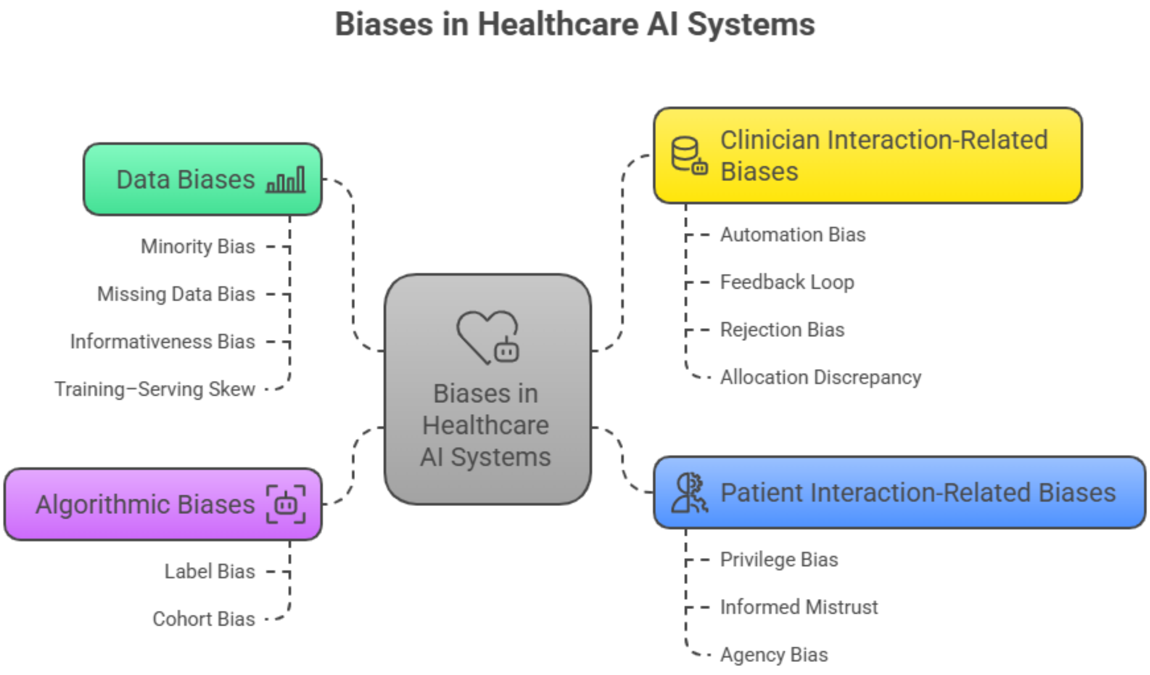
Healthcare is progressively depending on data-driven prediction models and algorithms that rely on patient-specific demographics and disease profile to predict the chance of a specific disease or outcome. Such algorithms have been created for many clinical conditions over many decades in different settings. The availability of big population datasets and computational power has enabled analysis using techniques like artificial intelligence (AI).

AI has the potential to transform health care by executing clinical functions quicker and with more accuracy than human beings. There have been recent concerns raised in studies about biases in AI-based predictive algorithms that undermine health equity and cause harm. Identifying sources of bias and also pointing out methods of addressing possible disparities are important processes towards promoting health equity and human rights.

This review seeks to identify key sources of bias in every step of the creation of AI algorithms in medicine and outline methods to decrease bias and disparities.

**Biases of AI in Healthcare:**

Generally, biases in AI may stem from two main sources: algorithmic bias, which results from the algorithm's inherent design or learning processes, and data bias, which results from the training data for the algorithm.  Yet, with the intricacies of human relationships and decision-making involved, other biases could manifest in the health context. These additional biases can be categorized into two types: those due to interactions between clinicians and AI and those due to interactions between patients and AI. Fig. 1 gives an overview of these biases.



**Mathematics and Bias of AI in Healthcare:**

Artificial Intelligence and Machine Learning have revolutionized medicine, enabling predictive diagnosis, individualized treatment, and medical imaging interpretation. Techniques like Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, KNN, Naïve Bayes, Convolutional Neural Networks, Recurrent Neural Networks, Gradient Boosting, and Reinforcement Learning are used for disease classification and patient risk prediction. However, these models can be biased, potentially affecting minorities, diagnostic databases, and clinical notes. Understanding the mathematical bases and biases of these algorithms is crucial for fairness, safety, and reliability in AI-based healthcare.

**Logistic Regression**

**Formula:**

$$P(y = 1 \mid x) = \sigma(\beta_0 + \beta^T x) = \frac{1}{1 + e^{-(\beta_0 + \sum_i \beta_i x_i)}}$$

**Interpretation:**

- Logistic regression calculates a probability that a patient has a condition by linearly combining features with weights. The sigmoid maps the linear score to [0,1].

**Use in healthcare:**

- Risk scoring (diabetes, heart disease), binary screening tests, sepsis risk flags.

**Causes of bias:**

- Feature exclusion / confounding: Omission of socioeconomic or access-to-care variables causes coefficients to soak up correlated effects that are associated with protected attributes.
- Training distribution imbalance: When training set overrepresents a subgroup, estimates capture that subgroup's relationships.
- Measurement error / heteroskedasticity: Laboratory values measured differently by site alter effective estimates.

- Thresholding effects: Single cut-off (e.g., 0.5) may lead to different TPR/FPR for groups with differing base rates.

**Mitigations:**

- Mean incorporate relevant social determinants or employ causal variable selection.
- Reweight or stratify the training data; employ group-specific thresholds or calibration.
- Calibrate probabilities by group (Platt scaling / isotonic) and provide subgroup metrics.

### Decision Trees (Gini & Entropy)

**Formula:**

$$\boldsymbol{Gini(S) = 1 - \sum_{i} p_i^2}$$

$$\boldsymbol{Entropy(S) = -\sum_{i} p_i \log_2 p_i}$$

$$\boldsymbol{IG(S, A) = Entropy(S) - \sum_{v} \frac{|S_v|}{|S|} Entropy(S_v)}$$

**Interpretation:**

- Trees recursively split data to maximize purity (minimize impurity) at a node; Gini and Entropy are measures of impurity that drive splits.

**Healthcare Applications:**

- Clinical decision rules, triage trees, transparent diagnostic flows.

**Sources of bias:**

- Feature-selection bias: Utility for features with large numbers of levels (zip code, hospital code) can proxy for race/SES.
- Small-sample leaves: Infrequent groups produce unstable splits (high variance).

- Greedy splits: Initial splits controlled by majority-group cues, suppresses subgroup-relevant splits.

**Mitigations:**

- Cap tree depth and impose minimum samples per leaf (per group).
- Apply fairness-aware splitting rules or constraint-based tree learning.
- Prune based on subgroup performance as a condition.

**Random Forest:**

**Formula:**

Train large numbers of trees on bootstrap samples and with random subsets of features; make prediction by majority vote or averaged probability:

$$\hat{\boldsymbol{y}} = \mathbf{mode}\{\boldsymbol{h_m}(\boldsymbol{x})\}$$

**or**

$$\hat{\boldsymbol{p}} = \frac{\mathbf{1}}{\boldsymbol{M}}\sum_{\boldsymbol{m}} \boldsymbol{h_m}(\boldsymbol{x})$$

**Interpretation:**

Ensemble diminishes variance of individual trees by averaging numerous randomized trees.

**Healthcare applications:**

Mortality/readmission prediction, diagnostics with mixed features.

**Causes of bias:**

- Bootstrap sampling bias: Rare group instances might be under-sampled across numerous trees.
- Aggregate masking: Ensemble's good overall performance conceals subgroup failures.
- Feature importance masking: Global importance conceals predictors related to subgroups

**Mitigations:**

- Bootstrapping stratified or balanced so each tree gets to see subgroup instances.
- Report group-wise metrics and calibrate subgroup-wise
- Use feature importance per-group and construct ensemble elements trained per-group.

**Support Vector Machines (SVM)**

**Formula:**

Linear SVM margin optimization:

$$\boldsymbol{min_{w,b}} ½||\boldsymbol{w}||^{\boldsymbol{2}}\boldsymbol{s.t.y}_{\mathrm{i}}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{i}}+\boldsymbol{b})\geq\mathbf{1}$$

$$\boldsymbol{Kernels}:\boldsymbol{K}(\boldsymbol{x}_{\mathrm{i}},\boldsymbol{x}_{\mathrm{j}})=\boldsymbol{\varphi}(\boldsymbol{x}_{\mathrm{i}})^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}_{\mathrm{j}})$$

**Interpretation:**

Identifies the hyperplane that maximizes margin between classes; support vectors define boundary.

**Healthcare applications:**

Imaging features, gene-expression classification, EEG/ECG pattern classification.

**Causes of bias:**

- Support-vector shortage: Minoritized groups offer few support vectors, thus boundary privileges majority.
- Kernel sensitivity: Kernel and scaling distinguish group distributions (different variances → different influence)
- Global cost (C): Global accuracy optimal single cost parameter can exacerbate subgroup errors.

**Mitigations**

- Employ class-/group-weighted SVM (greater weight for minority errors).
- Scale features per-group or employ robust kernels; analyse support-vector distribution by group.

- Cross-validate by group-aware folds.

### k-Nearest Neighbours (KNN)

**Formula:**

Prediction by among the k nearest neighbours under **distance** $\boldsymbol{d(x, y)}$.

**Interpretation:**

Labels a new patient by the labels of nearest neighbour historical patients.

**Applications in healthcare:**

Case-based reasoning, patient-similarity retrieval, small-sample tasks.

**Causes of bias:**

- Distance metric mismatch: Varying feature distributions between groups render distances non-comparable; neighbours can over-represent other groups.
- Sparse minority neighbourhoods: Minority patients have fewer neighbours in common → noisy predictions.
- High-dimension problems: Redundant features overwhelm distance in high-dim biomedical data.

**Mitigation:**

- Apply metric learning or Mahalanobis distance (determine how far a point is from a distribution — taking into account the **correlations between variables**) which is adjusted to maintain clinically important similarity.
- Dimensionality reduction that maintains subgroup structure (e.g., supervised PCA).
- Employ weighted-k (distance-weighted votes) and optimize k per-group.

## Naïve Bayes

### Formula:

$$P(C \mid x) \propto P(C)\prod_{i} P(x_i \mid C)$$

(Assumes conditional independence of features given class)

### Interpretation:

Estimates posterior probabilities through class priors and product of feature likelihoods.

### Healthcare applications:

Clinical notes text classification, basic disease likelihood calculators.

### Sources of bias:

- Independence assumption violation: Varying feature correlations by subgroup cause mis-estimated posteriors.
- Background/expectation mismatch: If class prior at deployment varies by subgroup from training, predictions bias.
- Textual/vocabulary bias: Clinical terminology varies by clinician or patient population; token frequency distributions differ.

### Mitigations

- Apply hierarchical Bayesian models with group-specific priors.
- Re-train priors with local prevalence; increase text corpora with varied clinical narratives.
- Apply embeddings reducing lexical bias.

## Convolutional Neural Networks (CNNs)

### Formula:

Convolution:

$$(F * K)(i,j) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} F(i+m, j+n)\ K(m,n)$$

followed by non-linearity.

### Interpretation:

CNNs learn hierarchical spatial filters learning visual patterns (edges → textures → high-level features).

### Healthcare applications:

Radiology (X-ray, CT), pathology slides, dermatology images, retinal scans.

### Causes of bias:

- Dataset skew: Lack of representation in skin tones, ages, or scanners → filters not generalizable to those subpopulations.
- Shortcut learning / spurious correlates: CNNs learn non-disease features (scanner watermark, hospital tags) that are associated with disease labels.
- Scanner/device/site shift: Covariate shift across acquisition devices correlated with demographics.

### Mitigations:

- Collect representative multi-site, multi-device image datasets (include stratification by skin tone).
- Apply domain adaptation and augmentation to mimic real-world variation.
- Apply and monitor saliency/attribution maps by subgroup to identify shortcuts.
- Group-wise calibration of predicted probabilities.

## Recurrent Neural Networks / LSTMs

**Formula:**

$$\boldsymbol{h_t} = \boldsymbol{\phi}\backslash \boldsymbol{big}(\boldsymbol{W_h h_{t-1}} + \boldsymbol{W_x x_t} + \boldsymbol{b}\backslash \boldsymbol{big})$$

LSTM adds gates $(input\ \ i_t, forget\ \ f_t, output\ o_t\ )$ with gating equations.

**Interpretation:**

- RNN/LSTM models learn temporal dependencies in time-series (physiological signals, EHR sequences).
- Healthcare applications
- ICU deterioration/sepsis prediction, ECG/EEG waveform classification, longitudinal outcome prediction.

**Sources of bias:**

- Variable sampling/missingness: Varying frequencies of recording between groups (sparser records among underserved patients) bias learned temporal dynamics.
- Treatment-path confounding: Treatment patterns vary by group and are learned as signals for outcomes.
- Censoring / label timing discrepancy: Varying follow-up patterns introduce label noise between groups.

**Mitigations:**

- Model missingness explicitly (mask) and incorporate time-interval features; employ models resilient to irregular sampling (e.g., GRU-D).
- Employ causal methods to separate treatment effects from physiology.
- Assess temporal generalization between hospitals and subgroups.

### Gradient Boosting Machines (XGBoost, LightGBM, CatBoost)

**Formula:**

Additive Model: $\widehat{\boldsymbol{y}^{(t)}} = \boldsymbol{y}^{(\boldsymbol{t}-\mathbf{1})} + \boldsymbol{\eta} \boldsymbol{f}_{\boldsymbol{t}}(\boldsymbol{x})$ where each $\boldsymbol{f}_{\boldsymbol{t}}$ is a tree fitted to the gradient of the loss.

**Interpretation:**

Sequentially fit trees to adjust residuals of previous ensemble; effective for tabular data.

**Healthcare Uses:**

Risk scoring, readmission prediction, sepsis early warning.

**Causes of bias:**

- Global loss optimization: Minimizing total loss can mask subgroup errors; model trades minority performance for reduced average loss.
- Interaction bias: Generalized interactions capture majority behaviour (small-group interactions overlooked).
- Overfitting on spurious correlations: Expressive trees can overfit proxies that correlate with protected attributes.

**Mitigations:**

- Apply sample reweighting or fairness-aware loss terms that penalize subgroup gaps.
- Track per-group metrics across boosting rounds and early-stop based on fairness constraints.
- Construct subgroup-specific models or multi-task architectures when interactions vary.

## Clustering (k-means, hierarchical)

### Formula:

K-means objective:

$$\min_{\{C_k\}} \sum_{k} \sum_{x \in C_k} \backslash lVert\, x - \mu_k \backslash rVert^2$$

### Interpretation:

Clustering patients into groups with minimal within-group variance — valuable for identifying subtypes.

### Healthcare Uses:

Disease subtyping (cancer molecular subtypes), patient phenotyping in tailored care.

### Causes of Bias:

- Distance metric bias: Attributes with varying variances or distributions between groups might cause clusters to separate based on protected features instead of clinical characteristics.
- Density imbalance: Clinically significant but small subgroups become absorbed into larger groups and lose distinctiveness.
- Label transfer bias: Human label assignment/interpretation may capture annotator bias.

### Mitigations:

- Standardize attributes and employ clinically informed distance measures; think about fairness-aware clustering algorithms.
- Employ cluster validation by subgroup and ensure the utilization of rare cluster detection strategies.
- Engage various clinical specialists in cluster interpretation.

## Natural Language Processing (Transformers):

### Self-Attention (core):

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}!\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{\boldsymbol{d}_{\boldsymbol{k}}}}\right)\boldsymbol{V}$$

### Interpretation:

Models calculate contextualized word embeddings through pairwise attention, encapsulating semantics in clinical text.

### Healthcare Uses:

EHR information extraction, clinical summarization, triage chatbots.

### Causes of bias:

- Annotator & documentation bias: Clinical documentation captures provider biases (words used to talk about compliance, pain, behaviour), which models learn.
- Vocabulary/dialect mismatch: Cultural and linguistic differences lead to poorer performance for specific patient populations
- Pretraining corpus bias: Web-pretrained models inherit social stereotypes and incorrect information.

### Mitigations:

- Fine-tune on highly annotated, diverse clinical corpora; debias embeddings (post-hoc neutralization)
- Employ fairness-aware downstream objectives; audit outputs for adverse correlations.
- Use diverse annotators and clinical environments in datasets.

## Reinforcement Learning (RL)

### Q-learning Update:

$$\boldsymbol{Q(s,a) \leftarrow Q(s,a) + \alpha \backslash Big(R + \gamma \max_{a'} Q\,(s',a') - Q(s,a) \backslash Big)}$$

### Interpretation:

Learns policies that optimize cumulative reward — beneficial for sequential treatment decisions.

### Healthcare Uses:

Personalized dosing (e.g., insulin), adaptive treatment policies, planning of repeated interventions.

### Causes of bias:

- Reward-definition bias: Reward based on cost minimization instead of fair health outcomes produces discriminatory policies.
- Logged-policy bias: RL learned from logged historical data simulates clinician biases in logged actions.
- State coverage shortage: Minority-group states underrepresented → unsafe/poor policies for them.

### Mitigations:

- Design reward functions such as equity penalties or per-group minimums.
- Employ off-policy correction and counterfactual estimation; guarantee strong coverage through simulation prior to deployment.
- Restrict policy to guarantee minimal subgroup performance.

**Cross-cutting (how bias appears mathematically & operationally):**

**Coefficient shifts / residual bias:** Non-zero

$$\boldsymbol{E\backslash big[\hat{y} - y\ \backslash big|\ A = a\backslash big]}$$

shows systematic over/under-estimation appears across subgroups for regression models.

**Group error gaps:**

$$|\mathbf{TPR_0} - \mathbf{TPR_1}|$$

$$|\mathbf{FPR_0} - \mathbf{FPR_1}|$$

$$|\mathbf{PPV_0} - \mathbf{PPV_1}|$$

differences quantify classification imbalances.

**Calibration gaps:** Probability calibration varies across groups and results in miscalibration.

**Shortcut learning:** Models using spurious features that are highly correlated with protected attributes get high accuracy but don't generalize to new settings.

| Algorithm | How it works | Healthcare Applications | Causes of Bias | Possible Solutions |
|---|---|---|---|---|
| **Logistic Regression** | Estimates probability of disease from patient features (age, BMI, blood pressure). | Risk prediction for diabetes, heart disease, cancer. | Missing key variables, imbalanced patient groups, errors in hospital records. | Collect balanced datasets, reweight minorities, apply fairness-aware models. |
| **Decision Trees** | Splits patients into decision rules (“if blood sugar > X → diabetes”). | Diagnosis support, disease screening. | Prefers features linked to socioeconomic status, ignores rare diseases, unstable for small subgroups. | Fairness-based splitting, oversampling minorities, pruning with fairness checks. |
| **Random Forests** | Combines many decision trees and predicts by majority vote. | Predicting sepsis, ICU mortality, readmission risk. | Inherits tree biases, favours majority class, systemic healthcare data imbalance. | Balanced sampling, fairness-aware forests, subgroup-specific calibration. |
| **Support Vector Machines (SVMs)** | Finds a boundary separating healthy vs diseased patients. | Cancer detection, imaging-based diagnosis. | Majority groups dominate the model, kernels ignore minority patterns, sensitive attributes leak into features. | Weighted SVMs, fairness-constrained optimization. |

| Algorithm | How it works | Healthcare Applications | Causes of Bias | Possible Solutions |
|---|---|---|---|---|
| **Convolutional Neural Networks (CNNs)** | Learns patterns in images (edges, textures, shapes). | Tumour detection, pneumonia from X-rays, skin cancer, diabetic retinopathy. | Training data lacks diversity (e.g., mostly lighter skin), spurious correlations, underrepresented rare cases. | Build diverse datasets, domain adaptation, explainable AI verification. |
| **NLP Models** | Reads clinical notes and extracts useful information. | Symptom extraction, medical chatbots, triage systems. | Clinical notes use biased language, dialect/language misinterpretation, stereotypes in training data. | Use balanced medical corpora, debias embeddings, apply fairness checks. |

**Discussions and Recommendations:**

The results of this study show that while AI holds extraordinary promise for improving healthcare, the risks from algorithmic bias cannot be underestimated. Bias in AI models often surfaces subtly: in how data is collected, in choices related to model design, and even in post-deployment feedback loops. In healthcare, these biases impact clinical decision-making, risk stratification, and patient prioritization and thus may result in unfair or unsafe outcomes. For example, diagnostic models trained primarily on data from a certain ethnic or socioeconomic group tend to underperform in underrepresented populations, which increases health disparities rather than closing them.

These issues can be addressed only by developing a multi-level strategy that integrates technical, ethical, and regulatory approaches. On the technical front, fairness-aware methods include bias detection, subgroup performance evaluation, and model calibration; data preprocessing methods, such as reweighting, resampling, and feature balancing, can correct unequal distributions. During training, adversarial debiasing and fairness-constrained optimization can work to ensure that models yield equal performance for different demographic groups. Moreover, post-hoc approaches-such as explainable AI methods and interpretable visualizations-allow the practitioner to audit how predictions are formed and whether they rely on spurious correlations.

Fairness cannot be based on algorithmic solutions but needs governance frameworks that enshrine transparency, accountability, and inclusiveness from an ethical and organizational standpoint. Governance policies around data should necessitate a range of representative datasets, ethical review of model development, and informed consent processes for the use of patient data. Similarly, institutions should form interdisciplinary fairness committees comprising clinicians, data scientists, and ethicists who can assess new AI tools prior to deployment.

At the regulatory level, compliance with international standards such as GDPR, HIPAA, and the proposed EU Artificial Intelligence Act can provide a common ethical grounding for AI in health. These frameworks advance patient autonomy, data privacy, and fairness in automated decision-making. National and regional health authorities should also promote AI model registries-public databases in which developers disclose dataset sources, fairness audits, and model performance across populations.

Finally, education and awareness play a crucial role. Clinicians must be trained not only to use AI tools but also to understand their limitations and potential biases. Developers, in turn, need clinical context to design systems that serve all patient groups equitably. A continuous feedback mechanism between the AI model's outcomes and clinical observations can foster adaptive, bias-aware systems that evolve with real-world healthcare needs. In sum, fairness in AI healthcare systems demands a holistic ecosystem: technological innovation intertwined with ethical reflection and institutional responsibility. Only then could AI become a transformative force, which improves healthcare outcomes without compromising equity or trust.

**References:**

[1] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[3] D. S. Char, N. H. Shah, and D. Magnus, “Implementing machine learning in health care — addressing ethical challenges,” *New England Journal of Medicine*, vol. 378, no. 11, pp. 981–983, 2018.

[4] A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” *New England Journal of Medicine*, vol. 380, pp. 1347–1358, Apr. 2019.

[5] S. R. Patil, J. M. Glover, and L. M. Becker, “Bias recognition and mitigation strategies in artificial intelligence for healthcare,” *npj Digital Medicine*, vol. 8, no. 1, pp. 1–13, Jan. 2025. [Online]. Available: https://www.nature.com/articles/s41746-025-01503-7

[6] N. Sharma, P. D. Choudhury, and K. S. Lee, “A scoping review and evidence gap analysis of clinical AI fairness,” *npj Digital Medicine*, vol. 8, no. 2, pp. 1–17, Feb. 2025. [Online]. Available: https://www.nature.com/articles/s41746-025-01667-2

[7] A. K. Gupta, T. D. Li, and M. R. Hassan, “AI-Driven Healthcare: A Review on Ensuring Fairness and Mitigating Bias,” *PLOS Digital Health*, vol. 3, no. 2, pp. 1–14, 2024. [Online]. Available: https://journals.plos.org/digitalhealth/article?id=10.1371%2Fjournal.pdig.0000864

[8] E. M. Hutton, D. J. Kim, and S. A. Lee, “Algorithmic individual fairness and healthcare: A scoping review,” *JAMIA Open*, vol. 8, no. 1, pp. 1–10, 2023. [Online]. Available: https://academic.oup.com/jamiaopen/article/8/1/ooae149/7934945

[9] R. T. Morales, A. P. Singh, and L. K. Peterson, “Systematic review of artificial intelligence biases across racial and ethnic disparities in healthcare,” *Artificial Intelligence in Medicine*, vol. 158, pp. 102840, Sept. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0720048X24005837