Bias Assessment Checklist for Healthcare AI

Pre-Development Phase

Problem Formulation

- [ ] Is the problem clearly defined?
- [ ] Have we consulted diverse stakeholders in defining the problem?
- [ ] Could this problem be solved without AI?
- [ ] Have we considered whether AI might perpetuate existing health disparities?
- [ ] Is there potential for the AI to be misused or cause harm?

Use Case Analysis

- [ ] What is the intended use of this AI system?
- [ ] Who will be impacted by this AI system?
- [ ] What populations might be disproportionately affected?
- [ ] Are there vulnerable or marginalized groups to consider?
- [ ] What are potential unintended consequences?

Data Collection and Preparation

Data Source Assessment

- [ ] What are the sources of training data?
- [ ] Are data sources representative of the target population?
- [ ] Are there known biases in data collection methods?
- [ ] Have historical inequities affected the data?
- [ ] Is the data collection process documented?

Demographic Representation

- [ ] Does training data include diverse demographic groups?
    - [ ] Age groups (pediatric, adult, geriatric)
    - [ ] Gender (male, female, non-binary)
    - [ ] Race and ethnicity
    - [ ] Socioeconomic status
    - [ ] Geographic location (urban, suburban, rural)
    - [ ] Language and culture
    - [ ] Disability status
    - [ ] Insurance status
- [ ] Calculate representation statistics:
    - Sample size for each demographic group
    - Percentage of total dataset
    - Comparison to target population distribution

Data Quality by Group

- [ ] Is data completeness similar across demographic groups?
- [ ] Are there systematic differences in missing data?
- [ ] Is label quality consistent across groups?
- [ ] Are measurement methods consistent across groups?
- [ ] Document any group-specific data quality issues

Historical Bias Analysis

- [ ] Have past inequities in healthcare affected the data?
    - [ ] Differential access to care
    - [ ] Differential quality of care
    - [ ] Diagnostic bias
    - [ ] Treatment bias
    - [ ] Social determinants of health
- [ ] Have we consulted domain experts about historical biases?
- [ ] Can historical biases be quantified in the data?
- [ ] Have we documented known historical biases?

Representation Bias

- [ ] Are any groups underrepresented (< 5% of dataset)?
- [ ] Are any groups overrepresented?
- [ ] Do we have sufficient samples for rare conditions in all groups?
- [ ] Are edge cases represented for all demographic groups?
- [ ] Have we documented representation gaps?

Measurement Bias

- [ ] Are measurement tools equally valid across groups?
    - [ ] Do pulse oximeters work equally well across skin tones?
    - [ ] Are diagnostic criteria culturally appropriate?
    - [ ] Are scales and assessments validated for all groups?
- [ ] Are outcome definitions appropriate for all groups?
- [ ] Could proxy variables introduce bias?
- [ ] Have we validated measurements across demographic groups?

Label Bias

- [ ] Who labeled the data?
- [ ] Could labelers have introduced bias?
- [ ] Is inter-rater reliability similar across groups?

- [ ] Are label definitions culturally sensitive?
- [ ] Have we assessed label quality by demographic group?

Model Development

Feature Selection

- [ ] Could features encode protected attributes?
- [ ] Are proxies for protected attributes included?
    - [ ] ZIP code (proxy for race, socioeconomic status)
    - [ ] First name (proxy for gender, ethnicity)
    - [ ] Language preference
    - [ ] Insurance type
- [ ] Are features equally predictive across groups?
- [ ] Have we justified inclusion of sensitive features?
- [ ] Have we tested models without potentially biasing features?

Model Architecture

- [ ] Does the model architecture allow for fairness constraints?
- [ ] Can the model learn different patterns for different groups?
- [ ] Have we considered simpler, more interpretable models?
- [ ] Can the model provide uncertainty estimates?
- [ ] Is the model architecture documented?

Training Process

- [ ] Are we monitoring for overfitting to majority groups?
- [ ] Are we using stratified sampling or reweighting?
- [ ] Have we implemented fairness-aware training?
- [ ] Are we validating on diverse held-out sets?
- [ ] Is the training process reproducible?

Hyperparameter Tuning

- [ ] Are hyperparameters optimized for fairness as well as accuracy?
- [ ] Have we tuned separately for different demographic groups?
- [ ] Are we using fairness-aware hyperparameter search?
- [ ] Is hyperparameter selection documented?

Model Evaluation

Overall Performance

- [ ] What is the model's overall accuracy?
- [ ] What are precision, recall, F1 scores?

- [ ] What are false positive and false negative rates?
- [ ] What is the model's calibration?
- [ ] What is the model's AUC-ROC?

Disaggregated Performance

Calculate all metrics separately for each demographic subgroup:

By Gender:

- [ ] Male
- [ ] Female
- [ ] Non-binary/Other

By Race/Ethnicity:

- [ ] White
- [ ] Black or African American
- [ ] Asian
- [ ] Hispanic or Latino
- [ ] Native American
- [ ] Other

By Age Group:

- [ ] Pediatric (0-18)
- [ ] Young adult (19-39)
- [ ] Middle age (40-64)
- [ ] Senior (65+)

By Socioeconomic Status (if available):

- [ ] Low income
- [ ] Middle income
- [ ] High income

By Geography:

- [ ] Urban
- [ ] Suburban
- [ ] Rural

Intersectionality Analysis

- [ ] Have we examined intersectional groups?
  - Example: Black women, elderly Hispanics, rural low-income
- [ ] Are there groups with notably different performance?
- [ ] Have we identified the most vulnerable combinations?
- [ ] Is performance adequate for smallest subgroups?

Fairness Metrics

Select appropriate metrics based on use case:

- [ ] **Demographic Parity**: $P(\hat{Y}=1|A=0) = P(\hat{Y}=1|A=1)$
  - Equal positive prediction rates across groups
- [ ] **Equalized Odds**: $P(\hat{Y}=1|Y=y,A=0) = P(\hat{Y}=1|Y=y,A=1)$ for $y \in \{0,1\}$
  - Equal true positive and false positive rates across groups
- [ ] **Equal Opportunity**: $P(\hat{Y}=1|Y=1,A=0) = P(\hat{Y}=1|Y=1,A=1)$
  - Equal true positive rate (recall) across groups
- [ ] **Predictive Parity**: $P(Y=1|\hat{Y}=1,A=0) = P(Y=1|\hat{Y}=1,A=1)$
  - Equal precision across groups
- [ ] **Calibration**: $P(Y=1|\hat{Y}=p,A=0) = P(Y=1|\hat{Y}=p,A=1)$ for all p
  - Predicted probabilities match outcomes across groups
- [ ] Document which fairness metrics are prioritized and why

Disparate Impact Analysis

- [ ] Calculate disparate impact ratio: (Positive rate for group A) / (Positive rate for group B)
- [ ] Is ratio between 0.8 and 1.25? (80% rule)
- [ ] If not, is the disparity justified and necessary?
- [ ] Have we documented disparate impacts?

Error Analysis by Group

- [ ] What types of errors occur in each group?
- [ ] Are error patterns different across groups?
- [ ] Are certain groups more likely to experience specific harms?
- [ ] Have we conducted case reviews of errors in each group?
- [ ] Is there a plan to address differential errors?

Calibration Analysis

- [ ] Plot calibration curves for each demographic group
- [ ] Is the model well-calibrated for all groups?
- [ ] Are there groups with systematic over/under-prediction?
- [ ] Have we applied calibration corrections?

Bias Mitigation

Pre-processing Techniques

- [ ] Resampling to balance demographic groups
- [ ] Reweighting examples from underrepresented groups

- [ ] Data augmentation for minority groups
- [ ] Removing biased features
- [ ] Transforming features to reduce correlation with protected attributes

In-processing Techniques

- [ ] Adversarial debiasing
- [ ] Fairness constraints in optimization
- [ ] Multi-objective optimization (accuracy + fairness)
- [ ] Regularization for fairness
- [ ] Separate models for different groups

Post-processing Techniques

- [ ] Threshold optimization by group
- [ ] Prediction calibration by group
- [ ] Reject option classification
- [ ] Equalized odds post-processing

Mitigation Effectiveness

- [ ] Have we measured bias before and after mitigation?
- [ ] What is the trade-off between accuracy and fairness?
- [ ] Is the trade-off acceptable for the use case?
- [ ] Have we validated on held-out test data?
- [ ] Is mitigation approach documented?

Deployment Considerations

Target Population

- [ ] Does the deployment population match the training population?
- [ ] Are there new demographic groups in deployment?
- [ ] Have we tested on representative samples from deployment setting?
- [ ] Are there differences in data distribution?
- [ ] Have we planned for population shifts?

Monitoring Plan

- [ ] Real-time monitoring of predictions by demographic group
- [ ] Alerts for performance degradation in any group
- [ ] Regular fairness metric calculation (weekly/monthly)
- [ ] Feedback mechanism for bias reports
- [ ] Scheduled retraining and reevaluation

Deployment Safeguards

- [ ] Human review for high-stakes decisions
- [ ] Uncertainty thresholds for deferring to humans
- [ ] Override mechanisms for clinicians

- [ ] Clear communication of AI limitations
- [ ] Incident response plan for bias incidents

Documentation and Transparency

Model Card / Data Sheet

- [ ] Intended use and users
- [ ] Training data characteristics
- [ ] Demographic composition of training data
- [ ] Performance metrics (overall and disaggregated)
- [ ] Fairness metrics
- [ ] Known limitations
- [ ] Recommended usage
- [ ] Out-of-scope uses
- [ ] Bias mitigation approaches

Transparency Reporting

- [ ] Public documentation of fairness assessment
- [ ] Disclosure of known biases
- [ ] Explanation of mitigation efforts
- [ ] Regular updates on deployment performance
- [ ] Channels for stakeholder feedback

Stakeholder Review

Clinical Review

- [ ] Clinicians from diverse backgrounds reviewed the model
- [ ] Clinical validity confirmed for all demographic groups
- [ ] Potential harms identified and assessed
- [ ] Clinical guidelines developed for AI use

Ethics Review

- [ ] Ethics committee review completed
- [ ] Fairness assessment approved
- [ ] Vulnerable populations considered
- [ ] Ethical concerns addressed

Community Engagement

- [ ] Patient advocates consulted
- [ ] Community representatives involved
- [ ] Feedback from diverse patients gathered
- [ ] Concerns documented and addressed

Continuous Improvement

Regular Audits

- [ ] Schedule quarterly fairness audits
- [ ] Annual comprehensive bias assessment
- [ ] Post-incident reviews
- [ ] Updates based on new research

Feedback Loops

- [ ] Clinician feedback on biased predictions
- [ ] Patient complaint process
- [ ] Performance monitoring dashboards
- [ ] Regular stakeholder consultations

Retraining Strategy

- [ ] Triggers for model retraining defined
- [ ] New data collection to address gaps
- [ ] Bias reassessment after retraining
- [ ] Version control and comparison

Sign-off

I certify that a thorough bias assessment has been conducted and documented for this AI system.

Data Scientist: _____ Date: _____

Clinical Lead: _____ Date: _____

Ethics Officer: _____ Date: _____

Bias Audit Completed: _____ (Date)

Next Audit Due: _____ (Date)

Resources

- Fairness Indicators Toolkit
- AI Fairness 360 Toolkit
- What-If Tool
- Model Card Toolkit

This checklist is for educational purposes. Organizations should adapt it to their specific context with input from domain experts, ethicists, and affected communities.