Detecting and Classifying Bias Indicators in Clinical AI: A Framework for Primary and Secondary Feature Flagging and Comparative Fairness Analysis of OptAB

Master Thesis

To obtain the degree of a Master of Science (M.Sc.) in Web and Data Science

Presented by

Christy Abraham Kuriakose

[223100203]

Koblenz, in September 8, 2025

Erstgutachter:	Prof. Dr. Andreas Mauthe
	(Institut für Wirtschafts- und Verwaltungsinformatik, FG Mauthe)
Zweitgutachter:	Christopher Latz
	(Institut für Wirtschafts- und Verwaltungsinformatik, FG Mauthe)

Affidavit

I assure that I have written the present work independently and have not used any sources and aids other than those indicated and that the work in the same or similar form has not yet been available to any other examination authority and has been accepted by it as a part of an examination performance.

I agree with the setting of the work in the Library. Yes ☐ No ☐

I agree to the publication of this work on the internet. Yes ☐ No ☐

. . . . . .

(Place, Date) (Sighn)

Zusammenfassung

Bias in klinischer Künstlicher Intelligenz (KI) bleibt ein zentrales Problem, da Modelle häufig bestehende Ungleichheiten aus den Trainingsdaten übernehmen und dadurch das Risiko ungleicher Ergebnisse fÃ¼r vulnerable Gruppen verstÃ¤rken [1–3]. Während ein Großteil der bisherigen Forschung die Fairness erst nach der Implementierung bewertet, wurde bislang weniger untersucht, welche Datensätzeigenschaften als **primäre Indikatoren** für Bias (z. B. Geschlecht, Alter, Ethnie) und welche als **sekundäre Indikatoren** (z. B. Versicherungsstatus, Bildungsniveau, Wohnort) wirken [4–7].

Diese Arbeit adressiert diese Forschungslücke anhand von drei Zielen. Erstens wird eine systematische Methode entwickelt, um primäre und sekundäre Bias-Indikatoren in weit verbreiteten ICU-Datensätzen wie MIMIC-III, MIMIC-IV und eICU zu klassifizieren [8–10]. Zweitens wird der Einsatz interpretierbarer KI/ML-Modelle untersucht, die Attribute automatisch nach ihrem Bias-Risiko kennzeichnen, wobei statistische Tests mit literaturgestützter Begründung kombiniert werden [11–13]. Drittens wird ein Fairness-Audit von **OptAB**, einem aktuellen KI-Framework zur Antibiotikawahl bei Sepsispatienten, durchgeführt [14, 15].

Das Audit bewertet die Leistungsfähigkeit in Subgruppen, die nach Geschlecht, Alter, Nieren- und Leberfunktion definiert sind, unter Verwendung etablierter Fairness-Metriken wie Demographic Parity, Disparate Impact Ratio, Equal Opportunity, Equalized Odds, AUC Disparität und Brier Score [16–18]. Statistische Tests und Visualisierungsmethoden werden eingesetzt, um Ungleichheiten sichtbar zu machen [2, 3].

Der erwartete Beitrag ist eine kombinierte Taxonomie von Bias-Indikatoren, ein prototypisches Flagging-Tool sowie ein Fairness-Evaluationsrahmen für klinische KI. Zusammen sollen diese Ergebnisse zu gerechteren Anwendungen von KI in der Intensivmedizin beitragen und praktische Handlungsempfehlungen für eine verantwortungsvolle Implementierung liefern [19–21].

Abstract

Bias in clinical artificial intelligence (AI) remains a critical concern, as models often inherit inequities from training data and risk producing uneven outcomes for vulnerable groups [1–3]. While most research evaluates fairness after deployment, less attention has been paid to identifying which dataset features act as **primary indicators** of bias (e.g., race, sex, age) and which act as **secondary indicators** (e.g., insurance type, education, location) [4–7].

This thesis addresses that gap through three objectives. First, it develops a systematic method to classify primary and secondary bias indicators using widely used ICU datasets such as MIMIC-III, MIMIC-IV, and eICU [8–10]. Second, it explores the use of interpretable AI/ML models to automatically flag attributes by bias risk, combining statistical tests with literature guidance [11–13]. Third, it applies a fairness audit to **OptAB**, a recent AI framework for antibiotic selection in sepsis patients [14, 15].

The audit evaluates subgroup performance across sex, age, renal, and liver function using fairness metrics including demographic parity, disparate impact ratio, equal opportunity, equalized odds, AUC disparity, and Brier scores [16–18]. Statistical testing and visualization methods will be used to highlight disparities [2,3].

The expected contribution is a combined taxonomy of bias indicators, a prototype flagging tool, and a fairness evaluation framework for clinical AI. Together, these outputs aim to support more equitable applications of AI in critical care and provide practical insights for responsible deployment [19–21].

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

The integration of Artificial Intelligence (AI) into clinical practice has transformed modern healthcare, enabling improvements in diagnostics, prognostics, and personalized treatment strategies. Leveraging vast amounts of electronic health record (EHR) data, medical imaging, and sensor-derived information, AI-driven models have demonstrated superior performance compared to traditional statistical methods across tasks such as mortality prediction, disease classification, and clinical decision support [1, 2].

Despite these advancements, the rapid adoption of AI in healthcare raises pressing concerns about fairness and equity. Clinical AI models often inherit biases present in their training data, which may arise from historical inequities, imbalanced demographic representation, or systemic healthcare disparities [3]. When deployed, such biases can lead to disparate outcomesâoverestimating or underestimating risk in vulnerable subpopulations, misallocating clinical resources, or amplifying pre-existing gaps in healthcare delivery.

Global regulatory bodies, including the U.S. Food and Drug Administration (FDA), European Commission, and World Health Organization (WHO), have emphasized the necessity of addressing these challenges by enforcing ethical principles such as fairness, transparency, and accountability throughout the AI development lifecycle [22]. Consequently, research into bias detection, mitigation frameworks, and fairness evaluation metrics has become essential not only for improving model performance but also for safeguarding patient trust and ensuring equitable healthcare outcomes.

1.2 Problem Statement

Most fairness research focuses on auditing models after deployment rather than addressing bias at the feature level. There is limited systematic guidance on identifying which dataset attributes are primary(directly sensitive attributes like race, age) versus secondary indicators(indirect proxies like family history, zip code) of bias.This distinction is critical for proactive fairness interventions.

1.3 Research Questions

1. Which dataset attributes constitute primary versus secondary indicators of bias in clinical AI?
2. Can we build a model that flags these indicators automatically across multiple datasets?
3. What fairness disparities exist in OptAb an existing model for Antibiotics prediction for patients with Sepsis, and how do they vary across demographic metrics?

2 Objectives

1. Conduct an extensive review and statistical analysis to define primary and secondary bias indicators.
2. Build an AI/ML-based classifier to flag dataset attributes by bias risk level.
3. Perform a fairness audit on OptAb an existing model for Antibiotics prediction for patients with Sepsis using multiple fairness metrics across demographic subgroups.

3 Literature Review

3.1 Bias in Clinical Models

Bias in clinical models refers to systematic deviations in predictive outcomes that disadvantage certain patient groups based on attributes such as race, gender, age, or socioeconomic status. In healthcare AI, this issue is particularly critical as biased predictions can perpetuate or amplify existing disparities in access, diagnosis, and treatment outcomes. The literature highlights that bias can emerge at every stage of the AI lifecycleâfrom data collection and feature selection to algorithm training, deployment, and post-market surveillance. For instance, [1] emphasize that biases frequently stem from imbalanced or incomplete datasets and the influence of historical inequities embedded in clinical records, where minority populations are often underrepresented or mischaracterized. Similar observations arise in studies leveraging large-scale datasets like MIMIC-IV, where [2]reveal that deep learning mortality prediction models rely disproportionately on demographic features, raising fairness concerns even when predictive accuracy appears high.

3.1.1 Primary Indicators of Bias

Primary indicators of bias are direct, quantitative measures of performance disparities between subgroups. Metrics such as *sensitivity*, *specificity*, *predictive value*, and *calibration* stratified by protected attributes (e.g., race, sex, age) are widely used. In fairness research, group-based metrics like *equal opportunity* (ensuring similar true positive rates) and *equalized odds* (ensuring comparable false positive and false negative rates) are standard benchmarks [3]. For example, studies of clinical mortality prediction models show lower AUC scores in high-mortality subgroups, suggesting unequal reliability of predictions across patient categories [2].

3.1.2 Secondary Indicators of Bias

Secondary indicators are indirect or qualitative markers that reveal structural inequities or potential for disparate impact. These include *dataset imbalances*, such as underrepresenta-

tion of darker skin tones in dermatology imaging datasets, which lead to poorer diagnostic performance for those groups [3]. Another key secondary indicator is the presence of *proxy variables*, where socio-economic factors like healthcare cost or insurance type inadvertently act as surrogates for sensitive attributes, embedding systemic disparities into model predictions. Studies also point to *annotation inconsistencies* and *measurement variability* across institutions (e.g., differing imaging protocols) as sources of bias that may not surface in initial performance evaluations but become evident during real-world deployment [1, 22].

3.2 OptAB – An Optimal Antibiotic Selection Framework

The management of sepsis remains one of the most pressing challenges in critical care, with approximately 48.9 million cases and 11 million deaths globally in 2017, accounting for nearly 20% of all deaths worldwide [14]. A crucial determinant of patient outcomes is the timely and appropriate selection of initial antibiotic therapy. However, this decision is complicated by the heterogeneity of pathogens, the delayed availability of microbiological cultures, and the risks of drug-induced toxicities [14, 15]. In this context, artificial intelligence (AI) has emerged as a promising tool to support precision medicine in sepsis management, although most prior work has concentrated on early detection rather than therapeutic optimization [14].

OptAB (Optimal Antibiotic Selection Framework) represents a novel AI-based approach designed specifically to address the challenge of antibiotic selection in sepsis. It is the first fully data-driven and online-updateable antibiotic selection framework that accounts for both therapeutic efficacy and potential side effects [14, 15]. Unlike earlier AI models, such as the Artificial Intelligence Clinician, which optimized fluid and vasopressor administration without incorporating antibiotics [14, 15], OptAB focuses explicitly on antibiotic regimens. The system integrates clinical, demographic, and laboratory data with the Sepsis-related Organ Failure Assessment (SOFA) score—a widely accepted measure of disease severity [14]—to forecast disease progression under different antibiotic strategies.

OptAB leverages the *Treatment-Effect Controlled Differential Equation (TE-CDE)* model, a state-of-the-art neural controlled differential equation framework capable of handling irregular measurements, missing values, and time-dependent confounding [15]. This enables real-time assimilation of patient data and iterative refinement of predictions as new clinical measurements become available. Specifically, OptAB forecasts not only SOFA score trajectories but also laboratory markers of nephrotoxicity (creatinine) and hepatotoxicity (bilirubin, alanine transaminase), thereby balancing treatment success with safety considerations [14].

Initial evaluations of OptAB using the *MIMIC-IV* dataset (over 26,000 sepsis cases) demonstrated that its recommended antibiotic regimens achieved faster reductions in SOFA scores

compared to actual clinical treatments administered, with notable improvements observable within the first 18–48 hours [14, 15]. Importantly, OptAB was able to anticipate contraindications and identify high-risk patients for vancomycin-induced acute kidney injury and ceftriaxone-associated hepatotoxicity, offering safer alternative regimens in many cases [14]. External validation on the *AmsterdamUMCdb* dataset further suggested that the model generalized across different patient populations, although prediction horizons and accuracy were somewhat reduced, highlighting challenges associated with distribution shifts in clinical data [15].

Beyond population-level findings, OptAB has shown utility in individual-level decision support by simulating counterfactual treatment scenarios. For example, case studies revealed instances where OptABâs suggested regimen provided equal or better therapeutic efficacy with lower predicted risks of organ toxicity compared to physician's choices [14, 15]. Such capabilities position OptAB as a step toward **personalized sepsis therapy**, where antibiotic decisions can be dynamically tailored to each patientâs evolving clinical state.

Despite these promising results, limitations remain. Counterfactual predictions cannot be directly validated since each patient receives only one factual treatment, and unknown confounders in observational data may bias estimates [14]. Moreover, the current model is limited to three antibiotics (vancomycin, ceftriaxone, and piperacillin/tazobactam), reflecting both data availability and clinical relevance; broader antibiotic coverage and dose optimization represent critical areas for future research [15]. Additionally, explainability and uncertainty quantification are ongoing challenges for building clinician trust in AI-driven decision support [14].

In summary, OptAB is a pioneering framework for **AI-guided antibiotic selection in sepsis**, demonstrating improved treatment efficacy, safety monitoring, and adaptability across diverse patient datasets. While still at an early stage of validation, it addresses a crucial gap in sepsis management by moving beyond detection and monitoring toward **therapeutic decision optimization**, potentially contributing to reduced mortality, shorter ICU stays, and minimized adverse drug effects.

3.3 Frameworks for Bias Detection and Mitigation

The reviewed literature outlines several frameworks designed to address bias systematically across the AI pipeline. A dominant approach is the *bias lifecycle framework*, which embeds fairness checks at every developmental phaseâconception, data collection, preprocessing, in-processing (algorithm training), and post-deployment monitoring. [1] advocate for continuous bias auditing aligned with principles of diversity, equity, and inclusion, emphasizing

stakeholder participation, including underrepresented patient groups, in model design and evaluation.

Complementing lifecycle approaches are *fairness taxonomies* described by [3], categorizing mitigation strategies into *pre-processing, in-processing, and post-processing techniques*. Pre-processing methods involve balancing or reweighting datasets; in-processing methods incorporate fairness constraints or adversarial debiasing during training; post-processing techniques adjust outputs or thresholds for different subgroups.

Additionally, frameworks like the *responsible dataset rubric* [22] evaluate datasets on three axes: fairness (diversity and inclusivity), privacy (risk of sensitive information leakage), and regulatory compliance (consent, institutional approval, and data correction mechanisms). This rubric exposes trade-offs, such as the fairness–privacy paradox, where collecting sensitive attributes to improve fairness can inadvertently heighten privacy risks.

Emerging studies also integrate *interpretability and fairness analyses*, using feature attribution methods (e.g., SHAP, Integrated Gradients) to uncover disproportionate reliance on demographic features and guide corrective measures [2]. Together, these frameworks underscore a shift from static fairness assessments to *dynamic, context-specific governance models* that evolve alongside clinical and societal expectations.

3.4 Fairness Metrics

Standard metrics used in clinical fairness audits include:

1. **Demographic Parity Difference (DPD)** Measures the difference in positive prediction rates between two demographic groups.

$$\text{DPD} = P(\hat{y} = 1 \mid A = 0) - P(\hat{y} = 1 \mid A = 1)$$

2. **Disparate Impact Ratio (DIR)** A ratio comparing favorable outcomes across groups; ideal value is near 1.

$$\text{DIR} = \frac{P(\hat{y} = 1 \mid A = 1)}{P(\hat{y} = 1 \mid A = 0)}$$

3. **Equal Opportunity Difference (EOD)** Measures the difference in true positive rates (TPR) between groups.

$$\text{EOD} = \text{TPR}_{A=0} - \text{TPR}_{A=1}$$

4. **Equalized Odds (EO)** Measures differences in both true positive rate (TPR) and false positive rate (FPR) between groups.

$$\text{EO} = (\text{TPR}_{A=0} - \text{TPR}_{A=1}) + (\text{FPR}_{A=0} - \text{FPR}_{A=1})$$

5. **AUC Disparity** Difference in the area under the ROC curve across demographic groups.

6. **Brier Score by Group** Evaluates calibration fairness by comparing Brier scores across groups.

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

These metrics are widely used in clinical fairness audits [17].

4 Methodology

This chapter outlines the detailed methodologies for achieving the three objectives of the thesis: (1) identifying primary and secondary bias indicators, (2) developing a bias-flagging AI/ML model, and (3) conducting a fairness audit on an existing sepsis prediction model.

4.1 Objective 1: Identifying Primary and Secondary Bias Indicators

1. Overview

This methodology outlines a systematic approach to identify and classify variables in clinical datasets into *primary* and *secondary* indicators of potential bias. This classification supports downstream modeling for bias detection and mitigation. The datasets considered in this study include **MIMIC-III**, **MIMIC-IV**, and the **eICU Collaborative Research Database**, all of which are publicly available, de-identified datasets widely used in clinical machine learning research [8–10].

Unlike traditional fairness metrics such as demographic parity or disparate impact ratios, this methodology constructs a typology of bias indicators based on their structural relationship to social inequities in healthcare systems.

Primary indicators of bias are defined as variables that directly reflect protected characteristics (e.g., race, gender), which are well-documented sources of systemic disparities in medical care. **Secondary indicators of bias** are variables that indirectly encode social disadvantage or vulnerability (e.g., marital status, education level, insurance type), often acting as mediators or amplifiers of primary inequities.

4.1.1 Dataset Selection and Preparation

We utilize publicly available ICU datasets such as MIMIC-IV [23] and eICU [24]. These datasets provide detailed demographic, clinical, and outcome data essential for bias analysis. Data preprocessing involves:

- Handling missing values via imputation (mean for continuous, mode for categorical variables).
- Standardizing categorical variables (e.g., unifying race categories like "Caucasian" and White).
- Normalizing continuous features for comparability.

4.1.2 Literature-Guided Feature Preselection

Variables are categorized based on literature in public health, sociology, and ethics:

- **Primary Bias Indicators**: Attributes such as *race/ethnicity, gender/sex, age, language proficiency* are considered primary indicators. These are strongly associated with institutional biases and historical discrimination in healthcare delivery [4, 5, 25].
- **Secondary Bias Indicators**: Variables like *marital status, education, income level, insurance type, and geographic location* are categorized as secondary. While not protected attributes per se, these are proxies for socioeconomic status and often interact with primary variables to magnify disparities [6, 7].

4.1.3 Statistical Analysis

We perform statistical tests to classify indicators:

- Chi-square tests for categorical variables (e.g., race vs. mortality outcome) [11].
- ANOVA/Kruskal-Wallis tests for continuous variables (e.g., age distribution differences).
- Correlation analysis for proxy detection (e.g., ZIP code correlating with race).

4.1.4 Classification and Reasoning

Features are classified as:

- **Primary Indicators**: Strong direct link to bias (e.g., race directly affecting mortality predictions).
- **Secondary Indicators**: Indirect link (e.g., ZIP code correlates with socioeconomic status).

Documented reasoning combines statistical significance with literature evidence.

Table 4.1: Primary and Secondary Indicators of Biasness

Feature	**Type**	**Reason for Category**
Race/Ethnicity	Primary	Strongly associated with systemic disparities, major regulatory mandate, affects outcomes/access
Age	Primary	Key determinant in health risk and access, widely reported source of bias in outcomes
Sex/Gender	Primary	Systematic disparities in healthcare, often a required fairness audit dimension
Insurance Type (SES)	Primary	Proxy for socioeconomic status, impacts access, environment, treatment options
Location/Region	Primary	Reflects health system, urban/rural disparity, environmental and service variation
Family History	Secondary	Risk factor often mediated by SES/race; relevance via inherited conditions
Marital Status	Secondary	Influences social support, adherence, but less direct/strong effect than primary features
Language Preference	Secondary	Affects communication, access, often associated with but mediated by ethnicity/immigration
Comorbidities	Secondary	Reflect cumulative effects of upstream disparities, not independent axes of bias
Education Level	Secondary	SES proxy, affects health literacy, but typically operates via SES and insurance type

4.2 Objective 2: Developing Bias-Flagging AI/ML Model

4.2.1 Training Data Construction

Each feature (attribute) is represented as an instance with properties:

- Representation imbalance (percentage differences across groups).
- Correlation with sensitive attributes.
- Statistical significance from Objective 1.

- Literature-derived bias flag.

Labels are assigned as "Primary" or "Secondary" based on Objective 1 classification.

4.2.2 Model Selection and Training

We select interpretable models like Decision Trees or Logistic Regression to ensure transparency in healthcare contexts [12, 18]. Training involves:

- Splitting data into 80/20 train-test sets.
- Optimizing for accuracy and F1-score.

4.2.3 Explainability and Validation

SHAP values [13] are used to interpret why features are classified as primary or secondary, validating alignment with clinical literature.

4.2.4 Output

The final tool can process any clinical dataset and automatically classify bias indicators, providing explainability for each decision.

4.3 Objective 3: Fairness Audit of OptAB

4.3.1 OverView

This study evaluates the fairness of *OptAB*, an online-updateable, data-driven antibiotic selection framework for sepsis patients that forecasts disease trajectories (SOFA score) and side-effect indicators (creatinine, total bilirubin, ALT) under candidate antibiotics and their combinations. Prior work has established OptAB's overall forecasting accuracy and its capacity to propose treatments that reduce SOFA while respecting contraindications. Here, we focus specifically on *fairness*: whether prediction quality, the allocation of favorable/safe recommendations, and expected clinical improvements are equitably distributed across demographic and clinical subgroups.

We operationalize fairness along three dimensions:

1. **Prediction Fairness**: parity of forecast errors (SOFA and labs) across groups.
2. **Treatment Recommendation Fairness**: parity in access to favorable (effective and safe) recommendations and in avoidance of contraindicated drugs.

3. **Outcome Fairness**: parity in expected benefit (SOFA reduction) from OptAB's recommended therapy across groups.

4.3.2 Dataset

We use **MIMIC-IV (v2.2)**, a large U.S.-based ICU EHR resource. Variables required for this work are available in MIMIC-IV:

- *Demographics*: sex, age, height, weight (ethnicity is available but may be imbalanced).
- *Vitals/Labs*: SOFA components (for SOFA reconstruction), serum creatinine, total bilirubin, alanine transaminase (ALT), platelets, blood pressures, etc.
- *Therapies*: antibiotic administrations including Vancomycin, Ceftriaxone, and Piperacillin/-Tazobactam.

Inclusion adheres to Sepsis-3: ICU stays with SOFA ≥ 2 within the specified infection window; index time (*sepsis onset*) is defined per Sepsis-3 convention.

4.3.3 Subgroup Definitions

We stratify patients into subgroups to assess within-cohort fairness using MIMIC-IV alone. Unless otherwise noted, *baseline* refers to measurements closest prior to or at antibiotic initiation.

Sex. Male vs. Female.

Age. Young $<$ 60 years; Older $\geq$ 60 years. (If data permit, a three-way split such as $<$60, 60–75, $>$75 may be explored; analyses below generalize.)

Renal function.

- **Normal**: baseline creatinine $< 2.0 mg/dL$ and no *Stage-1 AKI*.
- **Impaired**: baseline creatinine $\geq 2.0 mg/dL$ *or* Stage-1 AKI, where Stage-1 AKI is any of:
    1. increase in creatinine $\geq 0.3 mg/dL$ within 48 h,
    2. increase to $\geq 1.5\times$ baseline creatinine within 7 days,
    3. urine output $< 0.5 mL/kg/h$ for 6–12 h.

Liver function.

- **Normal**: bilirubin $\leq 2.2mg/dL$ (female) or $\leq 2.4mg/dL$ (male), *and* ALT $\leq 280U/L$.
- **Impaired**: bilirubin above sex-specific thresholds *or* ALT $> 280U/L$.

4.3.4 Fairness Metrics

We formalize three classes of metrics. Throughout, let g index a subgroup, N_g denote its number of patients, and for patient i let $\{y_{i,t}\}_{t=1}^{T_i}$ be observed targets (SOFA or lab value) and $\{\hat{y}_{i,t}\}_{t=1}^{T_i}$ the corresponding OptAB forecasts aligned to observation times.

Prediction Fairness

Goal: ensure that forecasting errors are comparable across subgroups so that no group is systematically disadvantaged by poorer predictions.

Mean Squared Error (MSE) per subgroup.

$$\mathrm{MSE}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} \left(\frac{1}{T_i} \sum_{t=1}^{T_i} \left(\hat{y}_{i,t} - y_{i,t}\right)^2 \right). \tag{4.1}$$

Error Fairness Gap. For two subgroups g_1, g_2 (e.g., Female vs. Male),

$$\Delta_{\mathrm{MSE}}(g_1, g_2) = \left|\mathrm{MSE}_{g_1} - \mathrm{MSE}_{g_2}\right|. \tag{4.2}$$

Smaller Δ_{MSE} indicates greater prediction parity.

Treatment Recommendation Fairness

Goal: ensure equitable access to *favorable* (effective and safe) antibiotic recommendations and equitable avoidance of contraindicated drugs across groups.

Operational definition of *favorable treatment*. A recommendation is *favorable* if (i) it yields a lower projected SOFA at the chosen horizon (24/48 h) than the factual or alternatives, and (ii) it does not violate safety thresholds (e.g., avoid Vancomycin when renal impaired; avoid Ceftriaxone when hepatotoxic thresholds are exceeded).

Disparate Impact Ratio (DIR). Let $\mathcal{F}$ be the event that OptAB recommends a favorable treatment. Then

$$\mathrm{DIR}(g_1, g_2) = \frac{\Pr(\mathcal{F} \mid g_1)}{\Pr(\mathcal{F} \mid g_2)}. \tag{4.3}$$

Values close to 1 suggest parity; values below 0.8 or above 1.25 may indicate disparity (context-dependent).

Contraindication Avoidance Rate (CAR). For subgroup g,

$$\mathrm{CAR}_g = \frac{\#\{\text{patients in } g \text{ where a contraindicated drug was } \textit{not} \text{ recommended}\}}{\#\{\text{patients in } g \text{ with a relevant contraindication risk}\}}. \tag{4.4}$$

Examples: avoiding Vancomycin in renal impairment; avoiding Ceftriaxone in liver impairment.

Outcome Fairness

Goal: ensure subgroups derive comparable *benefit* from OptAB's recommendations (even if predictions and safety are comparable, benefits could still differ).

Patient-level SOFA Improvement. Let $\mathrm{SOFA}_{i,\mathsf{factual}}(H)$ be the observed SOFA at horizon H (e.g., 48 h) under factual care; let $\mathrm{SOFA}_{i,\mathsf{OptAB}}(H)$ be OptAB's predicted SOFA at H under its recommended regimen. Define

$$\Delta\mathrm{SOFA}_i(H) = \mathrm{SOFA}_{i,\mathsf{factual}}(H) - \mathrm{SOFA}_{i,\mathsf{OptAB}}(H). \tag{4.5}$$

Group-level benefit and disparity.

$$\overline{\Delta\mathrm{SOFA}}_g(H) = \frac{1}{N_g} \sum_{i=1}^{N_g} \Delta\mathrm{SOFA}_i(H), \tag{4.6}$$

$$\Delta_{\mathrm{SOFA}}(g_1, g_2; H) = |\overline{\Delta\mathrm{SOFA}}_{g_1}(H) - \overline{\Delta\mathrm{SOFA}}_{g_2}(H)|. \tag{4.7}$$

Smaller Δ_{SOFA} indicates outcome parity.

4.3.5 Statistical Testing

We test whether observed disparities are beyond sampling noise:

- **Two-sample t-tests** for continuous metrics (e.g., compare MSE_g between two groups, or $\overline{\Delta\mathrm{SOFA}}_g$).

- **ANOVA** when comparing > 2 groups (e.g., age <60, 60–75, >75).
- **Chi-square tests** for proportions (e.g., $\Pr(\mathcal{F} \mid g)$ for DIR; CAR_g).
- **Bootstrap CIs (95%)** for key gaps/ratios: resample patients within groups to obtain confidence intervals for Δ_{MSE}, DIR, CAR, and Δ_{SOFA}.

Alongside p-values, we will report effect sizes and confidence intervals to convey *practical* significance.

4.3.6 Visualization

The results will be presented to highlight fairness transparently:

- **Heatmaps** of MSE_g by subgroup and by horizon (e.g., 1 h, 6 h, 12 h, 24/48 h).
- **Bar charts** of $\overline{\Delta\mathrm{SOFA}}_g(H)$ with 95% CIs to compare expected benefit across groups.
- **DIR plots** (Eq. 4.3) with CIs to visualize allocation parity of favorable recommendations.
- **Scatter/forest plots** for CAR_g (Eq. 4.4) across renal/liver strata, stratified further by sex/age when powered.

These plots aid clinical interpretability (e.g., ensuring high-risk groups are equally protected).

4.3.7 Expected Outcomes

This MIMIC-IV–only analysis is designed to determine whether:

1. **Prediction accuracy** (Eq. 4.1) is comparable across sex, age, renal, and liver function groups (Eq. 4.2 close to zero).
2. **Recommendation allocation** is equitable: DIR near 1 (Eq. 4.3) and consistently high CAR across groups (Eq. 4.4).
3. **Expected clinical benefit** (SOFA reduction) is similar across groups (Eq. 4.7 small).

If disparities are found, they will be quantified and discussed, and potential mitigations will be outlined (e.g., subgroup-aware reweighting, calibration, or constraint-based objectives) for future development. Regardless of outcome, this evaluation strengthens the *trustworthiness* of OptAB by making its subgroup behavior explicit.

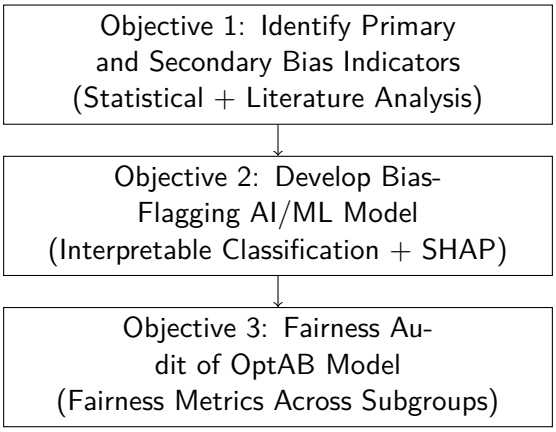

Figure 4.1: Flow of Methodology: From Bias Identification to Fairness Audit

4.4 Flow of Methodology

The flow of methodology for this study begins with Objective 1, which focuses on identifying both primary and secondary bias indicators through a combination of statistical analysis and literature review. This step ensures that the investigation is grounded in both empirical evidence and prior scholarly findings. Objective 2 involves developing a bias-flagging AI/ML model using interpretable classification methods and SHAP (Shapely Additive explanations) values. The emphasis here is on creating a model that not only detects potential biases but also provides transparency in its decision-making process. Finally, Objective 3 addresses the fairness audit of a OptAB model, where fairness metrics are evaluated across various subgroups to assess and ensure equitable performance. Together, these objectives form a structured progression from bias identification to a comprehensive fairness audit, as illustrated in Figure 4.1

5 Timeline

Table 5.1: Six-Month Timeline for Thesis Activities

Month	Tasks
1	**Literature Review & Dataset Setup** - Review literature on bias in clinical AI and fairness metrics. - Collect datasets (e.g., MIMIC-IV, eICU). - Clean and preprocess data (handle missing values, standardize features).
2	**Identify Primary vs Secondary Bias Indicators** - Perform statistical analysis (chi-square, ANOVA) to detect features strongly linked to bias. - Use literature evidence to classify features as primary or secondary. - Document reasoning for each classification.
3	**Develop Bias-Flagging Model** - Build an interpretable AI model (decision tree/logistic regression) to classify attributes. - Train and validate using statistical properties of features. - Apply explainability tools (e.g., SHAP) to interpret results.
4	**Prepare OptAb Model for Audit** - Setup OptAb model. - Define demographic subgroups (age, sex, race). - Set up fairness metrics for evaluation
5	**Conduct Fairness Analysis** - Compute fairness metrics across subgroups for the model. - Compare performance disparities across groups. - Identify features contributing to observed disparities.
6	**Integrate Results and Write Thesis** - Combine bias-indicator framework and fairness audit findings. - Draft final thesis (methods, results, discussion, limitations). - Prepare defense/presentation materials.

6 Significance

This thesis provides a structured taxonomy and conceptual framework for distinguishing between different types of bias indicators at the feature level in clinical datasets [16, 26]. Building on this foundation, it delivers an automated flagging tool capable of identifying high-risk features across diverse datasets, thereby supporting scalable bias detection [19]. Furthermore, the work contributes a detailed empirical fairness analysis of clinical AI models, exemplified through a real-world case study involving a sepsis prediction system [20]. Finally, it advances the field by proposing practical strategies for bias mitigation, addressing key challenges in the responsible deployment of clinical AI technologies [21, 25].

Bibliography

[1] F. Hasanzadeh, C. B. Josephson, G. Waters, D. Adedinsewo, Z. Azizi, and J. A. White, "Bias recognition and mitigation strategies in artificial intelligence healthcare applications," *npj Digital Medicine*, vol. 8, no. 154, 2025.

[2] C. Meng, L. Trinh, N. Xu, J. Enouen, and Y. Liu, "Interpretability and fairness evaluation of deep learning models on mimic-iv dataset," *Scientific Reports*, vol. 12, no. 7166, 2022.

[3] M. Liu, Y. Ning, S. Teixayavong *et al.*, "A scoping review and evidence gap analysis of clinical ai fairness," *npj Digital Medicine*, vol. 8, no. 360, 2025.

[4] Z. D. Bailey, N. Krieger, M. AgÃ©nor, J. Graves, N. Linos, and M. T. Bassett, "Structural racism and health inequities in the usa: evidence and interventions," *The Lancet*, vol. 389, no. 10077, pp. 1453–1463, 2017.

[5] D. R. Williams *et al.*, "Social determinants of health: the solid facts," *World Health Organization*, 2010.

[6] P. Braveman and L. Gottlieb, "The social determinants of health: coming of age," *Annual review of public health*, vol. 32, pp. 381–398, 2011.

[7] J. Robards, M. Evandrou, J. Falkingham, and A. Vlachantoni, "Marital status, health and mortality," *Maturitas*, vol. 73, no. 4, pp. 295–299, 2012.

[8] A. E. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[9] A. E. Johnson, T. J. Pollard, L. A. Celi, and R. G. Mark, "Mimic-iv (version 2.2)," https://physionet.org/content/mimiciv/, 2021.

[10] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eicu collaborative research database, a freely available multi-center database for critical care research," *Scientific data*, vol. 5, p. 180178, 2018.

[11] L. Zeng *et al.*, "Bias evaluation in machine learning for healthcare," *arXiv preprint arXiv:2310.14109*, 2023.

[12] N. Mehrabi *et al.*, "A survey on bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2021.

[13] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[14] P. Wendland, C. Schenkel-HÃ¤ger, I. Wenningmann, and M. Kschischo, "An optimal antibiotic selection framework for sepsis patients using artificial intelligence," *npj Digital Medicine*, vol. 7, p. 343, 2024. [Online]. Available: https://doi.org/10.1038/s41746-024-01350-y

[15] M. Kschischo, P. Wendland, C. Schenkel-HÃ¤ger, and I. Wenningmann, "Optab – an optimal antibiotic selection framework for sepsis patients using artificial intelligence," Preprint, Research Square, 2024, version 1, posted July 11, 2024. [Online]. Available: https://doi.org/10.21203/rs.3.rs-4598166/v1

[16] e. a. Matos, "Critical appraisal of fairness metrics in clinical predictive ai," *arXiv preprint arXiv:2506.17035*, 2025.

[17] M. Research, "Fairlearn: A toolkit for fairness in ai," 2023. [Online]. Available: https://fairlearn.org

[18] E. Chang *et al.*, "Explainable ai for fair sepsis mortality prediction," *arXiv preprint arXiv:2404.13139*, 2024.

[19] A. Rajkomar, M. Hardt, and M. D. Howell, "Ensuring fairness in machine learning to advance health equity," *Annals of Internal Medicine*, vol. 169, no. 12, pp. 866–872, 2018.

[20] C. W. Seymour, V. X. Liu, T. J. Iwashyna *et al.*, "Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 762–774, 2017.

[21] A. Chouldechova and A. Roth, "Frontiers in algorithmic fairness," *arXiv preprint arXiv:1810.08810*, 2018.

[22] S. Mittal, K. Thakral, R. Singh, M. Vatsa, T. Glaser, C. C. Ferrer, and T. Hassner, "On responsible machine learning datasets emphasizing fairness, privacy and regulatory

norms with examples in biometrics and healthcare," *Nature Machine Intelligence*, vol. 6, pp. 936–949, 2024.

[23] A. e. a. Johnson, "Mimic-iv (v2.2)," *PhysioNet*, 2023. [Online]. Available: https://physionet.org/content/mimiciv/2.2/

[24] T. Pollard *et al.*, "eicu collaborative research database," *Scientific Data*, 2018. [Online]. Available: https://eicu-crd.mit.edu/

[25] 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, 2019.

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