Appendix:

eTable 1: Summary of Fairness Mitigation Stages and Methods

Mitigation Method/Limitations	Purpose	When to Use	Mathematical Formulation
Pre-processing: modifies training data to reduce bias before model training, but it may distort the data distribution, remove predictive signals, and decrease accuracy; it is best suited when one has complete control over the dataset and seeks a model-agnostic approach.			
SMOTE	To address class imbalance by generating synthetic samples for minority classes.	When minority classes are underrepresented and the model is sensitive to class imbalance, it is suitable to address this issue before training any classifier.	$x_{\text{new}} = x_i + \delta \cdot (x_{zi} - x_i), \quad \delta \sim U(0,1)$
Reweighing	To balance the representation of protected groups and outcomes by assigning weights to instances.	When the dataset has unequal group-outcome representation, and you want a model-agnostic fairness adjustment.	$w(a,y) = \frac{P(A=a) \cdot P(Y=y)}{P(A=a, Y=y)}$
DIR	To reduce correlation between features and protected attributes while preserving feature rank order.	When you want to preprocess features to remove group-dependent bias without altering relative ordering.	$\tilde{X}_j = F^{-1}\left(\frac{1}{\lvert A\rvert}\sum_{a\in A} F_{j\lvert A=a}(X_j)\right)$
In-Processing: integrates fairness constraints into the learning objective during model training, requiring specialized implementation and customization of the training pipeline; it is best applied when fairness needs to be enforced during optimization, and compatibility with pre-trained models is not a priority.			
AD	To learn representations that predict the outcome while minimizing the ability to predict the protected attribute.	When the model training process can be customized, and you want to enforce fairness during learning.	$\min_{\theta}\max_{\phi} \mathbb{E}[\mathcal{L}_Y(f_\theta(X), Y)] - \lambda \mathbb{E}[\mathcal{L}_A(g_\phi(f_\theta(X)), A)]$
EGR	To satisfy fairness constraints by solving a constrained optimization problem.	When fairness can be expressed as linear constraints on classifier decisions, it is best suited for binary classification tasks.	$h^* = \sum_{t=1}^{T} \alpha_t h^{(t)}, \quad \alpha_t = \frac{1}{T}$
Post-Processing: adjusts model predictions after training to meet fairness criteria, but the original score distribution limits it, does not address upstream bias, and may lead to accuracy loss; it is most appropriate when working with black-box models where retraining is not feasible.			

ROC	To adjust predictions for samples near the decision boundary in favor of disadvantaged groups.	When you have a trained model and can identify a confidence band around the threshold, it is suitable when predictions near the threshold are less specific.	$P(\hat{Y}=1 \mid Y=1, A=a) = P(\hat{Y}=1 \mid Y=1, A=a')$ (Equal TPR) $P(\hat{Y}=1 \mid Y=0, A=a) = P(\hat{Y}=1 \mid Y=0, A=a')$ (Equal FPR)
EOP	To achieve equalized odds, ensure equal TPR and FPR across groups.	When group-specific thresholds can be applied, this is particularly useful when the goal is parity in error rates across sensitive groups.	$P(\hat{Y}=1 \mid Y=1, A=a) = P(\hat{Y}=1 \mid Y=1, A=a') \quad \forall a, a'$
Calibrated EOP	To achieve equalized odds while preserving probability calibration.	When equalized odds are desired, but calibrated probability estimates must be maintained for downstream decision-making.	$P(Y=1 \mid \hat{R}=r, A=a) = P(Y=1 \mid \hat{R}=r, A=a')$

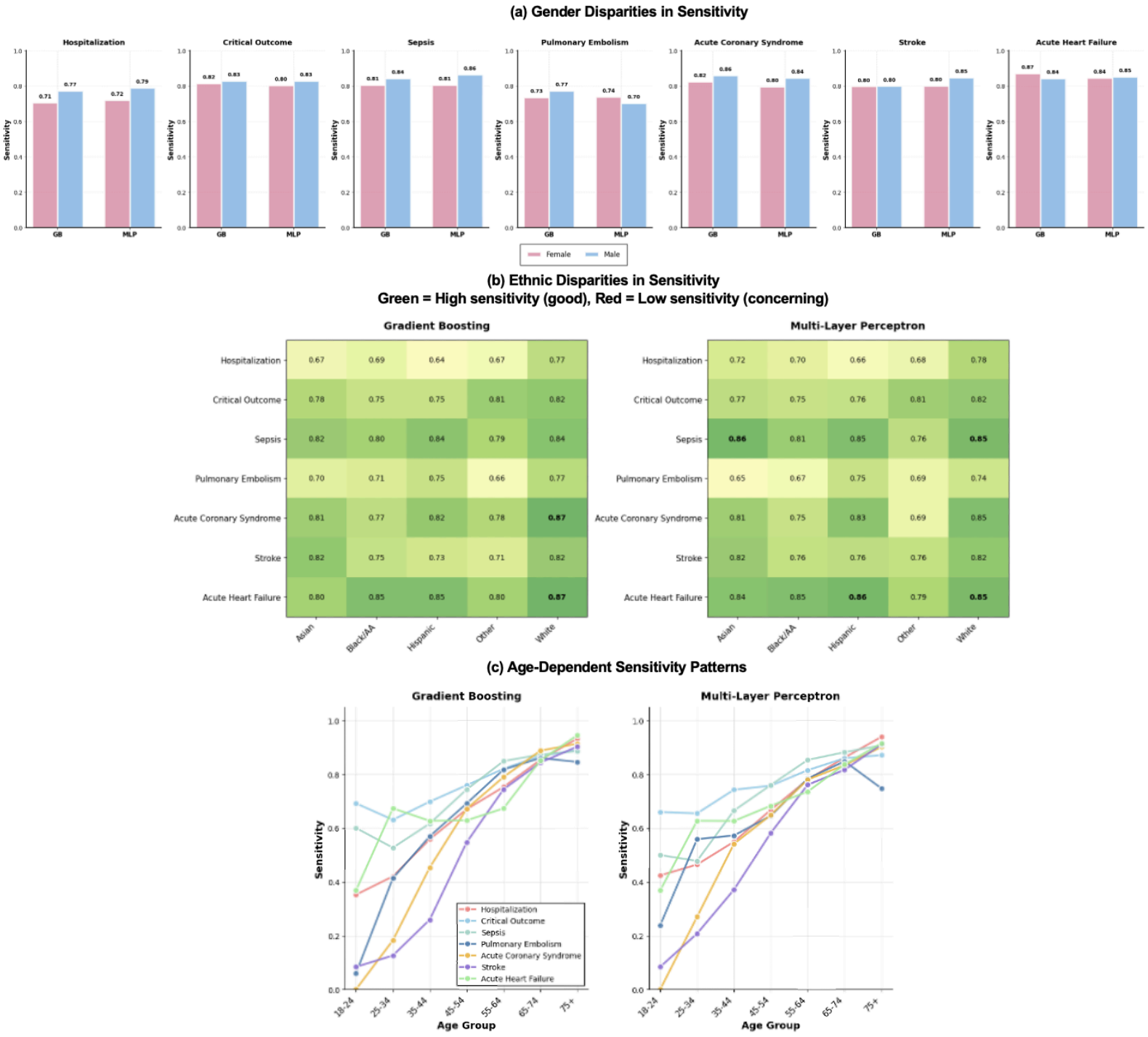
This table outlines the key fairness mitigation techniques applied across pre-processing, in-processing, and post-processing stages. For each method, the purpose, recommended use case, and mathematical formulation are summarized. Pre-processing methods adjust the training data, in-processing methods integrate fairness during model optimization, and post-processing methods modify predictions after training to equalize group performance.

eTable 2: Baseline Characteristics of BIDMC and Stanford Cohorts

	Cohort	BIDMC Train Set	BIDMC Test Set	Stanford Train Set	Stanford Test Set
Sample Size (N)		353,150	88,287	94,679	23,670
Gender	*Male*	161,244 (45.7%)	40,399 (45.8%)	43,201 (45.6%)	10,876 (45.9%)
	Female	191,906 (54.3%)	47,888 (54.2%)	51,478 (54.4%)	12,794 (54.1%)
Age	*18-24*	36,846 (10.4%)	9,408 (10.7%)	6,935 (7.3%)	1,686 (7.1%)
	25-34	50,897 (14.4%)	12,746 (14.4%)	14,415 (15.2%)	3,646 (15.4%)
	35-44	41,706 (11.8%)	10,483 (11.9%)	13,553 (14.3%)	3,398 (14.4%)
	45-54	53,672 (15.2%)	13,413 (15.2%)	12,922 (13.6%)	3,224 (13.6%)
	55-64	59,583 (16.9%)	14,887 (16.9%)	14,632 (15.5%)	3,748 (15.8%)
	65-74	49,549 (14.0%)	12,290 (13.9%)	13,668 (14.4%)	3,444 (14.6%)
	75+	60,897 (17.2%)	15,060 (17.1%)	17,529 (18.5%)	4,293 (18.1%)
Race/Ethnicity	*Asian*	7,757 (2.2%)	1,890 (2.1%)	15,399 (16.3%)	3,867 (16.3%)
	Black/ African Americans	50,910 (14.4%)	12,829 (14.5%)	6,031 (6.4%)	1,464 (6.2%)
	Hispanic/Latino	17,797 (5.0%)	4,429 (5.0%)	26,017 (27.5%)	6,473 (27.3%)
	White	141,526 (40.1%)	35,223 (39.9%)	35,204 (37.2%)	8,830 (37.3%)
	Others	135,160 (38.3%)	33,916 (38.4%)	12,028 (12.7%)	3,036 (12.8%)
Diagnoses	*Hospitalization*	167,165 (47.3%)	41,811 (47.4%)	23,499 (24.8%)	5,842 (24.7%)
	Critical Outcome	21,026 (6.0%)	5,119 (5.8%)	2,191 (2.3%)	567 (2.4%)
	ICU Transfer within 12 Hours	19,791 (5.6%)	4,816 (5.5%)	2,181 (2.3%)	563 (2.4%)
	Acute Coronary Syndrome	6,599 (1.9%)	1,493 (1.7%)	746 (0.8%)	179 (0.8%)
	Acute Heart Failure	18,501 (5.2%)	4,466 (5.1%)	793 (0.8%)	198 (0.8%)
	Pulmonary Embolism	2,958 (0.8%)	740 (0.8%)	359 (0.4%)	114 (0.5%)
	Sepsis	8,531 (2.4%)	2,046 (2.3%)	674 (0.7%)	169 (0.7%)
	Stroke	15,045 (4.3%)	3,725 (4.2%)	1,088 (1.1%)	266 (1.1%)

This table presents demographic and clinical characteristics of the final analysis cohorts from BIDMC and Stanford after applying the inclusion/exclusion criteria in Figure 1. The table reports the distributions of gender, age, race/ethnicity, and diagnosis for each site's 80% training and 20% test sets. These population differences, such as Stanford's higher proportions of Asian and Hispanic/Latino patients and BIDMC's higher hospitalization rate, provide essential context for interpreting cross-institutional fairness results.

eFigure 1: Fairness evaluation across demographic subgroups in the BIDMC cohort (Gradient Boosting and Multi-Layer Perceptron)

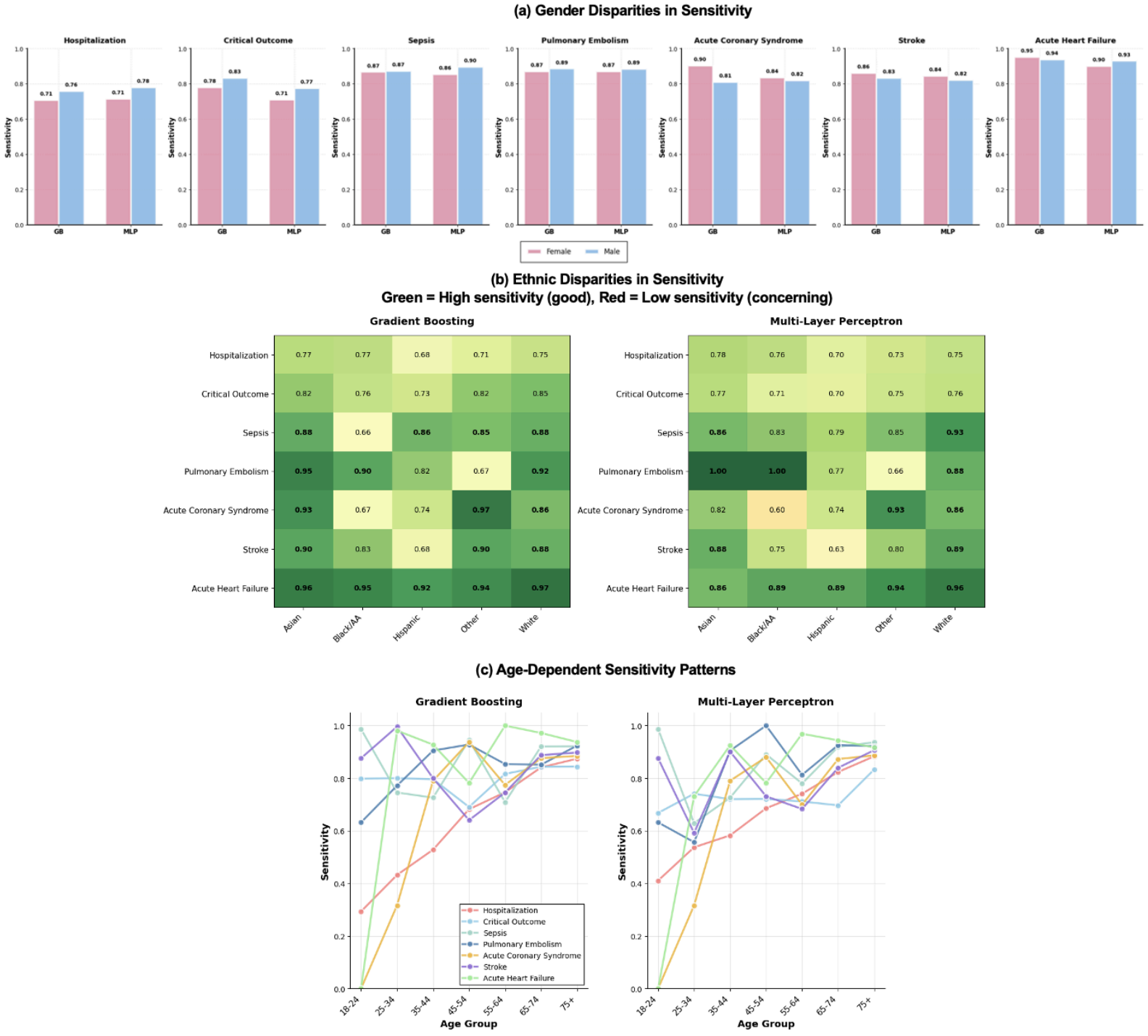


(a) Gender disparities in sensitivity were modest across outcomes and models, with minimal differences between male and female patients.

(b) Ethnic disparities in sensitivity (green = higher, red = lower) were generally small, though slightly lower recall was observed for some minority groups in pulmonary embolism and acute coronary syndrome.

(c) Age-dependent sensitivity patterns showed increasing sensitivity with age, with greater variability among younger adults, indicating performance differences likely driven by prevalence and sample size rather than algorithmic bias.

eFigure 2: Fairness evaluation across demographic subgroups in the Stanford cohort (Gradient Boosting and Multi-Layer Perceptron)

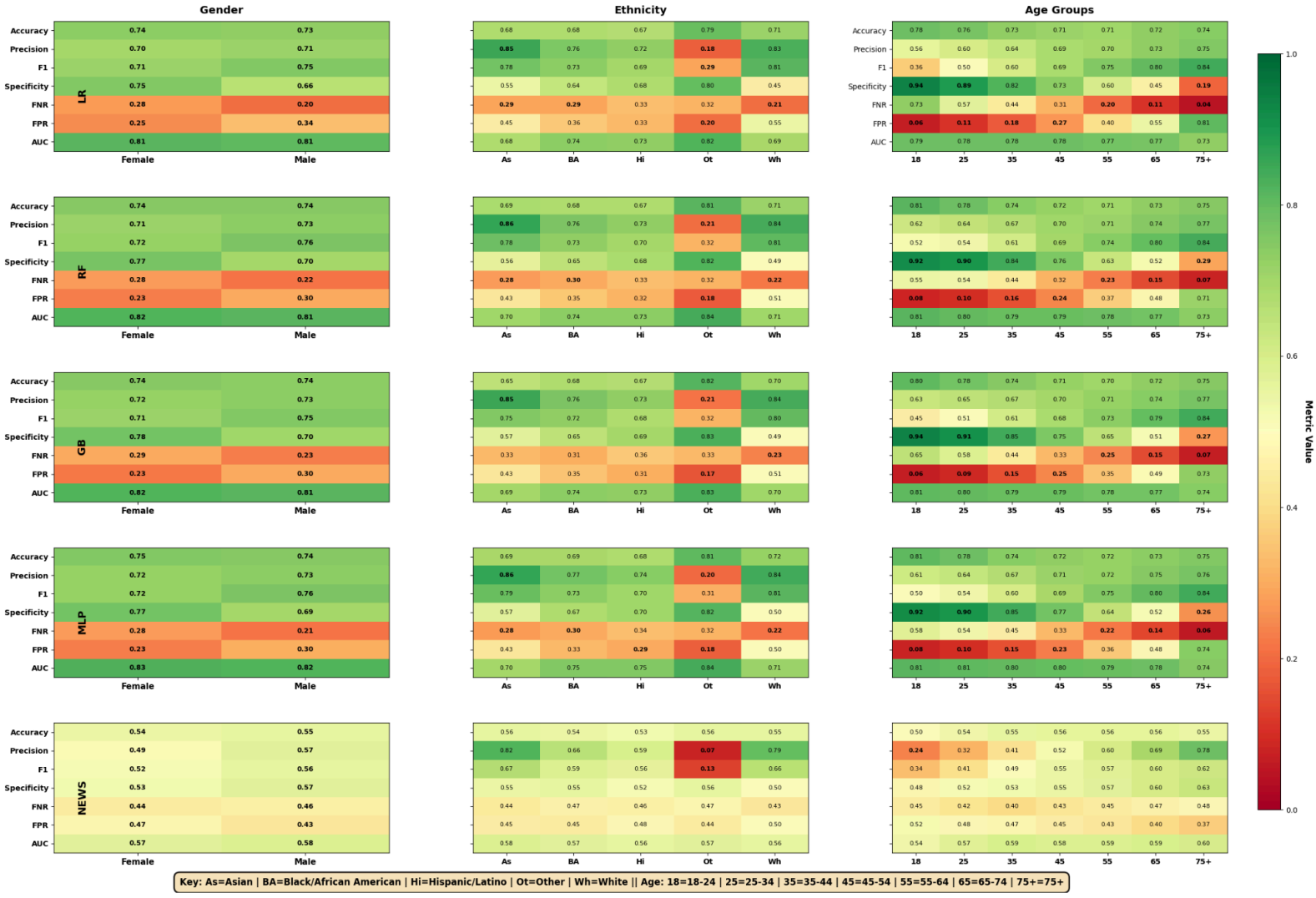


(a) Gender disparities in sensitivity were minimal across models and outcomes, with no consistent advantage for either gender.

(b) Ethnic disparities in sensitivity (green = higher, red = lower) were generally small, though slightly reduced performance was observed for sepsis prediction among Hispanic patients.

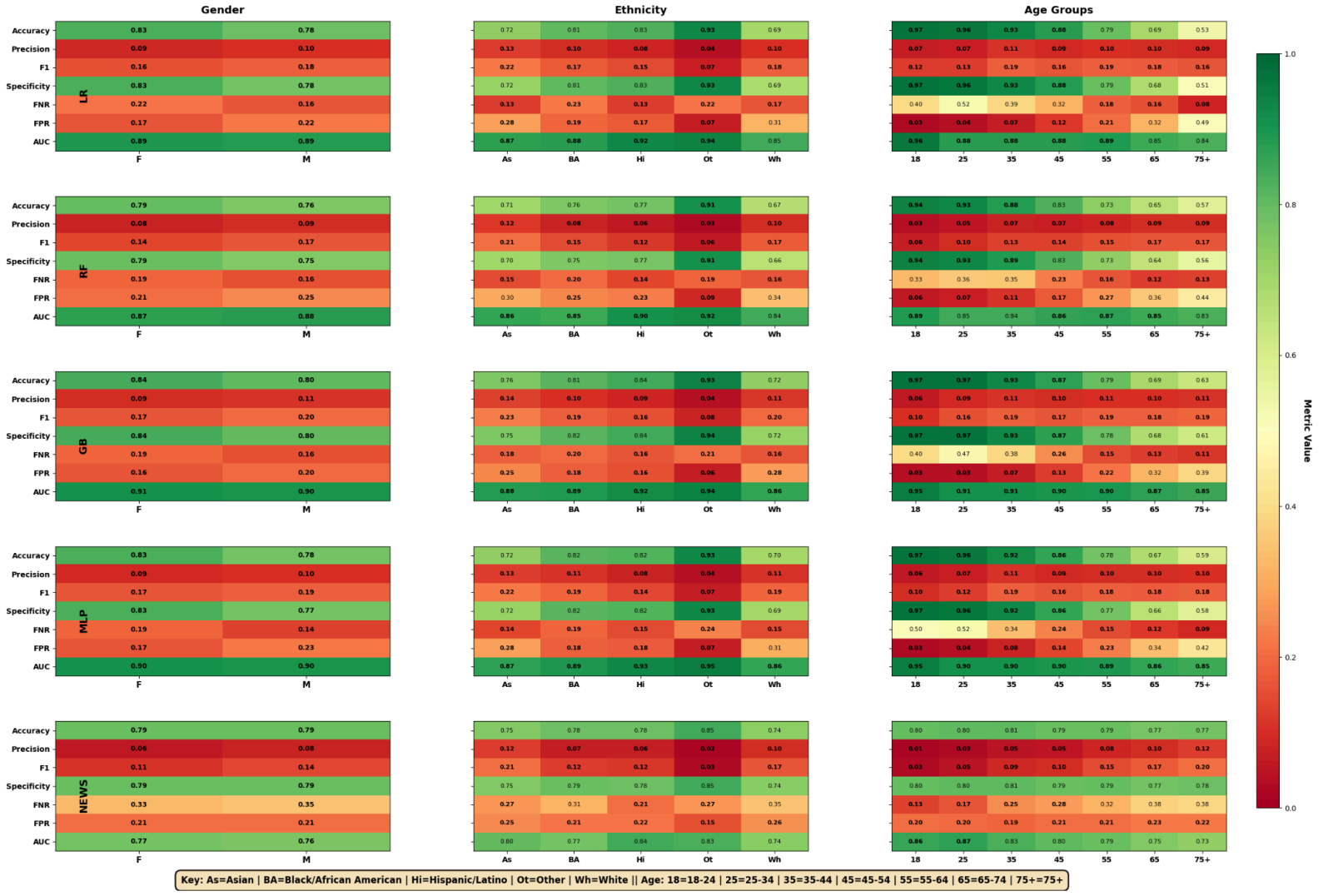
(c) Age-dependent sensitivity patterns showed increasing sensitivity with advancing age and greater variability among younger adults, suggesting that age-related prevalence and data imbalance drive most observed differences rather than model bias.

eFigure 3: BIDMC cohort hospitalization performance metrics across demographic subgroups



These heatmaps display performance metrics (accuracy, precision, specificity, FPR, and AUC) for hospitalization prediction by gender, ethnicity, and age group across models. Green indicates higher performance, and red indicates lower performance. Limited variability was observed between demographic subgroups, with lower specificity and higher FPR among older adults.

eFigure 4: BIDMC cohort sepsis performance metrics across demographic subgroups



These heatmaps show the accuracy, precision, specificity, FPR, and AUC for sepsis prediction across gender, ethnicity, and age groups. Greater variability has been shown compared to hospitalization, with reduced precision and specificity among younger and minority groups. The color intensity reflects relative performance (green = higher, red = lower).

eTable 3: Clinical Outcome Definitions and ICD-9/10 Coding Criteria

Outcomes	Definitions
Hospitalization	(1) Admission to a general ward; (2) Admission to an intensive care unit (ICU)
Critical Outcome	(1) ICU admission; (2) Death in the emergency department (ED); (3) Hospital admission with ICU transfer within 12 hours
Sepsis	(1) ICD-9: 99591, 99592, 67024; (2) ICD-10: A41.x, R65.20–R65.21, A4189; O03.87, O04.87, O03.37; T81.44XA; A40.x
Pulmonary embolism	(1) ICD-9:415.11, 415.19; (2) ICD-10: I26.x
Acute Coronary Syndrome	(1) ICD-9: 411.0–411.1, 411.81, 411.89; (2) ICD-10: I20.x, I21.x, I22.x, I23.x, I24.x; 410.x
Stroke	(1) ICD-9:430–436; 433.01, 434.01, 434.11, V12.54; (2) ICD-10: I60–I66, I63–I64; G45.x; Z86.73
Acute Heart Failure	(1) ICD-9:428.0, 428.21, 428.23, 428.31, 428.33, 428.41, 428.43; (2) ICD-10:I50.21, I50.23, I50.31, I50.33, I50.41, I50.43, I50.811, I509