
INTERSECTIONAL CONSEQUENCES FOR MARGINAL FAIRNESS IN PREDICTION MODELS FOR EMERGENCY ADMISSIONS

Elle Lett^{1,2,3}, Shakiba Shahbandegan^{3,4}, Yuval Barak-Corren⁵,
Andy Fine^{6,7}, William G. La Cava^{3,7,*}

¹Center for Anti-Racism and Community Health, University of Washington School of Public Health, Seattle, WA

²Health Systems and Population Health, University of Washington School of Public Health, WA

³Computational Health Informatics Program, Boston Children's Hospital, Boston, MA

⁴Department of Computer Science and Engineering, Michigan State University, East Lansing, MI

⁵Division of Cardiology, The Children's Hospital of Philadelphia, Philadelphia, PA

⁶Division of Emergency Medicine, Boston Children's Hospital, Boston, MA

⁷Harvard Medical School, Boston, MA

*william.lacava@childrens.harvard.edu

ABSTRACT

1 **Background:** Fair clinical prediction models are crucial for achieving equitable health outcomes.
2 Recently, intersectionality has been applied to develop fairness algorithms that address discrimination
3 among intersections of protected attributes (e.g., Black women rather than Black persons or women
4 separately). Still, the majority of medical AI literature applies marginal de-biasing approaches, which
5 constrain performance across one or many isolated patient attributes. We investigate the extent to
6 which this modeling decision affects model equity and performance in a well-defined use case in
7 emergency medicine.

8 **Methods:** The study focused on predicting emergency room admissions using electronic health record
9 data from two large U.S. hospitals, Beth Israel Deaconess Medical Center (MIMIC-IV-ED, n=160,016)
10 and Boston Children's Hospital (BCH, n=22,222), covering both adult and pediatric populations. In a
11 comprehensive experiment over fairness definitions, modeling methods, we compared the performance
12 of single- and multi-attribute, marginal de-biasing approaches to intersectional de-biasing approaches.

13 **Results:** Intersectional de-biasing produces greater reductions in subgroup calibration error (MIMIC-
14 IV: 21.2%; BCH: 27.2%) than marginal de-biasing (MIMIC-IV: 10.6%; BCH: 22.7%), and also
15 lowers subgroup false negative rates on MIMIC-IV an additional 3.5% relative to marginal de-

Intersectional Consequences for Marginal Fairness

16 biasing. These fairness gains were achieved without a significant decrease in model accuracy between
17 baseline and intersectionally-debiased models (MIMIC-IV: AUROC=0.85±0.00, both models; BCH:
18 AUROC=0.88±0.01 vs 0.87±0.01). Intersectional de-biasing more effectively lowered subgroup
19 calibration error and FNRs in low-prevalence groups in both datasets compared to other de-biasing
20 conditions.

21 **Conclusion:** Intersectional de-biasing better mitigates performance disparities across intersecting
22 groups compared to marginal approaches for emergency admission prediction. These strategies
23 meaningfully reduce group-specific error rates without compromising overall accuracy. These
24 findings highlight the importance of considering interacting aspects of patient identity in model
25 development, and suggest that intersectional de-biasing would be a promising gold standard for
26 ensuring equity in clinical prediction models.

27 **Keywords:** algorithmic fairness; intersectionality; clinical decision-making; emergency department hospital admissions

28 **INTRODUCTION**

29 Emergency departments (EDs) are dynamic environments where patients present with varying acuity, requiring tailored
30 and efficient treatment plans that prioritize achieving desired health outcomes while optimizing clinician workflow
31 and hospital resource utilization. EDs often function as safety-net care for marginalized populations with reduced
32 economic resources or healthcare access, particularly among minoritized ethnoracial groups in the United States¹.
33 These populations also experience the most severe health inequities in disease burden, mortality, and morbidity²⁻⁴.
34 Racialized health inequities also manifest throughout the ED workflow; Black and Hispanic/Latino patients are subject
35 to longer wait times for initial evaluation by a physician in the ED⁵, despite data suggesting that Black and Hispanic
36 individuals account for a disproportionate amount of ED visits and are more likely to be repeat visitors¹. After initial
37 triage, Black patients who are designated for admission also experience longer ED boarding times (stays in the ED
38 before entering an inpatient service)⁶, with such delays associated with adverse health outcomes including intensive
39 care unit (ICU) mortality rates⁷ and ventilator-associated pneumonia⁸.

40 **Machine Learning Models Capacity for Improving Emergency Department Patient Management**

41 A key challenge in ED workflow contributing to wait time inequities is coordinating admissions for patients needing
42 inpatient care, as hospital beds are a limited resource. Bed coordination - the assignment of patients to care teams and
43 beds - can create bottlenecks, increasing ED boarding times and delaying treatment when demand exceeds capacity
44 or allocation is inefficient. Machine learning (ML) models can help accelerate this process by identifying potential
45 admissions early during triage and initial work-up, before the formal decision to admit is made (Fig. 1). Our study
46 builds on previous ED admission prediction models that have shown strong performance in adult⁹ and pediatric¹⁰
47 settings, improving and complementing the assessments of patient disposition made by attending physicians¹¹.

Intersectional Consequences for Marginal Fairness

48 Fairness and Intersectionality

49 Previous ED admission prediction models focused on optimizing overall performance without addressing differences in
 50 subgroup outcomes. Given existing inequities in ED wait and board times, a “fairness-agnostic” model could narrow,
 51 maintain, or even widen disparities between privileged and marginalized groups. Therefore, we develop “fairness-aware”
 52 models that optimize both overall accuracy and equitable performance across groups defined by demographic traits.
 53 Prior work in fair ML has described the common limitation of many fairness approaches to focusing on groups defined
 54 by a single demographic trait such as race, or considering multiple demographic traits in isolation (i.e. race and gender
 55 separately)¹². We refer to these approaches as “marginal”, as they focus on the marginal distribution of one or more
 56 protected attributes while ignoring groups defined by their intersections. Marginal fairness approaches are subject to
 57 “fairness gerrymandering”^{13,14}, wherein models that are “fair” for groups defined by single attributes (i.e. Black people,
 58 or women, separately) still exhibit unfair performance for groups defined by intersections of protected attributes (i.e.
 59 Native American women, or Latino men). We provide a break-down of currently available fair ML algorithms and their
 60 support for intersecting subgroup definitions in Table 1.

Table 1: Properties of a number of algorithms proposed for fair machine learning, along with their properties and support for intersectional fairness definitions. DP: Demographic Parity; FNR: False Negative Rate; FPR: false positive rate. Model-Agnostic indicates that the algorithm supports many common base ML models. The algorithms in bold are the two used in this study.

Stage	Algorithm	Fairness Definition			Intersectional Groups	Model-Agnostic
		DP	FNR/FPR	Calibration		
Pre	Reweighting ¹⁵	✓				✓
	Fair Feature Selection ¹⁶	✓				✓
Train	Adversarial Debiasing ¹⁷		✓			
	Differential Fairness ¹⁸	✓			✓	
	Exponentiated Gradients ¹⁹	✓	✓			✓
	GerryFair ¹³	✓	✓		✓	✓
	FOMO ²⁰	✓	✓	✓	✓	✓
Post	Threshold Optimization ²¹	✓	✓		✓	
	Calibrated Equalized Odds ²²		✓	✓		✓
	MultiCalibration ²³			✓	✓	✓
	MultiAccuracy ²⁴		✓		✓	✓

61 Approaches to mitigate fairness gerrymandering are rooted in intersectionality, a framework established by legal
 62 scholar Kimberlé Crenshaw^{25,26} and sociologist Patricia Hill Collins²⁷, but with origins in 1830s social movements²⁸⁻³⁰.
 63 Intersectionality views systems of oppression such as racism and cis-sexism as co-occurring, emphasizing that analyzing
 64 a single axis of discrimination—such as race—fails to capture the harms experienced by individuals facing multiple
 65 forms of discrimination³¹. Our previous work shows how this framework applies to ML fairness throughout different
 66 stages of the prediction task, from defining to evaluation and updating the task¹².

67 In algorithmic fairness, this framework motivates what we refer to as “intersectional” fairness that constrains model
 68 performance across groups that are defined by the intersections of protected attributes, rather than what we refer to as
 69 “marginal” fairness that is only concerned with the groups defined by the marginal distributions of one or more protected

Intersectional Consequences for Marginal Fairness

70 attributes. Theoretically, intersectional fairness is clearly ideal; in practice it can be difficult to achieve computationally
71 due to scarce data on multiply-marginalized groups.

72 **De-biasing and Evaluating Fairness**

73 Fairness metrics must be selected based on specific context of the implementation environment and adapted to the
74 prediction task¹². Depending on the hospital's patient population, ED traffic, and operating practices, different metrics
75 may be most salient to optimizing care across groups. For example, in an ED with particularly high ethnoracial
76 inequities in boarding wait times, ensuring fair calibration would ensure that specific groups aren't systematically
77 deprioritized or over-prioritized by the algorithm via assigned risk scores. Ensuring low subgroup false negative rates
78 (FNRs), meanwhile, would help ensure that no one group is being falsely discharged at a higher rate. To cover the
79 breadth of potential use case scenarios we focus on two fundamental notions of fairness: *sufficiency*, i.e. patients with
80 the same risk score should experience outcomes at a rate that is independent of group membership; and *separation*, i.e.
81 patients with the same outcomes should receive risk scores that are independent of group membership³². For example,
82 if an ED admission model meets sufficiency, patients with a 90% risk score should have equal admission likelihoods
83 regardless of group membership. Conversely, if the model meets separation, risk scores for admitted patients should not
84 differ by group, meaning false negative rates (FNRs) and false positive rates (FPRs) should be the same across groups.
85 Both of these traits, sufficiency and separation, are important characteristics for fair prediction models, yet cannot be
86 simultaneously satisfied when admission rates differ among groups³³. Hence, we study both notions here by applying
87 two fairness algorithms: one that achieves sufficiency by de-biasing group-level calibration, and one that achieves
88 (a relaxation of) separation by de-biasing group-level FNRs. The first algorithm, multicalibration boosting³⁴, is a
89 post-processing algorithm that constrains the group-level calibration error. The second, fairness-oriented multiobjective
90 optimization (FOMO)²⁰, is a training algorithm we use to constrain group-level FNRs.

91 In our experiments, we evaluate the ED admission prediction task (Fig. 1) across adult and pediatric populations in
92 two Boston-based healthcare centers. With these models, we compare the performance of marginal and intersectional
93 de-biasing approaches with multicalibration and FOMO, specifically with 1) no de-biasing, 2) marginal de-biasing based
94 on single-attributes (ethnoracial group or gender) or multiple attributes concomitantly (ethnoracial group and gender),
95 and 3) intersectional de-biasing based on ethnoracial group and gender. We implement these de-biasing approaches on
96 both logistic regression and random forest base models. The overall goal of the present study is to measure the extent to
97 which optimization of algorithmic fairness on marginal groups transfers to intersectional patient groups, under different
98 definitions of fairness, models, and clinical settings.

Intersectional Consequences for Marginal Fairness

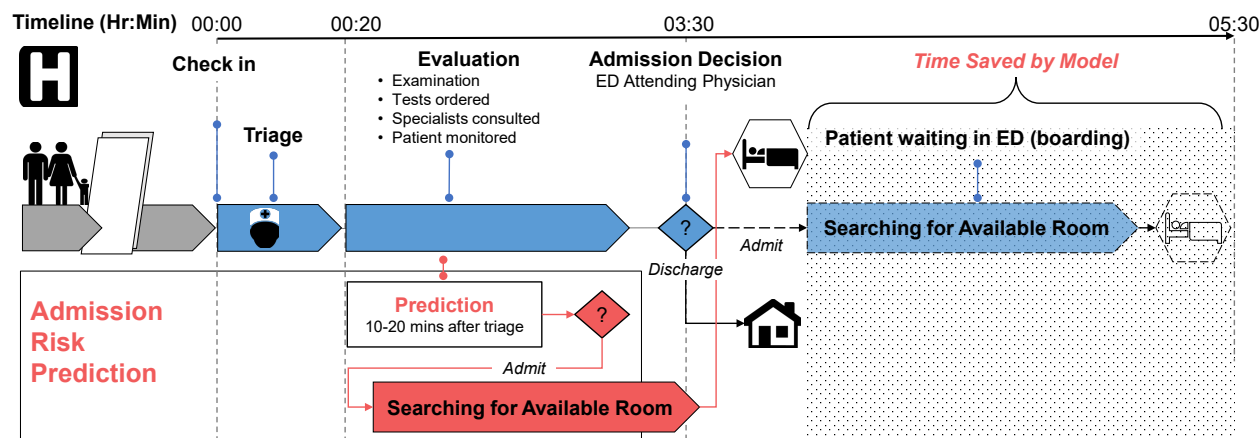


Figure 1: An illustration of the admission prediction task and its utility in the emergency department (ED) during the typical timeline of a patient visit. Normally, patients who will be admitted wait while care coordinators find an available room (known as boarding). Admission prediction algorithms flag high risk patients early in the visit so that the bed coordination can happen before the ED attending physician makes an admission decision for the patient.

99 METHODS

100 Data Curation

101 We base our experiments on the task of inpatient hospital admission prediction for patients visiting the ED. Recently,
102 multiple care centers have sought to develop, validate, and deploy ML models for this task, due to its significant impact
103 on patient flow.^{9-11,35} In our experiments we use data from two EDs that are described in detail in Table 2. The first is
104 from the Medical Information Mart for Intensive Care-IV Emergency Department (MIMIC-IV-ED) database³⁶, a freely
105 available data source on ED visits to Beth Israel Deaconess Medical Center between 2011 and 2019. The second is
106 collected from Boston Children’s Hospital (BCH) ED from 2017 to 2018. After data preprocessing (see Supplement for
107 more details), our analysis consists of 160,016 visits by 90,005 unique patients in the MIMIC-IV cohort and 22,222
108 visits by 17,938 unique patients in the BCH cohort.

109 Model Development

110 In both cohorts, we train a model to predict admission to an in-patient service among patients whose final disposition
111 has yet to be decided. We use data collected during check-in (e.g. chief complaint), triage (e.g. vitals), patient clinical
112 history (e.g. number of previous admissions) and demographic data. In the BCH cohort, we include additionally
113 available data collected during the first 60 minutes of a patient’s stay, including lab orders and medications. Table 3 lists
114 the full set of features used in both cohorts.

115 We test two baseline ML models: tree ensembles implemented in XGBoost (main tables and figures) and penalized
116 logistic regression models (see Supplement). The hyperparameters of these models were tuned via halving grid search.

Intersectional Consequences for Marginal Fairness

Table 2: Patient visit characteristics for the MIMIC-IV and BCH data. AI/AN: American Indian / Alaskan Native; AA: African American; NHPI: Native Hawaiian Pacific Islander; (N)HL: (Not) Hispanic/Latino; F: Female; M: Male.

MIMIC-IV ED, 2011 - 2019		Overall	No	Yes
Visits, n		160016	112733	47283
Patients, n		90005	51306	38699
Age at Visit in Years, mean (SD)		53.0 (19.3)	49.4 (18.5)	61.6 (18.4)
ED Triage Acuity Group, n (%)	1	8720 (5.5)	2530 (2.3)	6190 (13.7)
	2	47570 (30.2)	24743 (22.1)	22827 (50.4)
	3	90948 (57.7)	74755 (66.6)	16193 (35.7)
	4	9922 (6.3)	9809 (8.7)	113 (0.2)
	5	382 (0.2)	376 (0.3)	6 (0.0)
Ethnoracial Group, n (%)	AI/AN	427 (0.3)	275 (0.2)	152 (0.3)
	ASIAN	5979 (3.7)	3904 (3.5)	2075 (4.4)
	BLACK/AA	41944 (26.2)	36217 (32.1)	5727 (12.1)
	HL	16057 (10.0)	13826 (12.3)	2231 (4.7)
	WHITE	95609 (59.7)	58511 (51.9)	37098 (78.5)
Gender, n (%)	F	91774 (57.4)	68327 (60.6)	23447 (49.6)
	M	68242 (42.6)	44406 (39.4)	23836 (50.4)
BCH ED, 2017-2018		Overall	No	Yes
Visits, n		22222	18605	3617
Patients, n		17938	15533	3069
Age at Visit in Years, mean (SD)		8.2 (6.8)	8.0 (6.6)	9.5 (7.6)
ED Triage Acuity, n (%)	1	130 (0.6)	41 (0.2)	89 (2.5)
	2	4935 (22.2)	2905 (15.6)	2030 (56.1)
	3	10017 (45.1)	8553 (46.0)	1464 (40.5)
	4	6177 (27.8)	6143 (33.0)	34 (0.9)
	5	963 (4.3)	963 (5.2)	
Race, n (%)	AI/AN	28 (0.1)	17 (0.1)	11 (0.3)
	ASIAN	888 (4.0)	753 (4.0)	135 (3.7)
	BLACK/AA	4383 (19.7)	3936 (21.2)	447 (12.4)
	NHPI	35 (0.2)	29 (0.2)	6 (0.2)
	Other	8507 (38.3)	7509 (40.4)	998 (27.6)
	WHITE	8381 (37.7)	6361 (34.2)	2020 (55.8)
Ethnicity, n (%)	HL	6799 (30.6)	6082 (32.7)	717 (19.8)
	NHL	15423 (69.4)	12523 (67.3)	2900 (80.2)
Gender, n (%)	F	10639 (47.9)	8962 (48.2)	1677 (46.4)
	M	11583 (52.1)	9643 (51.8)	1940 (53.6)

117 **Fairness Approaches**

118 For all models, we experiment with multicalibration post-processing to improve subgroup calibration performance and
 119 fairness-oriented multiobjective optimization (FOMO) to improve subgroup FNRs.

120 **Multicalibration Postprocessing** Multicalibration post-processing^{23,34} allows for flexible specification of groups
 121 for marginal and intersectional fairness models. Briefly, assume we have sample data (\mathbf{x}_i, y_i) , where \mathbf{x}_i is a vector of
 122 features and y is a binary outcome for individual i , drawn from joint distribution \mathcal{D} . Let C represent a collection of
 123 subsets specified by protected attributes in \mathbf{x} (i.e., subgroups). An α -multicalibrated model fulfills the constraint that
 124 among all subsets in C and binned prediction intervals, the absolute difference between the expected outcome and
 125 expected prediction is at most α . Hébert-Johnson²³ showed that multicalibration is achieved without a fairness-utility
 126 tradeoff such that multicalibrated models have at least the same predictive power as the base model, which is ideal
 127 for our prediction task. The multicalibration algorithm updates model predictions until all groups defined by binned

Intersectional Consequences for Marginal Fairness

Table 3: Features used for Emergency admission prediction in the MIMIC-IV and BCH cohorts. The BCH data includes a larger set of predictors (n = 155, BCH; n = 60, MIMIC-IV) including indicators of laboratory tests and a larger set of reported symptoms beyond chief complaint. HR: heart rate; RR: respiratory rate; SBP: systolic blood pressure; DBP: diastolic blood pressure; BMI: body mass index.

Description	Features
MIMIC-IV	
Vitals	temperature, HR, RR, oxygen saturation, SBP, DBP
Triage Acuity	Emergency Severity Index ³⁷
Check-in Data	chief complaint, self-reported pain score
Health Record Data	no. previous visits, no. previous admissions
Demographic Data	ethnoracial group, gender, age, marital status, insurance, primary language
BCH	
Demographic Data	Gender, Race, Ethnicity, Age
Check-in Data	ED Day Of Week, ED Checkin Month, ED Checkin Year, ED Arrival Mode, Weekend, Miles Traveled, Patient State of Residence
Triage Data	Emergency Severity Index, ED Teams, ED Room Type, chief complaint, pain count, pain max, pain increase, CHEWS ³⁸ count, CHEWS max, CHEWS increase
Medications	acetaminophen, albuterol, dexamethasone, epinephrine/lidocaine/tetracaine topical, ibuprofen, ipratropium, ondansetron, Sodium Chloride 0.9%, gastro count, gastro max, gastro increase, time till first med, drugs count, medication route (inhaled, intravenous, nebulized, oral, topical)
Vitals	HR count, HR min, HR max, HR sd, HR mean, RR count, RR min, RR max, RR sd, RR mean, temp low, temp normal, temp high low, temp not taken
Lab Test Orders	labs count, time till first lab, Blood Culture Routine, Aerobic, Blood Culture, Aerobic and Anaerobic, Blood Gas, Venous, C-Reactive Protein, Calcium, Plasma, Chemistry Panel, Chemistry Extended Panel, Complete Blood Count with Differential, Differential, Automated, Drug Screen, Urine (drugs of abuse), Erythrocyte Sedimentation Rate, Lipase, Plasma, Liver Function Tests, Urinalysis, Dipstick, Urine Culture, time till first test, tests count, Chest X-ray
Symptoms	Abdominal pain, Attention deficit disorder with hyperactivity, Allergic rhinitis, Anxiety, Asthma, Atopic eczema, Autistic disorder, Chronic constipation, Constipation, Cough, Dental caries, Depression, Dermatitis, Developmental delay, Dysphagia, Epilepsy, Eustachian tube dysfunction, Feeding problem, Fever, Food allergy, Gastroesophageal reflux, Global developmental delay, Headaches, Hypotonia, Obesity, Obstructive sleep apnea, Oropharyngeal dysphagia, Other Problem, Patent ductus arteriosus, Patent foramen ovale, Pediatric BMI greater than or equal to 95th percentile, Prematurity, Recurrent acute otitis media, Seizure, Snoring, Speech delay, Tonsillar and adenoid hypertrophy, Tonsils hypertrophy, Vitamin D deficiency, Vomiting
Clinical History	problems count, prior visits, prior admissions, admission ratio

128 prediction intervals within collections in C with group probability greater than γ satisfy the calibration error constraint
 129 α . For our main results we used $\alpha=0.01$ (constrain calibration error to 0.01) and $\gamma=0.001$ (consider groups with 0.1%
 130 or higher probability). The supplement contains a sensitivity analysis of these hyperparameters.

131 **Fairness-Oriented Multiobjective Optimization** Achieving different notions of fairness in machine learning
 132 involves balancing the tradeoff between error and fairness, where increased fairness may lead to higher error rates, and
 133 vice versa. Traditionally, fair machine learning methods treat this as a single objective problem, introducing a parameter
 134 to weigh error against fairness. FOMO optimizes this tradeoff through multi-objective optimization, treating error and
 135 fairness as separate objectives²⁰.

Intersectional Consequences for Marginal Fairness

Table 4: Experimental setup for assessing algorithmic fairness under intersectional and marginal fairness scenarios.

Variable	Settings
Group Construction Scenarios	Base: no fairness optimization Gender, Race, Ethnicity, Ethnoracial: protect single attribute Marginal: marginally protect all sensitive attributes Intersectional: protect intersections of all sensitive attributes
Fairness Task (Algorithm)	Fair Calibration (Multicalibration Boosting) Fair False Negative Rate (Fairness-Oriented Multiobjective Optimization)
Base Classifier	Penalized Logistic Regression (penalty: $\{\ell_1, \ell_2\}$, C: $\{0.01 \dots 10\}$) Random Forest (n_estimators: 100, max_depth: 4)
Realizations	100 trials of independent 50/50 train/test splits

136 We use FOMO to jointly optimize the overall balanced accuracy of the models while minimizing the maximum FNRs
137 among intersectional subgroups. This fairness definition has two motivations: first, it assumes that false discharges from
138 the emergency room have the potential to cause more harm to a patient than a false admission. Second, unlike fairness
139 metrics that optimize for FNR parity among groups, which can be achieved e.g. by making the model worse for some
140 subgroups where it performs well, this metric focuses solely on improving the worst-case performance among patient
141 subgroups. Minimizing subgroup FNRs must be balanced with minimizing overall FNRs and overall FPRs, which cause
142 distributed harm to waiting patients due to overcrowding; hence, we jointly maximize for overall balanced accuracy.

143 **Protected Attributes and Intersectionality** The experiment in this study focuses on three protected attributes: race,
144 ethnicity, and gender (in the MIMIC-IV cohort, race and ethnicity are reported as a combined ethnoracial variable). We
145 observe stark differences in admission rates by intersections of race, ethnicity, and gender (See Table S1), suggesting
146 the importance of a performance-based fairness constraint (e.g., calibration or error rates) as opposed to demographic
147 parity, which would cause substantial deviations in subgroup admission rates.

148 **Statistical Tests** All reported p -values are the result of two-sided Mann-Whitney-Wilcoxon tests with Holm-
149 Bonferroni correction.

150 RESULTS

151 Fairness without accuracy tradeoffs

152 The prevailing understanding of fairness as derived from the notions of equalized odds and demographic parity is that
153 they require trade-offs with overall accuracy²². This trade-off is theoretically well-established, yet recent work has shown
154 that in practice, such trade-offs may be negligible³⁹. Our findings were consistent with the latter: across both data sets
155 (MIMIC-IV and BCH), and both fairness targets (calibration and FNR), de-biasing on gender, ethnoracial identity, both
156 concomitantly (marginally), and across intersectional groups, had nearly identical overall classification performance
157 (mean AUROC within ± 0.01 ; Fig. 2). When tasked with balancing FNRs on the BCH cohort, intersectionally de-biased
158 models exhibited slightly lower area under the precision recall curve (base scenario AUPRC: 0.67 ± 0.01 ; intersectional

Intersectional Consequences for Marginal Fairness

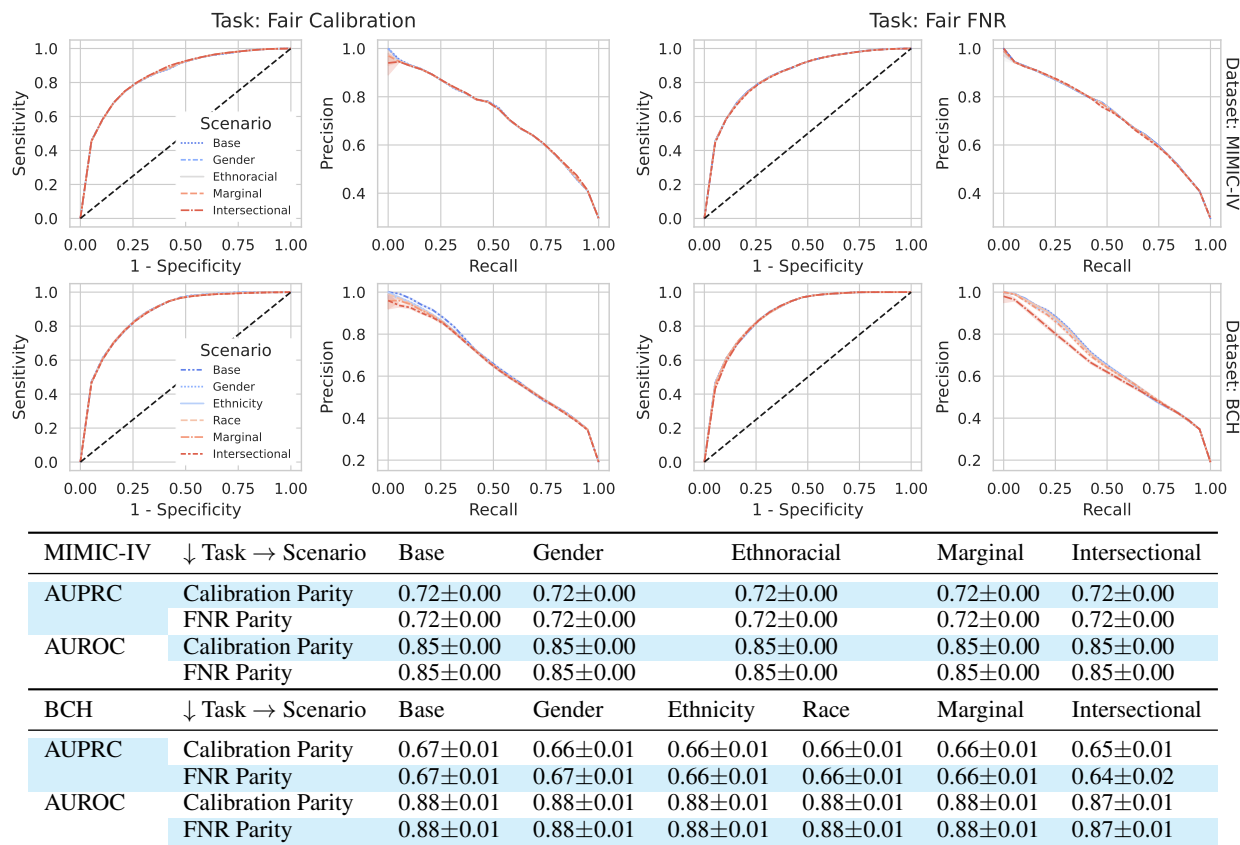


Figure 2: De-biased models perform as well as baseline models. (a) Receiver operating characteristic (ROC) curves and precision-recall curves for the prediction models on data from MIMIC-IV (top row) and BCH (bottom row). The left and right columns of subplots compare debiasing scenarios for fair calibration and fair false negative rates (FNR), respectively. (b) The mean (\pm standard deviation) area under the ROC curve (AUROC) and precision-recall curve (AUPRC) of prediction models by dataset, fairness task, and modeling scenario, corresponding to the curves above. In general, the fairness-aware models perform very similarly to the baseline models.

159 scenario AUPRC: 0.64 ± 0.02) due to lower precision in operating regimes with low recall/sensitivity, but nearly identical
 160 precision for model operating points with moderate to high sensitivity/recall that are desirable in this use case (Fig. 2,
 161 bottom right curve).

162 Fairness gains with intersectional de-biasing

163 To compare fairness-unaware, marginal single-attribute, marginal multi-attribute, and intersectional de-biasing ap-
 164 proaches at a high level, we compare the expected calibration error (ECE) and FNRs for the intersectional groups
 165 (ethnoracial group and gender cross-strata) across de-biasing conditions in Fig. 3. We observe that multi-attribute,
 166 marginal fairness de-biasing reduces ECE among intersectional groups on MIMIC-IV and BCH by 10.6% and 22.7%,
 167 whereas the fully intersectional approach reduces ECE by 21.2% and 27.2%, respectively (Fig. 3 left). In a similar vein,
 168 intersectional fairness de-biasing results in significantly lower FNRs among intersectional groups in the cohort compared
 169 to baseline (11% reduction, MIMIC-IV, $p < 1e-16$; 6.4% reduction, BCH, $p < 3e-6$). On MIMIC-IV, intersectional
 170 de-biasing reduces intersectional FNRs by an additional 3.5% compared to marginal fairness de-biasing ($p = 1e-5$). We

Intersectional Consequences for Marginal Fairness

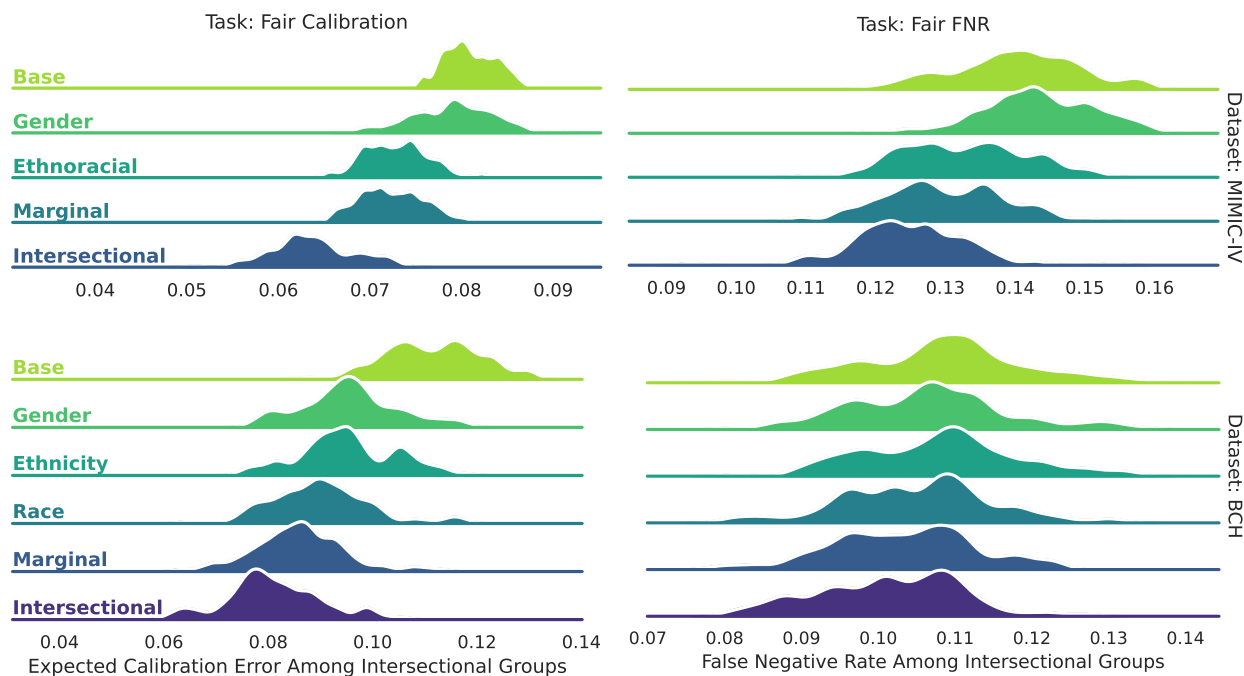


Figure 3: Intersectionally de-biased models improve fairness for intersectional groups beyond marginally de-biased models. Fairness measures under different de-biasing scenarios for MIMIC-IV (top) and BCH (bottom). Left plots report the expected calibration error (ECE) among intersectional groups when trying to ensure within-group calibration. Right plots report false negative rate among intersectional groups when optimizing for equal group-wise false negative rates. The scenarios (Base, Intersectional, etc.) are detailed in Table 4.

171 observe across the experimental results that de-biasing on ethnoracial group produces a larger singular reduction in
172 error rates among intersectional groups than de-biasing on gender alone, but that de-biasing using the intersectional
173 combination of ethnoracial group and gender yields better performance than considering either attribute alone, or
174 additively.

175 Intersectional de-biasing improves fairness for small and large groups

176 It is challenging to build models that both perform well on marginalized groups and minimize overfitting. This
177 is particularly concerning when evaluating intersectional fairness approaches, as with each additional attribute to
178 consider, the number of groups grows factorially while group size decreases. Therefore, we evaluate how the benefits of
179 intersectional de-biasing approaches are distributed across the groups of varying prevalence. In the MIMIC-IV ED,
180 intersectional de-biasing approaches minimize both the group-specific ECE (Fig. 4, top left) and the FNRs for the
181 lowest prevalence group (AIAN, M, prevalence=0.11%) and highest prevalence groups (White, F, prevalence=31.36%),
182 in contrast to no de-biasing, single-attribute de-biasing, and multi-attribute, marginal de-biasing. For intermediate
183 prevalence groups (0.16% to 28.39%), intersectional de-biasing either outperformed or equalled all other de-biasing
184 conditions in the MIMIC-IV data. Similar performance was noted in the pediatric setting across both fairness
185 optimization targets (Fig. 4, bottom left and right).

Intersectional Consequences for Marginal Fairness

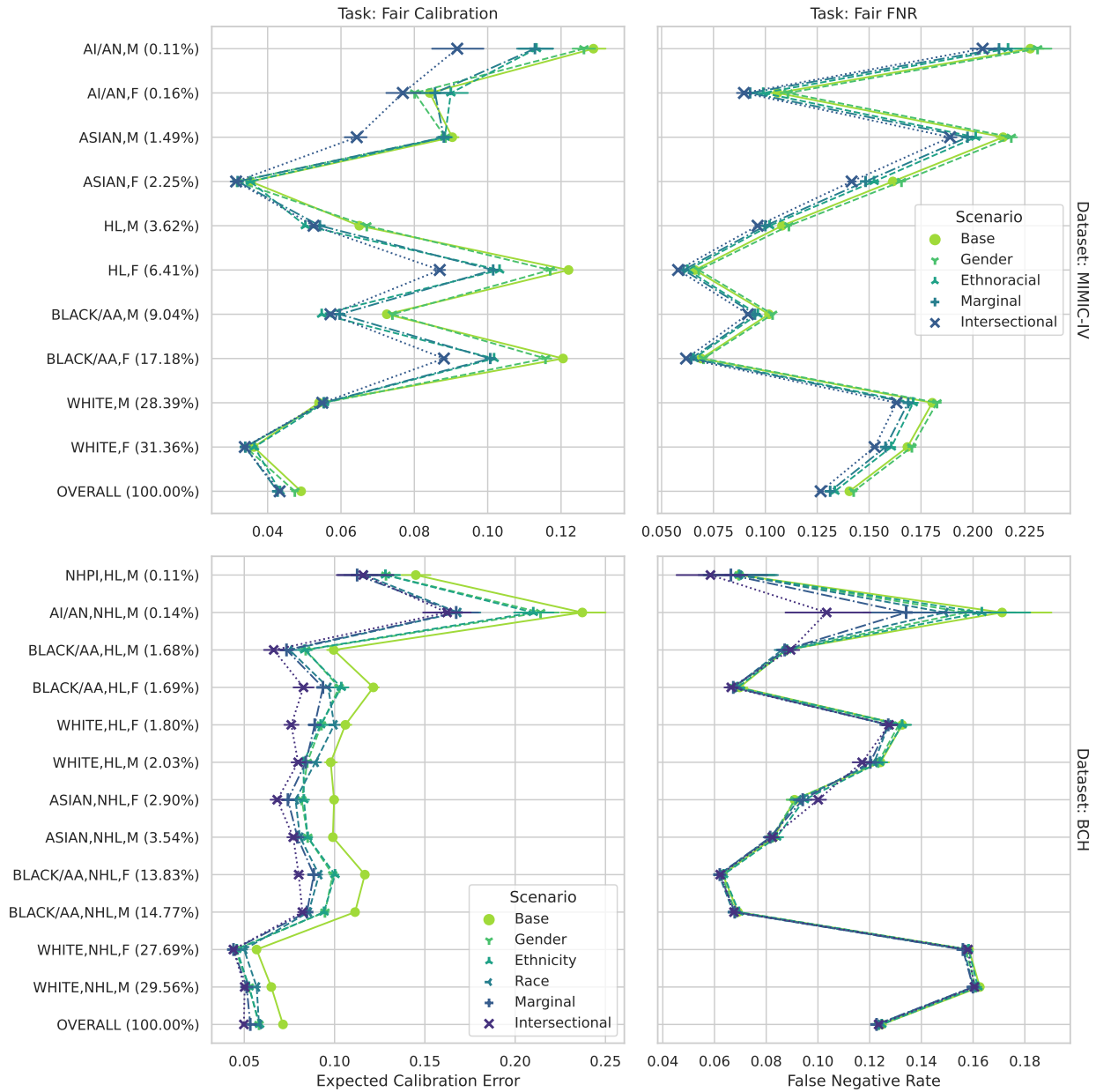


Figure 4: Model performance on each intersectional position (y-axis) according to dataset (top: MIMIC-IV, bottom: BCH), fairness consideration (left: expected calibration error, right: false negative rate), demarcated by scenario. Points indicate bootstrap-estimated median performance over trials and bars indicate the 95% confidence interval. AI/AN: American Indian / Alaskan Native; AA: African American; NHPI: Native Hawaiian Pacific Islander; (N)HL: (Not) Hispanic/Latino; F: Female; M: Male.

Intersectional Consequences for Marginal Fairness

186 **DISCUSSION**

187 **Exclusions and Limitations**

188 To date, most model bias is identified post-deployment⁴⁰, with few clinical prediction models incorporating fairness
189 notions in the development process. This study is among the first to implement an intersectional de-biasing approach
190 for clinical prediction models and demonstrate that 1) it can significantly improve the performance of a model on
191 subgroups versus the more common, marginal approaches; and 2) it can reduce unfairness with minor changes in
192 overall performance. In MIMIC-IV, intersectionally de-biased ML models exhibit a 27% reduction in subgroup ECE or
193 11% reduction in subgroup FNR with no change in AUROC or AUPRC; in BCH, models exhibit a 27% reduction in
194 subgroup ECE with no reduction in AUROC or AUPRC, and a 6.4% reduction in subgroup FNR for no reduction in
195 AUROC and a 3% reduction in mean AUPRC (concentrated at low sensitivity model thresholds).

196 A challenge of intersectional approaches using demographic traits is that as more protected attributes are added, group
197 sizes shrink. We limited our analysis to three attributes: race, ethnicity, and sex, and only considered intersectional
198 groups representing at least 0.1% of the population. While multicalibration handles small group sizes with a threshold,
199 other fairness methods use a prior probability for group outcomes. We tested both approaches in FOMO and found no
200 significant effect on results. Future studies could explore additional attributes and larger datasets to examine the limits
201 of fairness gains for smaller intersectional groups.

202 Our results are limited to one clinically relevant prediction problem, but it is a type of resource allocation problem that
203 is widely found in clinical settings. Further work should examine the extent to which our observations generalize to
204 other settings of interest, which may additionally have their own appropriate measures of fairness.

205 We do not attempt to answer whether subgroup calibration or subgroup FNRs are a more important fairness consideration
206 for this task; instead, we attempt to measure the importance of intersectional de-biasing of multiple scenarios. Calibration
207 is important for interpreting risk scores and doing risk stratification. FNRs are important for interpreting the risk of
208 missed interventions (in this case, hospital admissions). It is well known that FNRs, FPRs, and calibration cannot be
209 simultaneously equal when subgroups exhibit different prevalence of the outcome³³. Future studies could consider
210 two-way optimizations of these fairness metrics which are not covered here. Similarly, future prospective studies
211 depend on extended engagement with community collaborators to define which metrics are more important in clinical
212 decision support.

213 **Data Availability**

214 MIMIC-IV-ED is available from physionet.org/mimic-iv-ed. The full preprocessing code for the MIMIC-IV admissions
215 dataset is available from the repository github.com/cavalab/mimic-iv-admissions. The BCH pediatric dataset is
216 not publicly available under the terms of the BCH Institutional Review Board. Interested readers may contact the
217 corresponding author for additional details.

Intersectional Consequences for Marginal Fairness

218 **Code Availability**

219 The code for reproducing the experiments is available from github.com/cavalab/marginal-intersectional.

220 **Author Contributions**

221 EL and WGL conceived the study and designed the experiment. EL wrote the initial manuscript and contributed to
222 code and experimental evaluation. SS and WGL wrote methods and experimental code and ran the experiments. WGL
223 created the tables and figures and contributed to writing the manuscript. YBC developed and curated the BCH dataset.
224 YBC, AF and BYR provided feedback and guidance on the study design, clinical use case, and manuscript.

225 **Additional Information**

226 Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed
227 to william.lacava@childrens.harvard.edu.

228 **Acknowledgments**

229 This work was partially supported by National Institutes of Health grant no. R01LM014300 from the National Library
230 of Medicine.

231 **References**

- 232 [1] Layla Parast et al. “Racial/Ethnic Differences in Emergency Department Utilization and Experience”. In: *Journal*
233 *of General Internal Medicine* 37.1 (Jan. 2022), pp. 49–56. ISSN: 0884-8734, 1525-1497. DOI: [10.1007/s11606-](https://doi.org/10.1007/s11606-021-06738-0)
234 [021-06738-0](https://doi.org/10.1007/s11606-021-06738-0). URL: <https://link.springer.com/10.1007/s11606-021-06738-0> (visited on
235 06/17/2024) (cit. on p. 2).
- 236 [2] Farhad Islami et al. “American Cancer Society’s Report on the Status of Cancer Disparities in the United States,
237 2021”. In: *CA: A Cancer Journal for Clinicians* 72.2 (Mar. 2022), pp. 112–143. ISSN: 0007-9235, 1542-4863.
238 DOI: [10.3322/caac.21703](https://doi.org/10.3322/caac.21703). URL: [https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/](https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21703)
239 [caac.21703](https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21703) (visited on 06/17/2024) (cit. on p. 2).
- 240 [3] Jiang He et al. “Trends in Cardiovascular Risk Factors in US Adults by Race and Ethnicity and Socioeconomic
241 Status, 1999–2018”. In: *Jama* 326.13 (2021), pp. 1286–1298. URL: [https://jamanetwork.com/journals/](https://jamanetwork.com/journals/jama/article-abstract/2784659)
242 [jama/article-abstract/2784659](https://jamanetwork.com/journals/jama/article-abstract/2784659) (visited on 06/17/2024) (cit. on p. 2).

Intersectional Consequences for Marginal Fairness

- 243 [4] Mursal A. Mohamud et al. “20-Year Trends in Multimorbidity by Race/Ethnicity among Hospitalized Patient
244 Populations in the United States”. In: *International Journal for Equity in Health* 22.1 (July 24, 2023), p. 137.
245 ISSN: 1475-9276. DOI: [10.1186/s12939-023-01950-2](https://doi.org/10.1186/s12939-023-01950-2). URL: [https://equityhealthj.biomedcentral.
246 com/articles/10.1186/s12939-023-01950-2](https://equityhealthj.biomedcentral.com/articles/10.1186/s12939-023-01950-2) (visited on 06/17/2024) (cit. on p. 2).
- 247 [5] Frederick Q. Lu, Amresh D. Hanchate, and Michael K. Paasche-Orlow. “Racial/Ethnic Disparities in Emer-
248 gency Department Wait Times in the United States, 2013–2017”. In: *The American Journal of Emergency
249 Medicine* 47 (2021), pp. 138–144. URL: [https://www.sciencedirect.com/science/article/pii/
250 S0735675721002369](https://www.sciencedirect.com/science/article/pii/S0735675721002369) (visited on 06/17/2024) (cit. on p. 2).
- 251 [6] Jesse M. Pines, A. Russell Localio, and Judd E. Hollander. “Racial Disparities in Emergency Department
252 Length of Stay for Admitted Patients in the United States”. In: *Academic Emergency Medicine* 16.5 (May
253 2009), pp. 403–410. ISSN: 1069-6563, 1553-2712. DOI: [10.1111/j.1553-2712.2009.00381.x](https://doi.org/10.1111/j.1553-2712.2009.00381.x). URL:
254 <https://onlinelibrary.wiley.com/doi/10.1111/j.1553-2712.2009.00381.x> (visited on
255 06/17/2024) (cit. on p. 2).
- 256 [7] Donald B. Chalfin et al. “Impact of Delayed Transfer of Critically Ill Patients from the Emergency Depart-
257 ment to the Intensive Care Unit”. In: *Critical care medicine* 35.6 (2007), pp. 1477–1483. URL: [https://
258 //journals.lww.com/ccmjournal/fulltext/2007/06000/Data,_data_everywhere.
259 00004.aspx?casa_token=4dfPLn27crEAAAAA:9dFwJP23HIR95h3VT_d8gke-fuM9SeDC6Nnq2hd_
260 Hf0Z3zEG1L7MpoHHTVHcZAXPGPSY_FPrsd0VtTJfUVXo_1M](https://journals.lww.com/ccmjournal/fulltext/2007/06000/Data,_data_everywhere.00004.aspx?casa_token=4dfPLn27crEAAAAA:9dFwJP23HIR95h3VT_d8gke-fuM9SeDC6Nnq2hd_Hf0Z3zEG1L7MpoHHTVHcZAXPGPSY_FPrsd0VtTJfUVXo_1M) (visited on 06/17/2024) (cit. on p. 2).
- 261 [8] Brendan G. Carr et al. “Emergency Department Length of Stay: A Major Risk Factor for Pneumonia in Intubated
262 Blunt Trauma Patients”. In: *Journal of Trauma and Acute Care Surgery* 63.1 (2007), pp. 9–12. URL: [https://
263 //journals.lww.com/jtrauma/fulltext/2007/07000/Biomechanical_Considerations_in_Plate.
264 2.aspx](https://journals.lww.com/jtrauma/fulltext/2007/07000/Biomechanical_Considerations_in_Plate.2.aspx) (visited on 06/17/2024) (cit. on p. 2).
- 265 [9] Yuval Barak-Corren, Shlomo Hanan Israelit, and Ben Y Reis. “Progressive Prediction of Hospitalisation in the
266 Emergency Department: Uncovering Hidden Patterns to Improve Patient Flow”. In: *Emergency Medicine Journal*
267 34.5 (May 2017), pp. 308–314. ISSN: 1472-0205, 1472-0213. DOI: [10.1136/emered-2014-203819](https://doi.org/10.1136/emered-2014-203819). URL:
268 <https://emj.bmj.com/lookup/doi/10.1136/emered-2014-203819> (visited on 12/30/2021) (cit. on
269 pp. 2, 5).
- 270 [10] Yuval Barak-Corren, Andrew M. Fine, and Ben Y. Reis. “Early Prediction Model of Patient Hospitalization From
271 the Pediatric Emergency Department”. In: *Pediatrics* 139.5 (May 2017). ISSN: 1098-4275. DOI: [10.1542/peds.
272 2016-2785](https://doi.org/10.1542/peds.2016-2785). pmid: [28557729](https://pubmed.ncbi.nlm.nih.gov/28557729/) (cit. on pp. 2, 5).
- 273 [11] Yuval Barak-Corren et al. “Prediction of Patient Disposition: Comparison of Computer and Human Approaches
274 and a Proposed Synthesis”. In: *Journal of the American Medical Informatics Association* 28.8 (July 30, 2021),
275 pp. 1736–1745. ISSN: 1527-974X. DOI: [10.1093/jamia/ocab076](https://doi.org/10.1093/jamia/ocab076). URL: [https://academic.oup.com/
276 jamia/article/28/8/1736/6278435](https://academic.oup.com/jamia/article/28/8/1736/6278435) (visited on 12/10/2021) (cit. on pp. 2, 5).

Intersectional Consequences for Marginal Fairness

- 277 [12] Elle Lett and William G. La Cava. “Translating Intersectionality to Fair Machine Learning in Health Sciences”. In:
278 *Nature Machine Intelligence* (Apr. 28, 2023), pp. 1–4. ISSN: 2522-5839. DOI: [10.1038/s42256-023-00651-3](https://doi.org/10.1038/s42256-023-00651-3).
279 URL: <https://www.nature.com/articles/s42256-023-00651-3> (visited on 04/28/2023) (cit. on pp. 3,
280 4).
- 281 [13] Michael Kearns et al. “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness”. In:
282 *arXiv:1711.05144 [cs]* (Dec. 2018). arXiv: [1711.05144 \[cs\]](https://arxiv.org/abs/1711.05144). (Visited on 10/06/2020) (cit. on pp. 3, 22).
- 283 [14] Michael Kearns et al. “An Empirical Study of Rich Subgroup Fairness for Machine Learning”. Aug. 24, 2018.
284 arXiv: [1808.08166 \[cs, stat\]](https://arxiv.org/abs/1808.08166). URL: <http://arxiv.org/abs/1808.08166> (visited on 03/22/2019)
285 (cit. on p. 3).
- 286 [15] Faisal Kamiran and Toon Calders. “Data Preprocessing Techniques for Classification without Discrimination”.
287 In: *Knowledge and Information Systems* 33.1 (Oct. 2012), pp. 1–33. ISSN: 0219-3116. DOI: [10.1007/s10115-](https://doi.org/10.1007/s10115-011-0463-8)
288 [011-0463-8](https://doi.org/10.1007/s10115-011-0463-8). (Visited on 07/15/2020) (cit. on p. 3).
- 289 [16] Ayaz Ur Rehman, Anas Nadeem, and Muhammad Zubair Malik. *Fair Feature Subset Selection Using Multiob-*
290 *jective Genetic Algorithm*. Apr. 2022. arXiv: [2205.01512 \[cs\]](https://arxiv.org/abs/2205.01512). (Visited on 02/07/2023) (cit. on p. 3).
- 291 [17] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. “Mitigating Unwanted Biases with Adversarial
292 Learning”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’18. New York,
293 NY, USA: Association for Computing Machinery, Dec. 2018, pp. 335–340. ISBN: 978-1-4503-6012-8. DOI:
294 [10.1145/3278721.3278779](https://doi.org/10.1145/3278721.3278779). (Visited on 01/17/2021) (cit. on p. 3).
- 295 [18] Kamrun Naher Keya et al. “Equitable Allocation of Healthcare Resources with Fair Survival Models”. In:
296 *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM, 2021, pp. 190–198
297 (cit. on p. 3).
- 298 [19] Alekh Agarwal et al. “A Reductions Approach to Fair Classification”. In: *International Conference on Machine*
299 *Learning*. July 2018, pp. 60–69. (Visited on 12/05/2019) (cit. on p. 3).
- 300 [20] William G. La Cava. “Optimizing Fairness Tradeoffs in Machine Learning with Multiobjective Meta-Models”.
301 In: *Proceedings of the 2023 Genetic and Evolutionary Computation Conference (GECCO)*. ACM, Apr. 2023.
302 DOI: [10.1145/3583131.3590487](https://doi.org/10.1145/3583131.3590487). arXiv: [2304.12190 \[cs\]](https://arxiv.org/abs/2304.12190). (Visited on 04/28/2023) (cit. on pp. 3, 4, 7).
- 303 [21] Moritz Hardt et al. “Equality of Opportunity in Supervised Learning”. In: *Advances in Neural Information*
304 *Processing Systems* 29. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 3315–3323. URL: [http:](http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf)
305 [//papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf](http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf)
306 (visited on 07/15/2020) (cit. on p. 3).
- 307 [22] Geoff Pleiss et al. “On Fairness and Calibration”. In: *Advances in Neural Information Processing Systems* 30.
308 Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 5680–5689. (Visited on 07/15/2020) (cit. on pp. 3, 8).
- 309 [23] Ursula Hebert-Johnson et al. “Multicalibration: Calibration for the (Computationally-Identifiable) Masses”. In:
310 *Proceedings of the 35th International Conference on Machine Learning*. PMLR, July 2018, pp. 1939–1948.
311 (Visited on 11/09/2021) (cit. on pp. 3, 6).

Intersectional Consequences for Marginal Fairness

- 312 [24] Michael P. Kim, Amirata Ghorbani, and James Zou. “Multiaccuracy: Black-box Post-Processing for Fairness in
313 Classification”. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 247–254
314 (cit. on p. 3).
- 315 [25] Kimberle Crenshaw. “Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidis-
316 crimination Doctrine, Feminist Theory and Antiracist Politics”. In: *University of Chicago Legal Forum* 1989.1
317 (1989), p. 31 (cit. on p. 3).
- 318 [26] Kimberlé Williams Crenshaw. “Mapping the Margins: Intersectionality, Identity Politics, and Violence against
319 Women of Color”. In: *The Public Nature of Private Violence*. Routledge, 2013, pp. 93–118. URL: [https://api.taylorfrancis.com/content/chapters/edit/download?identifierName=doi&identifierValue=](https://api.taylorfrancis.com/content/chapters/edit/download?identifierName=doi&identifierValue=10.4324/9780203060902-6&type=chapterpdf)
320 [10.4324/9780203060902-6&type=chapterpdf](https://api.taylorfrancis.com/content/chapters/edit/download?identifierName=doi&identifierValue=10.4324/9780203060902-6&type=chapterpdf) (visited on 06/17/2024) (cit. on p. 3).
- 322 [27] Patricia Hill Collins. “Black Feminist Thought in the Matrix of Domination”. In: *Black feminist thought: Knowledge, consciousness, and the politics of empowerment* 138.1990 (1990), pp. 221–238. URL: [https://archive.cunyh umanitiesalliance.org/introsocspring20/wp-content/uploads/sites/50/](https://archive.cunyh umanitiesalliance.org/introsocspring20/wp-content/uploads/sites/50/2019/03/Collins.Black-Feminist-Thought.pdf)
323 [2019/03/Collins.Black-Feminist-Thought.pdf](https://archive.cunyh umanitiesalliance.org/introsocspring20/wp-content/uploads/sites/50/2019/03/Collins.Black-Feminist-Thought.pdf) (visited on 06/17/2024) (cit. on p. 3).
- 326 [28] Ange-Marie Hancock. *Intersectionality: An Intellectual History*. Oxford University Press, 2016. URL: [https://books.google.com/books?hl=en&lr=&id=H9bNCgAAQBAJ&oi=fnd&pg=PP1&dq=18.%09Hancock+AM.+Intersectionality:+An+Intellectual+History&ots=Pr-xFsrnFs&sig=PzkXEBrVjbI4FINVMRGK_](https://books.google.com/books?hl=en&lr=&id=H9bNCgAAQBAJ&oi=fnd&pg=PP1&dq=18.%09Hancock+AM.+Intersectionality:+An+Intellectual+History&ots=Pr-xFsrnFs&sig=PzkXEBrVjbI4FINVMRGK_a-iSq4)
327 [a-iSq4](https://books.google.com/books?hl=en&lr=&id=H9bNCgAAQBAJ&oi=fnd&pg=PP1&dq=18.%09Hancock+AM.+Intersectionality:+An+Intellectual+History&ots=Pr-xFsrnFs&sig=PzkXEBrVjbI4FINVMRGK_a-iSq4) (visited on 06/17/2024) (cit. on p. 3).
- 330 [29] Combahee River Collective. “The Combahee River Collective Statement: Black Feminist Organizing in the
331 Seventies and Eighties”. In: (*No Title*) (1986). URL: <https://cir.nii.ac.jp/crid/1130282270401842432>
332 (visited on 06/17/2024) (cit. on p. 3).
- 333 [30] Leonard Owens Iii, Tim Bishop, and Scott Ortolano. “Sojourner Truth, “Ain’t I a Woman?” (1851)”. In: *Starting the Journey: An Intro to College Writing* (). URL: [https://fsw.pressbooks.pub/enc1101/chapter/](https://fsw.pressbooks.pub/enc1101/chapter/sojourner-truth-aint-i-a-woman-1851/)
334 [sojourner-truth-aint-i-a-woman-1851/](https://fsw.pressbooks.pub/enc1101/chapter/sojourner-truth-aint-i-a-woman-1851/) (visited on 06/17/2024) (cit. on p. 3).
- 336 [31] Elle Lett et al. “Conceptualizing, Contextualizing, and Operationalizing Race in Quantitative Health Sciences
337 Research”. In: *The Annals of Family Medicine* 20.2 (2022), pp. 157–163. URL: <https://www.annfammed.org/content/20/2/157.abstract> (visited on 06/17/2024) (cit. on p. 3).
- 339 [32] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019.
340 253 pp. URL: fairmlbook.org (cit. on p. 4).
- 341 [33] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. “Inherent Trade-Offs in the Fair Determination of
342 Risk Scores”. In: *Proceedings of Innovations in Theoretical Computer Science (ITCS)* (2017). arXiv: [1609.05807](https://arxiv.org/abs/1609.05807)
343 (cit. on pp. 4, 12).
- 344 [34] Florian Pfisterer et al. “Mcboost: Multi-Calibration Boosting for R”. In: *Journal of Open Source Software* 6.64
345 (2021), p. 3453 (cit. on pp. 4, 6).

Intersectional Consequences for Marginal Fairness

- 346 [35] Yuval Barak-Corren et al. “Prediction across Healthcare Settings: A Case Study in Predicting Emergency
347 Department Disposition”. In: *npj Digital Medicine* 4.1 (1 Dec. 15, 2021), pp. 1–7. ISSN: 2398-6352. DOI:
348 [10.1038/s41746-021-00537-x](https://doi.org/10.1038/s41746-021-00537-x). URL: <https://www.nature.com/articles/s41746-021-00537-x>
349 (visited on 12/16/2021) (cit. on p. 5).
- 350 [36] Alistair Johnson et al. *MIMIC-IV-ED*. Version 1.0. PhysioNet, 2021. DOI: [10.13026/77Z6-9W59](https://doi.org/10.13026/77Z6-9W59). URL:
351 <https://physionet.org/content/mimic-iv-ed/1.0/> (visited on 09/29/2022) (cit. on p. 5).
- 352 [37] Paula Tanabe et al. “Reliability and Validity of Scores on The Emergency Severity Index Version 3”. In: *Academic
353 Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine* 11.1 (Jan. 2004), pp. 59–
354 65. ISSN: 1069-6563. DOI: [10.1197/j.aem.2003.06.013](https://doi.org/10.1197/j.aem.2003.06.013) (cit. on p. 7).
- 355 [38] Mary C. McLellan, Kimberlee Gauvreau, and Jean A. Connor. “Validation of the Cardiac Children’s Hospital
356 Early Warning Score: An Early Warning Scoring Tool to Prevent Cardiopulmonary Arrests in Children with Heart
357 Disease”. In: *Congenital Heart Disease* 9.3 (2014), pp. 194–202. ISSN: 1747-0803. DOI: [10.1111/chd.12132](https://doi.org/10.1111/chd.12132).
358 URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/chd.12132> (visited on 06/20/2024)
359 (cit. on p. 7).
- 360 [39] Kit T. Rodolfa, Hemank Lamba, and Rayid Ghani. “Empirical Observation of Negligible Fairness–Accuracy
361 Trade-Offs in Machine Learning for Public Policy”. In: *Nature Machine Intelligence* 3.10 (2021), pp. 896–904.
362 URL: <https://www.nature.com/articles/s42256-021-00396-x> (visited on 06/17/2024) (cit. on p. 8).
- 363 [40] Ziad Obermeyer et al. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations”.
364 In: *Science* 366.6464 (Oct. 25, 2019), pp. 447–453. ISSN: 0036-8075, 1095-9203. DOI: [10.1126/science.
365 aax2342](https://doi.org/10.1126/science.aax2342). pmid: 31649194. URL: <https://science.sciencemag.org/content/366/6464/447> (visited
366 on 02/17/2020) (cit. on p. 12).
- 367 [41] William G. La Cava, Elle Lett, and Guangya Wan. “Fair Admission Risk Prediction with Proportional Multicali-
368 bration”. In: *Proceedings of the Conference on Health, Inference, and Learning*. PMLR, June 2023, pp. 350–378.
369 URL: <https://proceedings.mlr.press/v209/la-cava23a.html> (visited on 06/20/2023) (cit. on p. 19).
- 370 [42] Kalyanmoy Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, July 2001.
371 ISBN: 978-0-471-87339-6 (cit. on pp. 19, 23).
- 372 [43] Fabian Pedregosa et al. “Scikit-Learn: Machine Learning in Python”. In: *Journal of Machine Learning Research*
373 12.Oct (2011), pp. 2825–2830. URL: <http://www.jmlr.org/papers/v12/pedregosa11a.html> (visited on
374 10/26/2016) (cit. on p. 19).

Intersectional Consequences for Marginal Fairness

375 **SUPPLEMENT**

376 **1 Additional Cohort Details**

Table S1: Fraction of emergency admissions (%) by intersectional position for patients in the MIMIC-IV (top) and BCH (bottom) cohorts. AI/AN: American Indian / Alaskan Native; AA: African American; NHPI: Native Hawaiian Pacific Islander; (N)HL: (Not) Hispanic/Latino; F: female; M: male. Subgroups with fewer than five samples are omitted.

MIMIC-IV ED				
Gender		Female	Male	Overall
Ethnoracial Group				
AI/AN		70 / 257 (27)	82 / 170 (48)	152 / 427 (36)
ASIAN		1,043 / 3,595 (29)	1,032 / 2,384 (43)	2,075 / 5,979 (35)
BLACK/AA		3,124 / 27,486 (11)	2,603 / 14,458 (18)	5,727 / 41,944 (14)
HL		1,063 / 10,262 (10)	1,168 / 5,795 (20)	2,231 / 16,057 (14)
WHITE		18,147 / 50,174 (36)	18,951 / 45,435 (42)	37,098 / 95,609 (39)
Overall		23,447 / 91,774 (26)	23,836 / 68,242 (35)	47,283 / 160,016 (30)
BCH ED				
Race	Gender Ethnicity	Female	Male	Overall
AI/AN	HL	-	-	-
	NHL	1 / 5 (20)	10 / 19 (53)	11 / 24 (46)
	Overall	1 / 7 (14)	10 / 21 (48)	11 / 28 (39)
ASIAN	HL	-	-	-
	NHL	61 / 398 (15)	74 / 484 (15)	135 / 882 (15)
	Overall	61 / 401 (15)	74 / 487 (15)	135 / 888 (15)
BLACK/AA	HL	22 / 232 (9)	25 / 231 (11)	47 / 463 (10)
	NHL	188 / 1,892 (10)	212 / 2,028 (10)	400 / 3,920 (10)
	Overall	210 / 2,124 (10)	237 / 2,259 (10)	447 / 4,383 (10)
NHPI	HL	2 / 9 (22)	1 / 15 (7)	3 / 24 (12)
	NHL	2 / 7 (29)	1 / 4 (25)	3 / 11 (27)
	Overall	4 / 16 (25)	2 / 19 (11)	6 / 35 (17)
Other	HL	261 / 2,812 (9)	308 / 2,963 (10)	569 / 5,775 (10)
	NHL	182 / 1,232 (15)	247 / 1,500 (16)	429 / 2,732 (16)
	Overall	443 / 4,044 (11)	555 / 4,463 (12)	998 / 8,507 (12)
WHITE	HL	53 / 247 (21)	45 / 280 (16)	98 / 527 (19)
	NHL	905 / 3,800 (24)	1,017 / 4,054 (25)	1,922 / 7,854 (24)
	Overall	958 / 4,047 (24)	1,062 / 4,334 (25)	2,020 / 8,381 (24)
Overall	HL	338 / 3,305 (10)	379 / 3,494 (11)	717 / 6,799 (11)
	NHL	1,339 / 7,334 (18)	1,561 / 8,089 (19)	2,900 / 15,423 (19)
	Overall	1,677 / 10,639 (16)	1,940 / 11,583 (17)	3,617 / 22,222 (16)

377 Table S1 shows a detailed breakdown of patient admission characteristics over combinations of race, ethnicity and
378 gender.

379 **2 Additional Experiment Details**

380 **Data preprocessing and cleaning** For numeric data in the MIMIC-IV-ED triage table (Table 3), we encoded outliers
381 as NaNs according to the following (min,max) ranges: temperature (95-105 F); heart rate (30-300 beats per minute);
382 respiratory rate (2-200 breaths per minute), oxygen saturation (50-100%); systolic blood pressure (30-400 mmHg),
383 diastolic blood pressure (30-300 mmHg); pain scores (0-20); acuity score (1-5).

Intersectional Consequences for Marginal Fairness

384 For both cohorts, chief complaint consists of brief strings of free text. For these data, we first applied simple
385 harmonization and cleaning heuristics and then one-hot- encoded the result, filtering out tokens occurring less than 1%
386 of the time. In our preliminary analysis we evaluated the use of pre-trained word embeddings for chief complaint but
387 did not find that they improved performance versus one-hot-encoding.

388 **Algorithm Implementation** We use a Python implementation of Multicalibration Boosting available from
389 github.com/cavalab/pmboost and derived from La Cava, Lett, and Wan [41]. Fairness-Oriented Multiobjective
390 Optimization (FOMO) is available from cavalab.org/fomo. FOMO serves as a generic interface between the multi-
391 objective optimization algorithms from [pymoo](https://pymoo.org) and ML methods that follow the [scikit-learn](https://scikit-learn.org) API while accepting sample
392 weights as an argument during training (i.e. in calls to `fit()`). Our experimental study focuses on utilizing the popular
393 NSGA2⁴² algorithm in conjunction with two widely used ML methods that support weighted classification: random
394 forests (implemented in [XGBoost](https://xgboost.ai)) and penalized linear regression (implemented in [scikit-learn](https://scikit-learn.org)⁴³). The code to run the
395 experiments is available from the repository github.com/cavalab/marginal-intersectional.

396 **Training** We ran 100 trials of each combination of dataset (MIMIC-IV, BCH), fairness task (fair calibration, fair
397 false negative rates), group construction scenario (Base, Race, Gender, Ethnicity, Marginal, Intersectional), and base
398 model (penalized logistic regression, random forests), as shown in Table 4. Each trial utilized a unique random seed that
399 resulted in a random shuffle of the data which was split into 50% train/ 50% test sets. Splits were stratified by outcome
400 (admission), gender, and race to maintain appropriate representation in each. For the runs using FOMO and MIMIC-IV
401 data, the training set was further reduced to 10% (approximately 16k patients) to reduce computation time.

402 **3 Additional Experiments**

403 In this section we report additional experiments meant to characterize the sensitivity of the studied fairness algorithms
404 to hyperparameters and design variables. For both multicalibration boosting and FOMO, we analyze how the choice of
405 base ML model, group prevalence, and dataset affect the results. In the case of multicalibration boosting, we studied the
406 choice of α , a termination criteria that defines the group-specific calibration error threshold, and γ , a parameter that
407 controls the minimum prevalence of a group to be considered for updating. In the case of FOMO, we looked at the
408 effect of using a weighted subgroup FNR metric that accounts for prior probability of the groups, and the effect of a
409 fairness meta-model complexity.

410 **3.1 Multicalibration Boosting**

411 **Sensitivity Analysis** In Fig. S1, we visualize the expected calibration error of LR and RF models on MIMIC-IV
412 as a function of base model, α , γ , and modeling scenario. At higher levels of γ , low-prevalence groups are excluded
413 from fairness updating; hence, performance differences between scenarios tend to shrink. Relatedly, higher values of
414 α loosen the threshold needed for multicalibration to perform an update, and so model performance tends to become
415 similar between groups. Conversely, for very small values of α and γ , small groups have a larger impact on fairness

Intersectional Consequences for Marginal Fairness

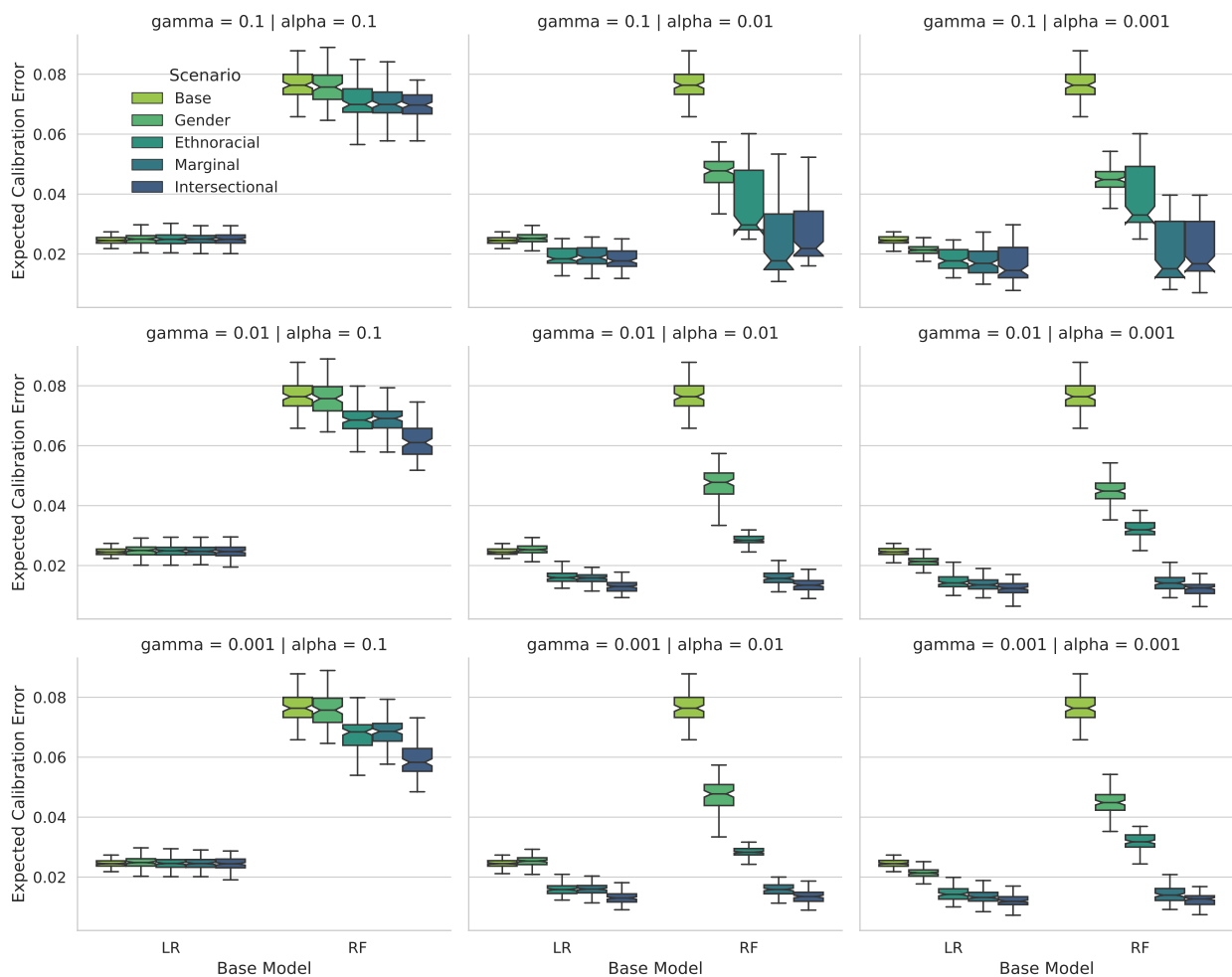


Figure S1: Intersectional group-wise Expected Calibration Error on MIMIC-IV as a function of γ (row), α (column), base ML model (x-axis), and optimization scenario (color). At high levels of α , the models remain unchanged, whereas at very low values of α and γ , performance on intersectional groups can suffer due to small sample sizes.

416 optimization, meaning intersectional modeling matters more for achieving low ECE among intersectional groups.
 417 Overfitting can occur when α is too stringent, leading to degradation of performance on intersectional groups on test
 418 set: see top middle and right panel of Fig. S1, RF models.

419 Fig. S2 sheds light on the interaction between group prevalence, α and γ under multicalibration boosting. Here we
 420 explicitly look at training and test set performance of the intersectional de-biasing approach relative to the baseline
 421 approach, illustrating how the constraints on calibration error (α) and minimum group probability (γ) interplay with
 422 group prevalence (x-axis). In general, we observe that groups that are less prevalent in the data tend to have higher
 423 expected calibration error (ECE). Therefore, when α and γ are set high relative to model performance on adequately
 424 sized groups (e.g., $\alpha = \gamma = 0.1$, top left panel), no de-biasing occurs. Conversely, if γ and α is set very low, de-biasing
 425 occurs over all groups in the training data but this does not fully generalize to test data (bottom right panel).

Intersectional Consequences for Marginal Fairness

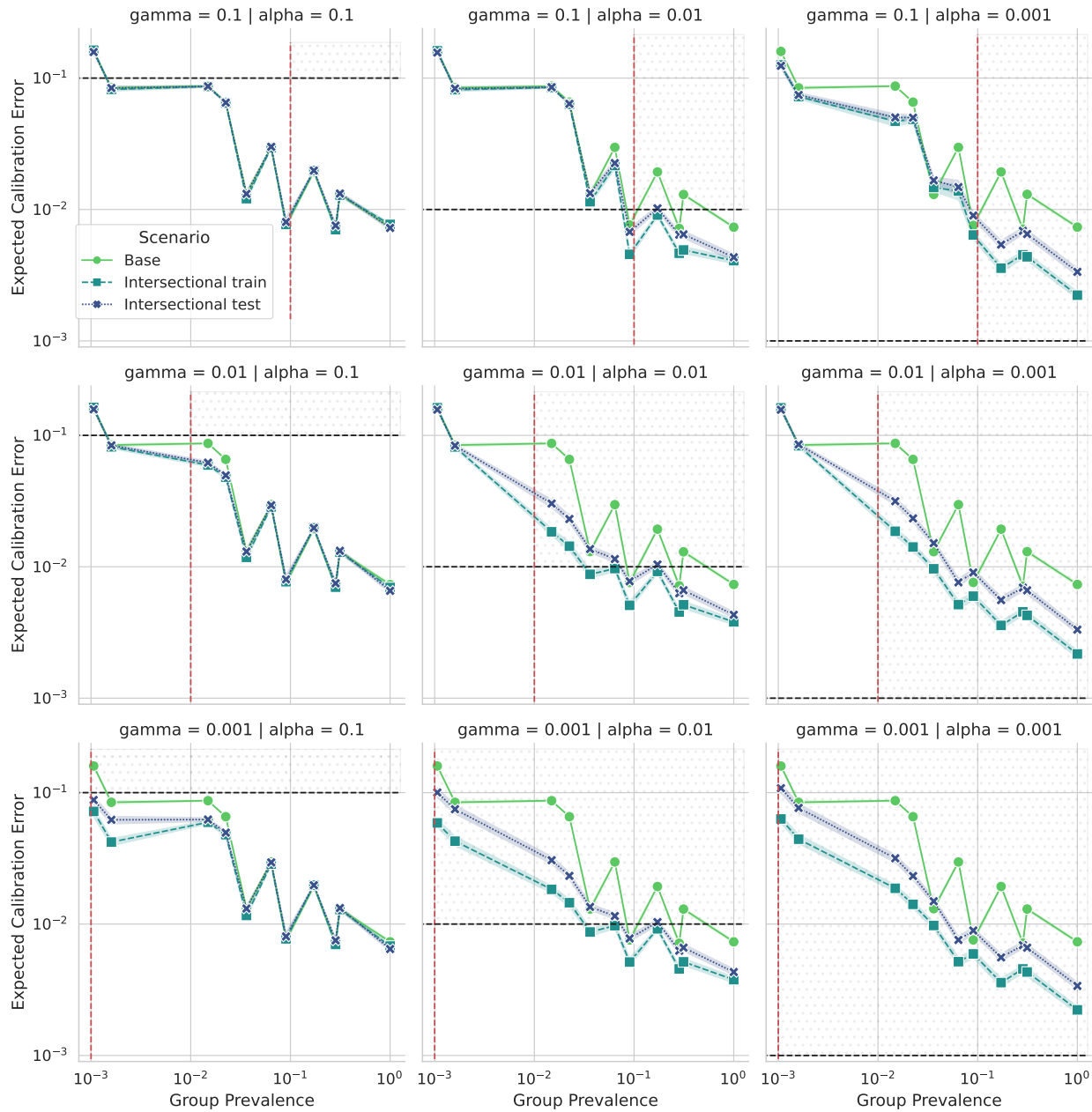


Figure S2: Expected calibration error (ECE) as a function of group prevalence for LR models trained on MIMIC-IV, under different combinations of α and γ . The shaded area indicates the region of model performance that is subjected to optimization by either having an ECE higher than the threshold, α , or a group prevalence higher than the cutoff, γ .

Intersectional Consequences for Marginal Fairness

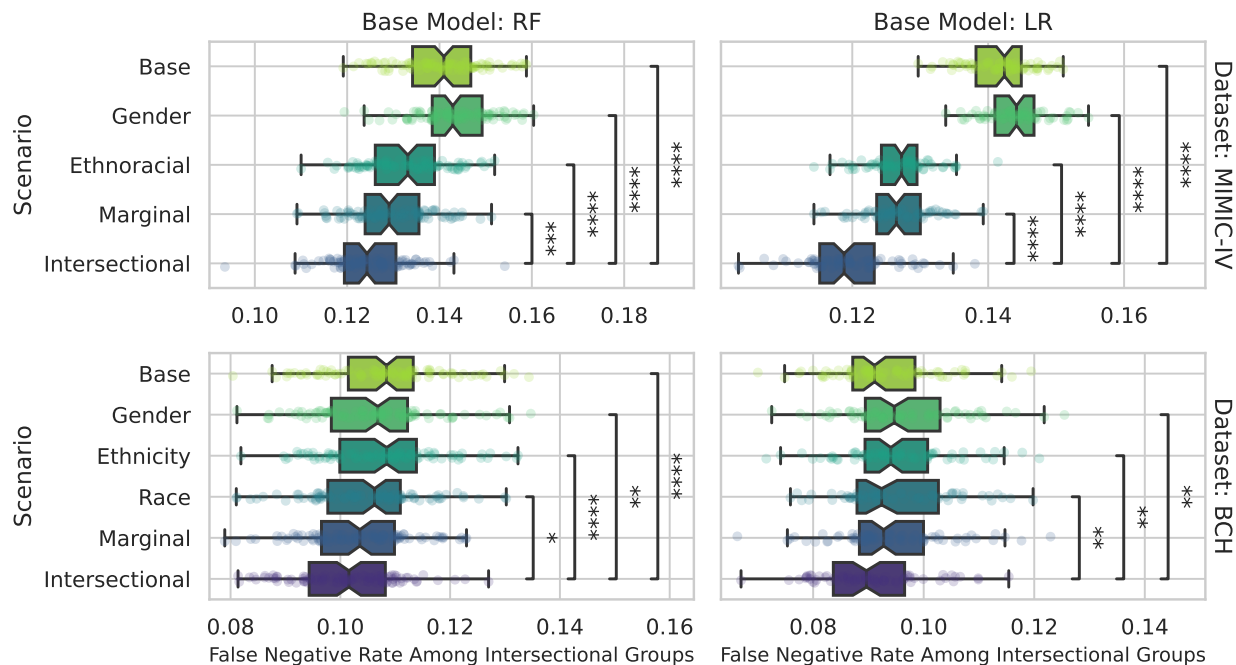


Figure S3: False negative rates (FNR) among intersectional groups under different base models (left: random forests (RF), right: penalized logistic regression (LR)) and FOMO de-biasing scenarios (y-axis) for MIMIC-IV (top) and BCH (bottom). Statistical tests are two-sided Mann-Whitney-Wilcoxon tests with Holm-Bonferroni correction (*: $1e-2 < p \leq 5e-2$; **: $1e-3 < p \leq 1e-2$; ***: $1e-4 < p \leq 1e-3$; ****: $p \leq 1.0e-4$).

426 3.2 Fairness-Oriented Multiobjective Optimization

427 **Sensitivity Analysis** We varied several parameters during our experimentation with FOMO: 1) The choice of ML
 428 model (penalized logistic regression or random forests); 2) whether the definition of subgroup fairness incorporates the
 429 prior probability of the group as in other work¹³; 3) the type of meta-model used to estimate the sample weights used
 430 to train the base models. Regarding 1), we saw similar trends in results when working with linear models, as shown
 431 in Fig. S3. Regarding 2), we did not observe a difference in performance when incorporating prior probabilities of
 432 the groups; our results here do not incorporate these adjustments for group size. Regarding 3), we did not observe a
 433 difference in performance with variations of the meta-model. In our results, we use a standard linear formulation to
 434 map patient attributes to training sample weights; when using the intersectional fairness implementation, we extend the
 435 linear model with interaction terms between the scenario's protected features. Our observations suggest that whether or
 436 not the group probability was factored into the fairness definition, it had minimal discernible impact on the outcomes
 437 for both RF and LR models across both datasets.

438 **Trade-off Visualization** Fig. S4 shows the set of models generated by FOMO as part of its optimization process,
 439 which characterizes the trade-off space (i.e. the Pareto frontier) between fairness and accuracy objectives.

Intersectional Consequences for Marginal Fairness

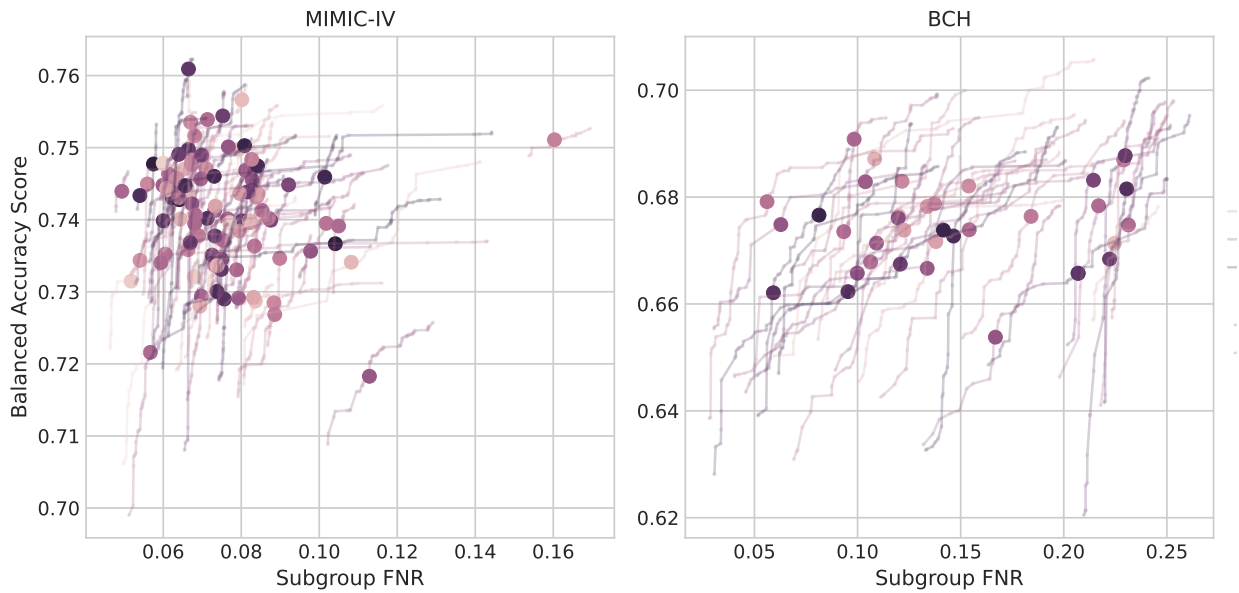


Figure S4: Accuracy-Fairness Tradeoffs and Model Selection. FOMO optimizes a Pareto frontier of solutions simultaneously in order to characterize the trade-off between accuracy and fairness objectives. These final frontiers are shown for MIMIC-IV (left) and BCH (right), with each line representing one realization of the experiment. In order to choose a final model (marked by large circles for each run), a multi-criteria decision making method known as Pseudo-Weights is used⁴². This method chooses the model that maximizes a weighted sum of the objectives. For each candidate model, the weights of each objective depend on the normalized distance to the worst solution for that objective. FNR: false negative rate.