Ensuring Fairness in Priority-Based Admissions with Uncertain Scores

Zhiqiang Zhang[1] **Pengyi Shi**[2] **Amy R. Ward**[1]
[1]The University of Chicago Booth School of Business
[2]Purdue University Mitch Daniels School of Business
{zqzhang0,amy.ward}@chicagobooth.edu shi178@purdue.edu

Abstract

Priority-based admission policies are widely used to determine who can access scarce resources. Such policies assign scores to arriving individuals and prioritize those with higher scores for admission. Ideally, the resources yield greater benefits for individuals with higher scores, while the scores also provide a mechanism for ensuring fairness in resource access, according to some agreed-upon metric. The core problem is that scores must be estimated from historical data, and so are prone to estimation error. As a result, well-intentioned interventions to promote fairness can backfire. Our contribution is to provide a framework for analytically adjusting these estimated scores, ensuring that fairness interventions are implemented with a high degree of confidence.

1 Introduction

Priority-based policies are widely used to manage resource-constrained admission decisions in many high-stakes programs, from hospital ICUs to criminal justice diversion programs [Chan et al., 2012, Zhang et al., 2025]. The core challenge in these programs is balancing the competing costs of admission and denial. Admitting an individual who subsequently fails incurs a significant cost, such as a patient's death in an ICU or re-offense during a diversion program. Denial also carries a distinct outsourcing cost, such as rerouting a patient to another hospital or resorting to incarceration. Since an individual's likelihood of failure is not known a priori, machine learning (ML) algorithms are typically used to estimate this risk and generate priority scores. To best utilize limited resources, it is therefore natural to use priority policies guided by these ML-generated scores. Yet, such prioritization can lead to unfair outcomes across demographic groups [Obermeyer et al., 2019, Angwin et al., 2016], necessitating a careful study of fairness in these priority-based admissions.

This challenge of ensuring fairness is compounded by the inevitable error in ML-generated scores. ML error not only leads to the risk of misprioritizing individuals, but more importantly, to the risk of misidentifying the disadvantaged group. A well-intentioned fairness policy might adjust priority scores to allocate more resources to a group it identifies as disadvantaged; however, this can backfire and exacerbate unfairness if that group is misidentified. This raises two critical **research questions**: When can we be confident that fairness-aware adjustments truly mitigate unfairness in the presence of ML error? Furthermore, if we are confident in the overall adjustment, which individual post-adjustment decisions can be trusted?

To address these questions, we first formulate a fairness-aware optimization model and characterize its solution (Section 2). We then present our core contribution: a two-fold uncertainty quantification framework (Section 3). The first fold quantifies the confidence in the overall fairness adjustment itself, while the second, building on Zhang et al. [2025], quantifies confidence in each post-adjustment individual decision.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

2 Fairness-Aware Priority Policies

2.1 Model Formulation

We consider an admission decision process for a program with limited resources, specifically N available slots. Individuals arrive at a rate λ and are described by a vector of covariates $\mathbf{X}$ (e.g., gender, race, age, etc). An individual with covariate $\mathbf{x}$ requests a stay of length $m(\mathbf{x})$. By Little's Law, the expected resource requirement for each covariate $\mathbf{x}$ is $\lambda\mathbb{P}(\mathbf{X}=\mathbf{x})m(\mathbf{x})$. For fairness analysis, we consider two protected groups (e.g., $i=1$ for male, $i=2$ for female), where the arrival rate for group i is denoted by λ_i (with $\lambda=\lambda_1+\lambda_2$). We use $\mathbb{E}[\cdot]$ and $\mathbb{E}_i[\cdot]$ to denote expectations over the covariate distribution across all groups and within group i, respectively. Individuals belong to one of two classes (e.g., $j=1$ for low-risk, $j=2$ for high-risk), which determines their probability of program failure q_j.[1] The probability that an individual with covariates $\mathbf{x}$ belongs to the low-risk class is $p_1(\mathbf{x})$, which is unknown and must be estimated from historical data. Let $\widehat{p}_1(\mathbf{x})$ denote the estimated probability, implying an estimated failure probability of $\widehat{q}(\mathbf{x}) := \sum_j \widehat{p}_j(\mathbf{x})q_j$.

Admission decisions have associated costs. Admitting an individual from class j incurs a total expected cost of c_j^T, incorporating program maintenance and potential failure costs. Denying them results in a cost of c_j^D, accounting for outsourcing expenses. We assume admitting always has a lower cost ($c_j^T \le c_j^D$), but resources are insufficient for everyone ($\lambda\mathbb{E}[m(\mathbf{X})] > N$), making the admission control necessary. Our decision variable is $a(\mathbf{x})$, the fraction of individuals with covariate $\mathbf{x}$ to admit. The optimization problem (1) is to minimize total costs while incorporating a penalty on the unfairness $\widehat{\mathcal{F}}(\mathbf{a})$ (will be defined in Section 2.2), weighted by a parameter $\eta \ge 0$ that captures the fairness-efficiency trade-off:

$$\min_{\mathbf{a}\in\mathcal{A}} \widehat{\mathcal{C}}(\mathbf{a}) + \eta\widehat{\mathcal{F}}(\mathbf{a}), \tag{1}$$

where $\mathcal{A} := \{\mathbf{a} : a(\mathbf{x}) \in [0,1] \text{ and } \lambda\mathbb{E}[a(\mathbf{X})m(\mathbf{X})] \le N\}$ is the feasible region, and $\widehat{\mathcal{C}}(\mathbf{a}) := \sum_j c_j^D\lambda\mathbb{E}[(1-a(\mathbf{X}))\widehat{p}_j(\mathbf{X})] + c_j^T\lambda\mathbb{E}[a(\mathbf{X})\widehat{p}_j(\mathbf{x})]$ is the total estimated cost.

2.2 Fairness Criteria

We define unfairness as the disparity of fairness metrics between the two groups, that is, $\widehat{\mathcal{F}}(\mathbf{a}) := |\widehat{\mathcal{F}}_1(\mathbf{a}) - \widehat{\mathcal{F}}_2(\mathbf{a})|$. For the group-specific fairness metric $\widehat{\mathcal{F}}_i(\mathbf{a})$, we incorporate various fairness definitions from the ML literature. By framing the admission decision as analogous to a positive prediction in classification, we can adapt common criteria such as Statistical Parity, False Positive Rate (FPR) Parity, and True Positive Rate (TPR) Parity [Chouldechova, 2017, Corbett-Davies et al., 2017, Verma and Rubin, 2018]. Specifically, we define a false positive as an admitted individual who subsequently fails the program and a true positive as one who successfully completes it. Additionally, we consider a resource-based criterion, $\mathbf{f}$-Resource Parity, which aims to allocate a target fraction f_i of resources to each group i. These criteria are summarized in Table 1.[2] Since outcomes for denied applicants are unobserved, FPR and TPR Parity must be estimated using $\widehat{p}_1(\mathbf{x})$ and thus involve uncertainty analyzed in Section 3.

Table 1: Fairness criteria and corresponding priority scores

Fairness Criterion	Fairness Metric $\widehat{\mathcal{F}}_i(\mathbf{a})$	Priority Score $\widehat{s}_{\textit{fair}}(\mathbf{x})$
Efficiency Only	N/A	$\widehat{s}(\mathbf{x}) := \frac{\sum_j \widehat{p}_j(\mathbf{x})(c_j^D - c_j^T)}{m(\mathbf{x})}$
Statistical Parity	$\mathbb{E}_i[a(\mathbf{X})]$	$\widehat{s}(\mathbf{x}) \pm \eta\frac{1}{\lambda_i m(\mathbf{x})}$
FPR Parity	$\frac{\mathbb{E}_i[a(\mathbf{X})\widehat{q}(\mathbf{x})]}{\mathbb{E}_i[\widehat{q}(\mathbf{x})]}$	$\widehat{s}(\mathbf{x}) \pm \eta\frac{\widehat{q}(\mathbf{x})}{\lambda_i m(\mathbf{x})\mathbb{E}_i[\widehat{q}(\mathbf{X})]}$
TPR Parity	$\frac{\mathbb{E}_i[a(\mathbf{X})(1-\widehat{q}(\mathbf{x}))]}{\mathbb{E}_i[1-\widehat{q}(\mathbf{x})]}$	$\widehat{s}(\mathbf{x}) \pm \eta\frac{1-\widehat{q}(\mathbf{x})}{\lambda_i m(\mathbf{x})\mathbb{E}_i[1-\widehat{q}(\mathbf{X})]}$
$\mathbf{f}$-Resource Parity	$\lambda_i\mathbb{E}_i[a(\mathbf{X})m(\mathbf{X})] - f_i N$	$\widehat{s}(\mathbf{x}) \pm \eta$

[1] The class-based dynamic is to reduce the dimension of model parameters from the number of covariates to the number of classes, given the limited dataset. In this paper, we focus on binary classes, groups, and outcomes for simplicity. All results derived in this paper can be generalized to multiple classes, groups, and outcomes.

[2] Some common criteria, such as Error Rate Balance and Predicted Parity, are not listed, where the former is to combine FPR Parity and TPR Parity, and for the latter, we do not have a theoretical result yet since the penalty is non-linear in decision variable $a(\mathbf{x})$.

2.3 Priority Policy Structure

Incorporating the penalty terms $\widehat{\mathcal{F}}(\mathbf{a})$ from Section 2.2, the optimal solution to the optimization problem (1) is a priority score policy. With no fairness concerns, the *efficiency-only policy* prioritizes individuals based on the score $\widehat{s}(\mathbf{x})$ in Table 1, which represents the expected cost reduction per unit of resource. However, the efficiency-only policy may result in an imbalanced allocation of resources based on the chosen fairness criterion. The fairness-aware score is adjusted to ensure that the disadvantaged group receives a higher score, thereby allocating more resources to that group.

Definition 1 *A* priority score policy $\mathbf{a}$*, with score function* $s : \mathbb{X} \to \mathbb{R}$*, admits the highest-scoring covariates with total resource requirement* $\lambda\mathbb{P}(\mathbf{X} = \mathbf{x})m(\mathbf{x})$ *until the resource limit* N *is met. The* decision boundary covariate *is partially admitted, denoted by* $\mathbf{x}_0(\mathbf{s})$.

Definition 2 *A group is defined as* advantaged *if, under the efficiency-only policy, it has higher fairness metric* $\widehat{\mathcal{F}}_i(\mathbf{a})$ *under the chosen fairness criterion.*[3] *The other group is* disadvantaged.

Proposition 1 *For sufficiently small* η,[4] *a priority score policy* $\widehat{a}_{\textit{fair}}(\mathbf{x})$ *with score function* $\widehat{s}_{\textit{fair}}(\mathbf{x})$ *as defined in Table 1 solves* (1)*. A positive score adjustment is applied to the disadvantaged group, and a negative adjustment is applied to the advantaged group.*

The key to the proof is to rewrite (1) into a fractional knapsack problem. The solution to such a problem is known to be a greedy policy based on a score function [Goodrich and Tamassia, 2001, chap. 5.1.1]. Each fairness criterion in Table 1 corresponds to a distinct adjustment to the efficiency-only score, designed to allocate resources more equitably toward the disadvantaged group. For example, $\mathbf{f}$-Resource Parity uses a constant shift. Statistical Parity assigns a larger shift for individuals with a shorter length of stay $m(\mathbf{x})$, rapidly changing the admission fraction with minimal impact on individual decisions. For similar reasons, FPR Parity and TPR Parity assign larger shifts to individuals with a higher failure rate $\widehat{q}(\mathbf{x})$ and a higher success rate $1 - \widehat{q}(\mathbf{x})$, respectively.

However, the entire adjustment mechanism—from identifying the disadvantaged group to calculating the score shifts—relies on the uncertain probability estimate $\widehat{p}_1(\mathbf{x})$. Is it possible that the adjustment exacerbates true unfairness, due to the misguidance from this estimation?

3 Two-fold Uncertainty Quantification Framework

To investigate the uncertainty in adjusted priority decisions, our analysis is two-fold. First, we characterize whether the fairness adjustment is truly mitigating unfairness (i.e., applied in the correct direction). Second, if the direction is confident, we characterize the confidence of individual decisions.

3.1 Confidence in Adjustment Mechanism

Recall from Section 2.2 that FPR and TPR are subject to ML error in $\widehat{p}_1(\mathbf{x})$. Consequently, the true advantaged group may differ from the estimated one. We must therefore determine if the same group identification holds under the true probability $p_1(\mathbf{x})$. This verification is critical, as a misidentification would cause the subsequent fairness adjustment to exacerbate, rather than mitigate, unfairness.

Mathematically, assume that the efficiency-only policy with $\widehat{s}(\mathbf{x})$ identifies group 1 as the advantaged group in estimation, that is, $\widehat{\mathcal{F}}_1(\mathbf{a}) > \widehat{\mathcal{F}}_2(\mathbf{a})$. We want to verify if $\mathcal{F}_1(\mathbf{a}) > \mathcal{F}_2(\mathbf{a})$ still holds under true probability $p_1(\mathbf{x})$. We will focus on the FPR Parity for illustration, approaching this by comparing confidence intervals (CIs) for the FPRs in the two groups.

Since the global resource limit constraint makes all admission decisions interdependent, our analysis must control for estimation error across all covariates simultaneously. Therefore, our analysis requires simultaneous CIs $I_p(\mathbf{x})$, which provide a joint guarantee $\mathbb{P}(p_1(\mathbf{x}) \in I_p(\mathbf{x}), \forall \mathbf{x}) \geq 1 - \epsilon \in (0, 1)$. Such simultaneous CIs can be derived from bootstrapping methods [Mandel and Betensky, 2008] or conformal predictors [Gazin et al., 2024]. From these, we apply the FPR formula from Table 1 to construct CIs $I_i^{\textit{FPR}}$ for the true FPR in both groups $i = 1, 2$.

[3] Advantaged group is assigned more resources—for the four criteria in Table 1, it means: higher admission fraction, higher FPR, higher TPR, and more resources allocated compared to the target $f_i N$.

[4] For large η, the solution becomes the perfectly fair policy that achieves $\widehat{\mathcal{F}}(\mathbf{a}) = 0$.

Proposition 2 *If $I_1^{FPR} \cap I_2^{FPR} = \emptyset$, then group 1 is the true advantaged group under $p_1(\mathbf{x})$ with probability at least $1 - \epsilon$.*

When the identification of disadvantaged group is confident (i.e., the CIs do not overlap), the fairness adjustments proposed in Table 1 can proceed with confidence. However, when the identification lacks confidence, the unreliable adjustment could potentially exacerbate the unfairness that the policy aims to mitigate. In such cases, further study is needed before proceeding with any score adjustments.

3.2 Confidence in Individual Decisions

If we are confident that the adjustment mechanism is in the correct direction, we proceed to characterize the confidence in individual decisions under the adjusted policy with score $\widehat{s}(\mathbf{x}) \pm \eta\widehat{t}(\mathbf{x})$, where the adjustment term $\widehat{t}(\mathbf{x})$ depends on the selected fairness criterion in Table 1. Our goal is to characterize which covariates have a correct decision guarantee within confidence level $1 - \epsilon$ and which remain uncertain. Utilizing the priority structure, we can characterize confident or uncertain covariates by comparing individual CIs with the decision boundary's CI in the following three steps.

CI for Individual Scores. From the same simultaneous CIs in Section 3.1, applying the transformation of the adjusted score function $\widehat{s}(\mathbf{x}) \pm \eta\widehat{t}(\mathbf{x})$, we construct CIs $\widetilde{I}(\mathbf{x}) = [\widetilde{L}(\mathbf{x}), \widetilde{U}(\mathbf{x})]$ for the true fair score with negative adjustment in group 1 and with positive adjustment in group 2.

CI for Decision Boundary. The score of the decision boundary is monotonic with respect to each individual score. This critical property, enabled by the priority structure, allows us to construct a CI for the decision boundary $[T_L, T_U] := [\widetilde{L}(\mathbf{x}_0(\widetilde{L})), \widetilde{U}(\mathbf{x}_0(\widetilde{U}))]$, where the bounds are the decision boundaries when all individual scores are simultaneously set to their lower or upper confidence limits.

Decision Confidence Guarantee. We certify the individual admission decisions by comparing each covariate's individual CI $\widetilde{I}(\mathbf{x})$ with the decision boundary's CI $[T_L, T_U]$, as shown in Figure 1.

Proposition 3 *If covariate $\mathbf{x} \in \mathbb{X}$ satisfies $\widetilde{L}(\mathbf{x}) > T_U$ (if admitted) or $\widetilde{U}(\mathbf{x}) < T_L$ (if denied), then the decision $\widehat{a}_{fair}(\mathbf{x})$ is correct with probability at least $1 - \epsilon$.*

The proof follows the idea in the indifference-zone selection from the R&S literature [Hong et al., 2021]. Uncertain covariates that require additional human oversight fall into two categories: (1) those with scores close to the decision boundary, where small estimation errors can flip the decision (e.g., $\mathbf{x}_4$ in Figure 1); and (2) those with high estimation uncertainty (e.g., due to small sample size or poor predictive power), even if their score estimate is far from the boundary (e.g., $\mathbf{x}_3$ in Figure 1).

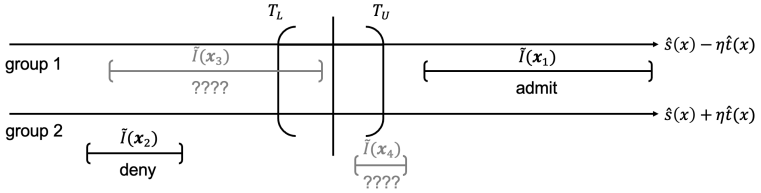

Figure 1: Confident (colored in black) and uncertain (colored in grey) admission decisions.

4 Conclusion

This paper integrates an operations research (OR) model for fair admission control with ML tools for uncertainty quantification. By leveraging the priority structure inherent to the OR model, our framework translates the uncertainty from ML estimation to the final admission decisions. Within this analytical framework, we address the research questions posed in the introduction as follows.

The effect of ML error on fairness is nuanced—it can sometimes exacerbate unfairness and at other times reduce it. We provide a diagnostic to determine when the adjustment direction for mitigating unfairness is statistically reliable. Building on this, we characterize which individual post-adjustment decisions can be trusted and flag those with insufficient confidence for targeted human review. Future research could focus on approaches to ensure fairness even when the adjustment direction is uncertain, as well as on developing more refined human-in-the-loop mechanisms.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias risk assessments in criminal sentencing. *ProPublica*, May 23, 2016.

Carri W Chan, Vivek F Farias, Nicholas Bambos, and Gabriel J Escobar. Optimizing intensive care unit discharge decisions with patient readmissions. *Operations Research*, 60(6):1323–1341, 2012.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

Ulysse Gazin, Gilles Blanchard, and Etienne Roquain. Transductive conformal inference with adaptive scores. In *International Conference on Artificial Intelligence and Statistics*, pages 1504–1512. PMLR, 2024.

Michael T Goodrich and Roberto Tamassia. *Algorithm Design: Foundations, Analysis, and Internet Examples*. John Wiley & Sons, 2001.

L Jeff Hong, Weiwei Fan, and Jun Luo. Review on ranking and selection: A new perspective. *Frontiers of Engineering Management*, 8(3):321–343, 2021.

Micha Mandel and Rebecca A Betensky. Simultaneous confidence intervals based on the percentile bootstrap approach. *Computational statistics & data analysis*, 52(4):2158–2165, 2008.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018.

Zhiqiang Zhang, Pengyi Shi, and Amy Ward. Admission decisions under imperfect classification: An application in criminal justice. *Available at SSRN 5214197*, 2025.