A Burden Shared is a Burden Halved: A Fairness-Adjusted Approach to Classification

Bradley Rava[1], Wenguang Sun[2], Gareth M. James[3] and Xin Tong[4]

Abstract

We investigate the fairness issue in classification, where automated decisions are made for individuals from different protected groups. In high-consequence scenarios, decision errors can disproportionately affect certain protected groups, leading to unfair outcomes. To address this issue, we propose a fairness-adjusted selective inference (FASI) framework and develop data-driven algorithms that achieve statistical parity by controlling the false selection rate (FSR) among protected groups. Our FASI algorithm operates by converting the outputs of black-box classifiers into R-values, which are both intuitive and computationally efficient. These R-values serve as the basis for selection rules that are provably valid for FSR control in finite samples for protected groups, effectively mitigating the unfairness in group-wise error rates. We demonstrate the numerical performance of our approach using both simulated and real data.

Keywords: Calibration by group; Fairness in machine learning; False selection rate; Selective Inference; Statistical parity.

[1]University of Sydney Business School.
[2]Center for Data Science, Zhejiang University. Author for correspondence: wgsun@zju.edu.cn
[3]Goizueta Business School, Emory University.
[4]Faculty of Business and Economics, University of Hong Kong.

We are grateful to the associate editor and two referees whose meticulous and constructive feedback has substantially improved the clarity, presentation and theory of our manuscript. We would also like to thank Matteo Sesia, Zinan Zhao and Wangcheng Li for their valuable discussion and suggestions on methodology and theory.

1 Introduction

In a broad range of applications, artificial intelligence (AI) systems are rapidly replacing human decision-making. Many of these scenarios are sensitive in nature, where the AI's decision, correct or not, can directly impact one's social or economic status. A few examples include a bank determining credit card limits, stores using facial recognition systems to detect shoplifters, and hospitals attempting to identify which of their patients has a specific disorder. Unfortunately, despite their supposedly unbiased approach to decision-making, there has been increasing evidence that AI algorithms often fail to treat equally people of different genders, races, religions, or other protected attributes. Whether this is due to the historical bias in one's training data, or otherwise, it is important, for both legal and policy reasons, that we make ethical use of data and ensure that decisions are made fairly for everyone regardless of their protected attributes.

Despite the significant efforts in developing supervised learning algorithms to improve the prediction accuracy, making reliable and fair decisions in the classification setting remains a critical and challenging problem for two main reasons. Firstly, AI algorithms are often required to make classifications on all new observations without a careful assessment of associated uncertainty or ambiguity. This limitation highlights the need for a more flexible framework to handle intrinsically difficult classification tasks where a definitive decision carries high stakes. Such a framework should enable decision-makers to wait and gather additional information with greater confidence before making a final decision. Secondly, modern machine learning models, such as neural networks, are often highly complex, making it challenging, if not impossible, to explicitly quantify the uncertainty associated with their outputs or to provide guarantees on the fairness of the decisions. Therefore, developing methods that can ensure both risk control and fairness is crucial for AI systems to be reliable and trustworthy.

This article develops a "fairness-adjusted selective inference" (FASI) framework to address the critical issues of uncertainty assessment, error rate control and statistical parity in classification. We provide an *indecision* option for observations which cannot be selected into any classes with confidence. These observations will then be separately evaluated. This practice often aligns with the policy objectives in many real world scenarios. For example, incorrectly classifying a low-risk individual as a recidivist or rejecting a well-deserving candidate for the loan request is much more expensive than turning the case over for a more careful review. A mis-classification is an error, the probability of which must be controlled to be small as its consequence can be severe. By contrast, the cost of an indecision is usually much less. For example, the ambiguity can be mitigated by collecting additional contextual knowledge of the convicted individual or requesting more information from the loan applicant. Under the selective inference (Benjamini 2010) framework, we only make definitive decisions on a *selected subset* of all individuals; the less consequential indecision option is considered as a wasted opportunity rather than an error. A natural error rate notion under this framework is the *False Selection Rate* (FSR), which is defined as the expected fraction of erroneous classifications among the selected subset of individuals. The goal is to develop decision rules that aim to control and equalize the FSR across protected groups, while minimizing the total wasted opportunities.

A critical issue is that a classification rule that controls the overall FSR may have disparate impacts on different protected groups. We illustrate the point using the COMPAS data set (Angwin et al. 2016, Dieterich et al. 2016). The COMPAS algorithm has been widely used in the US to help inform courts about a defendant's recidivism likelihood, i.e., the likelihood of a convicted criminal recommitting a crime, so any prediction errors could have significant implications. The left panel of Figure 1 shows the *False Selection Proportions* (FSP), i.e. the fraction of individuals who did not recommit a crime among those who were classified as recidivists.

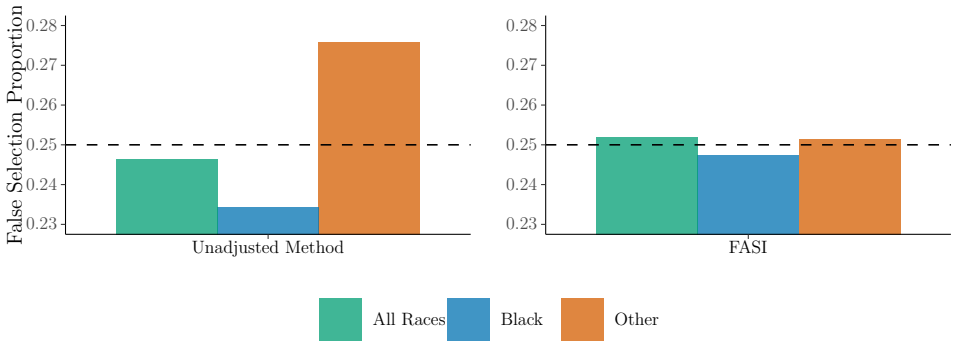

Figure 1: The selection of recidivists from a pool of criminal defendants (Broward County, Florida). The target FSR is 25%. Left: the unadjusted approach. Right: the proposed FASI approach.

The classification rule was constructed via a Generalized Additive Model (GAM) [1] (Hastie et al. 2009, James et al. 2023) to achieve the target FSR of 25%. We first split the COMPAS data into distinct training and test sets. The GAM was fitted using the training data set, and subsequently applied to the test set to predict whether a defendant was a recidivist.

We can see that the green bar, which provides the overall FSP for all races, is close to the target value. Moreover, the rule appears to be "fair" for all individuals, regardless of their protected attributes, in the sense that the *same* threshold has been applied to the confidence scores (i.e. estimated class probabilities) produced by the *same* GAM fit. However, the blue and orange bars show that the FSPs for different racial groups differ significantly from 25%, which is clearly not a desirable situation.

This article introduces a new notion of fairness that requires parity in FSR control across various protected groups. This aligns with the social and policy goals in a range of decision-making scenarios such as selecting recidivists or determining risky loan applicants, where the burden of erroneous classifications should be shared equally among different genders and races. However, the development of effective and fair FSR rules is challenging. First, controlling the error rate associated with a classifier, such as one built around the GAM procedure, critically depends on the accuracy of the scores. However, the assessment of the accuracy/uncertainty of these scores largely remains unknown. Second, we wish to provide practitioners with theoretical guarantees on the parity and validity for FSR control, regardless of the algorithm being used, including complex black-box classifiers.

To address these issues, we develop a data-driven FASI algorithm specifically designed to control the FSRs of protected groups below a user-specified level α. The right panel of Figure 1 illustrates the FSPs of FASI on the recidivism data. All individual FSPs are controlled at 25% approximately. FASI works by converting the confidence scores from a black-box algorithm to an R-value, which is intuitive, easy to compute, and comparable across different protected groups. We then show that selecting all observations with R-value no greater than α will result in an FSR of approximately α. Hence, we can directly use this R-value to assign new observations a class label or, for observations with high R-values, assign them to the *indecision* class.

This paper makes several contributions. Firstly, we introduce a novel notion of fairness within the selective inference framework, incorporating an indecision option. In high-consequence situations, it is sensible to exercise caution, by either withholding or separately evaluating such cases until additional evidence is gathered. This reduces the risk of making definitive decisions without sufficient support, thus promoting cautious and fair decisions in these complex scenarios. Secondly, a data-driven FASI Algorithm is developed based on the utilization of the

[1] Although a GAM was utilized for illustration purposes, we emphasize that the same issue can arise regardless of the specific machine learning algorithm employed.

R-value. This algorithm, which can be deployed with user-specified learning algorithms (e.g. random forest, neural networks), is intuitively appealing and easy to interpret. Thirdly, rigorous theoretical justifications are provided for the FASI algorithm. The theory on FSR control is established with mild assumptions on data exchangeability, accommodating scores generated by black-box algorithms. Finally, the empirical performance of FASI is investigated through extensive experimentation using simulated and real-world data sets, demonstrating the effectiveness and practical utility of the proposed approach.

The rest of the paper is structured as follows. In Section 2 we define the FSR and describe the problem formulation. Section 3 introduces the R-value and FASI algorithm. The numerical results for simulated and real data are presented in Sections 4 and 5, respectively. Section 6 concludes the main article with a discussion of related works and possible extensions. The Online Supplementary Material provides additional technical details about the methodology, proof of theorems, and supplementary numerical results.

2 Problem Formulation

Suppose we observe a data set $\{(X_i, A_i, Y_i) : i \in \mathcal{D}\}$, where $\mathcal{D} = [n] \equiv \{1, ..., n\}$ is an index set, $X_i \in \mathbb{R}^p$ is a p-dimensional vector of features, $A_i \in \mathcal{A}$ is an additional feature representing the protected or sensitive attribute, and Y_i is a class label taking values in $\mathcal{C} = \{1, \ldots, C\}$. The goal is to predict the classes for m new individuals indexed by $\mathcal{D}^{test} = \{n+1, ..., n+m\}$, with observed features $\{(X_j, A_j) : j \in \mathcal{D}^{test}\}$. Denote $\mathcal{D}_a^{test} = \{j \in \mathcal{D}^{test} : A_j = a\}$ for $a \in \mathcal{A}$. The predicted values for their class labels $\{Y_j : j \in \mathcal{D}^{test}\}$ are denoted by $\{\hat{Y}_j : j \in \mathcal{D}^{test}\}$.

2.1 Background: predictive parity in classification

We focus on scenarios where an individual's membership to a particular protected group is known. Group-fairness approaches, which explicitly enforce fairness across groups, have been widely applied across various disciplines, ranging from medicine to the criminal justice system. To provide context for our fairness notion, we start with the widely used predictive parity or sufficiency principle in classification, as discussed in Crisp (2003), Barocas et al. (2017) and Chouldechova (2017). According to this principle, the probability of misclassifying an individual to class c should be equal across all protected groups:

$$\mathbb{P}(Y \neq \hat{Y} | \hat{Y} = c, A = a) \text{ are the same for all } a \in \mathcal{A}. \tag{1}$$

We highlight three primary issues related to machine learning methods developed under the sufficiency principle (Zeng et al. 2022, Pleiss et al. 2017, Zafar et al. 2017). First, the calibration by group method (Barocas et al. 2017), a popular approach for ensuring fair outcomes for subgroups, does not offer a theoretical guarantee on controlling the misclassification rate at a user-specified level. This lack of a guarantee can be particularly problematic in high-stakes decision-making situations. Second, current classification methods only focus on the accuracy of individual classifications, neglecting the complexities that arise when multiple individuals are classified simultaneously. This oversight regarding multiplicity can lead to severe inflation of misclassification errors. Finally, concurrent state-of-the-art classifiers typically exhibit high complexity and analytical intractability, making it difficult to quantify the uncertainties around their predictions. Even when such theoretical analyses are feasible, they often involve strong assumptions about the underlying model and the accuracy of its outputs, which may not hold in practice. In response to these challenges, we propose a comprehensive approach comprising a selective classification framework (Section 2.2), a modified error rate criterion (Section 2.3),

and a novel class of model-free algorithms with strong theoretical guarantees (Sections 3.1-3.2). Together, these components provide a highly effective solution to the identified issues.

2.2 A selective inference framework for binary classification

This article focuses on binary classification problems. The extension to the general multi-class setting is discussed briefly in Section 6.

Consider an application scenario for predicting mortgage default, where $Y = 2$ indicates default and $Y = 1$ otherwise. A common practice is to produce *confidence scores*, denoted $\hat{S}^c(x, a)$ for $c \in \mathcal{C} \equiv \{1, 2\}$, which are generated from a user-specified classifier and risk assessment software. We focus on scores corresponding to the estimated class probabilities of $Y = c$ given the covariates $(X, A) = (x, a)$. The scores satisfy $\hat{S}^1(x, a) + \hat{S}^2(x, a) = 1$. Suppose we need to classify m individuals with confidence scores $\{\hat{S}_j^c \equiv \hat{S}^c(X_j, A_j) : c \in \mathcal{C};\ j \in \mathcal{D}^{test}\}$ into "high," "medium," or "low" risk classes. It is natural to consider a class of rules in the form of

$$\hat{Y}_j \;=\; \sum_{c \in \mathcal{C} \equiv \{1,2\}} c \cdot \mathbb{I}(\hat{S}_j^c > t_c) \;=\; \mathbb{I}(\hat{S}_j^2 < 1 - t_1) + 2 \cdot \mathbb{I}(\hat{S}_j^2 > t_2), \quad \text{for } j \in \mathcal{D}^{test}, \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and the thresholds t_c satisfy $t_c \in (0.5, 1]$ for $c \in \mathcal{C}$.

Remark 1. The constraint $t_c > 0.5$ provides two benefits. First, it enhances interpretability by ensuring that a definitive class assignment occurs only when the confidence score exceeds 50%. Second, it prevents overlapping selections: since $S_j^1 + S_j^2 = 1$, it ensures unique class assignments, effectively avoid overlapping selections.

The predicted label $\hat{Y}_j$ takes three possible values in the action space $\Lambda = \{1, 2, 0\}$, indicating that an individual has low ($\hat{Y}_j = 1$), high ($\hat{Y}_j = 2$), and medium ($\hat{Y}_j = 0$) risks of default, respectively. The value 0, referred to as an "indecision" or "reject option" in classification [cf. Herbei & Wegkamp 2006, Sun & Wei 2011, Lei 2014, Lee et al. 2021], is used to express "doubt," indicating insufficient confidence to make a definitive decision. For example, an individual with $\hat{Y}_j = 1$ will be approved for a mortgage, an individual with $\hat{Y}_j = 2$ will be rejected, while an individual with $\hat{Y}_j = 0$ will receive a pending decision and be asked to provide additional information before resubmitting the application.

Remark 2. We can interpret (2) as a *selective inference* procedure that assigns individuals with extreme scores to high- or low-risk classes while returning an indecision for the remainder. Notably, in this framework the state space $\mathcal{C} = \{1, 2\}$ differs from the action space $\Lambda = \{1, 2, 0\}$, contrasting with the standard classification setup which mandates $\mathcal{C} = \Lambda = \{1, 2\}$. This flexible framework provides a useful interface for practitioners: assigning individuals to Class 1 offers economic benefits by preventing the misallocation of resources to low-risk candidates, thereby reducing study costs associated with unnecessary follow-up. Moreover, the selection for Class 2 is essential for identifying high-risk cases that require further intervention. The selective inference perspective can be employed to handle various types problems including outlier detection and multinomial classification; further discussion is provided in Sections C and G of the Appendix.

2.3 False selection rate and the fairness issue

In practice, it is desirable to avoid erroneous selections, which often have negative social or economic impacts. In the mortgage example, approving an individual who will truly default (i.e., $\hat{Y} = 1$ but $Y = 2$) would increase the financial burden of the lender, while rejecting an individual who will not default (i.e., $\hat{Y} = 2$ but $Y = 1$) would lead to a loss of profit. In situations

where m is large, controlling the inflation of selection errors is a crucial task for policy makers. A practically useful notion is the false selection rate (FSR), which is defined as the expected fraction of erroneous decisions among all definitive decisions. We use the notation $\text{FSR}^{\mathcal{C}'}$, where $\mathcal{C}' \subset \mathcal{C} = \{1, 2\}$ is the set of class labels that we are interested in selecting.

Consider the two-class classification problem with selection rule (2). Denote $\mathcal{S} = \{j \in \mathcal{D}^{test} : \hat{Y}_j \neq 0\}$ the index set of the selected cases and $|\mathcal{S}|$ its cardinality. The FSR that combines the false selections from both classes is given by

$$\text{FSR}^{\{1,2\}} = \mathbb{E}\left[\frac{\sum_{j\in\mathcal{D}^{test}} \sum_{c\in\{1,2\}} \mathbb{I}(\hat{Y}_j = c, Y_j \neq c)}{|\mathcal{S}| \vee 1}\right], \tag{3}$$

where $x \vee y = \max\{x, y\}$, and the expectation $\mathbb{E}$ in Equation (3) [and later on in (4)-(6)] is taken over both the observed data $\{(X_i, A_i, Y_i) : i \in \mathcal{D}\}$ and test data $\{(X_j, A_j, Y_j) : j \in \mathcal{D}^{test}\}$. Let $\mathcal{S}^c = \{j \in \mathcal{D}^{test} : \hat{Y}_j = c\}$ denote the index set of the cases assigned to class c. The FSRs evaluated for individual classes are defined as:

$$\text{FSR}^{\{c\}} = \mathbb{E}\left[\frac{\sum_{j\in\mathcal{D}^{test}} \mathbb{I}(\hat{Y}_j = c, Y_j \neq c)}{|\mathcal{S}^c| \vee 1}\right], \quad c \in \mathcal{C}. \tag{4}$$

Incorporating the option of indecision facilitates the development of a decision rule that can control the FSRs at a user-specified level. However, attaining this objective is challenging within the conventional classification framework, which requires definitive decisions for every individual. As demonstrated in Meinshausen & Rice (2006) and Cai & Sun (2017), if the minimum condition on the classification boundary is not satisfied, it becomes impossible to simultaneously control both FSR^1 and FSR^2 at low levels.

The FSR is a general concept for selective inference that encompasses important special cases such as the misclassification rate, the false discovery rate (FDR, Benjamini & Hochberg 1995), among others. When both the state space and the action space are set to $\{1, 2\}$, thereby eliminating the possibility of indecision, then the FSR defined in (3) simplifies to the misclassification rate $\frac{1}{m}\mathbb{E}\left\{\sum_{j\in\mathcal{D}^{test}}(\hat{Y}_j \neq Y_j)\right\}$. The connection between our FSR framework, one-class classification, and the FDR is discussed in detail in Section C of the Appendix.

In practical scenarios, minimizing the number of indecisions is highly desirable. To quantify this concept, we introduce the expected proportion of indecisions:

$$\text{EPI} = (1/m)\mathbb{E}\Big\{\sum_{j\in\mathcal{D}^{test}} \mathbb{I}(\hat{Y}_j = 0)\Big\} = 1 - \mathbb{E}(|\mathcal{S}|)/m. \tag{5}$$

Under the same FSR level, a smaller EPI corresponds to greater statistical power.

Next we turn to the important fairness issue in selective inference. A major concern is that the rate of erroneous decisions might be unequally shared between the protected groups, as illustrated in the COMPAS example. To address this issue, it is desirable to control the FSR for each protected attribute in A. Therefore, we aim to find a selective classification rule obeying the following constraint on group-wise FSRs:

$$\text{FSR}_a^{\{c\}} = \mathbb{E}\left[\frac{\sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{Y}_j = c, Y_j \neq c)}{\left\{\sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{Y}_j = c)\right\} \vee 1}\right] \leq \alpha_c, \quad \text{for all } a \in \mathcal{A}, \tag{6}$$

where α_c is a user-specified tolerance level, $c \in \{1, 2\}$. The fairness-adjusted error rate constraint (6) equally bounds the fraction of erroneous decisions among protected groups. We aim to develop a selective classification rule that solves the following constrained optimization problem:

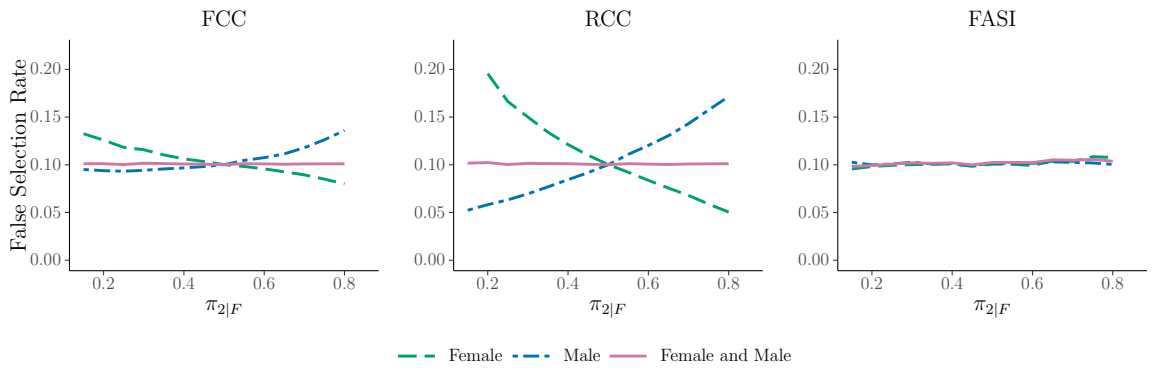

Figure 2: For FCC and RCC, the degree of unfairness increases as $\pi_{2|M}$ and $\pi_{2|F}$ become more disparate. FASI ensures that the group-wise FSRs are effectively controlled and approximately equalized.

$$\text{minimize the EPI subject to } \text{FSR}_a^{\{c\}} \leq \alpha_c, \quad \text{for } c \in \{1,2\} \text{ and } a \in \mathcal{A}. \tag{7}$$

Although our problem formulation (7) only sets upper bounds for group-wise FSR levels, minimizing the EPI enforces the exhaustion of allowable FSR levels for each group. This leads to algorithms that closely align all group-wise FSR levels with the designated nominal level, ensuring comparability of error rates across all protected groups.

Remark 3. Controlling group-wise FSRs (6) does not imply controlling the *overall* FSR (4), which combines individuals from all sensitive groups. While we did not observe violations of the overall FSR in our numerical experiments (see also Section H.2 in the Supplement), developing new methodologies and theories that equalize group-wise FSRs while controlling the overall FSR is challenging, and we leave this for future research.

2.4 The construction of fair classifiers: issues and roadmap

We investigate the important issue of what makes a "fair" classifier. In most classification tasks, the standard operation is to first construct a confidence score, and then secondly to turn this score into a decision by setting a threshold. Consider selection rule (2). We present two approaches for constructing confidence scores. The notation $S(x,a)$ is used instead of $\hat{S}(x,a)$ to indicate the ideal setting in which an oracle, possessing knowledge of the true data-generating model, computes the scores analytically without estimation.

The two approaches, respectively referred to as the "full covariate classifier" (FCC) and "reduced covariate classifier" (RCC), employs the following scores:

$$S_j^{c,FCC}(x,a) = \mathbb{P}\left(Y_j = c | X_j = x, A_j = a\right), \tag{8}$$

$$S_j^{c,RCC}(x) = \mathbb{P}\left\{Y_j = c | X_j = x\right\}, \tag{9}$$

for $c \in \{1,2\}$ and $j \in \mathcal{D}^{test}$. Consider the high-risk class $c = 2$. Then $S_j^{2,\text{FCC}}(x,a)$ denotes the (oracle) class probability of an individual belonging to class 2 based on all available covariates. In contrast, $S_j^{2,\text{RCC}}(x)$ is employed to estimate the same probability after removing the sensitive attribute from the covariate set. However, as will be demonstrated, both the FCC and RCC approaches may be inadequate for effectively addressing the fairness concern.

Consider the mortgage example where we simulate a data set that contains a sensitive attribute "gender". The goal is to select individuals into the high risk class with FSR control at 10%; the simulation setup is detailed in Section 4. We highlight here that the proportions

of individuals with label "2" are different across the protected groups: for the male group, the proportion of individuals with label "2", denoted as $\pi_{2|M}$, is fixed at 50%, whereas for the female group the proportion $\pi_{2|F}$ varies from 15% to 85%.

We apply the FCC approach and plot the overall FSR and group-wise FSRs as functions of $\pi_{2|F}$ on the left panel of Figure 2. We can see that FCC controls the overall FSR but not the group-wise FSRs. Hence thresholding rules based on (8) are harmful in the sense that the burden of erroneous decisions is not shared equally among the two gender groups. The RCC approach can be harmful as well, as illustrated in the middle panel of Figure 2. While the overall FSR is still controlled at 10%, the issue of unfairness is in fact aggravated rather than mitigated, with widened gaps in the group-wise FSRs. In addition, the RCC approach has two further drawbacks. Firstly, disregarding a sensitive attribute can result in significant power loss. Secondly, if feature X is highly predictive of sensitive attribute A, then the RCC approach can still lead to unfair decisions due to the issue of *surrogate encoding* (Kusner et al. 2017, Long & Albert 2021). Concretely, in fairness research, surrogate encoding pertains to the circumstance where the sensitive attribute A is absent from the list of predictors, but its information is encoded or concealed within other predictors, causing A to still influence the outcome Y.

We emphasize that the patterns in Figure 2 are not specific to any particular classification algorithm but indicate a systematic bias. In our simulation, where perfect scores are available, the unfairness depicted in Figure 2 still persists. In contrast, our proposed FASI algorithm, shown in the right panel of Figure 2, effectively controls the FSR and nearly equalizes the error rates across all protected groups.

3 Methodology

This section develops a fairness-adjusted selective inference (FASI) procedure for two-class classification with state space $\mathcal{C} = \{1, 2\}$ and action space $\Lambda = \{0, 1, 2\}$. We focus on the error rate $\text{FSR}_a^{\{c\}}$ defined in (6), which is more relevant for addressing fairness issues in high-stakes decision-making scenarios. The methodologies for the more complex tasks of controlling $\text{FSR}^{\{1,2\}}$ and performing multinomial classification are briefly discussed in Section 6 and Appendix G.

A major challenge in our methodological development is that many state-of-the-art machine learning algorithms are complex and offer no performance guarantees on their outputs. This limitation renders uncertainty quantification and error rate control challenging, if not intractable. To address this issue, we develop a model-free framework that is applicable to any black-box algorithm and relies solely on the exchangeability of the data points.

3.1 The R-value and FASI algorithm

We first introduce a significance index, called the R-value, for ranking individuals and then discuss how the R-values can be employed for selective classification.

The R-value is computed via the FASI algorithm, which consists of three steps: training, calibrating and thresholding. The observed data set $\{(X_i, A_i, Y_i) : i \in \mathcal{D}\}$ is randomly divided into a training set and a calibration set: $\mathcal{D} = \mathcal{D}^{train} \cup \mathcal{D}^{cal}$. Let $\mathcal{D}_a^{test} = \{i \in \mathcal{D} : A_i = a\}$ and $\mathcal{D}_a^{cal} = \{i \in \mathcal{D}^{cal} : A_i = a\}$ for $a \in \mathcal{A}$. Denote $n_a = |\mathcal{D}_a^{cal}|$ and $m_a = |\mathcal{D}_a^{test}|$.

In the first step, we train score functions $\hat{S}^c(x, a)$, $c \in \{1, 2\}$, using data $\{(X_i, A_i, Y_i) : i \in \mathcal{D}^{train}\}$. The scores, representing estimated class probabilities, can be generated from any user-specified classifier satisfying $\hat{S}^1(x, a) + \hat{S}^2(x, a) = 1$. We make no assumptions on the accuracy of these scores.

In the second step, we use the scores $\{\hat{S}_i^c := \hat{S}^c(X_i, A_i) : i \in \mathcal{D}^{cal} \cup \mathcal{D}^{test}\}$ to calculate

$$Q_k^c = \sum_{a\in\mathcal{A}} \mathbb{I}(A_k = a) \cdot \frac{\left\{\sum_{i\in\mathcal{D}_a^{cal}} \mathbb{I}\big(\hat{S}_i^c \geq \hat{S}_k^c, Y_i \neq c\big) + 1\right\}/(n_a+1)}{\left\{\sum_{i\in\mathcal{D}_a^{test}} \mathbb{I}\big(\hat{S}_i^c \geq \hat{S}_k^c\big)\right\}/m_a} \wedge 1, \tag{10}$$

for $k \in \mathcal{D}^{cal} \cup \mathcal{D}^{test}$. The operation $\wedge$ indicates that Q_k^c is set to 1 if it exceeds 1. As discussed in Section 3.2, Q_k^c represents the estimated fraction of false selections among all selections using the cutoff $\hat{S}_k^c$; a lower value of Q_k^c indicates greater confidence in classifying the kth individual to class c. To enhance the algorithm's stability, one may modify (10) to include both the calibration and test data in the denominator:

$$Q_k^c = \sum_{a\in\mathcal{A}} \mathbb{I}(A_k = a) \cdot \frac{\left\{\sum_{i\in\mathcal{D}_a^{cal}} \mathbb{I}\big(\hat{S}_i^c \geq \hat{S}_k^c, Y_i \neq c\big) + 1\right\}/(n_a+1)}{\left\{\sum_{i\in\mathcal{D}_a^{cal}\cup\mathcal{D}_a^{test}} \mathbb{I}\big(\hat{S}_i^c \geq \hat{S}_k^c\big) + 1\right\}/(n_a+m_a+1)} \wedge 1. \tag{11}$$

The subsequent steps of the algorithm are identical whether (10) or (11) is used, so we denote both by Q_k^c and provide a unified discussion.

Remark 4. The adjustment in Equation (11) is useful when m_a is small; numerical evidence in Section H.1 of the Appendix demonstrates the benefits of a larger sample size (i.e. $m_a + n_a$) in calibrating Q_k^c. Although (11) offers numerical advantages, it introduces additional theoretical complexity; accordingly, we develop separate theories for (10) and (11) in Theorem 1 below.

In practical scenarios, higher-scoring individuals may not consistently correspond to smaller Q_k^c values. To eliminate this inconsistency, we propose the following monotonicity adjustment:

$$\tilde{Q}_j^c = \sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a) \cdot \min_{\{k\in\mathcal{D}_a^{cal}\cup\mathcal{D}_a^{test}:\hat{S}_k^c \leq \hat{S}_j^c\}} Q_k^c, \quad j \in \mathcal{D}^{test}. \tag{12}$$

The R-value, with more explanations provided in Remark 5 below, is defined as

$$R_j^c = \max\left\{\mathbb{I}(\hat{S}_j^c \leq 0.5),\ \tilde{Q}_j^c\right\}, \quad \text{for } c \in \{1,2\} \text{ and } j \in \mathcal{D}^{test}. \tag{13}$$

In the third step, we compare the R-values against the designated level α_c:

$$\hat{Y}_j = \sum_{c\in\{1,2\}} c \cdot \mathbb{I}(R_j^c \leq \alpha_c), \quad j \in \mathcal{D}^{test}. \tag{14}$$

Remark 5. In binary classification, the jth individual is associated with two R-values, R_j^1 and R_j^2. Definition (13) provides a crucial adjustment to ensure that, in effect, only one R-value is used for decision-making. Specifically, if $\hat{S}_j^c \leq 0.5$, we set $R_j^c = 1$. Since $0 < \alpha_c < 1$ is a small constant, this adjustment guarantees that any $R_j^c = 1$ is effectively discarded, meaning the jth individual is never assigned to class c when $\hat{S}_j^c \leq 0.5$. Moreover, because $\hat{S}_j^1 + \hat{S}_j^2 = 1$, one of the two R-values in $\{R_j^c : c = 1, 2\}$ must equal 1, thereby preventing overlapping selections. Finally, if both R-values exceed α_c, we output an indecision, $\hat{Y}_j = 0$.

The FASI algorithm, summarized in Algorithm 1, offers several attractive properties. First, the R-value serves as an estimate of a proportion, making it easily interpretable and comparable across groups. Second, the FSR analysis based on R-values is straightforward: practitioners can directly make decisions by comparing the R-values with a user-specified FSR level. Third, fairness notion is integrated into the R-value. As demonstrated in Lemma 1 in Section D.1 of the Supplement, the decision rule in (14) involves finding the smallest group-wise threshold for $\hat{S}_j^c$ that satisfies $\mathrm{FSR}_a^{\{c\}} \leq \alpha_c$, thereby approximately aligning the group-wise FSR levels with

Algorithm 1 The FASI Algorithm

Input: $\{(X_i, A_i, Y_i) : i \in \mathcal{D}\}$, $\{(X_j, A_j) : j \in \mathcal{D}^{test}\}$, FSR levels $\{\alpha_c : c = 1, 2\}$.
Output: a selective classification rule $\{\hat{Y}_j \in \{0, 1, 2\} : j \in \mathcal{D}^{test}\}$.

1: Randomly split $\mathcal{D}$ into $\mathcal{D}^{train}$ and $\mathcal{D}^{cal}$.
2: Train a machine learning model on $\{(X_i, A_i, Y_i) : i \in \mathcal{D}^{train}\}$.
3: Predict confidence scores $\hat{S}_i^c$ for $i \in \mathcal{D}^{cal} \cup \mathcal{D}^{test}$.
4: Compute the R-values $\{R_j^c : c = 1, 2; j \in \mathcal{D}^{test}\}$ according to Equations (10) to (13).
5: Make decisions by comparing the R-values with α_c using (14), for all $j \in \mathcal{D}^{test}$.

the nominal level. Finally, FASI is model-free, providing a robust framework for FSR control, as discussed in the next subsection.

Remark 6. There are two potential strategies to achieve fairness across protected groups. The first strategy, as adopted in the FASI algorithm, involves modifying the current confidence scores to generate new scores (R-values) that are directly comparable across groups. The second strategy, on the other hand, involves retaining the original confidence scores and implementing group-adjusted thresholds. As demonstrated in Lemma 1 in Appendix D.1, this strategy is mathematically equivalent to the first. However, in practical applications, this approach may be seen as confusing or even controversial because it applies different thresholds to various protected groups. Such disparate treatment is difficult to interpret and could be perceived as introducing an alternative form of discrimination. In contrast, FASI employs a universal threshold for all individuals, with the R-value serving as a statistical wrapper that distills complex factors – such as error rate control and fairness – into a single, easy-to-use index.

3.2 Why FASI works?

We start by explaining why the R-value provides a sensible estimate of the FSR. To simplify the discussion, we focus on a specific group $A = a$ and consider a thresholding rule of the form $\{\mathbb{I}(\hat{S}_j^c \geq t) : j \in \mathcal{D}_a^{test}\}$. Consider the false selection proportion (FSP) process:

$$\mathrm{FSP}_a^{\{c\}}(t) = \frac{\sum_{j \in \mathcal{D}_a^{test}} \mathbb{I}\left(\hat{S}_j^c \geq t, Y_j \neq c\right)}{\left\{\sum_{j \in \mathcal{D}_a^{test}} \mathbb{I}(\hat{S}_j^c \geq t)\right\} \vee 1}, \tag{15}$$

with $\mathrm{FSP}(t) = 0$ if no individual is selected. The FSP cannot be computed from data because we do not observe the true states $\{Y_j : j \in \mathcal{D}_a^{test}\}$. The effectiveness of the FASI algorithm relies on the following exchangeability condition:

Assumption 1. *The data points* $\{(X_i, Y_i) : i \in \mathcal{D}_a^{cal} \cup \mathcal{D}_a^{test}\}$ *are exchangeable for all* $a \in \mathcal{A}$.

As FASI uses the same fitted model to compute the scores (Assumption 1), the confidence scores are exchangeable. Consequently, the unobserved process $\sum_{j \in \mathcal{D}_a^{test}} \mathbb{I}\left(\hat{S}_j^c \geq t, Y_j \neq c\right)$ is strongly resembled by its "mirror process" in the calibration data $\sum_{i \in \mathcal{D}_a^{cal}} \mathbb{I}\left(\hat{S}_i^c \geq t, Y_i \neq c\right)$. Constructing a mirror process and exploiting its symmetry for inference is a powerful idea that has been explored in recent works (cf. Barber & Candès 2015, Weinstein et al. 2017, Lei & Fithian 2018, Leung & Sun 2022, Du et al. 2023). To account for the unequal sample sizes between $\mathcal{D}_a^{cal}$ and $\mathcal{D}_a^{test}$, we derive the mirror FSP process as follows:

$$\widehat{\mathrm{FSP}}_a^{\{c\}}(t) = \frac{\left\{\sum_{i\in\mathcal{D}_a^{cal}} \mathbb{I}(\hat{S}_i^c \geq t, Y_i \neq c) + 1\right\}/(n_a+1)}{\left\{\sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{S}_j^c \geq t)\right\}/m_a}. \tag{16}$$

This provides insight into why (10) (and similarly (11)) has been employed in constructing the R-value. When computing the R-values, (16) is only evaluated at values in $\{S_j^c : j \in \mathcal{D}^{test}\}$; hence the denominator is always greater or equal to 1.

Remark 7. When inspecting the numerator in (16), we observe that a "+1" has been included in $\sum_{j\in\mathcal{D}_a^{cal}} \mathbb{I}\big(\hat{S}_j^c \geq t, Y_j \neq c\big)$ and n_a. This technical adjustment has only a minor impact on the empirical performance of FASI but guarantees that (16) corresponds to a martingale, which is crucial for proving the theory.

The FSP process (15) and its mirror process (16) together provide an intuitive interpretation of the R-value. Roughly speaking, the R-value represents the smallest estimated FSP at which the i^{th} individual is just selected. In other words, if we set the threshold at $R = r$ and select all individuals with R-values less than or equal to r into class c, then we expect that, for every group $a \in \mathcal{A}$, approximately $100r\%$ of the selections will be incorrect decisions. The fairness notion is inherently integrated into the R-value, enabling the calibration of a universal threshold to align all group-wise FSRs with the nominal level. Moreover, our interpretation resembles the q-value (Storey 2003) in FDR analysis; further details are provided in Section C of the Supplement. We emphasize that while Storey's q-value relies on the empirical distribution of p-values, our R-value is derived from calibration data through a carefully designed mirror process.

3.3 Theory on FSR control

We now present a theorem establishing the validity of FASI for FSR control. Our theory differs from existing work in that we make no assumptions about the accuracy of $\hat{S}_i^c$. Instead, the accuracy of the scores influences only the power of FASI, leaving its validity (for FSR control) unaffected. Practical guidelines on constructing more accurate confidence scores (and, hence, effective R-values) are provided in Section 3.5 and Section B.4 of the Supplement.

Theorem 1. *Define* $\gamma_{c,a} = \mathbb{E}\left(p_{c,null}^{test,a}/p_{c,null}^{cal,a}\right)$, *where* $p_{c,null}^{test,a}$ *and* $p_{c,null}^{cal,a}$ *are the empirical proportions of individuals in group* a *that do not belong to class* c *in the test and calibration data, respectively. Then under Assumption 1, for all* $a \in \mathcal{A}$, *we have:*

(a). The FASI algorithm with R-value defined via (10), (12) and (13) satisfies $FSR_a^{\{c\}} \leq \gamma_{c,a}\alpha_c$.

(b). The stable version of the FASI algorithm with R-value defined via (11) – (13) satisfies

$$FSR_a^{\{c,*\}} \leq \gamma'_{c,a}\alpha_c + \frac{\alpha}{2}\mathbb{E}\left|\textit{RES}(\tau_a^c) - 1\right|, \tag{17}$$

where $\gamma'_{c,a}$ *is a constant,* τ_a^c *is a stopping time, both defined in Appendix D.2,*

$$FSR_a^{\{c\},*} = \mathbb{E}\left[\frac{\sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{Y}_j = c, Y_j \neq c)}{\sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{Y}_j = c) + 1}\right], \quad \textit{and} \tag{18}$$

$$\textit{RES}(\tau_a^c) = \frac{\sum_{i\in\mathcal{D}_a^{cal}} \mathbb{I}(\hat{S}_i^c \geq \tau_a^c, Y_i = c)}{\sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{S}_j^c \geq \tau_a^c, Y_j = c) + 1} \cdot \frac{\sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{S}_j^c \geq \tau_a^c, Y_j \neq c)}{\sum_{i\in\mathcal{D}_a^{cal}} \mathbb{I}(\hat{S}_i^c \geq \tau_a^c, Y_i \neq c) + 1}. \tag{19}$$

Assumption 1 on exchangeability implies that $\gamma_{c,a}$ is typically close to 1, allowing nearly exact control in Part (a), as confirmed by our numerical studies (Section H.3 of the Supplement). Moreover, $\gamma'_{c,a}$ in Part (b) is also close to 1 (cf. Remark 13 in Appendix D.2). Assumption 1 implies that both terms in the product on the right-hand side of (19) are stochastically close to 1. To provide a more rigorous characterization of the upper bound on the FSR level in Part (b), Section D.3 of the Supplement presents an asymptotic analysis that specifies sufficient conditions for the strong convergence of τ_a^c. Specifically, let $\tau^* \in (0,1)$ be a constant. If $\tau_a^c \xrightarrow{a.s.} \tau^*$, one can show that $\lim_{(n_a,m_a)\to\infty} \mathbb{E}\,|\mathtt{Res}(\tau_a^c) - 1| = 0$. Hence, (17) indicates that the stable version of FASI controls the FSR at $\alpha + o(1)$.

Remark 8. In the modified FSR definition (18), the "+1" adjustment is utilized. A similar modification was employed in Theorem 1 of Barber & Candès (2015), but for different reasons. The difference between $\mathrm{FSR}_a^{\{c\},*}$ and $\mathrm{FSR}_a^{\{c\}}$ is usually negligible. Moreover, in Section A, we present a corollary demonstrating that a conservative version of the R-value guarantees FSR control below α, eliminating $\gamma_{c,a}$ from the bound. However, the conservative version tends to result in a higher proportion of indecisions. Hence we recommend the R-value (13), which is more powerful while providing almost exact control in practical situations.

Three major challenges in proving Theorem 1 are (i) handling the dependence between the scores $\hat{S}_i^c$ (since the same training data were used to compute (16)), and (ii) evaluating the FSR without any knowledge about the quality of the scores. Inspired by elegant ideas in the FDR literature (Storey et al. 2004, Barber & Candès 2015), we have carefully designed the R-values so that the corresponding FSP process (16) is stochastically dominated by a supermartingale. We then apply the optional stopping theorem and leverage the exchangeability assumption to establish an upper bound for the FSR. We stress that, in Theorem 1, part (a) guarantees validity in finite samples, and both parts (a) and (b) do not rely on any assumptions regarding the underlying models or the quality of the scores.

3.4 Connections to existing work

This section explores the connections and distinctions between FASI and existing methods developed under the sufficiency principle in fairness research. Additionally, we provide insights about recent developments in conformal inference relevant to FASI.

Our formulation in (6) is closely related to the sufficiency principle (1) in the fairness literature, but it overcomes several of its limitations. First, (6) operates within a selective classification framework by offering an indecision option for cases requiring further review, thereby enabling effective error rate control at user-specified levels. Second, we define the FSR notion to aggregate decision errors over m new individuals, which addresses the sufficiency principle's limitation of only pertaining to the error rate of an individual decision. Lastly, many algorithms developed under the sufficiency principle are complex and computationally intensive, lacking finite sample guarantees when applied to outputs from black-box models. In contrast, the FASI algorithm effectively controls the FSR in finite samples without relying on assumptions about the underlying model, classification algorithm, or score accuracy. A detailed comparison with related works, including Zeng et al. (2022) and Lee et al. (2021), is provided in Section F of the Supplement.

The R-value has a compelling interpretation under the conformal inference framework (Vovk et al. 2005, Lei & Wasserman 2014). In Section C of the Supplement, we show that a variation of our R-value corresponds to the Benjamini-Hochberg (BH) adjusted q-value of the conformal p-values (Mary & Roquain 2022, Bates et al. 2023) under the one-class classification setting (Moya & Hush 1996, Khan & Madden 2009, Kemmler et al. 2013). The connection to conformal

inference and the BH method provides valuable insights into why the FASI algorithm is model-free and offers effective FSR control in finite samples, as claimed in Theorem 1.

The theory presented in Bates et al. (2023) encounters a complication similar to ours, as the conformal p-values are dependent. To address this, Bates et al. (2023) first shows that the conformal p-values satisfy the condition of positive regression dependence on a subset (PRDS) and then applies the theory in Benjamini & Yekutieli (2001) to establish the validity of FDR control. While we conjecture that the PRDS approach may be relevant, its extension to our specific context is non-trivial because our R-values do not explicitly utilize conformal p-values under the binary classification setup. Therefore, our martingale-based theory appears to be a suitable and equally effective alternative. Moreover, incorporating conformal p-values—which rely on one-class classifiers—directly into our binary classification problem would entail discarding labeled outliers and consequently lead to information loss; this issue has been explored in a recent study by Liang et al. (2024).

Our mirror process leverages a calibration set containing data from both classes, unlike the counting knockoff approach (e.g., Weinstein et al. 2017, Bates et al. 2023), which relies solely on null training data (see Section C of the Supplement for further discussion). Using data from both classes eliminates the need for Storey's adjustment, which is required by both the counting knockoff and conformal BH methods (Bates et al. 2023, Jin & Candes 2023) to mitigate the conservativeness of the BH procedure. Additionally, our method addresses fairness in FSR control – a topic that the aforementioned conformal methods have not explored.

3.5 Theoretical R-value and optimality theory

We briefly discuss the theoretical R-value and its optimality theory, which extends the work of Sun & Cai (2007) and Cai et al. (2019) from multiple testing to selective binary classification. Details are deferred to Section B of the Supplement due to space limitations. Despite being developed under an idealized setup, the theory offers practical insights for training score functions to construct more powerful R-values that aim to minimize the number of indecisions while controlling the FSR rates for all sensitive groups. We emphasize two key messages.

First, the choice of an optimal score function indicates that, during the training stage, we should utilize all features, including the sensitive attribute A, to best capture individual-level information. Scores trained without the sensitive attribute are suboptimal. Fairness adjustments should not be made during the training stage but rather in the calibration stage, where the fully informative scores can be converted into R-values to adjust the disparity in error rates across groups. This strategy shares the same spirit as the selection-by-prediction or learn-then-test framework advocated by Jin & Candes (2023) and Angelopoulos et al. (2025).

Second, the optimal selection rule equalizes group-wise error rates. To minimize the EPI, the pre-specified marginal FSR (mFSR, defined in the Appendix, Equation B.7) must be exhausted in every group, making the mFSRs equal to the nominal level. In other words, the constrained optimization formulation (6) leads to asymptotic equality of error rates. Our numerical studies support this claim, although a complete analysis is hindered by the dependence among scores, which we leave for future research.

4 Simulations

This section presents two simulations under the binary classification setup. The objective is to compare the performance of FASI against the Full Covariate Classifier (FCC). We did not include the Restricted Covariate Classifier (RCC) in these simulations, as RCC has consistently demonstrated larger deviations from the target group-wise FSR levels. We demonstrate that

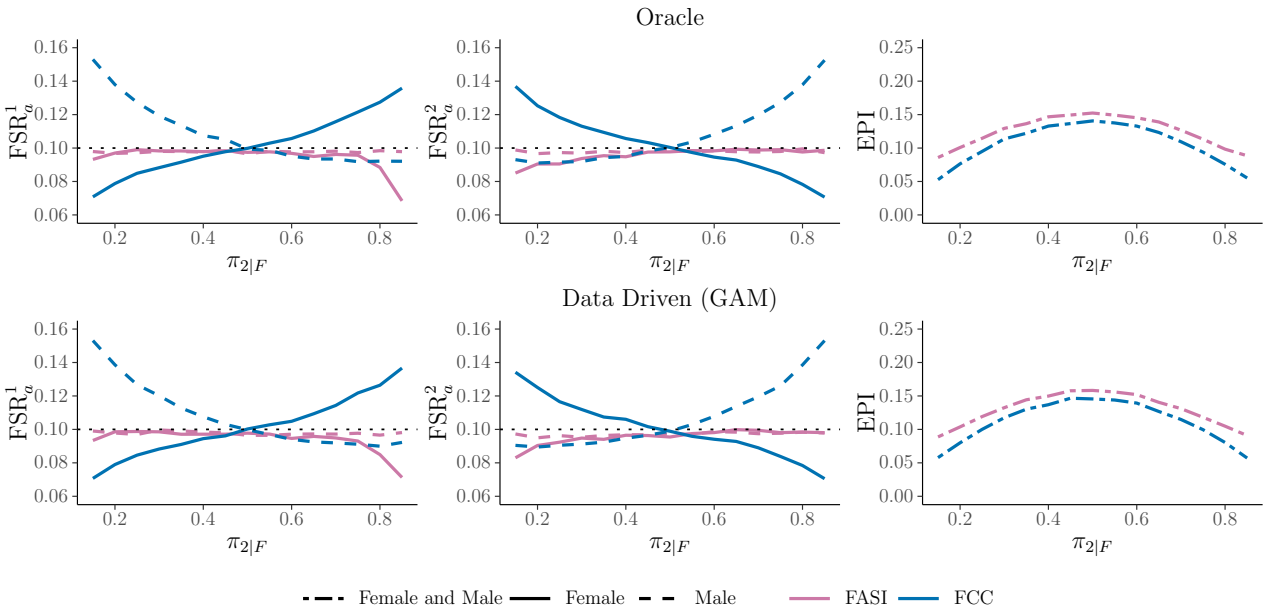

Figure 3: Simulation 1. Top row: the oracle procedure. Bottom row: the data-driven procedure using GAM. Left and middle columns: FSR_a^1 and FSR_a^2 levels for both females and males. Right column: the EPI levels.

both the oracle and data-driven versions of FASI can control the group-wise FSRs, while RCC fails to do so. The oracle versions of FASI and FCC use the exact class probabilities, defined in Equation 8, while the data-driven procedures employ the softmax scores via the GAM method (Hastie et al. 2009, James et al. 2023, Chen & Guestrin 2016).

In all simulations, we set $|\mathcal{D}^{train}| = 1{,}500$, $|\mathcal{D}^{cal}| = 1{,}000$ and $|\mathcal{D}^{test}| = 1{,}000$. Gender is our protected attribute taking two values $A = F$ (females) and $A = M$ (males). The feature vectors $\boldsymbol{X} \in \mathbb{R}^3$ are simulated according to the following model:

$$F(\cdot) = \pi_M\{\pi_{1|M}F_{1,M}(\cdot) + \pi_{2|M}F_{2,M}(\cdot)\} + \pi_F\{\pi_{1|F}F_{1,F}(\cdot) + \pi_{2|F}F_{2,F}(\cdot)\}, \tag{20}$$

where $\pi_a = \mathbb{P}(A = a)$, $\pi_{c|a} = \mathbb{P}(Y = c|A = a)$ and $F_{c,a}$ is the conditional distribution of $\boldsymbol{X}$ given $Y = c$ and $A = a$. Let $\pi_M = 0.5$, and $\pi_F = 1 - \pi_M = 0.5$. The Supplement (Section H.2) includes a setup with markedly imbalanced group sizes (e.g., $\pi_M \gg \pi_F$). Although only the GAM method is employed in our simulation, we report that our findings remain consistent regardless of the specific learning algorithms utilized. For a comparison of different machine learning algorithms, please refer to Section H.4 of the Supplement. We consider two scenarios.

In the first scenario, the conditional distributions of $\boldsymbol{X}$ given class Y are assumed to be multivariate normal and are identical for males and females:

$$F_{1,M} = F_{1,F} = \mathcal{N}(\boldsymbol{\mu}_1, 2 \cdot \mathbf{I}_3), \quad F_{2,M} = F_{2,F} = \mathcal{N}(\boldsymbol{\mu}_2, 2 \cdot \mathbf{I}_3),$$

where $\mathbf{I}_3$ is a 3×3 identity matrix, $\boldsymbol{\mu}_1 = (0, 1, 6)^\top$ and $\boldsymbol{\mu}_2 = (2, 3, 7)^\top$. The only difference between the group-wise distributions lies in the conditional proportions: we fix $\pi_{2|M} = \mathbb{P}(Y = 2|A = M) = 0.5$, while varying $\pi_{2|F} = \mathbb{P}(Y = 2|A = F)$ from 0.15 to 0.85. We shall see that in the asymmetric situation (i.e., when $\pi_{2|F}$ is very large or small), the unadjusted FCC rule leads to unfair policies (i.e. we observe disparate FSRs across the male and female groups).

We simulate 1,000 data sets and apply both the FCC and FASI methods at an FSR level of 0.1 to these simulated data sets. The FASI method is implemented with R-values defined

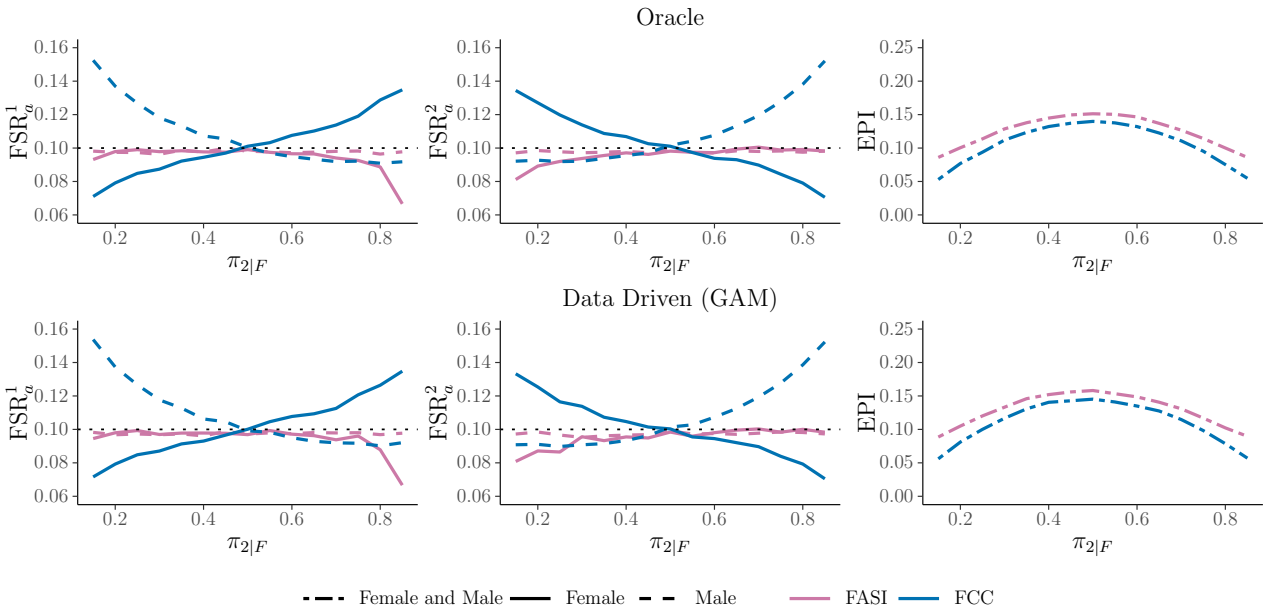

Figure 4: Simulation 2. Comparable setup to Simulation 1 except that the female and male distributions now differ from each other.

in (11)–(13). For the FCC method, the protected attributes are ignored when computing the fractions in (11), and these fractions are denoted as $Q_k^{c,\text{FCC}}$. Then, the $Q_k^{c,\text{FCC}}$ values are adjusted according to (13) to obtain the R-values, denoted as $R_j^{c,\text{FCC}}$. The corresponding selection rule is

$$\hat{Y}_j^{\text{FCC}} = \sum_{c\in\{1,2\}} c \cdot \mathbb{I}\big(R_j^{c,\text{FCC}} \le 0.1\big), \quad j \in \mathcal{D}^{test}.$$

The FSR levels are computed by averaging the respective false discovery proportions (FSPs) from 1,000 replications. The simulation results are summarized in Figure 3. The first and second rows respectively correspond to the oracle and data-driven versions of each method. The first two columns respectively plot the group-wise FSRs for class 1 and class 2 as functions of $\pi_{2|F}$. The final column plots the EPI (5), obtained by averaging the results from 1,000 replications. The following patterns can be observed.

- FCC fails to control the group-wise FSRs. As $\pi_{2|F}$ moves away from $\pi_{2|M} = 0.5$, the gap between the FSR control for Females and Males dramatically widens due to the asymmetry in the proportions of the signals (true class 2 observations) in the male and female groups.
- Both the oracle and data-driven FASI procedures consistently control the FDR at the nominal level. However, when $\pi_{2|F}$ is high, the number of selections decreases, resulting in a reduced total number of selections from both groups. Consequently, both methods exhibit increased conservativeness. This pattern can be attributed to the conservative nature of the R-value, which includes a "+1" adjustment and functions as an estimate of the true false selection proportion: the level of conservativeness becomes more pronounced as the proportion $\pi_{2|F}$ become close to either 0 or 1.
- Both oracle and data-driven FASI algorithms are able to roughly equalize the group-wise FSRs between the Female and Male groups. The data-driven FASI is able to closely mirror the behavior of the oracle method.

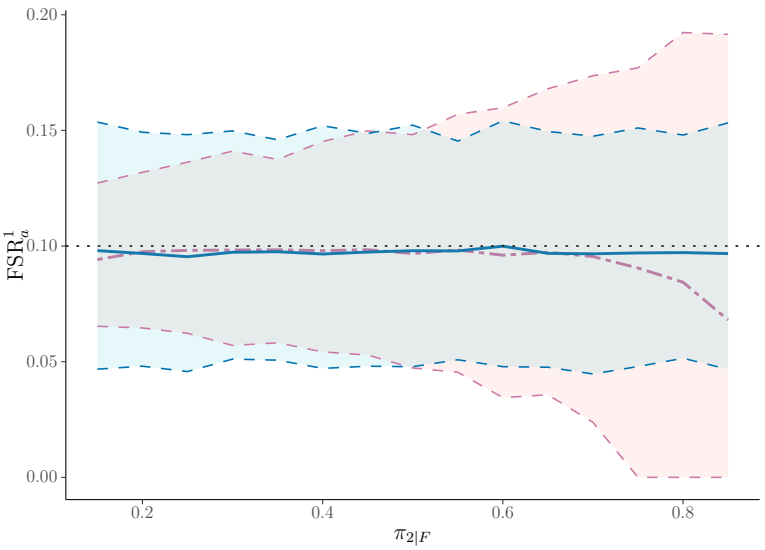

Figure 5: 90% quantiles of the FSPs. Red region: female group; Blue region: male group.

- The parity in FSR control is achieved at the price of slightly higher EPI levels.

Our second simulation considers the setting where $F_{c,M} \neq F_{c,F}$. Denoting the mean for class c and protected attribute a as $\boldsymbol{\mu}_{c,a}$, the data is generated from $F_{c,a} = \mathcal{N}(\boldsymbol{\mu}_{c,a}, 2 \cdot \mathbf{I}_3)$, with components $\boldsymbol{\mu}_{1,M} = (0,1,6)^\top$, $\boldsymbol{\mu}_{2,M} = (2,3,7)^\top$, $\boldsymbol{\mu}_{1,F} = (1,2,7)^\top$ and $\boldsymbol{\mu}_{2,F} = (3,4,8)^\top$. In all other respects Simulations 1 and 2 are identical. The results for the second simulation scenario are provided in Figure 4. We notice very similar patterns to our first simulation setup. FASI controls the group-wise FSRs for all values of $\pi_{2|F}$ while the FCC fails to do so. The data-driven FASI closely emulates the oracle procedure, for both the FSR and EPI levels.

Finally, we examine the variability of the false discovery proportions (FSP), which can fluctuate across replications. Specifically, the FSR is derived as the average of the FSPs. While our theory ensures that the FSR can be controlled under the nominal level α, it is important to note that the FSP may deviate significantly from α. To investigate this variability, we focus on the same experimental setting in Simulation 1 used to generate Figure 3, and present the 90% quantiles of the group-wise FSPs. The summarized results are depicted in Figure 5.

The group-wise FSRs, represented by solid blue and dot-dashed red lines, are effectively controlled at the desired 10% level. The quantiles are visually depicted by blue/red regions, corresponding to the male/female groups, respectively. For the male group, where $\pi_{2|M}$ remains constant, the 90% quantiles range between 5% and 15%. In contrast, the FSP variability for the female group is more pronounced, with greater variability when $\pi_{2|F}$ is larger, as few selections are made from the female group.

5 Real Data Examples

This section demonstrates the application of FASI on two real data sets. Sections 5.1 and 5.2 respectively analyze the COMPAS data (Angwin et al. 2016) and US census data (Dua & Graff 2017). For the COMPAS and census data, we have employed GAM and Adaboost models, respectively, to construct confidence scores. It is important to note that users have the flexibility to choose the best model for their specific application by utilizing their own training data. To facilitate the implementation of FASI with user-specified models, the R package `fasi` has been developed and is readily available on CRAN.

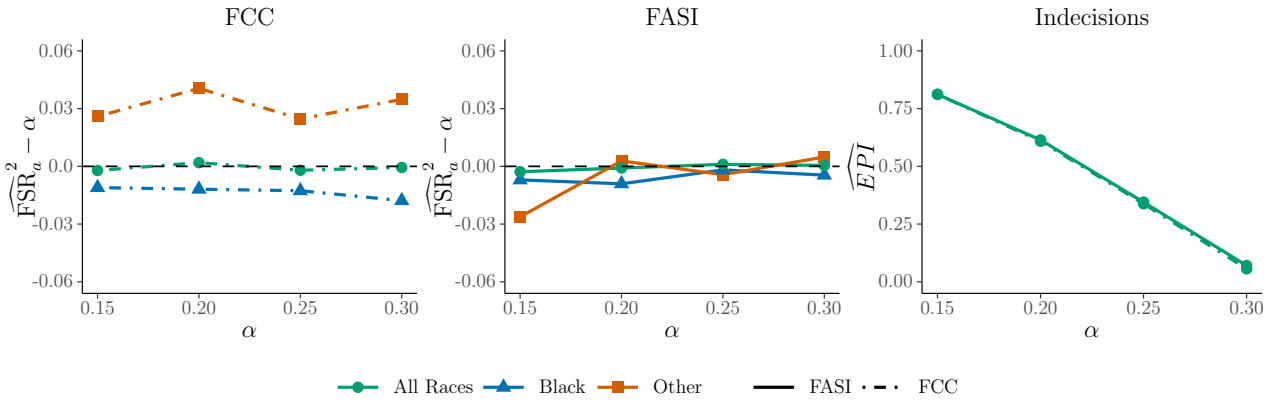

Figure 6: COMPAS data analysis for predicting recidivists. Left and Middle: False Selection Rate minus the desired control level for varying levels of α for the FCC and FASI method respectively. Right: The EPI for both the FCC and FASI method.

5.1 COMPAS data analysis

In 2016, ProPublica's investigative journalists curated a data set of 6,172 individuals, where 3,175 were Black and the remaining 2,997 belonged to other racial categories, who had been arrested in Broward County, Florida. These racial categories, Black and Other, serve as our protected attributes in this study. Within the data set, the "Black" group consisted of 1,773 individuals who were identified as having recidivated within the 2-year time frame considered in the study, while the "Other" group consisted of 1,217 individuals who also recidivated during this period. This 2-year window was chosen as a proxy for the true label of identifying recidivists.

All individuals were assigned a risk score by the COMPAS algorithm (a whole number between 1 and 10) developed by NorthPointe Inc. This score was used to inform the judge of each person's risk of recidivating during their bail hearing. The data set contains demographic information about each person including their race, age, number of previous offenses, sex, number of prior offenses, and their assigned COMPAS risk score.

In this analysis, our objective is to utilize FASI to address potential disparities in FSRs among different racial groups. The literature has extensively examined various fairness notions, such as disparate treatment (Zafar et al. 2017), as well as studies specifically related to the COMPAS data set (Angwin et al. 2016, Dieterich et al. 2016). While our approach is highly relevant and sensible in this context where high-stake consequences are involved, it is crucial to carefully evaluate and scrutinize its implementation, bearing in mind the societal trade-offs associated with different definitions of fairness.

We performed 100 random splits of the data set, where for each protected group and class label Y (our proxy for recidivism), 90% of the data was assigned to $\mathcal{D}$ and the remaining 10% to $\mathcal{D}^{test}$. Furthermore, we evenly split $\mathcal{D}$ into $\mathcal{D}^{train}$ and $\mathcal{D}^{cal}$. To assess the performance, we present the results across a range of α values from 0.15 to 0.30. The first two columns in Figure 6 illustrate the difference between the true and target FSRs for the FCC and FASI algorithms, respectively. The last column of the figure plots the EPI levels.

While the FCC approach effectively controls the overall FSR, it falls short in controlling the FSRs across different racial groups. In the left panel of Figure 6, we can observe that the race-wise FSRs deviate substantially from the nominal level, and the FSR levels for the Black group are significantly lower compared to those of the Other group. This discrepancy persists consistently across all values of α. In contrast, the middle panel of Figure 6 demonstrates that by employing the FASI algorithm, the race-wise FSR levels are effectively controlled below the nominal level and are approximately equalized across the sensitive groups. Moreover, the right

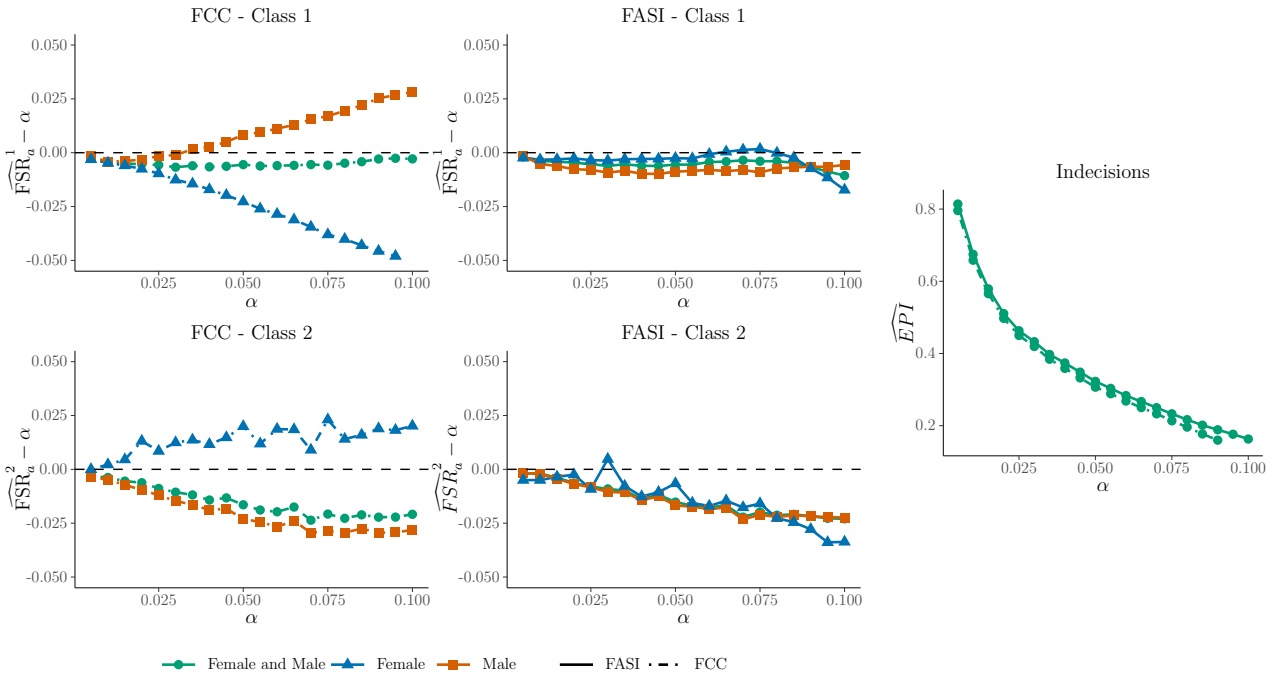

Figure 7: Census income prediction. Class 1 (top row) comprises individuals earning *less than* \$50,000 per year, while Class 2 (bottom row) includes individuals earning *more than* \$50,000 per year. Left and Middle: False Selection Rate minus varying levels of α. Right: The EPI levels.

panel illustrates that FASI achieves a nearly identical EPI level as the FCC approach.

5.2 1994 census income data analysis

The US census is a primary source of information for generating data concerning the American population. Consequently, the data they collect plays a direct role in informing future policy decisions, such as allocating resources for programs that offer economic assistance to vulnerable populations. These resources encompass necessities such as food, healthcare, job training, housing, and other forms of economic aid, which rely on accurate estimates of income levels within the population. The potential consequences of making unfair decisions when predicting income levels can be significant, as these predictions contribute to determining how hundreds of billions of dollars in federal funding will be allocated over the next decade. In this case study, we utilize the 1994 US Census Data set from the UCI Machine Learning Repository to predict whether an individual earns above or below \$50,000 per year, with Class 1 representing individuals earning less than \$50,000 and Class 2 representing those earning more than \$50,000. To avoid overlapping selections, we utilize the two-stage procedure described in Section G of the Supplement.

The data set in this study comprises 32,561 observations on 14 variables, predominantly demographic factors such as education level, age, and hours worked per week, among others. The protected attributes under consideration are Female and Male. Specifically, the Female group consists of a total of 10,771 observations, with 1,179 individuals earning over \$50,000 per year. Similarly, the Male attribute encompasses the remaining 21,790 observations, with 6,662 individuals earning over \$50,000 per year.

We applied the FCC and FASI algorithms at different FSR levels, ranging from 0.05% to 10%. We performed 100 random splits of the dataset, where for each gender and class label, 70% of the data was randomly assigned to $\mathcal{D}$, and the remaining 30% was assigned to $\mathcal{D}^{test}$. Furthermore, $\mathcal{D}$ was evenly divided into $\mathcal{D}^{train}$ and $\mathcal{D}^{cal}$. The left and middle panels of Figure 7

respectively show the FSR levels for both the FCC and FASI.

From the left column, we can observe that the group-wise FSR levels of FCC consistently deviate from the nominal level α, resulting in unfair decisions for the sensitive groups. This pattern is observed in both Class 1 and Class 2, although in opposite directions. The disparity in group-wise FSR levels becomes more pronounced as α increases. In contrast, the middle column demonstrates that for Class 1, the group-wise FSR levels of FASI remain close to α. For Class 2, the group-wise FSR levels of FASI exhibit conservativeness but are roughly equalized across the two sensitive groups. The conservativeness can be attributed to the R-value, which provides a conservative estimate of the true FSP. Furthermore, the right column highlights that FASI effectively achieves approximate parity, ensuring that the burden is roughly equally shared across the two genders, with only a slight increase in the EPI level.

6 Discussions and Extensions

This section concludes the article by discussing additional fairness notions and extensions of the FSR concept, highlighting limitations in existing research and suggesting future directions.

Fairness in machine learning presents a complex challenge. Multiple studies focus on addressing representation or sampling bias, which arises when data are collected in a non-representative fashion (Mehrabi et al. 2021). By contrast, algorithmic bias emerges when the model itself introduces bias beyond the inherent biases in the input data. This article addresses the issue of algorithmic bias, with the objective of ensuring an equitable distribution of erroneous decisions across different groups. FASI is model-free, allowing for deployment with any user-specified model. It achieves fairness by aligning group-wise FSRs to the same designated level, requiring only mild conditions on data exchangeability.

In addition to the sufficiency principle, the *separation principle* (Barocas et al. 2017) has been widely used. It requires that $P(Y \neq \hat{Y}|Y = c, A = a)$ are the same for all $a \in \mathcal{A}$. This principle differs from the sufficiency principle (1), whereby $\hat{Y}$ and Y interchange positions in the conditional probability expression. A third notion on fairness, in the context of prediction intervals, has been considered in Romano, Barber, Sabatti & Candès (2020). Rather than conditioning on either Y or $\hat{Y}$, this fairness criterion is concerned with the joint probabilities of $(\hat{Y}, Y)$, requiring that the misclassification rates are equalized across all protected groups $P(Y \neq \hat{Y}|A = a)$ are the same for all $a \in \mathcal{A}$. The fourth notion, known as *demographic parity* (Jiang et al. 2020) requires that $P(\hat{Y} \neq c|A = a)$ are the same for all $a \in \mathcal{A}$. Other popular fairness notions include *equalized odds* (Hardt et al. 2016, Romano, Bates & Candès 2020) and *equalized risks* (Corbett-Davies et al. 2023).

Zafar et al. (2017) proposed the use of cost-sensitive classifiers with group-specific costs (Menon & Williamson 2018) to address a fairness issue comparable to our work. However, their technique forces a decision to be made on all individuals, whereas our approach is a selective inference procedure that only makes confident judgments on a subset of subjects. Given human intervention, FASI can achieve higher accuracy than cost-sensitive classifiers, as practitioners are aware of the undecided cases that merit additional scrutiny, ultimately reducing erroneous decisions with potentially extensive societal costs.

Our fairness criterion, as described in Equation 6, constitutes a group fairness notion that presupposes full knowledge of the protected groups. This approach is widely adopted in the literature and finds applications across diverse domains, including medicine and the criminal justice system (Manrai et al. 2016, Angwin et al. 2016), often facilitated by specialized software tools (Bellamy et al. 2018, Saleiro et al. 2018). However, situations may arise where the protected groups lack clear delineation, such as when the sensitive attribute pertains to age or income.

New ideas, such as individual fairness and counterfactual fairness, provide useful alternatives. Specifically, individual fairness aims to ensure that comparable individuals receive commensurate outcomes (Mukherjee et al. 2020), while counterfactual fairness posits that fairness should not be exclusively contingent on observable attributes but should also consider potential counterfactual factors. Given the substantial complexities associated with individual and counterfactual fairness algorithms, we leave exploration of this promising avenue in future research.

A highly contentious matter is that disparate fairness criteria often yield distinct algorithms and different decisions in practice. For instance, the sufficiency and separation principles can be incompatible with one another (Kleinberg et al. 2017, Friedler et al. 2021), and classification parity or group calibration can potentially harm the very groups that these algorithms are intended to protect (Corbett-Davies et al. 2023). Despite growing awareness of fairness concerns in decision-making, a consensus is yet to be reached on the best approaches for achieving fairness in machine learning. While we do not claim that FASI is ubiquitously superior to competing approaches, adjusting group-wise FSRs appears to be an effective and suitable fairness criterion for high-stake applications, overcoming several limitations of the widely used sufficiency principle. Much research is still needed for understanding the trade-offs and applicability of different fairness notions across diverse contexts and applications.

The selective inference framework and FSR concepts can be extended beyond the binary classification setting. Denote the collection of all class labels by $\mathcal{C} = \{1, \ldots, C\}$. The case with $C = 1$ corresponds to the one-class classification problem, as recently explored in Guan & Tibshirani (2022) and Bates et al. (2023), which can be encompassed within our general framework. For situations with $C \geq 2$, denote the set of classes to be selected by $\mathcal{C}'$, and assume $\mathcal{C}' \subset \mathcal{C}$. With indecisions being allowed, the action space is given by $\Lambda = \{0, \mathcal{C}'\}$. Denote the selection rule $\{\hat{Y}_i : i \in \mathcal{D}^{test}\} \in \Lambda^m$. Then the FSR with respect to subset $\mathcal{C}'$ is defined as the expected fraction of erroneous selections among all selections:

$$\mathrm{FSR}^{\mathcal{C}'} = \mathbb{E}\left[\frac{\sum_{j \in \mathcal{D}^{test}} \mathbb{I}(\hat{Y}_j \in \mathcal{C}', \hat{Y}_j \neq Y_j)}{\left\{\sum_{i \in \mathcal{D}^{test}} \mathbb{I}(\hat{Y}_j \in \mathcal{C}')\right\} \vee 1}\right].$$

The group-wise FSRs taking into account the protected attribute A can be defined analogously to (6) by restricting the selections to specific groups. The EPI definition remains the same. The extension of the R-value in this setup is briefly discussed in Section G of the Supplement, with some remaining issues left for future research.

References

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I. & Lei, L. (2025), 'Learn then test: Calibrating predictive algorithms to achieve risk control', *The Annals of Applied Statistics* **19**(2), 1641 – 1662.

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016), 'Machine bias: There's software used across the country to predict future criminals', *And it's biased against blacks. ProPublica* **23**, 77–91.

Barber, R. F. & Candès, E. J. (2015), 'Controlling the false discovery rate via knockoffs', *The Annals of Statistics* **43**(5), 2055–2085.

Barocas, S., Hardt, M. & Narayanan, A. (2017), 'Fairness in machine learning', *Nips tutorial* **1**, 2.

Bates, S., Candès, E., Lei, L., Romano, Y. & Sesia, M. (2023), 'Testing for outliers with conformal p-values', *The Annals of Statistics* **51**(1), 149 – 178.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R. & Zhang, Y. (2018), 'Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias'.

Benjamini, Y. (2010), 'Simultaneous and selective inference: Current successes and future challenges', *Biometrical Journal* **52**(6), 708–721.

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *J. Roy. Statist. Soc. B* **57**, 289–300.

Benjamini, Y. & Hochberg, Y. (2000), 'On the adaptive control of the false discovery rate in multiple testing with independent statistics', *Journal of Educational and Behavioral Statistics* **25**, 60–83.

Benjamini, Y. & Yekutieli, D. (2001), 'The control of the false discovery rate in multiple testing under dependency', *Ann. Statist.* **29**(4), 1165–1188.

Cai, T. & Sun, W. (2017), 'Optimal screening and discovery of sparse signals with applications to multistage high-throughput studies', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(1), 197.

Cai, T. T. & Sun, W. (2009), 'Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks', *J. Amer. Statist. Assoc.* **104**, 1467–1481.

Cai, T. T., Sun, W. & Wang, W. (2019), 'CARS: Covariate assisted ranking and screening for large-scale two-sample inference (with discussion)', *J. Roy. Statist. Soc. B* **81**, 187–234.

Cao, H., Sun, W. & Kosorok, M. R. (2013), 'The optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing', *Biometrika* **100**(2), 495–502.

Chen, T. & Guestrin, C. (2016), XGBoost: A scalable tree boosting system, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '16, ACM, New York, NY, USA, pp. 785–794.

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

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R. & Goel, S. (2023), 'The measure and mismeasure of fairness', *Journal of Machine Learning Research* **24**(312), 1–117.

Crisp, R. (2003), 'Equality, priority, and compassion', *Ethics* **113**(4), 745–763.

Dieterich, W., Mendoza, C. & Brennan, T. (2016), 'Compas risk scales: Demonstrating accuracy equity and predictive parity', *Northpointe Inc* .

Du, L., Guo, X., Sun, W. & Zou, C. (2023), 'False discovery rate control under general dependence by symmetrized data aggregation', *Journal of the American Statistical Association* **118**(541), 607–621.

Dua, D. & Graff, C. (2017), 'Uci machine learning repository'. http://archive.ics.uci.edu/ml.

Friedler, S. A., Scheidegger, C. & Venkatasubramanian, S. (2021), 'The (im) possibility of fairness: different value systems require different mechanisms for fair decision making', *Communications of the ACM* **64**(4), 136–143.

Genovese, C. & Wasserman, L. (2002), 'Operating characteristics and extensions of the false discovery rate procedure', *J. R. Stat. Soc. B* **64**, 499–517.

Guan, L. & Tibshirani, R. (2022), 'Prediction and outlier detection in classification problems', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**(2), 524–546.

Hardt, M., Price, E. & Srebro, N. (2016), 'Equality of opportunity in supervised learning', *Advances in neural information processing systems* **29**, 3315–3323.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer series in statistics, Springer.

Herbei, R. & Wegkamp, M. H. (2006), 'Classification with reject option', *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* pp. 709–721.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2023), *An Introduction to Statistical Learning: with Applications in R*, Vol. 2, Springer.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H. & Chiappa, S. (2020), Wasserstein fair classification, *in* R. P. Adams & V. Gogate, eds, 'Proceedings of The 35th Uncertainty in Artificial Intelligence Conference', Vol. 115 of *Proceedings of Machine Learning Research*, PMLR, pp. 862–872.

Jin, Y. & Candes, E. J. (2023), 'Selection by prediction with conformal p-values', *Journal of Machine Learning Research* **24**(244), 1–41.

Kemmler, M., Rodner, E., Wacker, E.-S. & Denzler, J. (2013), 'One-class classification with gaussian processes', *Pattern recognition* **46**(12), 3507–3518.

Khan, S. S. & Madden, M. G. (2009), A survey of recent trends in one class classification, *in* 'Irish conference on artificial intelligence and cognitive science', Springer, pp. 188–197.

Kleinberg, J., Mullainathan, S. & Raghavan, M. (2017), Inherent Trade-Offs in the Fair Determination of Risk Scores, *in* C. H. Papadimitriou, ed., '8th Innovations in Theoretical Computer Science Conference (ITCS 2017)', Vol. 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, pp. 43:1–43:23.

Kusner, M. J., Loftus, J., Russell, C. & Silva, R. (2017), Counterfactual fairness, *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 30, Curran Associates, Inc.

Lee, J. K., Bu, Y., Rajan, D., Sattigeri, P., Panda, R., Das, S. & Wornell, G. W. (2021), Fair selective classification via sufficiency, *in* 'Proceedings of the 38th International Conference on Machine Learning', Vol. 139 of *Proceedings of Machine Learning Research*, PMLR, pp. 6076–6086.

Lei, J. (2014), 'Classification with confidence', *Biometrika* **101**(4), 755–769.

Lei, J. & Wasserman, L. (2014), 'Distribution-free Prediction Bands for Non-parametric Regression', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **76**(1), 71–96.

Lei, L. & Fithian, W. (2018), 'AdaPT: An Interactive Procedure for Multiple Testing with Side Information', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **80**(4), 649–679.

Leung, D. & Sun, W. (2022), 'ZAP: Z-Value Adaptive Procedures for False Discovery Rate Control with Side Information', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**(5), 1886–1946.

Liang, Z., Sesia, M. & Sun, W. (2024), 'Integrative conformal p-values for out-of-distribution testing with labelled outliers', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **86**(3), 671–693.

Long, K. D. & Albert, S. M. (2021), Use of zip code based aggregate indicators to assess race disparities in covid-19, *in* 'Use of Zip Code Based Aggregate Indicators to Assess Race Disparities in COVID-19', Vol. 31, Ethnicity & disease.

Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., Margulies, D. M., Loscalzo, J. & Kohane, I. S. (2016), 'Genetic misdiagnoses and the potential for health disparities', *New England Journal of Medicine* **375**(7), 655–665. PMID: 27532831.

Mary, D. & Roquain, E. (2022), 'Semi-supervised multiple testing', *Electronic Journal of Statistics* **16**(2), 4926 – 4981.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2021), 'A survey on bias and fairness in machine learning', *ACM Comput. Surv.* **54**(6).

Meinshausen, N. & Rice, J. (2006), 'Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses.', *Ann. Statist.* **34**, 373–393.

Menon, A. K. & Williamson, R. C. (2018), The cost of fairness in binary classification, *in* S. A. Friedler & C. Wilson, eds, 'Proceedings of the 1st Conference on Fairness, Accountability and Transparency', Vol. 81 of *Proceedings of Machine Learning Research*, PMLR, pp. 107–118.

Moya, M. M. & Hush, D. R. (1996), 'Network constraints and multi-objective optimization for one-class classification', *Neural networks* **9**(3), 463–474.

Mukherjee, D., Yurochkin, M., Banerjee, M. & Sun, Y. (2020), Two simple ways to learn individual fairness metrics from data, *in* 'Proceedings of the 37th International Conference on Machine Learning', ICML'20, JMLR.org.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. (2017), On fairness and calibration, *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 30, Curran Associates, Inc.

Romano, Y., Barber, R. F., Sabatti, C. & Candès, E. (2020), 'With malice toward none: Assessing uncertainty via equalized coverage'. https://hdsr.mitpress.mit.edu/pub/qedrwcz3.

Romano, Y., Bates, S. & Candès, E. J. (2020), Achieving equalized odds by resampling sensitive attributes, *in* 'Advances in Neural Information Processing Systems 33 (NIPS 2020)', Curran Associates, Inc. To appear.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T. & Ghani, R. (2018), 'Aequitas: A bias and fairness audit toolkit'.

Silverman, B. W. (1986), *Density estimation for statistics and data analysis / B.W. Silverman*, Chapman and Hall London ; New York.

Storey, J. D. (2002), 'A direct approach to false discovery rates', *J. Roy. Statist. Soc. B* **64**, 479–498.

Storey, J. D. (2003), 'The positive false discovery rate: a Bayesian interpretation and the q-value', *Ann. Statist.* **31**, 2013–2035.

Storey, J. D., Taylor, J. E. & Siegmund, D. (2004), 'Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach', *J. Roy. Statist. Soc. B* **66**(1), 187–205.

Sun, W. & Cai, T. T. (2007), 'Oracle and adaptive compound decision rules for false discovery rate control', *J. Amer. Statist. Assoc.* **102**, 901–912.

Sun, W. & Wei, Z. (2011), 'Large-scale multiple testing for pattern identification, with applications to time-course microarray experiments', *J. Amer. Statist. Assoc.* **106**, 73–88.

Vovk, V., Gammerman, A. & Shafer, G. (2005), *Algorithmic learning in a random world*, Vol. 29, Springer.

Weinstein, A., Barber, R. & Candes, E. (2017), 'A power and prediction analysis for knockoffs with lasso statistics'. arXiv preprint arXiv:1712.06465.

Zafar, M. B., Valera, I., Gomez Rodriguez, M. & Gummadi, K. P. (2017), Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment, *in* 'Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment', WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, p. 1171–1180.

Zeng, X., Dobriban, E. & Cheng, G. (2022), Fair bayes-optimal classifiers under predictive parity, *in* 'Advances in Neural Information Processing Systems', Vol. 35, Curran Associates, Inc., pp. 27692–27705.

Online Supplementary Material for "A Burden Shared is a Burden Halved: A Fairness-Adjusted Approach to Classification"

This supplement provides a comparison of various R-value notions (Section A), additional technical details of the methodology (Sections B-C), technical proofs (Sections D-E), discussion of related fairness algorithms and possible extensions (Sections F and G), and supplementary numerical results (Section H).

A Variants of the (empirical) R-value

The R-value (13) has been proposed as the basic operational unit of our FASI algorithm; we discuss its empirical variants in this section and its theoretical version in Section B.

First, while including both $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$ in the denominator of (11) enhances the algorithm's stability, the resulting FASI algorithm can only control a modified version of the FSR [cf. (18)] asymptotically. The simpler version, which only includes $\mathcal{D}^{test}$ in the denominator of (10), is particularly relevant for readers who prefer a validity theory in finite samples.

Secondly, one consideration, pertaining to the multiplicative factor $\gamma_{c,a} = \mathbb{E}\left(p^{test,a}_{c,null}/p^{cal,a}_{c,null}\right)$ in the theorem, is that FASI fails to provide precise FSR control due to the possible fluctuations in $\gamma_{c,a}$. While this issue is also minor (as under Assumption 1, this constant is approximately 1 and numerically negligible, cf. Section H.3 of this Supplement), we present a conservative version of the R-value next. We demonstrate that this multiplicative factor $\gamma_{c,a}$ can be eliminated from the theory when the conservative version is employed. It is important to note that the conservative R-value is primarily of theoretical interest, as it leads to a substantial loss in power in many practical scenarios.

We summarize related results in the subsequent corollary. The proof of the corollary follows directly from the proof of Theorem 1 and is therefore omitted.

Corollary 1. *Suppose we apply the FASI algorithm with the conservative R-value:*

$$R^{c,*}_j = \max\left\{\mathbb{I}(\hat{S}^c_j \le 0.5), \frac{n^{cal}_a + 1}{n^{cal,c}_{a,null} + 1}\tilde{Q}^c_j\right\} \tag{A.1}$$

for $j \in \mathcal{D}^{test}$, where $\tilde{Q}^c_j$ is defined via (10) and (12), $n^{cal}_a = \sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(A_i = a)$ and $n^{cal,c}_{a,null} = \sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(A_i = a, Y_i \neq c)$. Further define $n^{test}_a = \sum_{j\in\mathcal{D}^{test}} \mathbb{I}(A_j = a)$ and $n^{test,c}_{a,null} = \sum_{j\in\mathcal{D}^{test}} \mathbb{I}(A_j = a, Y_j \neq c)$. Then we have, for all $a \in \mathcal{A}$, $FSR^{\{c\}}_a \le \mathbb{E}\left(n^{test,c}_{a,null}/n^{test}_a\right)\alpha \le \alpha$.

The ratio $n^{test,c}_{a,null}/n^{test}_a$, in Corollary 1 is referred to as the null proportion in multiple testing, also appears in the classical Benjamini-Hochberg (BH) procedure for FDR control. In Section C.3 of this supplement, we will elaborate the connection between the FASI algorithm and the BH algorithm implemented with conformal p-values.

It is expected that the FASI algorithm with conservative R-values (A.1) can be enhanced by incorporating the unknown ratio $n^{test}_{a,null}/n^{test}_a$ into the analysis. This approach has been successfully adopted in various works, such as Benjamini & Hochberg (2000) and Storey (2002), to boost the power of the conservative BH algorithm in the context of FDR control. The FASI algorithm with R-value defined in (13) can be roughly regarded as such an approach. Specifically, the unknown ratio $n^{test,c}_{a,null}/n^{test}_a$ is initially estimated as $(n^{cal,c}_{a,null} + 1)/(n^{cal}_a + 1)$. This estimated ratio is then incorporated into the FASI algorithm by utilizing the conservative version of FASI

at the modified level of $(n_a^{cal}+1)/(n_{a,null}^{cal,c}+1)\alpha$. This practice leads to improved power at the expense of the additional factor $\gamma_{c,a}$ in Theorem 1.

B Theoretical R-value and Optimality Theory

In this section, we introduce the theoretical R-value and derive the optimal score function under a simplified setup. Our subsequent discussions are purely theoretical, where we assume an oracle with access to all distributional information and make several simplifying assumptions. Our primary goal is to develop a theoretical version of the R-value and an optimality theory for FSR control. This theoretical framework serves as a foundation for our practical algorithm and provides valuable insights into the properties of the R-value. Our theory provides practical insights for practitioners on how to train score functions to construct informative R-values.

B.1 The mixture model under an oracle setting

Denote $\pi_a = P(A_j = a)$ and $\pi_{c|a} = P(Y_j = c|A_j = a)$. We assume that (X_j, A_j) are independent observations obeying the following random mixture model:

$$F(x, A_j) = \sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a) \cdot \left\{\pi_{1|a}F_{1|a}(x) + \pi_{2|a}F_{2|a}(x)\right\}, \tag{B.2}$$

where $F_{1|a}(x)$ and $F_{2|a}(x)$ are the conditional CDFs of X_j coming from classes 1 and 2 given that $A_j = a$, respectively. Let $f_{c|a}(x)$ be the corresponding density function. For our analysis, we consider a class of oracle rules of the form

$$\hat{Y}_j \coloneqq \hat{Y}_j(X_j, A_j) \;=\; \sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a) \cdot \left\{ \sum_{c\in\{1,2\}} c \cdot \mathbb{I}(S_j^c > t_a^c) \right\}, \quad \text{for } j \in \mathcal{D}^{test}, \tag{B.3}$$

where the thresholds satisfy $t_a^c > 0.5$ to avoid overlapping selections. We assume that an oracle has knowledge of the conditional probabilities and conditional density functions defined above.

B.2 The conversion algorithm

In this section, we present a systematic approach for converting an arbitrary score $S^c(x, a)$ into a fair score $R^c(S^c)$, which we refer to as the theoretical R-value. Although the discussion is theoretical in nature, it highlights the existence of a fair score corresponding to every confidence score. This algorithm can be regarded as a method of *calibration by group*, a widely used technique in the fairness literature (see Barocas et al. (2017) for an example). Our discussion assumes that (a) the score function is known and (b) the distributional information of the scores is available; this is referred to as the oracle setup, which does not involve utilizing labeled training and calibration data $\mathcal{D}^{train}$ and $\mathcal{D}^{cal}$.

The conversion algorithm consists of three steps. In Step 1, we define and derive important quantities for $S_j^c \coloneqq S^c(X_j, A_j)$, where $j \in \mathcal{D}^{test}$. Under the random mixture model (B.2) (for the observed data points), the scores S_j^c obey the following mixture model:

$$G^c(s) \;=\; \sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a) \cdot G_a^c(s) \;=\; \sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a) \cdot \left\{\pi_{1|a}G_{1|a}^c(s) + \pi_{2|a}G_{2|a}^c(s)\right\},$$

where $G^c_{c'|a}(s)$ denotes the conditional CDF of S^c given $A = a$ and $Y = c'$, and $\pi_{c'|a} = \mathbb{P}(Y_i = c|A_i = a)$ are the conditional probabilities for $c' = 1, 2$.

Under the oracle setup, the conditional probabilities and conditional CDFs defined above are assumed to be known. In Step 2, we compute the conditional error probabilities for individuals from group a when the threshold for S^c_j is t^c_a using decision rule (B.3):

$$\begin{aligned}\texttt{err}^1_a(t^1_a) &= \mathbb{P}(Y = 2|S^1_j > t^1_a, A = a) = \frac{\pi_{2|a}\left\{1 - G_{2|a}(t^1_a)\right\}}{1 - G^1_a(t^1_a)};\\ \texttt{err}^2_a(t^2_a) &= \mathbb{P}(Y = 1|S^2_j > t^2_a, A = a) = \frac{\pi_{1|a}\left\{1 - G_{1|a}(t^2_a)\right\}}{1 - G^2_a(t^2_a)}.\end{aligned}$$

Finally, in Step 3 we compute a pair of fair scores for every individual from group a with observed scores $(S^1_j = s^1_j, S^2_j = s^2_j)$:

$$\texttt{TQ}^c_j \equiv \texttt{TQ}^c_j(s^c_j) = \sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a) \cdot \inf_{t\le s^c_j}\left\{\texttt{err}^c_a(t) = \mathbb{P}(Y_j \neq c|s^c_j > t, A = a)\right\}, \tag{B.4}$$

for $c \in \{1, 2\}$ and $j \in \mathcal{D}^{test}$. If the confidence score satisfies the monotone likelihood ration condition (MLRC, Sun & Cai 2007), then the infimum is achieved at s^c_j exactly. To avoid overlapping selections, we define the theoretical R-values as

$$\texttt{TR}^c_j = \max\left\{\mathbb{I}(S^c_j \le 0.5),\ \texttt{TQ}^c_j\right\}, \quad c \in \{1, 2\},\ j \in \mathcal{D}^{test}. \tag{B.5}$$

B.3 Theoretical R-value and fairness

Consider random mixture model (B.2). Let $\{S^c_j := S^c(X_j, A_j) : c \in \{1, 2\}, j \in \mathcal{D}^{test}\}$ be the confidence scores and $\{\texttt{TR}^c_j : c \in \{1, 2\}, j \in \mathcal{D}^{test}\}$ denote the corresponding theoretical R-values. The goal is to assign labels "0", "1" and "2" to new instances $\{(X_j, A_j) : j \in \mathcal{D}^{test}\}$. Consider the following classification rule:

$$\hat{Y}_j = \sum_{c\in\{1,2\}} c \cdot \mathbb{I}(\texttt{TR}^c_j \le \alpha_c). \tag{B.6}$$

Define the marginal FSR

$$\text{mFSR}^{\{c\}}_a = \frac{\mathbb{E}\left\{\sum_{j\in\mathcal{D}^{test}} \mathbb{I}(\hat{Y}_j = c, Y_j \neq c, : A_j = a)\right\}}{\mathbb{E}\left\{\sum_{j\in\mathcal{D}^{test}} \mathbb{I}(\hat{Y}_j = c, A_j = a)\right\}}. \tag{B.7}$$

We assume that the instances (X_j, A_j) are independent draws from an underlying random mixture model (B.2). It can be shown that, following arguments in Storey (2003) for FDR analysis,

$$\text{mFSR}^{\{c\}}_a = \mathbb{P}(Y_j \neq c|\hat{Y}_j = c, A = a), \tag{B.8}$$

which is the conditional probability required in the sufficiency principle [cf. Equation (1) in the main text].

Based on the work of Cai et al. (2019), we can similarly show that under mild conditions,

$$\text{FSR}^{\{c\}}_a = \text{mFSR}^{\{c\}}_a + o(1), \text{ when } m_a := |\{j \in \mathcal{D}^{test} : A_j = a\}| \to \infty. \tag{B.9}$$

The next proposition shows that thresholding the theoretical R-value leads to a fair selective

inference procedure.

Proposition 1. Consider the classification rule (B.6). Then we have

$$\mathbb{P}(Y_j \neq c|\hat{Y}_j = c, A = a) \leq \alpha_c \text{ for } c \in \{1, 2\} \text{ and } a \in \mathcal{A}. \tag{B.10}$$

We would like to make two important remarks. Firstly, the theoretical R-value, which may be viewed as the counterpart of the data-driven R-value, represents the minimum conditional probability required to ensure that an individual with score $S_j^c = s^c$ is selected into class c. Secondly, the theoretical R-value is a fundamental quantity that is closely linked to the sufficiency principle in the fairness literature. Proposition 1 highlights that by setting thresholds for the theoretical R-values, the thresholding procedure fulfills the sufficiency principle and controls the group-wise error rates.

B.4 Oracle theoretical R-value and the optimality theory

We present and prove an intuitive result that shows the class probability

$$S_{OR}^{c,j}(x, a) = \mathbb{P}(Y_j = c|X_j = x, A_j = a)$$

is the optimal choice of confidence score for calibrating the theoretical R-value. To simplify the arguments, we develop our optimality theory based on the mFSR, an asymptotically equivalent variation of the FSR. The relationship between the mFSR and FSR has been established in Equation (B.9).

We aim to construct a selection rule under the binary classification setting that solves the following constrained optimization problem:

$$\text{Minimize the EPI, subject to } \text{mFSR}_a^c \leq \alpha_c, \text{ for } c \in \{1, 2\} \text{ and } a \in \mathcal{A}. \tag{B.11}$$

Our strategy is to convert the oracle scores $S_{OR}^{c,j}$ to oracle theoretical R-values

$$\left\{\texttt{TR}_{OR}^{c,j} : c \in \{1, 2\}, j \in \mathcal{D}^{test}\right\}.$$

The process of conversion follows the general strategy outlined in Section B.2, and is described in more detail in the proof of Theorem 2 below.

Consider the selective classification problem outlined in (B.11). Define the oracle procedure $\boldsymbol{\delta}_{OR} = \{\delta_{OR}^j : j \in \mathcal{D}^{test}\}$, where

$$\delta_{OR}^j = \sum_{c \in \{1,2\}} c \cdot \mathbb{I}(\texttt{TR}_{OR}^{c,j} \leq \alpha_c). \tag{B.12}$$

The optimality of the oracle procedure is established in the next theorem.

Theorem 2. *Consider random mixture model (B.2). Let $\mathcal{D}_{\alpha_1,\alpha_2}$ denote the collection of selection rules of the form (B.6) that satisfy $mFSR_a^c \leq \alpha_c$ for $c = 1, 2$ and all $a \in \mathcal{A}$. Let $EPI_{\boldsymbol{\delta}}$ denote the EPI of an arbitrary decision rule $\boldsymbol{\delta}$ without overlapping selections. Then the oracle procedure (B.12) is optimal in the sense that $EPI_{\boldsymbol{\delta}_{OR}} \leq EPI_{\boldsymbol{\delta}}$ for any $\boldsymbol{\delta} \in \mathcal{D}_{\alpha_1,\alpha_2}$.*

The optimality theory indicates that, during the training stage, we should utilize *all features*, including the sensitive attribute A, to best capture individual level information.

B.5 R-value and Storey's Q-value

Excluding the sensitive attribute A in our analysis, the theoretical R-value is closely connected to the q-value, a useful tool in large-scale testing due to its intuitive interpretation and ease of use, as described in Storey (2003).

To test hypotheses $\{H_j : j \in \mathcal{D}^{test}\}$ with associated p-values $\{p_j : j \in \mathcal{D}^{test}\}$, let π be the proportion of non-nulls and $G(t)$ the alternative distribution of p-values. The q-value for hypothesis H_j is defined as

$$\inf_{t \geq p_j} \left\{ \text{pFDR}(t) := \frac{(1-\pi)t}{(1-\pi)t + \pi G(t)} \right\},$$

which roughly measures the fraction of false discoveries when H_j is rejected.

The q-value and R-value algorithms operate in the same manner. Conducting an FDR analysis at a given level α entails obtaining the q-value for hypothesis j and rejecting it if the q-value is less than or equal to α. Likewise, conducting an FSR analysis at level α involves obtaining the R-value for individual j and selecting it if the R-value is less than or equal to α.

C R-value and Conformal P-value

In this section, we adopt a multiple testing perspective to analyze the R-value. Although motivated differently, we show that the R-value is equivalent to the (BH) q-value of the conformal p-values (Bates et al. 2023) in a one-class classification scenario. For comparability considerations, we exclude the sensitive attribute A in the following discussions.

C.1 Selective inference for one-class classification: a multiple testing perspective

The selective inference perspective described in Section 2.2 provides a flexible framework accommodating various types of classification rules. For instance, if the interest lies solely in pinpointing high-risk individuals, the action space is defined as $\Lambda = \{0, 2\}$, and one may employ the following rule for screening:

$$\hat{Y}_j = 2 \cdot \mathbb{I}(S_j^2 > t_2), \quad t_2 > 0.5, \quad \text{for } j \in \mathcal{D}^{test}. \tag{C.13}$$

This setup is closely related to semi-supervised multiple testing or outlier detection in conformal inference (Mary & Roquain 2022, Bates et al. 2023, Liang et al. 2024).

Next, we explain the connection between the FSR and FDR from a multiple testing perspective. Consider testing m hypotheses:

$$H_{j0} : Y_j = 1 \quad \text{vs.} \quad H_{j1} : Y_j \neq 1 \text{ (i.e., } Y_j = 2), \quad \text{for } j \in \mathcal{D}^{test}. \tag{C.14}$$

In the context of the selective inference framework of Section 2.2, this multiple testing problem has the state space $\mathcal{C} = \{1, 2\}$. A multiple testing procedure represented by $\hat{Y}_j \in \{0, 2\}$, $j \in \mathcal{D}^{test}$ corresponds to the selection rule in (C.13). Here, the action space $\Lambda = \{0, 2\}$ differs from the state space $\mathcal{C} = \{1, 2\}$, with $\hat{Y}_j = 2$ indicating that H_{j0} is rejected, and $\hat{Y}_j = 0$ signifying insufficient evidence to reject H_{j0}. Consequently, $\text{FSR}^{\{2\}}$, as defined in (4), is equivalent to the widely used FDR, i.e., the expected proportion of false rejections among all rejections.

C.2 A brief review of conformal p-values

The problem of one-class classification, also known as outlier detection or out-of-distribution testing in conformal inference, can be formulated within the framework of selective inference. Consider the observed data $\{(X_i, Y_i) : i \in \mathcal{D}\}$ originating from two classes, $Y_i = 1$ and $Y_i = 2$. We divide the set $\mathcal{D}$ into two subsets, $\mathcal{D}^c = \{i : Y_i = c\}$, $c = 1, 2$, where $\mathcal{D}^1$ and $\mathcal{D}^2$ represent the index sets of inliers and outliers, respectively.

In the context of one-class classification, the objective is to accurately identify outliers (individuals with label $Y = 2$) in a set of unlabeled test data $\{X_j : j \in \mathcal{D}^{test}\}$, while maintaining strict control over the error rate. By considering individuals in class "1" as the null cases, we can formulate an equivalent multiple testing problem (C.14).

Remark 9. In the context of outlier detection, the standard practice is to only consider the labeled inliers $\mathcal{D}^1$ when computing conformal p-values. It is worth noting that recent research by Liang et al. (2024) has revealed that this approach may result in potential efficiency loss. Nevertheless, for the sake of comparability, we adhere to the conventional practice and exclude $\mathcal{D}^2$ in our investigation.

The construction of split-conformal p-values (Bates et al. 2023) involves partitioning $\mathcal{D}^1$ into two subsets: $\mathcal{D}^{train}$ for training a score function and $\mathcal{D}^{cal}$ for calibrating a significance index. Treating $\hat{S}^2(\cdot)$ as a conformity score function that indicates the likelihood of belonging to class 2, the conformal p-value for testing H_{j0} can be expressed using our notation as:

$$\hat{u}_j \equiv \hat{u}(X_j) = \frac{\sum_{i \in \mathcal{D}^{cal}} \mathbb{I}\{\hat{S}^2(X_i) \geq \hat{S}^2(X_j)\} + 1}{n^{cal} + 1}. \tag{C.15}$$

Remark 10. To avoid confusion, note that in our framework a higher score indicates a higher likelihood of being an outlier. This is opposite to the convention in Bates et al. (2023), where a lower score reflects stronger evidence. To align the definitions, we have replaced the expression "$S \leq t$" in the conformal p-value definition of Bates et al. (2023) with "$S \geq t$" in our formulation (C.15). This adjustment ensures that both formulations are equivalent.

C.3 R-value is the BH q-value of conformal p-values (for outlier detection problems)

For the outlier detection problem (C.14), we consider using the thresholding rule (C.13) instead of the selection rule (14) as we are only interested in selecting the high-risk class ($Y = 2$).

To see the connection of our R-value to the conformal p-value (C.15), recall the definition of Storey's q-value

$$\hat{q}^{ST}\{\hat{u}(t)\} = (1 - \pi)\hat{u}(t)/G\{\hat{u}(t)\},$$

where π is the proportion of non-null cases in $\mathcal{D}^{test}$ and $G(\cdot)$ is the cumulative distribution function (CDF) of the p-values. Now recall $m = |\mathcal{D}^{test}|$, let $\hat{G}(t)$ denote the empirical process of the scores $\{\hat{S}_j^2 : j \in \mathcal{D}^{test}\}$:

$$\hat{G}(t) = \frac{1}{m} \sum_{j \in \mathcal{D}^{test}} \mathbb{I}\left\{\hat{u}(\hat{S}_j^2) \leq \hat{u}(t)\right\} = \frac{1}{m} \sum_{j \in \mathcal{D}^{test}} \mathbb{I}\left(\hat{S}_j^2 \geq t\right), \tag{C.16}$$

where the last equality holds because, by (C.15), a larger score corresponds to a smaller conformal p-value. Next we consider a modification of Storey's q-value, referred to as the BH q-value, which

ignores the $(1-\pi)$ term and substitutes $\hat{G}$ in place of G in Storey's q-value:

$$\hat{q}_i^{BH} = \frac{\hat{u}(\hat{S}_i^2)}{\hat{G}(\hat{S}_i^2)} \quad \text{for } i \in \mathcal{D}^{cal} \cup \mathcal{D}^{test}. \tag{C.17}$$

Combining (C.15) – (C.17), we have

$$\begin{aligned} \hat{q}^{BH}(t) &= \frac{m}{n^{cal}+1} \cdot \frac{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}\left(\hat{S}_i^2 \geq t\right) + 1}{\sum_{j\in\mathcal{D}^{test}} \mathbb{I}\left(\hat{S}_j^2 \geq t\right)} \\ &= \frac{m}{n^{cal}+1} \cdot \frac{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}\left(\hat{S}_i^2 \geq t, Y_i = 1\right) + 1}{\sum_{j\in\mathcal{D}^{test}} \mathbb{I}\left(\hat{S}_j^2 \geq t\right)}. \end{aligned} \tag{C.18}$$

The last equality holds because under the one-class classification setup, $\mathcal{D}^{cal}$ is a "pure" training set in which all observations are from the null class "1". Let t take values in $\{S_k : k \in \mathcal{D}^{cal} \cup \mathcal{D}^{test}\}$ and denote the corresponding values $\{\hat{q}_k^{BH} : k \in \mathcal{D}^{cal} \cup \mathcal{D}^{test}\}$.

We also need to apply a monotonicity adjustment to ensure that the q-value function is non-decreasing in the conformity score. Let

$$q_j^{BH} = \min_{k\in\mathcal{D}^{cal}\cup\mathcal{D}^{test}:\hat{S}_k^2 \leq \hat{S}_j^2} \hat{q}_k^{BH}, \quad \text{for } j \in \mathcal{D}^{test}. \tag{C.19}$$

Since there is no overlapping selection, the adjustment in (13) is unnecessary. This precisely recovers the R-value defined by (10) and (12) (excluding the sensitive attribute).

C.4 Discussion

We emphasize that the fundamental connection between the R-value and conformal q-values only holds under the one-class classification setup. The BH q-value (C.18) will be different from the R-value (16) under the binary classification setup that we have considered in this article. Specifically, the cardinalities of the calibration sets will be different under the two setups, and the equality (C.18) does not hold. Our R-value does not explicitly utilize conformal p-values under the binary classification setup.

The conformal p-value approach by Bates et al. (2023) remains applicable for selective inference in the binary classification setup, specifically for the selection of cases from class 2. Nevertheless, it is noteworthy that the conformal p-value method utilizes a smaller data set, as the data set $\mathcal{D}^2$ is discarded, in comparison to our R-value approach. Consequently, this may lead to suboptimal information utilization and a reduction in statistical power. In addition, it is worth noting that the FASI algorithm may not be well-suited for the outlier detection problem, as it presumes that the test data and calibration data are exchangeable, which is unlikely to hold in practical scenarios. Therefore, both the conformal p-value and FASI approaches would require modification to address the outlier detection problem with labeled outliers. Related issues have gone beyond the scope of this study and will be pursued in future research.

D Proof of Theorem 1

We begin by presenting the proof of part (a) of Theorem 1 in Section D.1, followed by the more involved proof of part (b) and its corresponding lemmas in Section D.2. Since the non-asymptotic

theory in Theorem 1(b) provides only an upper bound on the FSR, we also offer an asymptotic analysis in Section D.3 to demonstrate that, under the standard regularity conditions commonly used in multiple testing, the upper bound from Theorem 1(b) converges to 0, thereby establishing the asymptotic validity of the stable version of FASI algorithm.

D.1 Proof of part (a)

D.1.1 An equivalent expression of the FASI algorithm

Consider a class of decision rules that select subjects into class c if the confidence scores exceed a threshold t. For the ath group, $a \in \mathcal{A}$, the estimated false discovery proportion (FSP), as a function of t, can be described as the following empirical process:

$$\widehat{\mathrm{FSP}}_a^c(t) = \frac{\left\{\sum_{i\in\mathcal{D}_a^{cal}} \mathbb{I}(\hat{S}_i^c \geq t, Y_i \neq c) + 1\right\}/(n_a+1)}{\left\{\sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{S}_j^c \geq t) \vee 1\right\}/m_a}. \tag{D.20}$$

Let $\mathbf{S}_a^c = \{\hat{S}_i^c : i \in \mathcal{D}_a^{cal} \cup \mathcal{D}_a^{test}\}$ for $a \in \mathcal{A}$.

Consider a selection procedure represented by the process given in (D.20). We aim to find the smallest threshold, denoted as τ_a^c, for which the estimated FSP is less than α:

$$\tau_a^c = \left[\min\left\{t \in \mathbf{S}_a^c : \widehat{\mathrm{FSP}}_a^c(t) \leq \alpha_c\right\}\right] \vee 0.5. \tag{D.21}$$

The adjustment "$(\cdot) \vee 0.5$", which shares similar ideas to the adjustment in (13), indicates that we never assign the jth individual to class c when the confidence score $\hat{S}_j^c \leq 0.5$ [cf. the equivalent rule based on $\hat{S}_j^c$ given in Equation (D.22) below]; this also effectively avoids overlapping selections; see Remark 5 for related discussions. Note that while the thresholds for the R-values from different groups are identical at α_c, the thresholds for the scores $\hat{S}_j^c$, denoted $\{\tau_a^c : a \in \mathcal{A}\}$, vary depending on $A_j = a$, the group membership of the j-th subject.

The R-value defined in (13) of the main text can be written as:

$$\tilde{Q}_j^c = \sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a) \cdot \min_{\{t\in\mathbf{S}_a^c : t\leq\hat{S}_j^c\}} \widehat{\mathrm{FSP}}_a^c(t), \quad R_j^c = \max\left\{\mathbb{I}(\hat{S}_j^c \leq 0.5), \tilde{Q}_j^c\right\}, \quad \text{for } j \in \mathcal{D}^{test}.$$

The following lemma shows that the decision rule based on thresholding the R-value can be equivalently represented using a decision rule based on thresholding the scores $\{\hat{S}_j^c, j \in \mathcal{D}^{test}\}$.

Lemma 1. *Consider τ_a^c defined in (D.21). Then the following two rules are equivalent:*

$$\delta_j = \mathbb{I}(R_j^c \leq \alpha_c) \quad \Longleftrightarrow \quad \delta_j' = \sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a) \cdot \mathbb{I}(\hat{S}_j^c \geq \tau_a^c), \quad j \in \mathcal{D}^{test}. \tag{D.22}$$

The proof of this lemma is provided in Section D.1.5. It follows that

$$\hat{Y}_j = \sum_{c\in\{1,2\}} c \cdot \mathbb{I}(R_j^c \leq \alpha_c) = \sum_{c\in\{1,2\}} c \cdot \left\{\sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a) \cdot \mathbb{I}(\hat{S}_j^c \geq \tau_a^c)\right\}.$$

We note that the thresholding rule for the R-value incorporates the sensitive attribute, whereas the thresholding rule based on confidence scores does not; see Lemma 12 in Section D.1.5 for further discussions.

D.1.2 Upper bounding the FSP process by martingales

We now describe the true FSP process of the FASI algorithm using the confidence scores, where the FSP process is outlined in (D.20) and the algorithm is given by the second representation in (D.22). Suppose our selection procedure chooses threshold t in the test set $\mathcal{D}_a^{test}$. Let

$$\begin{aligned} V_a^t(t) &= \sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{S}_j^c \geq t, Y_j \neq c), \quad W_a^t(t) = \sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{S}_j^c \geq t, Y_j = c) \quad \text{and} \\ K_a^t(t) &= \sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{S}_j^c \geq t) \;=\; V^t(t) + W^t(t) \end{aligned}$$

denote the counts of false selections, correct selections and total selections, respectively. For the calibration set $\mathcal{D}_a^{cal}$, define

$$\begin{aligned} V_a^c(t) &= \sum_{i\in\mathcal{D}_a^{cal}} \mathbb{I}(\hat{S}_i^c \geq t, Y_i \neq c), \quad W_a^c(t) = \sum_{i\in\mathcal{D}_a^{cal}} \mathbb{I}(\hat{S}_i^c \geq t, Y_i = c), \\ K_a^c(t) &= \sum_{i\in\mathcal{D}_a^{cal}} \mathbb{I}(\hat{S}_i^c \geq t) \;=\; V_a^c(t) + W_a^c(t) \end{aligned}$$

as the corresponding counts of selections. Consider the data-driven thresholds τ_a^c defined in (D.21). Then the group-wise FSPs of the proposed FASI algorithm, as defined by the second representation in (D.22), can be computed as

$$\mathrm{FSP}_a^{\{c\}}(\tau_a^c) = \frac{V_a^t(\tau_a^c)}{K_a^t(\tau_a^c) \vee 1}, \qquad \text{for } a \in \mathcal{A} \;\text{ and }\; c \in \{1,2\}.$$

The operation of the FASI algorithm implies that

$$\begin{aligned} \mathrm{FSP}_a^{\{c\}}(\tau_a^c) &= \frac{V^t(\tau_a^c)}{V^c(\tau_a^c)+1} \cdot \frac{V^c(\tau_a^c)+1}{K^t(\tau_a^c) \vee 1} \\ &= \widehat{\mathrm{FSP}}_a^c(\tau_a^c) \cdot \frac{|\mathcal{D}_a^{cal}|+1}{|\mathcal{D}_a^{test}|} \cdot \frac{V^t(\tau_a^c)}{V^c(\tau_a^c)+1} \\ &\leq \alpha \cdot \frac{|\mathcal{D}_a^{cal}|+1}{|\mathcal{D}_a^{test}|} \cdot \frac{V^t(\tau_a^c)}{V^c(\tau_a^c)+1}, \end{aligned} \tag{D.23}$$

where the last two steps utilize definitions (D.20) and (D.21), respectively.

D.1.3 Martingale arguments

The ratio appearing in (D.23) motivates us to consider the following process

$$V_a^t(t)/\{V_a^c(t)+1\}, \tag{D.24}$$

which we show is a martingale. We start with the following continuous-time filtration:

$$\begin{aligned} \mathcal{F}_t^a &= \sigma\{V_a^t(s), V_a^c(s), W_a^t(s), W_a^c(s) : t_a^l \leq s \leq t\} \\ &= \sigma\{V_a^t(s), V_a^c(s), K_a^t(s), K_a^c(s) : t_a^l \leq s \leq t\}, \end{aligned}$$

where t_a^l represents the lower limit of the threshold. That is, if t_a^l is employed, then all subjects in group a are classified into class c.

In our proof, it is sufficient to consider a discrete-time filtration since FASI only selects

thresholds from $\mathbf{S}_a^c$. Let $m_a^* = |\mathcal{D}_a^{cal}| + |\mathcal{D}_a^{test}|$ denote the total number of selections in both $\mathcal{D}_a^{cal}$ and $\mathcal{D}_a^{test}$ when the threshold is t_a^l. We consider a σ-field that contains all information of the entire selection process. Specifically, let $\{s_k : k = m_a^*, \cdots, 1\}$ denote a sequence of thresholds (times), where s_k is the threshold when exactly k subjects, including those from both $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$, are selected into class c, and k takes values in the order of $m_a^*, m_a^* - 1, \cdots, 1$ (backward in time). This leads to the following discrete-time filtration:

$$\mathcal{F}_k^a = \sigma\left\{V_a^t(s_j), V_a^c(s_j), W_a^t(s_j), W_a^c(s_j) : j = m_a^*, m_a^* - 1, \ldots, k\right\}. \tag{D.25}$$

We can see that $\mathcal{F}_k^a$ is a backward-running filtration as for $k_1 < k_2$, $\mathcal{F}_{k_2}^a \subset \mathcal{F}_{k_1}^a$. Note that at time s_k, only one of the four following events is possible:

$$\begin{aligned}
A_1 &= \{V_a^t(s_{k-1}) = V_a^t(s_k) - 1, V_a^c(s_{k-1}) = V_a^c(s_k), W_a^t(s_{k-1}) = W_a^t(s_k), W_a^c(s_{k-1}) = W_a^c(s_k)\},\\
A_2 &= \{V_a^t(s_{k-1}) = V^t(s_k), V_a^c(s_{k-1}) = V_a^c(s_k) - 1, W_a^t(s_{k-1}) = W_a^t(s_k), W_a^c(s_{k-1}) = W_a^c(s_k)\},\\
A_3 &= \{V_a^t(s_{k-1}) = V_a^t(s_k), V_a^c(s_{k-1}) = V_a^c(s_k), W_a^t(s_{k-1}) = W_a^t(s_k) - 1, W_a^c(s_{k-1}) = W_a^c(s_k)\},\\
A_4 &= \{V_a^t(s_{k-1}) = V_a^t(s_k), V_a^c(s_{k-1}) = V_a^c(s_k), W_a^t(s_{k-1}) = W_a^t(s_k), W_a^c(s_{k-1}) = W_a^c(s_k) - 1\}.
\end{aligned}$$

According to Assumption 1, and the fact that FASI uses same fitted model to compute the scores, we have $\mathbb{P}(A_1|\mathcal{F}_k^a)/\mathbb{P}(A_2|\mathcal{F}_k^a) = V_a^t(s_k)/V_a^c(s_k)$. Moreover, we must have $\mathbb{P}(A_1|\mathcal{F}_k^a) + \mathbb{P}(A_2|\mathcal{F}_k^a) + \mathbb{P}(A_3|\mathcal{F}_k^a) + \mathbb{P}(A_4|\mathcal{F}_k^a) = 1$. It follows that there exists a γ_k, such that

$$\mathbb{P}(A_1|\mathcal{F}_k^a) = \gamma_k \cdot \frac{V_a^t(s_k)}{V_a^t(s_k) + V_a^c(s_k)}, \quad \mathbb{P}(A_2|\mathcal{F}_k^a) = \gamma_k \cdot \frac{V_a^c(s_k)}{V_a^t(s_k) + V_a^c(s_k)},$$

and $\mathbb{P}(A_3|\mathcal{F}_k^a) + \mathbb{P}(A_4|\mathcal{F}_k^a) = 1 - \gamma_k$. It will soon become evident that the value of γ_k does not matter in the theory, as it will be canceled out in the calculations.

To see why (D.24) is a martingale wrt $\mathcal{F}_k^a$, note that

$$\begin{aligned}
\mathbb{E}\left\{\frac{V_a^t(s_{k-1})}{V_a^c(s_{k-1}) + 1}|\mathcal{F}_k^a\right\} &= \frac{V_a^t(s_k) - 1}{V_a^c(s_k) + 1} \cdot \gamma_k \cdot \frac{V_a^t(s_k)}{V_a^t(s_k) + V_a^c(s_k)}\\
&\quad + \frac{V_a^t(s_k)}{V_a^c s_k} \cdot \gamma_k \cdot \frac{V_a^c(s_k)}{V_a^t(s_k) + V_a^c(s_k)} + \frac{V_a^t(s_k)}{V_a^c(s_k) + 1}(1 - \gamma_k)\\
&= \frac{V_a^t(s_k)}{V_a^c(s_k) + 1} \cdot (\gamma_k + 1 - \gamma_k) = \frac{V_a^t(s_k)}{V_a^c(s_k) + 1}.
\end{aligned}$$

D.1.4 FSR Control

The threshold τ_a^c defined by (D.21) is a stopping time with respect to the filtration $\mathcal{F}_k^a$ since $\{\tau_a^c \le s_k\} \in \mathcal{F}_k^a$. In other words, the event whether the kth selection occurs completely depends on the information prior to time s_k (including s_k).

Let $\mathcal{D}_a^{test,0}$ and $\mathcal{D}_a^{cal,0}$ be the index sets for subjects in $\mathcal{D}_a^{test}$ and $\mathcal{D}_a^{cal}$ that do not belong to class c, respectively. In the final step of our proof, we shall apply the optional stopping theorem to the filtration $\{\mathcal{F}_k^a\}$. Recall that t_l^a is lower limit of the threshold, and m_a^* is the total number

of misclassifications in both $\mathcal{D}_a^{cal}$ and $\mathcal{D}_a^{test}$ when the threshold is t_l^a. The group-wise FSR is

$$\begin{aligned}
\text{FSR}_a^{\{c\}} &= \mathbb{E}\{\text{FSP}_a^{\{c\}}(\tau_a^c)\} && \text{(D.26)}\\
&\leq \alpha \cdot \mathbb{E}\left[\cdot\mathbb{E}\left\{\frac{|\mathcal{D}_a^{cal}|+1}{|\mathcal{D}_a^{test}|}\frac{V_a^t(\tau_a^c)}{V_a^c(\tau_a^c)+1}|\mathcal{F}_{m_a^*}\right\}\right]\\
&= \alpha \cdot \mathbb{E}\left[\frac{|\mathcal{D}_a^{cal}|+1}{|\mathcal{D}_a^{test}|}\cdot\frac{V_a^t(t_l)}{V_a^c(t_l)+1}\right]\\
&= \alpha \cdot \mathbb{E}\left\{\frac{|\mathcal{D}_a^{cal}|+1}{|\mathcal{D}_a^{test}|}\cdot\frac{|\mathcal{D}_a^{test,0}|}{|\mathcal{D}_a^{cal,0}|+1}\right\} && \text{(D.27)}\\
&\leq \gamma_{c,a}\alpha,
\end{aligned}$$

To get Equation (D.27) we have used the fact that when t_l^a is used then all subjects are classified to class c. This completes the proof.

Remark 11. We provide a remark to explain the $\mathbb{E}$ operator in (D.26). As indicated by (D.22), the FASI algorithm is equivalent to a thresholding rule based on $\hat{S}_i^c$. The data-driven threshold (or stopping time), τ_a^c, is a random variable that varies across different realizations or data sets. The FSP, denoted as $\text{FSP}_a^{\{c\}}(\tau_a^c)$, is a random variable that differs across data sets. The FSR, defined as the expectation of the FSP, integrates the randomness across the training, calibration and test data.

D.1.5 Proof of Lemma 1

First, it is easy to see that the two decision rules are equivalent (i.e. $\delta_j = \delta_j' = 0$) if $\hat{S}_j^c \leq 0.5$. We only consider the situation where $\hat{S}_j^c > 0.5$.

Next, suppose that $\delta_j' = 1$ holds for some $j \in \mathcal{D}^{test}$. Without loss of generality, assume that $A_j = a$. It follows that $\mathbb{I}(\hat{S}_j^c > \tau_a^c) = 1$, indicating that the stopping time τ_a^c must satisfy

$$\tau_a^c \in \{t \in \mathbf{S}_a^c : t \leq \hat{S}_j^c\}. \tag{D.28}$$

We conclude that

$$R_j^c = \tilde{Q}_j^c = \min_{\{t\in\mathbf{S}_a^c : t\leq\hat{S}_j^c\}} \widehat{\text{FSP}}_a^c(t) \leq \widehat{\text{FSP}}_a^c(\tau_c) \leq \alpha.$$

where the first two equalities are due to the definition of R-value and the monotonicity adjustment, whereas the first inequality follows from (D.28) and the second inequality follows from the definition of τ_a^c. Hence $\mathbb{I}\{R_j^c \leq \alpha\} = 1$, proving the first direction of the equivalence.

Conversely, suppose that $\mathbb{I}\{R_j^c \leq \alpha\} = 1$. Without loss of generality, assume that $A_j = a$. By the definition of the R-value, we have

$$R_j^c = \tilde{Q}_j^c = \min_{\{t\in\mathbf{S}_a^c : t\leq\hat{S}_j^c\}} \widehat{\text{FSP}}_a^c(t) \leq \alpha.$$

That is, there exists a threshold $t \leq \hat{S}_j^c$ in $\mathbf{S}_a^c$ such that $\widehat{\text{FSP}}_a^c(t) \leq \alpha$. It follows that

$$\tau_a^c = \min_{t\in\mathbf{S}_a^c}\left\{t : \widehat{\text{FSP}}_a^c(t) \leq \alpha\right\} \leq \min_{t\in\mathbf{S}_a^c, t\leq\hat{S}_j^c}\left\{t : \widehat{\text{FSP}}_a^c(t) \leq \alpha\right\} \leq \hat{S}_j^c,$$

implying that $\delta_j' = \sum_{\kappa\in\mathcal{A}} \mathbb{I}(A_j = \kappa)\cdot\mathbb{I}(\hat{S}_j^c \geq \tau_\kappa^c) = \mathbb{I}(A_j = a)\mathbb{I}(\hat{S}_j^c \geq \tau_a^c) = 1$.

Combining the two arguments above, we have $\delta_j = 1 \Longleftrightarrow \delta_j' = 1$. We can similarly show that $\delta_j = 0 \Longleftrightarrow \delta_j' = 0$, establishing (D.22).

Remark 12. In sensitive applications, there may be reservations about procedures that employ different thresholds based on membership in a protected group, even if the resulting fairness guarantees are equivalent. Therefore, it is crucial to communicate our fairness guarantee in a manner that avoids misinterpretation associated with the use of multiple thresholds. A key advantage of the FASI procedure based on the R-value [the first thresholding rule in (D.22)] is its simplicity in conveying fairness while remaining user-friendly in operation. By employing a single universal threshold, we provide a tool that is more interpretable from a fairness perspective. For practitioners, this approach is significantly easier to understand compared to managing multiple thresholds (when multiple thresholds are involved, it can become challenging to explain to the public how the resulting classification algorithm is fair).

D.2 Proof of part (b)

The proof is more complicated but follows essentially the same strategy of the proof for part (a). Details are provided for new arguments and omitted for repeated arguments.

D.2.1 Preliminaries and notations

Consider the R-value (13) that utilizes both $\mathcal{D}_a^{cal} \cup \mathcal{D}_a^{test}$ via (11). The estimated FSP in group a for a given threshold t is:

$$\widehat{\mathrm{FSP}}_a^c(t) = \frac{\left\{\sum_{i\in\mathcal{D}_a^{cal}} \mathbb{I}(\hat{S}_i^c \geq t, Y_i \neq c) + 1\right\}/(n_a+1)}{\left\{\sum_{i\in\mathcal{D}_a^{test}\cup\mathcal{D}_a^{cal}} \mathbb{I}(\hat{S}_i^c \geq t) + 1\right\}/(n_a+m_a+1)}. \tag{D.29}$$

Now employing the stable version of the FSP estimate (D.29), define

$$\tau_a^c = \left[\min\left\{t \in \mathbf{S}_a^c : \widehat{\mathrm{FSP}}_a^c(t) \leq \alpha\right\}\right] \vee 0.5,$$

Similar to the previous proof, Lemma 1 indicates that the FASI algorithm is equivalent to the following thresholding rule based on confidence scores:

$$\hat{Y}_j = \sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a)\left\{\sum_{c\in\{1,2\}} c \cdot \mathbb{I}(\hat{S}_j^c \geq \tau_a^c)\right\}, j \in \mathcal{D}^{test}.$$

Consider the modified FSP definition in Theorem 1: $\mathrm{FSR}_a^{\{c\},*} = \mathbb{E}\left[\mathrm{FSP}_a^{\{c\},*}(\tau_a^c)\right]$, where

$$\mathrm{FSP}_a^{\{c\},*}(\tau_a^c) = \frac{\sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{S}_i^c \geq \tau_a^c, Y_j \neq c)}{\sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{S}_i^c \geq \tau_a^c) + 1}.$$

It follows from the definition (D.29) and the operation of FASI algorithm that

$$\mathrm{FSP}_a^{\{c\},*}(\tau_a^c) \leq \alpha \cdot \frac{n_a+1}{n_a+m_a+1} \cdot \frac{\sum_{i\in\mathcal{D}_a^{test}\cup\mathcal{D}_a^{cal}} \mathbb{I}(\hat{S}_i^c \geq \tau_a^c) + 1}{\sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{S}_i^c \geq \tau_a^c, Y_j \neq c) + 1} \cdot \frac{\sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{S}_i^c \geq \tau_a^c, Y_j \neq c)}{\sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(\hat{S}_i^c \geq \tau_a^c) + 1}.$$

The product of the last two terms can be reorganized as

$$\begin{aligned}
&\left\{1+\frac{\sum_{i\in\mathcal{D}_a^{cal}}\mathbb{I}(\hat{S}_i^c\geq\tau_a)}{\sum_{j\in\mathcal{D}_a^{test}}\mathbb{I}(\hat{S}_i^c\geq\tau_a)+1}\right\}\cdot\frac{\sum_{j\in\mathcal{D}_a^{test}}\mathbb{I}(\hat{S}_i^c\geq\tau_a,Y_j\neq c)}{\sum_{j\in\mathcal{D}_a^{cal}}\mathbb{I}(\hat{S}_i^c\geq\tau_a,Y_j\neq c)+1}\\
=&\left\{1+\frac{\sum_{i\in\mathcal{D}_a^{cal}}\mathbb{I}(\hat{S}_i^c\geq\tau_a,Y_i=c)+\sum_{i\in\mathcal{D}_a^{cal}}\mathbb{I}(\hat{S}_i^c\geq\tau_a,Y_i\neq c)}{\sum_{j\in\mathcal{D}_a^{test}}\mathbb{I}(\hat{S}_i^c\geq\tau_a,Y_i=c)+\sum_{j\in\mathcal{D}_a^{test}}\mathbb{I}(\hat{S}_i^c\geq\tau_a,Y_i\neq c)+1}\right\}\cdot M_1(\tau_a^c)\\
\leq& M_1(\tau_a^c)+\max\left\{M_2(\tau_a^c),M_3(\tau_a^c)\right\},
\end{aligned}\tag{D.30}$$

where in the above equation, we have three martingales respectively defined as:

$$\begin{aligned}
M_1(\tau_a^c) &= \frac{\sum_{j\in\mathcal{D}_a^{test}}\mathbb{I}(\hat{S}_i^c\geq\tau_a^c,Y_j\neq c)}{\sum_{j\in\mathcal{D}_a^{cal}}\mathbb{I}(\hat{S}_i^c\geq\tau_a^c,Y_j\neq c)+1};\\
M_2(\tau_a^c) &= \frac{\sum_{i\in\mathcal{D}_a^{cal}}\mathbb{I}(\hat{S}_i^c\geq\tau_a^c,Y_i\neq c)}{\sum_{i\in\mathcal{D}_a^{cal}}\mathbb{I}(\hat{S}_i^c\geq\tau_a^c,Y_i\neq c)+1};\\
M_3(\tau_a^c) &= \frac{\sum_{i\in\mathcal{D}_a^{cal}}\mathbb{I}(\hat{S}_i^c\geq\tau_a^c,Y_i=c)}{\sum_{i\in\mathcal{D}_a^{test}}\mathbb{I}(\hat{S}_i^c\geq\tau_a^c,Y_i=c)+1}\cdot\frac{\sum_{i\in\mathcal{D}_a^{test}}\mathbb{I}(\hat{S}_i^c\geq\tau_a^c,Y_i\neq c)}{\sum_{i\in\mathcal{D}_a^{cal}}\mathbb{I}(\hat{S}_i^c\geq\tau_a^c,Y_i\neq c)+1}.
\end{aligned}$$

In the derivation of (D.30), we have used the following inequality:

$$\frac{x+y}{z+(t+1)}\cdot\frac{z}{x+1}\leq\max\left\{\frac{x}{x+1},\frac{yz}{(t+1)(x+1)}\right\}$$

for any positive integers x, y, z, and t. The proof for this inequality is elementary and hence omitted.

D.2.2 The main proof

Noting that $M_2(\tau_a^c)\leq 1$ holds trivially true, and utilizing the fact $\max(x,y)=\frac{(x+y)}{2}+\frac{|x-y|}{2}$, we can easily derive the following upper bound:

$$\mathbb{E}\left[\mathrm{FSP}_a^{\{c\},*}(\tau_a^c)\right]\leq\mathbb{E}\left[\frac{\alpha(n_a+1)}{n_a+m_a+1}\cdot\left\{M_1(\tau_a^c)+\frac{M_3(\tau_a^c)+1}{2}+\frac{|M_3(\tau_a^c)-1|}{2}\right\}\right].\tag{D.31}$$

To characterize the FSP process, recall the counts in Section D.1.3:

$$\begin{aligned}
V_a^t(t) &= \sum_{j\in\mathcal{D}_a^{test}}\mathbb{I}(\hat{S}_j^c\geq t,Y_j\neq c),\ W_a^t(t)=\sum_{j\in\mathcal{D}_a^{test}}\mathbb{I}(\hat{S}_j^c\geq t,Y_j=c),\ K_a^t(t)=\sum_{j\in\mathcal{D}_a^{test}}\mathbb{I}(\hat{S}_j^c\geq t),\\
V_a^c(t) &= \sum_{i\in\mathcal{D}_a^{cal}}\mathbb{I}(\hat{S}_i^c\geq t,Y_i\neq c),\ W_a^c(t)=\sum_{i\in\mathcal{D}_a^{cal}}\mathbb{I}(\hat{S}_i^c\geq t,Y_i=c),\ K_a^c(t)=\sum_{i\in\mathcal{D}_a^{cal}}\mathbb{I}(\hat{S}_i^c\geq t),
\end{aligned}$$

defined respectively for the test set and calibration set. Moreover, we employ the same discrete-time filtration as in Section D.1.3:

$$\mathcal{F}_k^a=\sigma\left\{V_a^t(s_j),V_a^c(s_j),W_a^t(s_j),W_a^c(s_j):j=m_a^*,m_a^*-1,\ldots,k\right\}.$$

Again, the threshold τ_a^c of the FASI algorithm with modified mirror process (D.29) is a stopping time conditional on $\mathcal{F}_k^a$.

In the next subsections (Sections D.2.3 and D.2.4), we demonstrate that both $M_1(t)$ and

$M_3(t)$ are super-martingales adapted to $\mathcal{F}_k^a$. By applying the optional stopping theorem, we obtain that

$$\mathbb{E}\big\{M_1(\tau_a^c)\big\} \le \frac{m_a^0}{n_a^0+1} \quad \text{and} \quad \mathbb{E}\big\{M_3(\tau_a^c)\big\} \le \frac{n_a^1 m_a^0}{(m_a^1+1)(n_a^0+1)}, \tag{D.32}$$

where $m_a^0 = \sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(Y_j \neq c)$, $m_a^1 = \sum_{j\in\mathcal{D}_a^{test}} \mathbb{I}(Y_j = c)$, $n_a^0 = \sum_{i\in\mathcal{D}_a^{cal}} \mathbb{I}(Y_j \neq c)$, and $n_a^1 = \sum_{i\in\mathcal{D}_a^{cal}} \mathbb{I}(Y_j = c)$. As we have focused on selection individuals from class c, for simplicity, we have suppressed c in the above notations. It follows from (D.31) and (D.32) that

$$\begin{aligned}
\mathbb{E}\left[\mathrm{FSP}_a^{\{c\},*}(\tau_a^c)\right] &= \mathbb{E}\left[\frac{\alpha(n_a+1)}{n_a+m_a+1}\cdot\left\{\frac{m_a^0}{n_a^0+1}+\frac{1}{2}+\frac{n_a^1\cdot m_a^0}{2(m_a^1+1)(n_a^0+1)}+\frac{|M_3(\tau_a^c)-1|}{2}\right\}\right] \\
&\le \alpha_c\gamma'_{c,a} + \frac{\alpha_c}{2}\mathbb{E}|M_3(\tau_a^c)-1|,
\end{aligned}$$

proving the desired result. □

Remark 13. Under the exchangeability condition, some simple calculations (considering only the leading terms) show that

$$\begin{aligned}
\gamma'_{c,a} &= \mathbb{E}\left[\frac{(n_a+1)}{n_a+m_a+1}\cdot\left\{\frac{m_a^0}{n_a^0+1}+\frac{1}{2}+\frac{n_a^1\cdot m_a^0}{2(m_a^1+1)(n_a^0+1)}\right\}\right] \\
&= \frac{1}{2}\mathbb{E}\left\{\frac{n_a(m_a^0+n_a^0)}{n_a^0(m_a+n_a)}+\frac{n_a m_a^0}{n_a^0 m_a}\cdot\frac{m_a(m_a^1+n_a^1)}{m_a^1(m_a+n_a)}\right\}+o(1).
\end{aligned}$$

We can see that $\gamma'_{c,a}$ is very similar to $\gamma_{c,a}$ (defined in Part (a) of the theorem), as both are essentially related to the empirical proportions. Moreover, under the exchangeability condition, both $\gamma'_{c,a}$ and $\gamma_{c,a}$ are very close to 1.

Moreover, in Section D.3, we discuss sufficient conditions under which the data-driven threshold satisfies the almost sure convergence $\tau_a^c \xrightarrow{a.s.} \tau^*$ for some constant $\tau^* \in (0,1)$. In the asymptotic regime, we assume that $m_a^0 \asymp m_a^1 \asymp m_a$ and $n_a^0 \asymp n_a^1 \asymp n_a$, with all of these quantities diverging to infinity. Specifically, we establish the strong convergence of the data-driven threshold τ_a^c. It follows that

$$\lim_{n_a,m_a\to\infty}\mathbb{E}|M_3(\tau_a^c)-1| = \left|\frac{\mathbb{P}(S\ge\tau^*,Y=c)}{\mathbb{P}(S\ge\tau^*,Y\neq c)}\cdot\frac{\mathbb{P}(S\ge\tau^*,Y\neq c)}{\mathbb{P}(S\ge\tau^*,Y=c)}-1\right| = 0.$$

Hence the stable version of the FASI algorithm controls the FSR asymptotically:

$$\mathbb{E}\left[\mathrm{FSP}_a^{\{c\},*}(\tau_a^c)\right] \le \alpha_c + o(1).$$

D.2.3 Martingale arguments for $M_1(t)$

Consider the events A_1 to A_4 defined in Section D.1.3. The conditional probabilities of these events along the filtration $\mathcal{F}_k^a$ are given by

$$\begin{aligned}
\mathbb{P}(A_1|\mathcal{F}_k^a) &= \gamma_k\cdot\frac{V_a^t(s_k)}{V_a^t(s_k)+V_a^c(s_k)}, \quad \mathbb{P}(A_2|\mathcal{F}_k^a) = \gamma_k\cdot\frac{V_a^c(s_k)}{V_a^t(s_k)+V_a^c(s_k)}, \\
\mathbb{P}(A_3|\mathcal{F}_k^a) &= (1-\gamma_k)\cdot\frac{W_a^t(s_k)}{W_a^t(s_k)+W_a^c(s_k)}, \quad \mathbb{P}(A_4|\mathcal{F}_k^a) = (1-\gamma_k)\cdot\frac{W_a^c(s_k)}{W_a^t(s_k)+W_a^c(s_k)}.
\end{aligned}$$

It is easy to verify that $M_1(t) = \frac{V_a^t(t)}{V_a^c(t)+1}$ is a backward-running martingale by noting that:

$$\begin{aligned}
&\mathbb{E}\left\{M_1(s_{k-1})|\mathcal{F}_k^a\right\} \\
= \;& \frac{V_a^t(s_k)-1}{V_a^c(s_k)+1}\cdot\gamma_k\cdot\frac{V_a^t(s_k)}{V_a^t(s_k)+V_a^c(s_k)} + \frac{V_a^t(s_k)}{V_a^c(s_k)}\cdot\gamma_k\cdot\frac{V_a^c(s_k)}{V_a^t(s_k)+V_a^c(s_k)} + \frac{V_a^t(s_k)}{V_a^c(s_k)+1}(1-\gamma_k) \\
= \;& \frac{V_a^t(s_k)}{V_a^c(s_k)+1} \;=\; M_1(s_k). \quad \square
\end{aligned}$$

D.2.4 Martingale arguments for $M_3(t)$

We consider the same events, same probabilities and same filtration as before. Write

$$M_3(t) = \frac{W_a^c(t)}{W_a^t(t)+1}\cdot\frac{V_a^t(t)}{V_a^c(t)+1}.$$

To see why $M_3(t)$ is a martingale, note that

$$\begin{aligned}
&\mathbb{E}\left\{M_3(s_{k-1})|\mathcal{F}_k^a\right\} \\
= \;& \frac{W_a^c(s_k)}{W_a^t(s_k)+1}\cdot\frac{V_a^t(s_k)-1}{V_a^c(s_k)+1}\cdot\gamma_k\cdot\frac{V_a^t(s_k)}{V_a^t(s_k)+V_a^c(s_k)} + \frac{W_a^c(s_k)}{W_a^t(s_k)+1}\cdot\frac{V_a^t(s_k)}{V_a^c(s_k)}\cdot\gamma_k\cdot\frac{V_a^c(s_k)}{V_a^t(s_k)+V_a^c(s_k)} \\
&+\frac{W_a^c(s_k)}{W_a^t(s_k)}\cdot\frac{V_a^t(s_k)}{V_a^c(s_k)+1}\cdot(1-\gamma_k)\cdot\frac{W_a^t(s_k)}{W_a^t(s_k)+W_a^c(s_k)} \\
&+\frac{W_a^c(s_k)-1}{W_a^t(s_k)+1}\cdot\frac{V_a^t(s_k)}{V_a^c(s_k)+1}\cdot(1-\gamma_k)\cdot\frac{W_a^c(s_k)}{W_a^t(s_k)+W_a^c(s_k)}.
\end{aligned}$$

Some simple calculations yield

$$\mathbb{E}\left\{M_3(s_{k-1})|\mathcal{F}_k^a\right\} = \frac{W_a^c(s_k)}{W_a^t(s_k)+1}\cdot\frac{V_a^t(s_k)}{V_a^c(s_k)+1} \;=\; M_3(s_k). \quad \square$$

D.3 Asymptotic analysis of upper bounds

To rigorously establish the asymptotic validity of the stable version of FASI, we conduct an asymptotic analysis of the residual term from Theorem 1(b). This analysis precisely shows that

$$\lim_{(n,m)\to\infty}\mathbb{E}\left|\texttt{Res}(\tau)-1\right| = 0,$$

thereby corroborating our numerical results and formally confirming our intuition that FASI effectively controls the FSR. In contrast to the finite-sample theory presented in the main text, this component of our theory utilizes classical limit theorems, which require standard regularity conditions commonly employed in statistics. It is important to emphasize that these conditions are necessitated by the available theoretical tools and do not imply that the FASI method is inherently dependent on them.

D.3.1 Preliminaries and assumptions

Since our focus is on group-wise FSR control and given that the theoretical analysis can be applied to each group to establish the properties of the FASI algorithm, we restrict our analysis to a particular group. The asymptotic regime assumes that $m_a^0 \asymp m_a^1 \asymp m_a$ and $n_a^0 \asymp n_a^1 \asymp n_a$,

with all of these quantities diverging to infinity. In what follows, we omit the notation a for group membership in order to reduce notational complexity.

Suppose our objective is to select cases with $Y_j = 2$. To simplify the discussion, we assume that the scores are defined as $S_i := 1 - \hat{S}^{c=2}(X_i, A_i) \in (0, 1)$. Consequently, we select individuals when their scores are small. This slightly modified notation system aligns more closely with the framework of multiple testing while utilizing the same decision rule as before.

Denote the estimated FSP as

$$\widehat{\mathrm{FSP}}(t) = \frac{\left\{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(S_i < t, Y_i \neq c) + 1\right\}/(n+1)}{\left\{\sum_{i\in\mathcal{D}^{test}\cup\mathcal{D}^{cal}} \mathbb{I}(S_i < t) + 1\right\}/(n+m+1)}.$$

Then the data-driven threshold of the FASI algorithm is given by

$$\tau = \left\{t \in [0, 1] : \widehat{\mathrm{FSP}}(t) \leq \alpha\right\}. \tag{D.33}$$

Our analysis focuses on the scenario where S_i are continuous random variables, such as softmax score outputs from machine learning algorithms. The data points (S_i, Y_i) are assumed to be i.i.d., which is reasonable because: (i) the score functions are trained using an independent dataset, and (ii) the De Finetti Theorem can be employed to identify a latent variable, enabling analysis conditional on that latent variable. This assumption leads to the random mixture model defined in (B.2), as considered in Section B.3. Omitting the group membership a notation, the CDF of the scores (within a specific group) is given by:

$$F(t) = \mathbb{P}(S_i < t, Y = 1) + \mathbb{P}(S_i < t, Y = 2) = \pi_1 F_1(t) + \pi_2 F_2(t), \quad i \in \mathcal{D}^{\mathrm{cal}} \cup \mathcal{D}^{\mathrm{test}}, \tag{D.34}$$

where $\pi_c = \mathbb{P}(Y_i = c)$ for $c \in \{1, 2\}$. We assume that the conditional CDFs $F_1(t)$ and $F_2(t)$ are strictly greater than 0 on the interval $(0, 1)$.

Define the marginal false selection rate as $\mathrm{mFSR}(t) = \pi_1 F_1(t)/F(t)$. The oracle threshold τ^* is defined as:

$$\tau^* = \sup\left\{t \in (0, 1) : \mathrm{mFSR}(t) \leq \alpha\right\}. \tag{D.35}$$

The marginal FSR is analogous to the marginal false discovery rate (mFDR) in multiple testing (Genovese & Wasserman 2002, Sun & Cai 2007).

We introduce the following assumption, which plays a key role in our asymptotic analysis to ensure the existence and convergence of the data-driven threshold. This critical assumption was also utilized in Jin & Candes (2023), albeit in a slightly different context.

Assumption 2. *Let $t \in [0, 1]$ be a threshold for the scores S_j. Denote α as the nominal FSR level, and mFSR(t) as the marginal FSR of the thresholding rule $\{I(S_i < t) : i \in [m]\}$. For any $\varepsilon > 0$, we have*

$$\exists \ \ \tau^* - \varepsilon < \tau' < \tau^*, \text{ such that } mFSR(\tau') < \alpha. \tag{D.36}$$

Remark 14. Assumption 2 can be derived from the Monotone Likelihood Ratio Condition (MLRC, Sun & Cai 2007). Under this condition, it can be demonstrated that mFSR(t) is monotonically increasing in t (Cao et al. 2013). For i.i.d. scores S_i following the random mixture model (D.34), the MLRC naturally implies (D.36). The MLRC is a common (and often implicit) condition in the FDR literature. For instance, it simplifies to the assumption of concavity of the p-value CDF (cf. Genovese & Wasserman 2002, Storey 2002). Therefore, while Assumption 2 might appear complex, it is actually less restrictive than many standard assumptions commonly used in the FDR literature.

D.3.2 Key lemmas

We first state three lemmas. The first two, which are similar to the large-sample theories in Storey et al. (2004), are consequences of the well-known Glivenko-Cantelli Theorem, and are therefore stated without proofs.

Again, we omit the group notation a; let $n = |\mathcal{D}^{\text{cal}}|$ and $m = |\mathcal{D}^{\text{test}}|$.

Lemma 2. *Consider the random mixture model (D.34). The empirical CDFs obey the following strong uniform convergence:*

$$\sup_{t\in(0,1)} \left| \frac{1 + \sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(S_i < t)}{n+1} - F(t) \right| \xrightarrow{a.s.} 0; \quad \sup_{t\in(0,1)} \left| \frac{1 + \sum_{j\in\mathcal{D}^{test}} \mathbb{I}(S_j < t)}{m+1} - F(t) \right| \xrightarrow{a.s.} 0.$$

Lemma 3. *Consider the random mixture model (D.34). The conditional empirical CDFs obey the following strong uniform convergence: for $c \in \{1, 2\}$,*

$$\begin{aligned} \sup_{t\in(0,1)} \left| \frac{1 + \sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(S_i < t, Y_i = c)}{n+1} - \pi_c F_c(t) \right| &\xrightarrow{a.s.} 0; \\ \sup_{t\in(0,1)} \left| \frac{1 + \sum_{i\in\mathcal{D}^{test}} \mathbb{I}(S_j < t, Y_j = c)}{m+1} - \pi_c F_c(t) \right| &\xrightarrow{a.s.} 0. \end{aligned}$$

The final lemma, which is integral to establishing the strong convergence of the data-driven threshold, follows from standard ϵ-N arguments; we provide a proof in Section D.3.5 for completeness.

Lemma 4. *Let $\{f_n\}_{n\geq 1} : [0, 1] \to (0, \infty)$ be a sequence of functions and $f : [0, 1] \to (0, \infty)$ be another function. For a given constant α, define*

$$\tau_n = \sup\{t \in [0, 1] : f_n(t) \leq \alpha\} \quad \text{and} \quad \tau^* = \sup\{t \in [0, 1] : f(t) \leq \alpha\}.$$

Assume the following conditions hold:

(i) For any $\epsilon > 0$, there exists some $t \in [\tau^* - \epsilon, \tau^*)$ such that $f(t) < \alpha$;

(ii) There exists a $\delta \in (0, \tau^*)$ satisfying $f(\delta) < \alpha$ such that $\sup_{t\in[\delta,1)} |f_n(t) - f(t)| \xrightarrow{a.s.} 0$.

Then we have $\tau_n \xrightarrow{a.s.} \tau^*$.

D.3.3 Uniform convergence of the threshold

Now we establish the strong convergence of τ as mentioned in Remark 13 in Appendix D.2.2. The next proposition establishes the strong convergence of the data-driven threshold.

Proposition 2. *Let τ [defined by (D.33)] and τ^* [defined by (D.35)] denote the data-driven and oracle thresholds, respectively. Under Model (D.34) and Assumption 2, we have $\tau \xrightarrow{a.s.} \tau^*$.*

Proof. To prove the desired result, we apply Lemma 4 by substituting $f(t)$ and $f_n(t)$ with $\text{mFSR}(t)$ and $\widehat{\text{FSP}}(t)$, respectively. Our primary task is to ensure that both Conditions (i) and (ii) of the lemma are met. Condition (i) is satisfied as a result of Condition (D.36) from Assumption 2, combined with the equation $\text{mFSR}(\tau^*) = \alpha$. Further, according to Lemmas 2 and 3, we can therefore find a $\kappa > 0$ such that the subsequent uniform strong convergence holds:

$$\sup_{t\in[\kappa,1)} \left| \widehat{\text{FSP}}(t) - \text{mFSR}(t) \right| \xrightarrow{a.s.} 0.$$

Thus, Condition (ii) is fulfilled. Recognizing the equal roles of τ and τ^* in both Lemma 4 and our proposition, by applying Lemma 4 we conclude that $\tau \xrightarrow{a.s.} \tau^*$. □

D.3.4 Asymptotic FSR control

To demonstrate that $\lim_{(n,m)\to\infty} \mathbb{E}\,|\texttt{Res}(\tau) - 1| = 0$, it suffices to prove the following lemma.

Lemma 5. *Consider Model (D.34). Suppose $\tau \xrightarrow{a.s.} \tau^* \in (0,1)$. Then, for $c \in \{1,2\}$, we have*

$$\begin{aligned} \mathbb{P}\left\{\lim_{n,m\to\infty} \frac{1+\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(S_i \le \tau, Y_i = c)}{n+1} = \pi_c F_c(\tau^*)\right\} &= 1; \\ \mathbb{P}\left\{\lim_{n,m\to\infty} \frac{1+\sum_{j\in\mathcal{D}^{test}} \mathbb{I}(S_j \le \tau, Y_j = c)}{m+1} = \pi_c F_c(\tau^*)\right\} &= 1. \end{aligned} \tag{D.37}$$

Proof. We only prove the first equality in (D.37), as the second can be established in a similar manner. The proof involves two simple decompositions. The first decomposition is

$$\begin{aligned} & \frac{1+\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(S_i \le \tau, Y_i = c)}{n+1} - \pi_c F_c(\tau^*) \\ = & \left[\frac{1+\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(S_i \le \tau, Y_i = c)}{n+1} - \frac{1+\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(S_i \le \tau^*, Y_i = c)}{n+1}\right] \\ & + \left[\lim_{n,m\to\infty} \frac{1+\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(S_i \le \tau^*, Y_i = c)}{n+1} - \pi_c F_c(\tau^*)\right] \\ = & \text{ I} + \text{II}. \end{aligned}$$

Term II converges to 0 almost surely due to the uniform strong convergence of the CDFs. To address Term I, we employ a second decomposition:

$$\begin{aligned} & \frac{1+\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(S_i \le \tau, Y_i = c)}{n+1} - \frac{1+\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(S_i \le \tau^*, Y_i = c)}{n+1} \\ = & \frac{1+\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(\tau^* < S_i \le \tau, Y_i = c)}{n+1} - \frac{1+\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(\tau < S_i \le \tau^*, Y_i = c)}{n+1}. \end{aligned} \tag{D.38}$$

We can see that (D.38) converges to 0 almost surely if S_i is continuous and $\tau \xrightarrow{a.s.} \tau^*$. This completes the proof of the lemma, thereby establishing asymptotic FSR control. □

D.3.5 Proof of Lemma 4

The lemma can be proven by combining the results from two directions.

Direction 1 (lower bound). We show that $\underline{\lim}_{n\to\infty} \tau_n \ge \tau^*$ almost surely. Let $\epsilon > 0$ be an arbitrarily small constant. Condition (i) implies that there exists $t_\epsilon \in [\tau^* - \epsilon, \tau^*)$ such that $f(t_\epsilon) < \alpha$. Since $t_\epsilon \ge \delta$ for sufficiently small ϵ, the uniform convergence in Condition (ii) applies on $[\delta, 1)$. Hence we can find N_1 such that for all $n \ge N_1$, $|f_n(t_\epsilon) - f(t_\epsilon)| < \alpha - f(t_\epsilon)$. This implies $f_n(t_\epsilon) < f(t_\epsilon) + \alpha - f(t_\epsilon) = \alpha$. Since $f_n(t_\epsilon) < \alpha$, we have $\tau_n \ge t_\epsilon \ge \tau^* - \epsilon$ for all $n \ge N_1$. As $\epsilon > 0$ is arbitrary, we conclude that $\underline{\lim}_{n\to\infty} \tau_n \ge \tau^*$ almost surely.

Direction 2 (upper bound). We now show the more complicated direction:

$$\overline{\lim}_{n\to\infty} \tau_n \le \tau^* \text{ almost surely.}$$

We argue by contradiction. If the upper bound does not hold, then we can find a subsequence $\{n'\}$ and $\epsilon > 0$ such that $\tau_{n'} \ge \tau^* + \epsilon$ holds with positive probability. Consider an arbitrary

$t_0 \in [\tau^*, 1)$. By the definition of τ^*, we must have $f(t_0) > \alpha$. Note that if we choose $\epsilon > 0$ sufficiently small, then $t_0 \geq \delta$, so we can apply the uniform convergence in Condition (ii). Specifically, we can find N_2 such that for all $n' \geq N_2$, $|f_{n'}(t_0) - f(t_0)| < \frac{f(t_0)-\alpha}{2}$. Hence,

$$f_{n'}(t_0) > f(t_0) - \frac{f(t_0) - \alpha}{2} > \alpha.$$

This indicates that for this subsequence $\{n'\}$, we must have $\tau_{n'} < t_0$ (and the inequality is strict); otherwise, it would contradict the definition of $\tau_{n'}$. Finally, since this argument holds for an arbitrary $t_0 \in [\tau^*, 1)$, we conclude that $\overline{\lim}_{n\to\infty} \tau_n \leq \tau^*$ almost surely. □

E Proof of Theorem 2

The theorem implies that the optimal confidence score for constructing R-values should be $S^c_{OR}(x, a) = \mathbb{P}(Y = c|X = x, A = a)$. A similar optimality theory has been developed in the context of multiple testing with groups (Cai & Sun 2009). However, the proof for the binary classification setup with the indecision option is much more complicated; we provide the proof here for completeness. We first establish an essential monotonicity property in Section E.1, then prove the optimality theory in Section E.2.

E.1 A monotonicity property

The oracle rule employs $\{S^{c,j}_{OR} : c \in \{1, 2\}, j \in \mathcal{D}^{test}\}$ as the confidence scores. The corresponding oracle theoretical R-values $\{\mathtt{TR}^{c,j}_{OR} : c \in \{1, 2\}, j \in \mathcal{D}^{test}\}$ can be obtained via the conversion algorithm in Appendix B.5 [cf. Equation B.5]. Let

$$\boldsymbol{t} = \left\{t^c_a : t^c_a \in [0.5, 1], c \in \{1, 2\}, a \in \mathcal{A}\right\}$$

be a collection of eligible thresholds.

Consider a class of thresholding rules of the form:

$$d^j_{OR}(\boldsymbol{t}) = \sum_{c\in\{1,2\}} c \cdot \left\{\sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a)\mathbb{I}(S^{c,j}_{OR} > t^c_a)\right\}, \quad j \in \mathcal{D}^{test}.$$

Denote the mFSR level in group a for selecting class c as $Q^c_{OR,a}(t^c_a)$. The next proposition characterizes the relationship between $Q^c_{OR,a}(t^c_a)$ and t^c_a.

Proposition 3. *$Q^c_{OR,a}(t^c_a)$ is monotonically decreasing in t^c_a.*

Proof of Proposition 3. Define $\tilde{Q}^c_{OR,a}(t^c_a) = 1 - Q^c_{OR,a}(t^c_a)$. We only need to show that $Q^c_{OR,a}(t^c_a)$ is monotonically increasing in t^c_a. Let $\mathcal{M}_a = \{j \in \mathcal{D}^{test} : A_j = a\}$. According to the definition of the mFSR and the definition of $S^c_{OR,j}$, we have

$$\mathbb{E}\left\{\sum_{j\in\mathcal{M}_a} \left\{S^c_{OR,j} - \tilde{Q}^c_a(t^c_a)\right\} \mathbb{I}(S^c_{OR,j} > t^c_a)\right\} = 0, \tag{E.39}$$

where the expectation is taken over $\mathcal{D}^{test}_a$. It is important to note that the oracle procedure, which assumes that all distributional information is known, does not utilize $\mathcal{D}^{train}$ and $\mathcal{D}^{cal}$. It is easy to see from Equation (E.39) that $\tilde{Q}^c_a(t) > t^c_a$ otherwise the summation on the LHS must be positive, leading to a contradiction.

Next we show that $t_1 < t_2$ implies $\tilde{Q}^c_{OR,a}(t_1) \leq \tilde{Q}^c_{OR,a}(t_2)$. We argue by contradiction. Assume instead that $\tilde{Q}^c_{OR,a}(t_1) > \tilde{Q}^c_{OR,a}(t_2)$, then we have

$$\begin{aligned}
&\sum_{j\in\mathcal{M}_a} \{S^{c,j}_{OR} - \tilde{Q}^c_{OR,a}(t_1)\}\mathbb{I}(S^{c,j}_{OR} > t_1) \\
= &\sum_{j\in\mathcal{M}_a} \{S^{c,j}_{OR} - \tilde{Q}^c_{OR,a}(t_2) + \tilde{Q}^c_{OR,a}(t_2) - \tilde{Q}^c_{OR,a}(t_1)\}\mathbb{I}(S^{c,j}_{OR} > t_1) \\
= &\sum_{j\in\mathcal{M}_a} \{S^{c,j}_{OR} - \tilde{Q}^c_{OR,a}(t_2)\}\mathbb{I}(S^{c,j}_{OR} > t_2) + \sum_{j\in\mathcal{M}_a} \{S^{c,j}_{OR} - \tilde{Q}^c_{OR,a}(t_2)\}\mathbb{I}(t_1 \leq S^{c,j}_{OR} \leq t_2) \\
&+ \sum_{j\in\mathcal{M}_a} \left\{\tilde{Q}^c_{OR,a}(t_2) - \tilde{Q}^c_{OR,a}(t_1)\right\}\mathbb{I}(S^{c,j}_{OR} > t_1) = I + II + III.
\end{aligned}$$

Taking expectations on both sides, it is easy to see that the LHS is zero. However, the RHS is strictly greater than zero. For term I, we have $\mathbb{E}(I) = 0$ according to the definition of mFSR. For term II, we have $\mathbb{E}(II) < 0$ as we always have $\tilde{Q}^c_{OR,a}(t) > t$. For term III, we have $\mathbb{E}(III) < 0$ since we assume $\tilde{Q}^c_{OR,a}(t_1) > \tilde{Q}^c_{OR,a}(t_2)$. It follows that the assumption $\tilde{Q}^c_{OR,a}(t_1) > \tilde{Q}^c_{OR,a}(t_2)$ cannot be true, and the proposition is proved. □

Remark 15. The proposition is essential for expressing the oracle procedure as a thresholding rule based on $S^{c,j}_{OR}$. Specifically, denote $Q^{c,-1}_{OR,a}(\cdot)$ the inverse of $Q^c_{OR,a}(\cdot)$. The monotonicity of $Q^c_{OR,a}(t)$ and the definition of the theoretical R-value together imply that for all $S^{c,j}_{OR} > 0.5$, we have for $j \in \mathcal{D}^{test}$, $S^{c,j}_{OR} = \sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a) \cdot Q^{c,-1}_{OR,a}(\mathtt{TR}^{c,j}_{OR})$. For notational convenience, let $T_j = \mathbb{P}(Y_j = 2|X_j = x, A_j = a)$. Then $S^{1,j}_{OR} = 1 - T_j$ and $S^{2,j}_{OR} = T_j$. Let $t^{c,a}_{OR} = \max\{0.5, (Q^{c,a}_{OR})^{-1}(\alpha_c)\}$. The oracle rule can be written as, for $j \in \mathcal{D}^{test}$,

$$\begin{aligned}
\delta^j_{OR}(X_j, A_j) &= \textstyle\sum_{c\in\{0,1\}} c \cdot \mathbb{I}(\mathtt{TR}^{c,j}_{OR} \leq \alpha_c) \\
&= \textstyle\sum_{c\in\{1,2\}} c \cdot \left\{\sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a)\mathbb{I}\left(S^{c,j}_{OR} \geq t^{c,a}_{OR}\right)\right\} \\
&= \textstyle\sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a)\left\{\mathbb{I}\left(T_j \leq 1 - t^{c,a}_{OR}\right) + 2\mathbb{I}\left(T_j \geq t^{c,a}_{OR}\right)\right\}.
\end{aligned}$$

E.2 Proof of the theorem

Define the expected number of true selections $\mathrm{ETS} = \sum_{j\in\mathcal{D}^{test}} \mathbb{I}(Y_j = c, \hat{Y}_j = c)$. Then it can be shown that minimizing the EPI subject to the FSR constraint is equivalent to maximizing the ETS subject to the same constraint.

According to Proposition 3, the oracle rule can be written as

$$\delta^j_{OR} \coloneqq \delta^j_{OR}(X_j, A_j) = \textstyle\sum_{a\in\mathcal{A}} \mathbb{I}(A_j = a)\left\{\mathbb{I}\left(T_j \leq 1 - t^{c,a}_{OR}\right) + 2\mathbb{I}\left(T_j \geq t^{c,a}_{OR}\right)\right\}.$$

The group-wise mFSR constraints for the oracle rule imply that, for all $a \in \mathcal{A}$:

$$\mathbb{E}\left\{\sum_{j\in\mathcal{M}_a} (T_j - \alpha_1)\mathbb{I}(\delta^j_{OR} = 1)\right\} = 0,\ \mathbb{E}\left\{\sum_{j\in\mathcal{M}_a} (1 - T_j - \alpha_2)\mathbb{I}(\delta^j_{OR} = 2)\right\} = 0. \tag{E.40}$$

Let $\boldsymbol{\delta} \in \{0, 1, 2\}^m$ be a general selection rule in $\mathcal{D}_{\alpha_1,\alpha_2}$. Then the mFSR constraints for $\boldsymbol{\delta}$ implies that, for all $a \in \mathcal{A}$,

$$\mathbb{E}\left\{\sum_{j\in\mathcal{M}_a} (T_j - \alpha_1)\mathbb{I}(\delta_j = 1)\right\} \leq 0,\ \mathbb{E}\left\{\sum_{j\in\mathcal{M}_a} (1 - T_j - \alpha_2)\mathbb{I}(\delta_j = 2)\right\} \leq 0. \tag{E.41}$$

The ETS of $\boldsymbol{\delta} = \{\delta_j : j \in \mathcal{D}^{test}\}$ is given by

$$\text{ETS}_{\boldsymbol{\delta}} = \mathbb{E}\left[\sum_{a\in\mathcal{A}}\sum_{j\in\mathcal{M}_a}\{\mathbb{I}(\delta_j = 1)(1-T_j) + \mathbb{I}(\delta_j = 2)T_j\}\right] = \sum_{a\in\mathcal{A}}(\text{ETS}^{1,a}_{\boldsymbol{\delta}} + \text{ETS}^{2,a}_{\boldsymbol{\delta}}).$$

The goal is to show that $\text{ETS}(\boldsymbol{\delta}^{OR}) \geq \text{ETS}(\boldsymbol{\delta})$. We only need to show $\text{ETS}^{c,a}_{\boldsymbol{\delta}^{OR}} \geq \text{ETS}^{c,a}_{\boldsymbol{\delta}}$ for all c and a. We will show $\text{ETS}^{1,a}_{\boldsymbol{\delta}^{OR}} \geq \text{ETS}^{1,a}_{\boldsymbol{\delta}}$ for a given a. The remaining inequalities follow similar arguments. According to (E.40) and (E.41), we have

$$\mathbb{E}\left[\sum_{j\in\mathcal{M}_a}(T_j - \alpha_1)\left\{\mathbb{I}(\delta^j_{OR} = 1) - \mathbb{I}(\delta_j = 1)\right\}\right] \geq 0. \tag{E.42}$$

Let $\lambda_{1,a} = (1 - t^{c,a}_{OR} - \alpha_1)/t^{c,a}_{OR}$. It can be shown that $\lambda_{1,a} > 0$. For $i \in \mathcal{M}_a$, we claim that the oracle rule can be equivalently written as

$$\delta^j_{OR} = \mathbb{I}\left\{\frac{T_j - \alpha_1}{1 - T_j} < \lambda_{1,a}\right\}.$$

Using the previous expression and techniques similar to the Neyman-Pearson lemma, we claim that the following result holds for all $j \in \mathcal{M}_a$:

$$\left\{\mathbb{I}(\delta^j_{OR} = 1) - \mathbb{I}(\delta_j = 1)\right\}\left\{T_j - \alpha_1 - \lambda_{1,a}(1 - T_j)\right\} \leq 0.$$

It follows that

$$\mathbb{E}\left[\sum_{j\in\mathcal{M}_a}\left\{\mathbb{I}(\delta^j_{OR} = 1) - \mathbb{I}(\delta_j = 1)\right\}\left\{T_j - \alpha_1 - \lambda_{1,a}(1 - T_j)\right\}\right] \leq 0. \tag{E.43}$$

According to (E.42) and (E.43), we have

$$\lambda_{1,a}\mathbb{E}\sum_{j\in\mathcal{M}_a}(1 - T_j)\left\{\mathbb{I}(\delta^j_{OR} = 1) - \mathbb{I}(\delta_j = 1)\right\} = \lambda_{1,a}\left(\text{ETS}^{1,a}_{\boldsymbol{\delta}^{OR}} - \text{ETS}^{1,a}_{\boldsymbol{\delta}}\right) \geq 0.$$

Note that $\lambda_{1,a} > 0$, the desired result follows. The theorem is proved by combining the results from all groups $a \in \mathcal{A}$.

F Related Fairness Algorithms

We discuss two closely related works developed based on the sufficiency principle, in order to emphasize the advantages of FASI.

Zeng et al. (2022) presents a group-wise thresholding rule that maximizes the classifier's power subject to the constraints imposed by the sufficiency principle. However, this method does not allow for indecisions, thereby rendering it impossible to control the error rate at user-specified levels. In contrast, Lee et al. (2021) proposes a selective classification procedure that satisfies the sufficiency principle and allows for indecisions. This enables fair decision-making with error rate control. However, the approach by Lee et al. (2021) relies on complex fitting algorithms and imposes stringent assumptions for theoretical development, which lacks reliable theoretical guarantees regarding output reliability in practical scenarios. Moreover, both methods fail to address the issue of inflated decision errors that arise when classifying multiple individuals simultaneously.

Table 1: Comparison of algorithms developed to fulfill the sufficiency principle.

	User Specified Error Rate	Finite Sample Theory	No Assumptions on Model Accuracy
Zeng et al. (2022): FairBayes-DPP	No	No	No
Lee et al. (2021): Fair Selective Classification Via Sufficiency	No	No	No
FASI	Yes	Yes	Yes

We emphasize that the choice of fairness definition should be contextual and informed by the specific automated decision-making scenario. FASI offers several advantages that make it a more practical choice for practitioners. Firstly, in high-stakes scenarios, the proposed selective inference framework with an indecision option effectively handles situations where the consequences of incorrect decisions are significant. This approach provides practitioners with guidance on which observations require further attention, rather than automatically making decisions when the accuracy may not be sufficient. Secondly, when multiple individuals need to be classified simultaneously, it is crucial to employ a suitable error criterion that can aggregate cumulative errors and control for multiplicity. FASI addresses this concern by providing the FSR, which generalizes the powerful and practical FDR criterion in large-scale testing problems. Lastly, in scenarios where complex or blackbox machine learning models are utilized, having a model-free algorithm like FASI becomes essential. This allows for the deployment of user-specified blackbox models while simultaneously ensuring provable validity in controlling the associated risks without imposing strong model assumptions.

G The setup of multinomial classification

Let $\mathcal{C}' \subset \mathcal{C}$ represent the set of classes to be selected. With indecisions being allowed, the action space is $\Lambda = \{0, \mathcal{C}'\}$. We denote the selection rule for m individuals in the test set as $\{\hat{Y}_j : j \in \mathcal{D}^{test}\} \in \Lambda^m$.

The FSR can be defined in two ways with respect to the subset $\mathcal{C}'$. The first definition evaluates the fraction of incorrect selections for each individual class separately:

$$\mathrm{FSR}_a^{\{c\}} = \mathbb{E}\left[\frac{\sum_{j\in\mathcal{D}^{test}} \mathbb{I}(\hat{Y}_j = c, Y_j \neq c, A_j = a)}{\left\{\sum_{j\in\mathcal{D}^{test}} \mathbb{I}(\hat{Y}_j = c, A_j = a)\right\} \vee 1}\right], \quad \text{for all } a \in \mathcal{A} \text{ and } c \in \mathcal{C}'.$$

By contrast, the second definition calculates an overall error rate by combining selections from all classes in $\mathcal{C}'$:

$$\mathrm{FSR}_a^{\mathcal{C}'} = \mathbb{E}\left[\frac{\sum_{j\in\mathcal{D}^{test}} \mathbb{I}(\hat{Y}_j \in \mathcal{C}', \hat{Y}_j \neq Y_j, A_j = a)}{\left\{\sum_{j\in\mathcal{D}^{test}} \mathbb{I}(\hat{Y}_j \in \mathcal{C}', A_j = a)\right\} \vee 1}\right], \quad \text{for all } a \in \mathcal{A}.$$

The second definition of FSR introduces several complicated issues. Firstly, it requires the employment of a new score function to achieve optimality under the oracle setting. Secondly, substantial adjustments must be made to the mirror process described in Section 3.2. Thirdly, the development of martingale theories becomes notably more intricate. Finally, when dealing with scenarios involving more than two classes, an additional layer of complexity arises. These various issues offer intriguing and crucial avenues for future exploration and research.

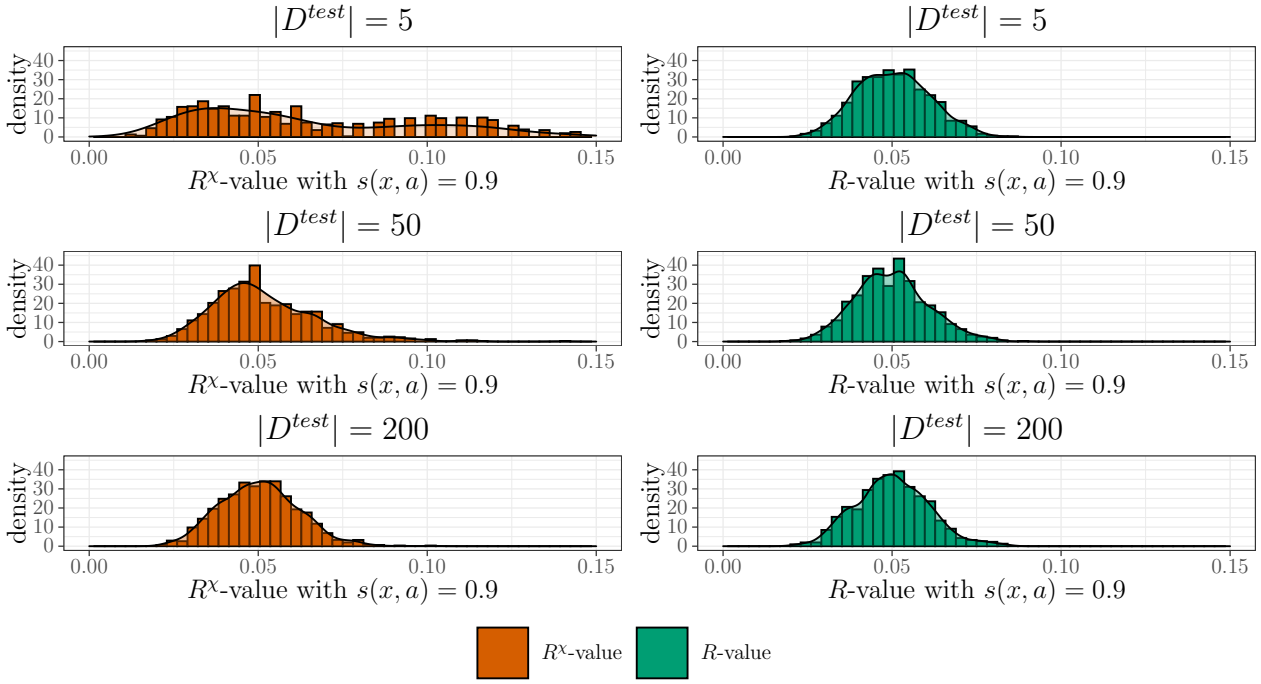

Figure 8: The comparison between the R^{χ}-value and R-value for varying sizes of the test data set. The left column shows the histograms of the R^{χ}-value (orange) and the right column shows the histograms of the R-value (green). The R^{χ}-values and R-values are computed for a fixed confidence score of $s(x, a) = 0.9$ based on $1,000$ randomly generated data sets.

H Additional Numerical Results

H.1 The stable version of the R-value

In this section, we present simulation results to demonstrate that when $|\mathcal{D}^{test}|$ is small, the stable version of the R-value [defined via (11), (12), and (13)] exhibits lower variability than the R-value defined via (10), (12), and (13). The only difference is that the stable version employs both test and calibration data to stabilize the denominator of the FSP estimate. For clarity and easy presentation, we refer to the version using more data as the R-value and the version using fewer data points as the R^{χ}-value in this subsection. However, in the main text, we do not give different names to these two R-values because their basic construction steps and underlying ideas are identical.

To illustrate important patterns in variability, we examine the distribution of the R-value corresponding to a fixed confidence score of $s(x, a) = 0.9$.

We consider the setting described in Section 4 with $F_{1,M} = F_{1,F} = \mathcal{N}(\boldsymbol{\mu}_1, 2 \cdot \mathbf{I}_3)$ and $F_{2,M} = F_{2,F} = \mathcal{N}(\boldsymbol{\mu}_2, 2 \cdot \mathbf{I}_3)$. We set $\pi_{2|F} = \pi_{2|M} = 0.8$, $\boldsymbol{\mu}_1 = (1, 1, 1)^\top$ and $\boldsymbol{\mu}_2 = (2, 2, 2)^\top$. The confidence scores are constructed as the oracle class probabilities $P(Y = c|X, A)$.

In Figure 11, we compute $1,000$ R^{χ}-values and R-values for a fixed score of $s = 0.9$ based on randomly generated $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$. The size of the calibration set is fixed at $|\mathcal{D}^{cal}| = 1,000$ and the test set has sizes $|\mathcal{D}^{test}| \in \{5, 50, 200\}$. The columns of Figure 11 show the histograms the R^{χ}-values (left) and R-values (right) with $\mathcal{D}^{test}$ increasing from 5 (first row) to 200 (last row).

When $|\mathcal{D}^{test}| = 5$, we notice that the R^{χ}-value has much more variability than the R-value. This is because the denominator of the R^{χ}-value only utilizes 5 observations when computing the total number of selections. By contrast, the R-value uses $1,005$ observations since it has access to data from both $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$. Moving further down the rows of Figure 11, the advantage of the R-value slowly disappears as $|\mathcal{D}^{test}|$ increases. This causes the variability of both R^{χ}-value

and R-value to become almost identical.

We conclude from this small simulation that the R-value [defined via (11)] is more desirable in settings where $|\mathcal{D}^{test}|$ is small since it can use more data to decrease its variability. However, while the R^X-value [defined via (10)] has more variability for small $|\mathcal{D}^{test}|$, this disadvantage can be quickly overcome through the introduction of a reasonably sized test set.

On the other hand, the R^X-value defined via (10) offers finite-sample guarantees for FSR control, whereas the R-value defined via (11) controls the FSR only asymptotically. In practice, however, the differences in FSR levels between the two versions of FASI are negligible.

H.2 Imbalanced group sizes

In this section, we revisit Simulations 1 and 2 presented in Section 4 to examine the impact of imbalanced group sizes on FASI's performance. In addition to evaluating group-wise FSRs, we also consider the overall FSR levels, as defined in Equation (4). It is important to note that our methodology is primarily designed to control group-wise FSRs; while controlling these does not automatically ensure overall FSR control, our simulations did not reveal violations of the overall FSR levels.

Our simulation setups are similar to those in Section 4, except that we now vary π_M, the proportion of the Male protected group, from 0.05 to 0.95, rather than fixing $\pi_M = \pi_F = 0.5$. We consider two settings: (i) $\pi_{2|F} = \pi_{2|M} = 0.5$, and (ii) $\pi_{2|F} = 0.5$ with $\pi_{2|M} = 0.2$. The results from these settings are illustrated in Figures 9 and 10, respectively. The following observations can be made:

- The FASI method controls both the group-wise and overall FSR at the nominal level across all values of π_M. However, when π_M is very small, the Male group-wise FSR control tends to be conservative due to the small sample size.
- When the conditional proportions are similar (e.g., $\pi_{2|F} = \pi_{2|M} = 0.5$), indicating minimal disparity between male and female distributions, the FCC method performs well in terms of FSR control. In contrast, when heterogeneity is more pronounced [i.e., Setting (ii) with $\pi_{2|F} > \pi_{2|M}$], the FCC method only controls the overall FSR but fails to control the group-wise FSRs.

H.3 Numerical investigations of the factor $\gamma_{c,a}$

In Theorem 1, we show that the FASI algorithm can control the FSR at level $\gamma_{c,a}\alpha_c$. This section investigates the deviations of $\gamma_{c,a}$ from 1. For simplicity, we only focus on $\gamma_{1,a}$. The setup of the simulations is identical to that in Section 4.

Figure 8 shows the estimates of $\gamma_{1,a}$ for both the Female (green solid line) and Male (orange dashed line) groups. We vary $\pi_{2|F}$ from 0.15 to 0.85 while fixing $\pi_{2|M} = 0.5$. The y-axis plots the estimate of $\gamma_{1,a}$ averaged over $1,000$ independent simulation runs. In both settings, $\gamma_{1,a}$ is nearly 1 across both the Female and Male groups. In the most extreme setting ($\pi_{1|F} = 0.85$), $\gamma_{1,a}$ deviates away from 1 by 0.01.

H.4 FASI deployed with other machine learning models

One of the attractive guarantees of our proposed selective inference framework is that we can have the guarantees of Theorem 1, regardless of the machine learning algorithm that is used to generate the confidence scores. In this section, Figure 12 replicates the results of Simulation 1 in Section 4, for a variety of machine learning models where the data has two protected groups,

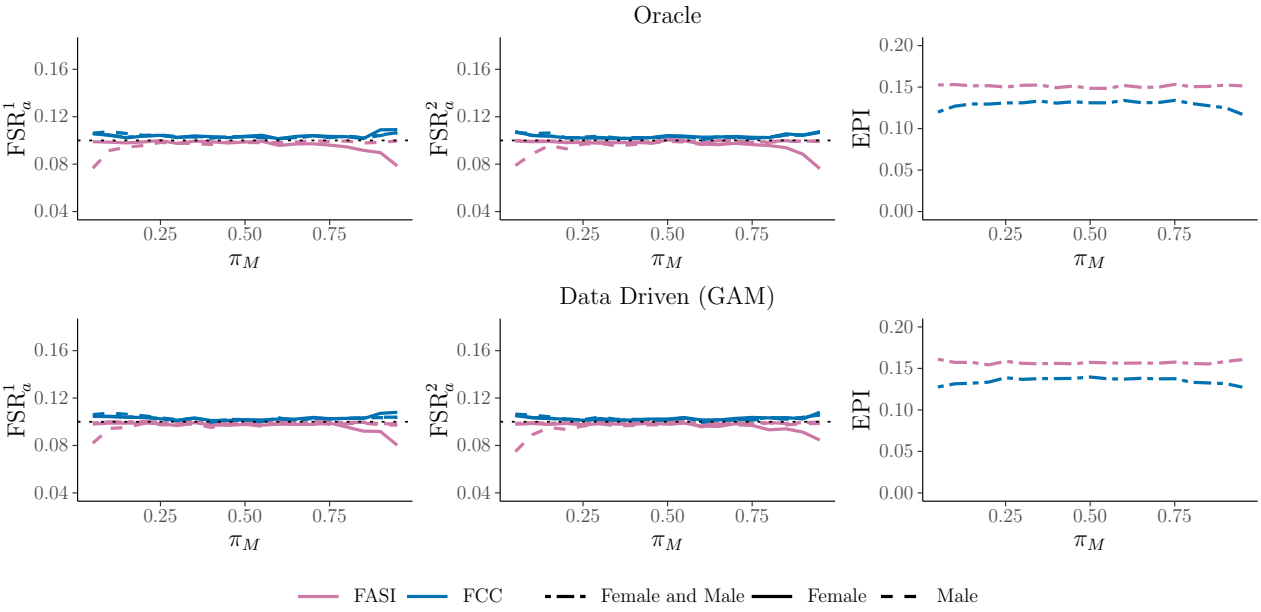

Figure 9: A similar setup to Simulation 1 (described in Section 4) but with $\pi_{2|F} = \pi_{2|M} = 0.5$, and π_M ranging from 0.05 to 0.95.

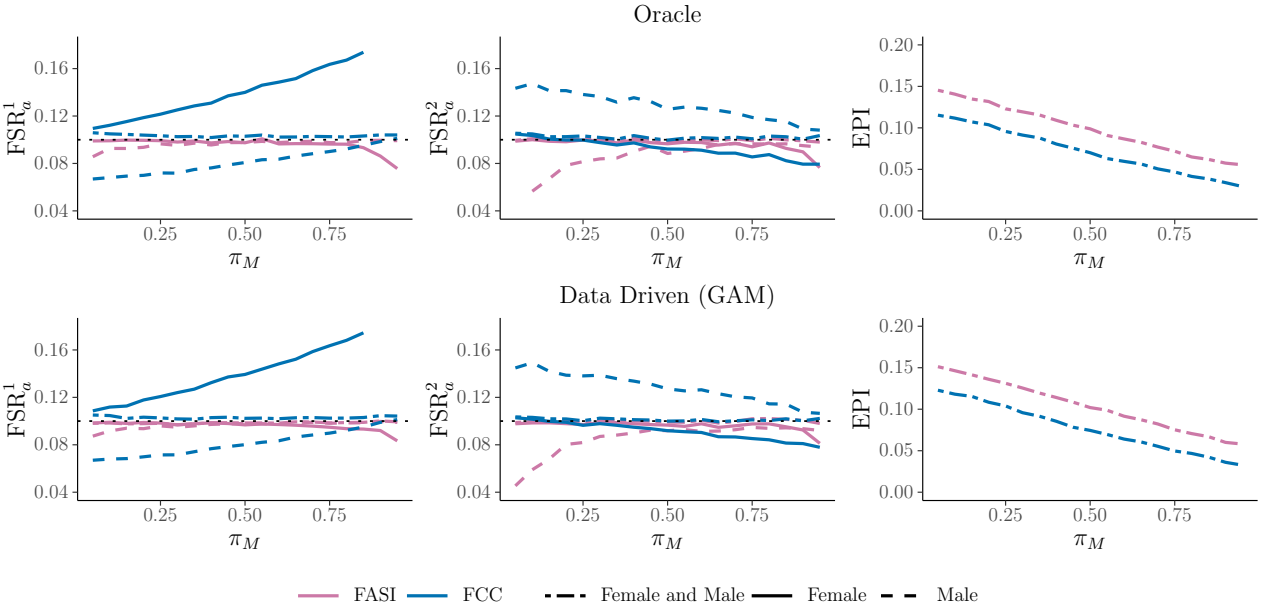

Figure 10: A similar setup to Figure 9 above in this response file and Simulation 1 (described in Section 4). However, now with $\pi_{2|F} = 0.5, \pi_{2|M} = 0.2$, and π_M ranging from 0.05 to 0.95. Overall FSR is controlled for all values of π_M.

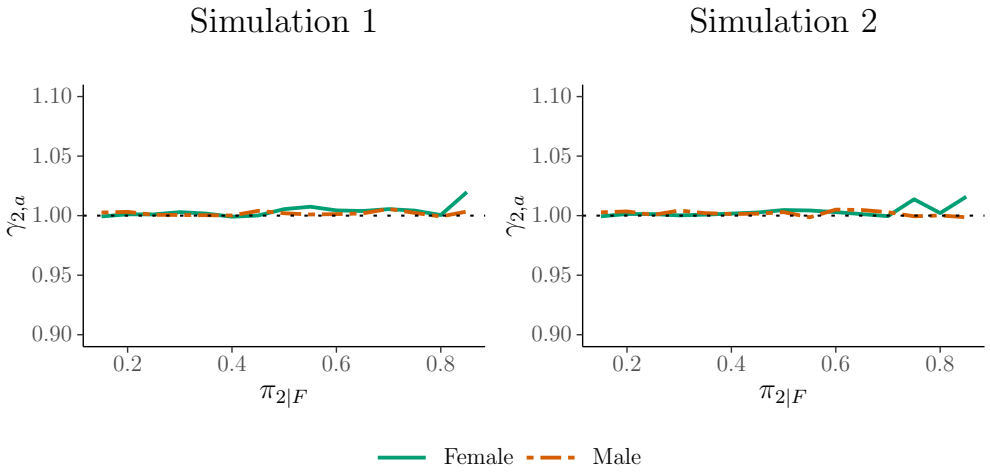

Figure 11: Estimates of $\gamma_{1,a}$ from the simulations in Section 4. The solid (green) line represents the estimate of $\gamma_{1,\mathrm{F}}$ for the Female protected group and similarly the orange (long-dashed) line for the Male protected group.

Female and Male. In this section we use, logistic regression, GAM, Nonparametric Naive Bayes, and XGBoost (James et al. 2023, Hastie et al. 2009, Silverman 1986, Chen & Guestrin 2016) to estimate the confidence scores that will be converted to the R-values for our FASI framework.

The left column of Figure 12 plots the FSR for classification group 2 against a varying proportion of signal $\pi_{2|F}$ from the Female protected group i.e. the true proportion of Females that belong to class 2. The right column shows the corresponding EPI for each ML model. The goal is to control FSR at the 10% level.

As we go down the rows, we notice that every model is able to effectively control the False Selection Rate (similar to Simulation 1), however each model has a different EPI. Here, it seems that Logistic Regression, GAM and Nonparametric Naive Bayes have a similar EPI that gets close to 20% in the most extreme case. However, XGBoost has a slightly higher EPI that gets closer to 30% in the worst case. This is a consequence of the accuracy that each ML model has when estimating the true conditional probability $P(Y = 2|X, A)$ for use in our FASI algorithm. However while some models are more or less accurate than others, they are all able to control the FSR at the desired level.

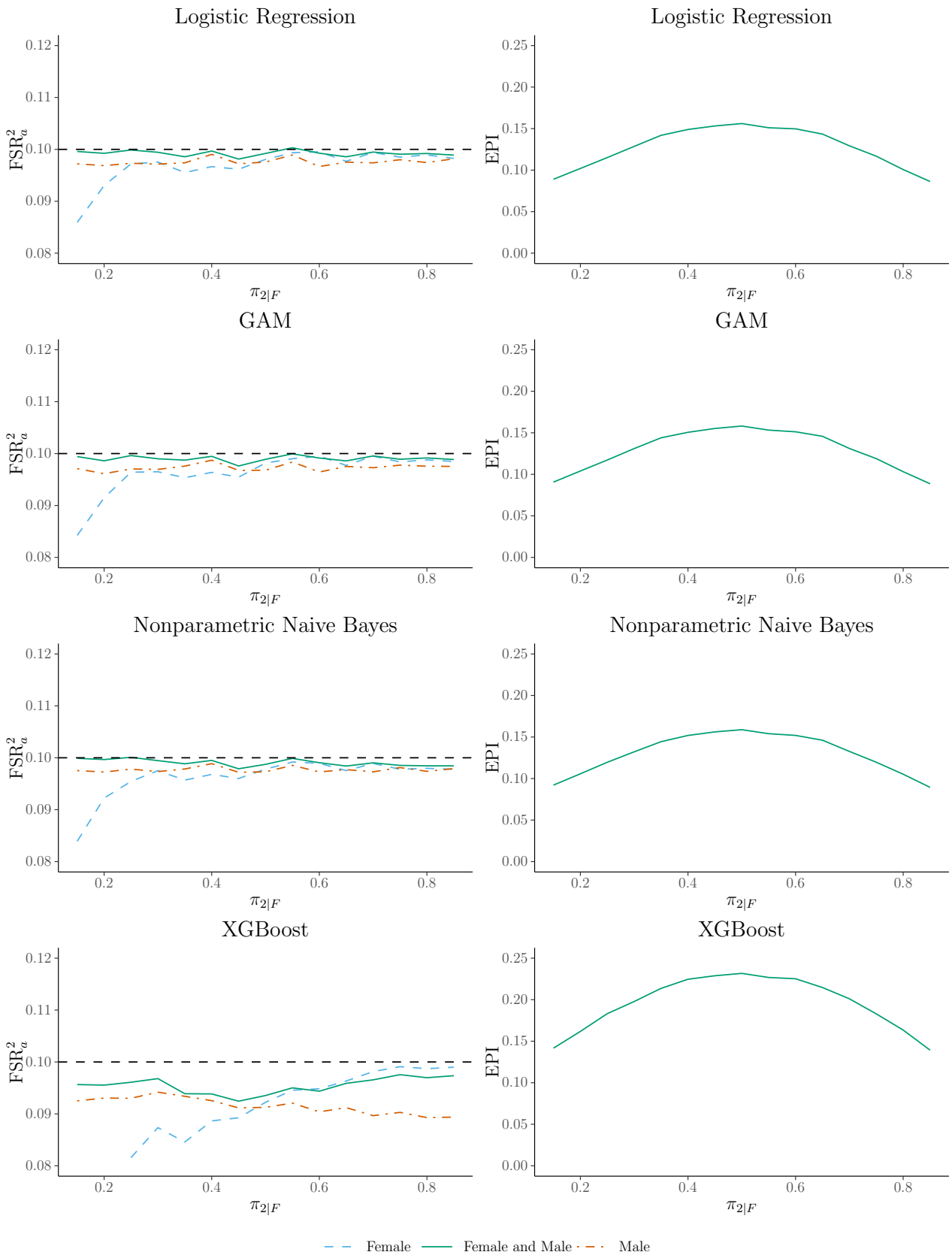

Figure 12: FSR control for the high risk classification. Left column: The resulting FSR from multiple different ML models that are used to estimate the confidence scores used to calculate the R-value. Right column: The corresponding EPI from different confidence scores. The overall FSR (green / solid) as well as both the Female (blue / dashed) and Male (orange / dot-dashed) protected group FSR's are controlled at the desired 10% level, for all ML algorithms. The x-axis varies the amount of true proportion of high risk observations from the Female protected group, while fixing the true proportion from the male group at 50%.