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 fairness 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 and equalizing 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. The selection rules based on R-values, which effectively mitigate disparate impacts on protected groups, are provably valid for FSR control in finite samples. We demonstrate the numerical performance of our approach through both simulated and real data.

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

[1]University of Sydney Business School.
[2]Center for Data Science and School of Management, Zhejiang University.
[3]Goizueta Business School, Emory University.
[4]Department of Data Sciences and Operations, University of Southern California.

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 develping 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 who cannot be selected into any classes with

confidence. These observations can 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 to ensure that the FSR is effectively controlled and equalized across protected groups, while trying to minimize the total wasted opportunities.

However, 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 hand plot of Figure 1 shows the *False Selection Proportions* (FSP), [1] i.e. the fraction of individuals who did not recommit a crime among those who were classified as recidivists. The classification rule was constructed via a Generalized Additive Model (GAM) [2] (James et al. 2023, Hastie et al. 2009) 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

[1] The term FSR is reserved to refer to the expected value of the FSP.

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

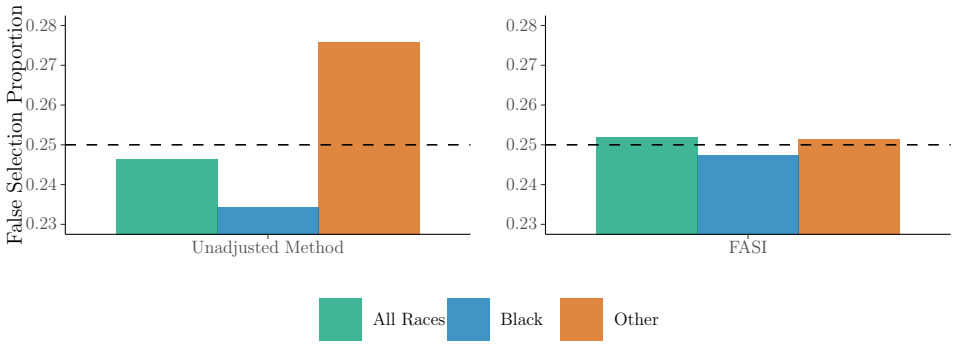

Figure 1: The selection of recidivists from a pool of criminal defendants (Broward County, Florida). The target FSR is 25%. Left: the unadjusted approach. Right: the proposed FASI approach.

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 base 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 various 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/or 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. If we build an algorithm around black-box models then it often becomes intractable to compute thresholds for controlling the FSR over multiple protected groups.

To address these issues, we develop a data-driven FASI algorithm, which is specifically de-

signed to control both the overall FSR, and the (protected) group-wise FSRs, below a user specified level of risk α. The right panel of Figure 1 illustrates the FSPs of FASI on the recidivism data. Now, not only is the overall FSP controlled at 25%, but so are the individual race FSPs. FASI works by converting the confidence scores from a black-box classifier 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. First, we introduce a new notion of fairness that involves controlling, not only the overall FSR, but also the FSR for designated sub-groups. Ours is not the only approach to fairness and we do not claim that it is universally superior relative to alternative approaches (Dwork et al. 2012, Hardt et al. 2016, Romano, Bates & Candès 2020). However, in high-consequence scenarios where decisions with high ambiguity are present, it may be beneficial to withhold or separately evaluate them, until more evidence is gathered, to reduce the risk of making definitive decisions. Therefore, controlling all sub-group FSRs is a reasonable approach for ensuring fairness in these scenarios. Second, we develop a data-driven FASI algorithm based on the R-value. The algorithm, which can be implemented with any user-specified classifier, is intuitively appealing and easy to interpret. Third, we provide theoretical results both justifying the use of R-values and the effectiveness of the FASI algorithm for FSR control. Our finite-sample theory is established with minimal assumptions: we allow the confidence scores used for classification to be generated from black-box classifiers, and make no assumptions about their accuracy. Finally, the strong empirical performance of FASI is demonstrated via both simulated and real data.

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 con-

cludes the main article with a discussion of related works and possible extensions. We establish theoretical properties of FASI in the Appendix Section A.

2 Problem Formulation

Suppose we observe a data set $\mathcal{D} = \{(X_i, A_i, Y_i) : i = 1, \ldots, n\}$, where $X_i \in \mathbb{R}^p$ 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 true class label taking values in $\mathcal{C} = \{1, \ldots, C\}$, with its predicted value denoted by $\hat{Y}_i$. The goal is to predict the classes for multiple individuals with instances (X_{n+j}, A_{n+j}), $j = 1, \ldots, m$.

2.1 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:

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

Several issues exist for machine learning approaches that satisfy the sufficiency principle (Zeng et al. 2022, Pleiss et al. 2017, Zafar et al. 2017). For example, the calibration by group method (Barocas et al. 2017) cannot control the misclassification rate at a user-specified level in high-stakes situations. Additionally, existing methods cannot tackle the issue of multiplicity, which is caused by the inflation of misclassification errors when multiple individuals are classified

simultaneously. Furthermore, sophisticated classifiers trained from black-box models tend to be complex and computationally intensive, posing significant challenges for analyzing uncertainty associated with their predictions. In particular, most analyses often involve strong assumptions about the underlying model and the accuracy of its outputs, which may not hold in practice.

Our proposed approach, detailed in the following sections, involves providing an indecision option for individuals who require further review before a definitive decision can be made. This enables effective error rate control at user-specified levels, and sets us apart from most algorithms that impose definitive decisions on all subjects. By integrating the fairness notion with the task of error rate control, we propose a modified version of the sufficiency principle:

$$\mathbb{P}(Y \neq \hat{Y} | \hat{Y} = c, A = a) \leq \alpha \text{ for all } a \in \mathcal{A}, \tag{2}$$

which ensures that the error rate, given that a subject is classified into class c, is controlled below a user-specified level α for all protected groups. We further adapt (2) to address the multiplicity issue by introducing a novel concept called the false selection rate (FSR, Section 2.3). The FSR framework, which defines the error rate by focusing only on subjects that are selected or classified into class c, is inspired by the widely used and powerful idea of false discovery rate (Benjamini & Hochberg 1995) in multiple testing. To achieve rigorous theoretical guarantees, we develop algorithms that effectively control the FSR at any user-specified level in finite samples, without making any assumptions about the underlying model, the classifier to be used, or the accuracy of the scores.

2.2 A selective inference framework for binary classification

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

Consider an application scenario that involves the prediction of mortgage default, where

$Y = 2$ indicates default and $Y = 1$ otherwise. The current practice is to use risk assessment software to produce a *confidence score*, which is used to classify an individual into "high", "medium" or "low" risk classes. Let $S(x, a)$ denote such a score that maps an instance (x, a) to a real value, with a higher value indicating a higher risk of default. Suppose we observe a new instance $(X^*, A^*) = (x, a)$. Consider a class of decision rules of the form: $\hat{Y} = \mathbb{I}\{S(x, a) < t_l\} + 2\mathbb{I}\{S(x, a) > t_u\}$, where t_l and t_u are thresholds chosen by the investigator to characterize the lower and upper limits of potential risks and $\mathbb{I}(\cdot)$ is the indicator function. $\hat{Y}$ takes three possible values in the action space $\Lambda = \{1, 2, 0\}$, respectively indicating that an individual has low, high and medium risks of default. The value "0", which is referred to as an *indecision* or *reject option* in classification (Herbei & Wegkamp 2006, Sun & Wei 2011, Lei 2014), is used to express "doubt" reflecting that there is not sufficient confidence to make a definitive decision. For example, an individual with $\hat{Y} = 1$ will be approved for a mortgage and an individual with $\hat{Y} = 2$ will be rejected. Whereas an individual with $\hat{Y} = 0$ will be asked to provide additional information and resubmit the application.

Now we turn to a classification task with m individuals whose confidence scores are given by $\boldsymbol{S}^{test} = (S_{n+1}, \cdots, S_{n+m})$. Consider the following decision rule $\hat{\boldsymbol{Y}} = (\hat{Y}_{n+1}, \cdots, \hat{Y}_{n+m})$, where

$$\hat{Y}_{n+j} = \mathbb{I}(S_{n+j} < t_l) + 2\mathbb{I}(S_{n+j} > t_u), \quad \text{for } 1 \le j \le m. \tag{3}$$

We can view (3) as a *selective inference* procedure, which selects individuals with extreme scores into the high and low risk classes, while returning an indecision on the remainder. The selective inference view provides a flexible framework that allows for various types of classification rules. For example, if it is only of interest to select high-risk individuals, then the action space is $\Lambda = \{0, 2\}$, and one can use the following rule,

$$\hat{\boldsymbol{Y}} = (\hat{Y}_{n+1}, \cdots, \hat{Y}_{n+m}), \quad \text{where } \hat{Y}_{n+j} = 2 \cdot \mathbb{I}(S_{n+j} > t_u), \quad 1 \le j \le m. \tag{4}$$

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 context of 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 truly 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 $\mathrm{FSR}^{\mathcal{C}'}$, where $\mathcal{C}' \subset \mathcal{C} = \{1, 2\}$ is the set of class labels that we are interested in selecting. To illustrate the definition, we consider two scenarios. In the first, we select individuals from both classes using rule (3). Denote $\mathcal{S} = \{1 \le j \le m : \hat{Y}_{n+j} \neq 0\}$ the index set of the selected cases and $|\mathcal{S}|$ its cardinality. Then we have

$$\mathrm{FSR}^{\{1,2\}} = \mathbb{E}\left[\frac{\sum_{j\in\mathcal{S}} \mathbb{I}(\hat{Y}_{n+j} \neq Y_{n+j})}{|\mathcal{S}| \vee 1}\right], \tag{5}$$

where $x \vee y = \max\{x, y\}$, and the exact meaning of $\mathbb{E}$ will become clear in Section 3 after we explicitly spell out our algorithm. In the second scenario, our goal is to select individuals in class $c = 2$ using rule (4). Then

$$\mathrm{FSR}^{\{2\}} = \mathbb{E}\left[\frac{\sum_{j=1}^{m} \mathbb{I}(\hat{Y}_{n+j} = 2, Y_{n+j} \neq 2)}{\left\{\sum_{j=1}^{m} \mathbb{I}(\hat{Y}_{n+j} = 2)\right\} \vee 1}\right]. \tag{6}$$

$\mathrm{FSR}^{\{1\}}$ can be defined similarly. Allowing for indecisions enables finding a decision rule that controls both $\mathrm{FSR}^{\{1\}}$ and $\mathrm{FSR}^{\{2\}}$ at a small user-specified level. However, this goal may be unattainable under the standard classification setup, which forces decisions to be made on all individuals. For instance, if the minimum condition on the *classification boundary* (Meinshausen & Rice 2006, Cai & Sun 2017) is not met, simultaneously achieving small FSR^{1} and FSR^{2} will

become impossible.

The FSR is a general concept for selective inference that encompasses important special cases such as the standard misclassification rate, the false discovery rate (FDR; Benjamini & Hochberg 1995) and beyond. If we set both the state space and action space to be $\{1, 2\}$, so there are no indecisions, then the FSR defined by (5) reduces to the (standard) misclassification rate $m^{-1}\mathbb{E}\left\{\sum_{j=1}^{m}(\hat{Y}_{n+j} \neq Y_{n+j})\right\}$. To see the connection to the FDR, consider a multiple testing problem with

$$H_{j0}: Y_{n+j} = 2 \quad vs. \quad H_{j1}: Y_{n+j} = 1, \quad j = 1, \ldots, m.$$

The state space is $\mathcal{C} = \{1, 2\}$. A multiple testing procedure $\hat{\boldsymbol{Y}} = (\hat{Y}_{n+1}, \cdots, \hat{Y}_{n+m}) \in \{0, 1\}^m$ corresponds to a selection rule that aims to select individuals from class 1 only. The action space $\Lambda = \{1, 0\}$ differs from the state space $\mathcal{C}$, with $\hat{Y}_{n+j} = 1$ indicating that H_{j0} is rejected, and $\hat{Y}_{n+j} = 0$ indicating that there is not enough evidence to reject H_{j0}. Then $\mathrm{FSR}^{\{1\}}$ precisely yields the widely used FDR, the expected fraction of false rejections among all rejections.

We use the expected proportion of indecisions (EPI) to describe the power concept (the smaller the EPI the larger the power):

$$\mathrm{EPI} = \tfrac{1}{m}\mathbb{E}\left\{\sum_{j=1}^{m}\mathbb{I}\left(\hat{Y}_{n+j} = 0\right)\right\} = 1 - \mathbb{E}(|\mathcal{S}|)/m. \tag{7}$$

Compared to erroneous decisions, the losses incurred due to indecisions are less consequential since they do not reflect a definitive decision. This leads to a constrained optimization problem where the goal is to develop a selective rule that satisfies $\mathrm{FSR} \leq \alpha$, while making the EPI as small as possible.

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 to

minimize the EPI (7) subject to the following constraint on group-wise FSRs:

$$\mathrm{FSR}_a^{\{c\}} = \mathbb{E}\left[\frac{\sum_{j=1}^m \mathbb{I}(\hat{Y}_{n+j} = c, Y_{n+j} \neq c, A_{n+j} = a)}{\left\{\sum_{j=1}^m \mathbb{I}(\hat{Y}_{n+j} = c, A_{n+j} = a)\right\} \vee 1}\right] \leq \alpha_c, \quad \text{for all } a \in \mathcal{A}. \tag{8}$$

We aim to develop a classification rule that fulfills the fairness criterion (8). This formulation, which adopts a fairness-adjusted error rate constraint, equally bounds the fraction of erroneous decisions among protected groups.

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 base confidence score, and then secondly to turn this score into a decision by setting a threshold. Let $S_{n+j}^c \in [0,1]$ denote a confidence score obtained from a machine learning model that predicts the conditional probability of individual $n+j$ being from class $c \in \mathcal{C}$. Consider a thresholding rule of the form $\hat{\boldsymbol{Y}} = (\hat{Y}_{n+1}, \cdots, \hat{Y}_{n+m})$, where

$$\hat{Y}_{n+j} = c \cdot \mathbb{I}(S_{n+j}^c > t), \quad 1 \leq j \leq m. \tag{9}$$

In binary classification problems, it is customary to set the threshold $t \geq 0.5$ in (9) to prevent overlapping selections.

We now present two approaches for constructing the score S_{n+j}^c in the ideal setting where an oracle has perfect knowledge of the underlying probabilities. The first method, referred to as the "full covariate classifier" (FCC), involves thresholding the following score:

$$S_{n+j}^{c,FCC}(x,a) = \mathbb{P}\left\{Y_{n+j} = c | X_{n+j} = x, A_{n+j} = a\right\}. \tag{10}$$

Here, $S_{n+j}^{c,FCC}(x,a)$ is used to assess the probability of an individual being in class c based on all covariates. The second approach, referred to as the "reduced covariate classifier" (RCC),

involves applying (9) by thresholding the following similar but less informative score:

$$S_{n+j}^{c,RCC}(x) = \mathbb{P}\left\{Y_{n+j} = c | X_{n+j} = x\right\}. \tag{11}$$

In this case, $S_{n+j}^{c,RCC}(x)$ is used to assess the same probability by eliminating the sensitive attribute from the covariate list. However, as we shall illustrate next, both the FCC and RCC approaches are inadequate for addressing the issue of fairness in our context.

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" (default) are different across the protected groups: for the male group, the default proportion p_M is fixed at 50%, whereas for the female group the default proportion p_F varies from 15% to 85%.

We apply the FCC approach and plot the overall FSR and group-wise FSRs as functions of p_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 (10) 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. Furthermore, there are two additional drawbacks of the RCC approach. First, ignoring an informative sensitive attribute can lead to substantial power loss. Second, the feature X can be highly predictive of the sensitive attribute A; hence the classifier is likely to form a *surrogate encoding* of the sensitive attribute based on other features, leading to unfair decisions in a similar fashion as if (10) were used.

It is worth emphasizing that the observations presented in Figure 2 are not tied to a particular classifier; rather, it is a phenomenon that is prevalent across many classifiers. Even if one has

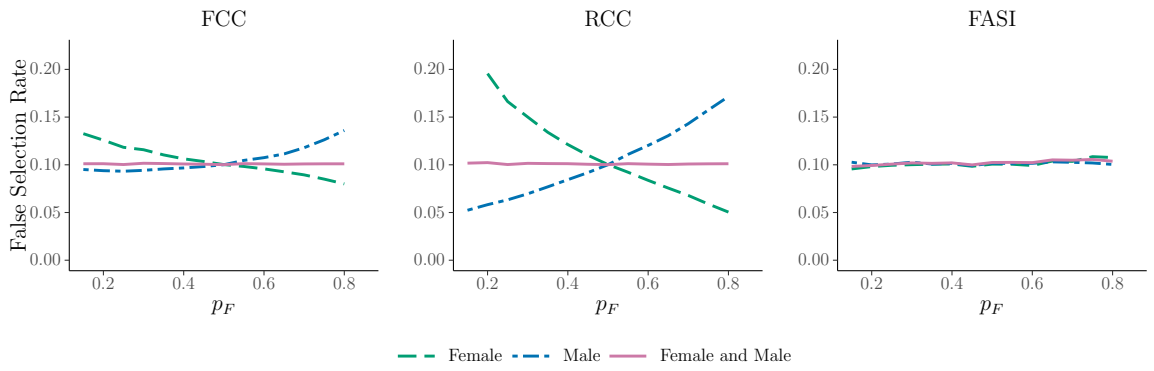

Figure 2: P_M is fixed at 50% and P_F ranges from 15% to 85%. For FCC and RCC, the degree of unfairness increases as p_M and p_F become more disparate. FASI ensures that the group-wise FSRs are effectively controlled and equalized.

access to perfect confidence scores (i.e., oracle posterior probabilities), the unfairness illustrated in Figure 2 would still persist.

To achieve fairness across protected groups, we can adopt two strategies. Strategy (a) entails creating new scores by modifying the current confidence scores such that the new scores are comparable across groups. When combined with thresholding rule (9), this strategy yields fair decisions in accordance with (8). Strategy (b), on the other hand, involves retaining the original confidence scores and setting varying group-adjusted thresholds to ensure compliance with (8).

Strategy (a), which applies a universal threshold to all individuals, is appealing because the decision making process would be straightforward once the adjusted scores are given to practitioners, given that the the new scores are comparable across the groups. By contrast, Strategy (b), although working equally effectively for addressing the fairness issue, can be less intuitive and nontrivial to implement. For practitioners that do not have a full understanding of the underlying algorithm, Strategy (b) can be confusing and even controversial as varied thresholds are being used for different protected groups, causing another level of concern about possible discrimination.

Accordingly, in subsequent sections we have employed Strategy (a) to ensure fairness by criterion (8). We define a score as *unfair* or *illegal* if the application of thresholding rule (9), which employs a universal cutoff, produces larger FSRs for one protected group compared to

the others. As highlighted in Figure 2, both (10) and (11) are unfair scores. Conversely, our proposed FASI algorithm, as demonstrated on the right panel of Figure 2, effectively controls and equalizes the FSRs across the protected groups, thereby fulfilling the fairness criterion (8).

3 Methodology

This section develops a fairness-adjusted selective inference (FASI) procedure for binary classification. We focus on the goal of controlling the FSR^c defined in (6). The methodologies for controlling the FSR of the form (5) and the case of multinomial classification will be briefly discussed in Section 6.

A major challenge in our methodological development is that most state-of-the-art classifiers are constructed based on complex models, which may not offer performance guarantees on their outputs. Consequently, this makes uncertainty quantification and error rate control challenging and even intractable. To overcome this challenge, this section develops a model-free framework that is applicable to any black-box classifier and relies only on the condition of exchangeability between observed and future data.

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 converted to a selection rule via thresholding.

The R-value is computed via the FASI algorithm, which consists of three steps: training, calibrating and thresholding. The observed data set $\mathcal{D} = \{(X_i, A_i, Y_i) : 1 \leq i \leq n\}$ is divided into a training set and a calibration set: $\mathcal{D} = \mathcal{D}^{train} \cup \mathcal{D}^{cal}$. The (future) test set is denoted $\mathcal{D}^{test}$. The first step trains a score function, denoted $\hat{\phi}^c(x, a)$, on $\mathcal{D}^{train}$. By convention a larger score indicates a higher probability of belonging to class c. The scores, often representing estimated class probabilities, can be generated from any user-specified classifier. We make no assumptions

on the accuracy of these scores.

In the second step, we first calculate scores $\hat{S}_i^c$ for $i \in \mathcal{D}^{cal} \cup \mathcal{D}^{test}$ using the previously trained function $\hat{\phi}^c(x, a)$, then calibrate an *R-value* for all $j \in \mathcal{D}^{test}$ by setting the threshold at $\hat{S}_{n+j}^c$:

$$\tilde{R}_{n+j}^c = \frac{\frac{1}{n_a^{cal}+1}\left\{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}\left(A_i = a, \hat{S}_i^c \geq \hat{S}_{n+j}^c, Y_i \neq c\right) + 1\right\}}{\frac{1}{m_a}\sum_{i\in\mathcal{D}^{test}} \mathbb{I}\left(A_i = a, \hat{S}_i^c \geq \hat{S}_{n+j}^c\right)} \wedge 1 \quad \text{if } A_{n+j} = a, \tag{12}$$

where $m_a = \sum_{i\in\mathcal{D}^{test}} \mathbb{I}(A_i = a)$ and $n_a^{cal} = \sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(A_i = a)$. As we will see, the R-value can be interpreted as a fraction. The $\wedge$ notation indicates that the R-value is set to 1 if it exceeds 1.

In situations where the test set is small, we consider the following modified R-value that utilizes both the test and calibration sets in the denominator:

$$\tilde{R}_{n+j}^{c,+} = \frac{\frac{1}{n_a^{cal}+1}\left\{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}\left(A_i = a, \hat{S}_i^c \geq \hat{S}_{n+j}^c, Y_i \neq c\right) + 1\right\}}{\frac{1}{m_a+n_a^{cal}+1}\left\{\sum_{i\in\mathcal{D}^{test}\cup D^{cal}} \mathbb{I}\left(A_i = a, \hat{S}_i^c \geq \hat{S}_{n+j}^c\right) + 1\right\}} \wedge 1 \quad \text{if } A_{n+j} = a, \tag{13}$$

for $1 \leq j \leq m$. Section E.1 in the supplement presents numerical results to show that the R^+-value (13) provides a more stable score than the R-value (12) when $|\mathcal{D}^{test}|$ is small.

While it may seem intuitive that individuals with higher scores would have smaller R-values, this is not always the case in practice. To address this issue, we have implemented the following adjustment: for all $j \in \mathcal{D}^{test}$, let

$$\hat{R}_{n+j}^c \equiv \min_{\{k\in\mathcal{D}^{test}:\hat{S}_{n+k}^c < \hat{S}_{n+j}^c\}} \tilde{R}_{n+k}^c, \quad \hat{R}_{n+j}^{c,+} \equiv \min_{\{k\in\mathcal{D}^{test}:\hat{S}_{n+k}^c < \hat{S}_{n+j}^c\}} \tilde{R}_{n+k}^{c,+}. \tag{14}$$

Next, we provide some intuition behind the interpretation of the R-value. Roughly speaking, the R-value corresponds to the smallest group-wise FSR such that the $(n+j)^{th}$ individual is *just selected*. In other words, if we make the cut at $\hat{R} = r$, e.g., selecting 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 are wrong decisions. This naturally incorporates our notion of fairness into the (group-adjusted) R-value, making it possible to calibrate a universal

threshold equalizing the FSRs across the groups. This interpretation is similar to the q-value (Storey 2003) in FDR analysis; the connection is elaborated in Section B in the Supplementary Material. While Storey's q-value is built upon the empirical distribution of p-values, our R-value is derived from a calibration data set and a carefully designed mirror process.

In the third thresholding step, we compare the R-value defined in (14) with a pre-specified FSR level α_c. For example, if we are interested in selecting individuals into class c, then the decision rule is

$$\hat{\boldsymbol{Y}} = \left[c \cdot \mathbb{I}(R^c_{n+1} \leq \alpha_c), \cdots, c \cdot \mathbb{I}(R^c_{n+m} \leq \alpha_c)\right], \tag{15}$$

where the threshold is simply the user-specified α_c. The R-value rule (15) ensures that the wasted opportunities caused by indecision will be minimized, subject to the constraints on group-wise FSRs. If the threshold is set higher than α_c, the error control would be compromised, and a threshold lower than α_c would lead to an increase in the wasted opportunities.

If we are interested in selecting both classes, then the decision rule is

$$\hat{\boldsymbol{Y}} = \textstyle\sum_{c=1}^{2} \left\{c \cdot \mathbb{I}(R^c_{n+1} \leq \alpha_c), \cdots, c \cdot \mathbb{I}(R^c_{n+m} \leq \alpha_c)\right\}.$$

To avoid assigning an individual to multiple classes, we classify the individual to the class with the smaller R-value when there is overlapping selection. The proposed FASI algorithm is summarized in Algorithm 1.

The FASI algorithm has several attractive properties. First, as we explain shortly, the R-value provides an estimate for a fraction (which is standardized between 0 and 1), it is comparable across protected groups, and it is easily interpretable. Second, the FSR analysis via R-values is straightforward: practitioners can make decisions directly with the R-values; the threshold is simply the user-specified FSR level. The fairness consideration is addressed properly / cleanly through the R-value definition (12, 13). Finally, the FASI algorithm is model-free and offers a powerful theory on FSR control; this is discussed in the next section.

Algorithm 1 Fairness Adjusted Selective Inference Algorithm

Input $\mathcal{D}$, $\mathcal{D}^{test}$, α_c

1: Randomly split $\mathcal{D}$ into $\mathcal{D}^{train}$ and $\mathcal{D}^{cal}$.
2: Train a machine learning model only on $\mathcal{D}^{train}$.
3: Predict base scores for all observations in $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$.
4: Compute the R-value for a specific classification group c, using Equation 12 or 13.
5: Threshold the R-value at a user specified level α_c, assigning an observation in $\mathcal{D}^{test}$ to class c if $R^c \leq \alpha_c$.

$$\hat{\boldsymbol{Y}} = \left\{ c \cdot \mathbb{I}(R_j^c \leq \alpha_c) : 1 \leq j \leq m \right\}$$

6: Repeat steps 5 and 6 for all classification groups $c \in \mathcal{C}$.
7: If an observation has multiple R^c-values less than α_c, classify that observation to the class where R^c is the smallest.
8: Return an indecision on all remaining observations where $R^c > \alpha_c$ for all $c \in \mathcal{C}$.

3.2 Why FASI works?

Now we explain why the FASI algorithm works. The effectiveness of our algorithm only leverages a condition on the exchangeability of data points.

Assumption 1. *The triples* $\{(X_i, A_i, Y_i) : i \in \mathcal{D}^{cal} \cup \mathcal{D}^{test}\}$ *are exchangeable.*

We start by explaining why the R-value provides a sensible estimate of the FSR. To simplify the discussion, we ignore the sensitive attribute A for the moment and consider a thresholding rule of the form $\hat{\boldsymbol{Y}} = \{\mathbb{I}(\hat{S}_{n+j}^c \geq t) : 1 \leq j \leq m\}$. Consider the false selection proportion (FSP) process for $\mathcal{D}^{test}$:

$$\text{FSP}(t) = \frac{\sum_{i \in \mathcal{D}^{test}} \mathbb{I}\left(\hat{S}_i^c \geq t, Y_i \neq c\right)}{\sum_{i \in \mathcal{D}^{test}} \mathbb{I}\left(\hat{S}_i^c \geq t\right)}, \tag{16}$$

with $\text{FSP}(t) = 0$ if no individual is selected. The FSP cannot be computed from data because we do not observe the true states $\{Y_i : i \in \mathcal{D}^{test}\}$. The good news is that under Assumption 1 on exchangeability, the unobserved process $\sum_{i \in \mathcal{D}^{test}} \mathbb{I}\left(\hat{S}_i^c \geq t, Y_i \neq c\right)$ will strongly resemble its "mirror process" in the calibration data $\sum_{i \in \mathcal{D}^{cal}} \mathbb{I}\left(\hat{S}_i^c \geq t, Y_i \neq c\right)$. Constructing a mirror process and exploiting the symmetry property to make inference is a powerful idea that has been explored in recent works (e.g. Barber & Candès 2015, Du et al. 2023, Leung & Sun 2022). Finally, adjusting for the possible unequal sample sizes between $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$, we obtain the

R-value process

$$\widehat{R}^c(t) = \frac{\frac{1}{n^{cal}+1}\left\{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}\left(\hat{S}_i^c \geq t, Y_i \neq c\right) + 1\right\}}{\frac{1}{m}\sum_{i\in\mathcal{D}^{test}} \mathbb{I}\left(\hat{S}_i^c \geq t\right)}, \tag{17}$$

where m and n^{cal} are respectively the cardinalities of $\mathcal{D}^{test}$ and $\mathcal{D}^{cal}$. Finally, the fairness-adjusted R-value defined in (12) can be recovered by restricting the R-value process to a specific group $a \in \mathcal{A}$ and substituting $\hat{S}_{n+j}^c$ in place of t in (17). The R^+-value defined by (13) can be conceptualized in a similar fashion.

Remark 1. Comparing (16) and (17), we note that "+1" has been incorporated into the count of false selections on $\mathcal{D}^{cal}$. This technical adjustment has virtually no impact on the empirical performance of FASI. However, it ensures that (17) effectively leads to a martingale, which is essential for proving the theory. It is natural to apply the same "+1" adjustment to n^{cal}, which makes the algorithm slightly more powerful.

Next we state a theorem that establishes the finite-sample property of FASI. Our theory fundamentally departs from those in existing works: we do not make assumptions regarding the accuracy of $\hat{S}_i^c$. The accuracy of the classifier only affects the power, not the validity for FSR control. See Section A.4 for practical guidelines on how to construct more informative classifiers/R-values.

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 proportions of individuals in group* a *that do not belong to class* c *in the test and calibration data, respectively. Then under Assumption 1, we have, for all* $a \in \mathcal{A}$,

1. *The FASI algorithm with R-value (12) satisfies* $FSR_a^{\{c\}} \leq \gamma_{c,a}\alpha_c$;

2. *The FASI algorithm with* R^+*-value (13) satisfies* $FSR_a^{\{c\},*} \leq \gamma_{c,a}\alpha_c$, *where*

$$FSR_a^{\{c\},*} = \mathbb{E}\left[\frac{\sum_{j=1}^m \mathbb{I}(\hat{Y}_{n+j} = c, Y_{n+j} \neq c, A_{n+j} = a)}{\sum_{j=1}^m \mathbb{I}(\hat{Y}_{n+j} = c, A_{n+j} = a) + 1}\right]. \tag{18}$$

Remark 2. In the modified FSR definition (18), the "+1" adjustment is used to account for the extra uncertainty in the approximation of the number of rejections. A similar modification, in the context of FDR analysis but for different reasons, has been used in Theorem 1 of Barber & Candès (2015).

Now we will explain the main idea behind the proof of Theorem 1. The discussion will focus on the R-value process (17), but it can be easily extended to the group-adjusted R-value (12). Three major challenges in the theoretical analysis include (a) how to handle the dependence between the scores $\hat{S}_i^c$ [as the same training data have been used to compute the scores in (17)], (b) how to evaluate the FSR in classification without knowledge about the theoretical properties of the scores, and (c) how to develop non-asymptotic guarantees on the performance of the FASI algorithm in finite samples.

Inspired by the elegant ideas in the FDR literature (Storey et al. 2004, Barber & Candès 2015), we have carefully constructed the R-values so that the corresponding FSP process (17) can be stochastically bounded above by a martingale. In the proof of Theorem 1, we first show that the threshold induced by the FSP process is a *stopping time*, and then apply Doob's optional stopping theorem to obtain an upper bound for the expectation of the martingale. Finally we leverage the exchangeability assumption to cancel out the cardinality adjustments and establish the upper bound for the FSR. We stress that our theory utilizes no assumptions on the underlying models or quality of scores, and the algorithm is provably valid in finite samples.

Under Assumption 1, $\gamma_{c,a}$ tends to be very close to 1, resulting in nearly exact control. This is verified in our simulation studies and real data analyses which can be found in the Appendix, Section E.2. However, due to the stochastic nature of the ratio, FASI may lead to FSRs that deviate from the nominal level. The next corollary shows that a conservative version of the R-value guarantees that the FSR level is controlled below α.

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

$$\tilde{R}^c_{n+j} = \frac{n^{cal}_a + 1}{n^{cal,c}_{a,null} + 1}\hat{R}^c_{n+j}, \quad \tilde{R}^{c,+}_{n+j} = \frac{n^{cal}_a + 1}{n^{cal,c}_{a,null} + 1}\hat{R}^{c,+}_{n+j}, \tag{19}$$

for $1 \le j \le m$, *where* $n^{cal}_a = \sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(A_i = a)$ *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 the group-wise FSRs satisfy (a)* $FSR^{\{c\}}_a \le \mathbb{E}\left(n^{test,c}_{a,null}/n^{test}_a\right)\alpha$ *for all* $a \in \mathcal{A}$ *when* $\tilde{R}^c_{n+j}$ *are used, and (b)* $FSR^{\{c\},*}_a \le \mathbb{E}\left(n^{test,c}_{a,null}/n^{test}_a\right)\alpha$ *for all* $a \in \mathcal{A}$ *when* $\tilde{R}^{c,+}_{n+j}$ *are used.*

Remark 3. Corollary 1 implies that the FSR levels are controlled strictly less than or equal to α. The ratio $n^{test,c}_{a,null}/n^{test}_a$, which is referred to as the null proportion in multiple testing, also appears in the (conservative) Benjamini-Hochberg (BH) procedure for FDR control. The connection between FASI and BH will be discussed shortly and elaborated on in the Appendix. It is anticipated that the FASI algorithm with conservative R-values (19) may be improved by methods that incorporate the unknown ratio $n^{test}_{a,null}/n^{test}_a$. This idea has been used in Benjamini & Hochberg (2000) and Storey (2002) to improve the power of BH in the context of FDR control. The FASI algorithm with original R-values may be viewed as such an approach in the sense that it can be recovered via firstly estimating the unknown ratio $n^{test,c}_{a,null}/n^{test}_a$ by $(n^{cal,c}_{a,null} + 1)/(n^{cal}_a + 1)$, and secondly applying the FASI algorithm with the conservative R-values at level $(n^{cal}_a + 1)/(n^{cal,c}_{a,null} + 1)\alpha$. This leads to power improvement with the price of the additional factor $\gamma_{c,a}$ in Theorem 1. The rest of this article does not focus on the conservative R-values since they usually result in a higher proportion of indecisions, while the original R-values are simple and intuitive, and offer almost exact control in practice.

3.3 Connection to conformal inference

The R-value has a compelling interpretation under the *conformal inference* framework. In Section B of the Supplementary Material, we show that a variation of our R-value corresponds to the

Benjamini-Hochberg (BH) adjusted q-value (Storey 2003) of the *conformal p-values* (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 also dependent. To address this, Bates et al. (2023) first show that the conformal p-values satisfy the PRDS condition and then apply 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 seems to be non-trivial. As discussed in Section B of the Supplement, our R-values do not explicitly utilize conformal p-values under the binary classification setup. Therefore, our theory via martingales appears to be a simple and equally effective alternative. Furthermore, implementing conformal p-values, which are based on one-class classifiers, directly in our binary classification problem involves discarding labeled outliers and would result in a loss of information. Related issues have recently been investigated in Liang et al. (2022).

3.4 Theoretical R-value and optimality theory

We introduce the theoretical R-value and derive the optimal score function under a simplified setup, following the works of Sun & Cai (2007) and Cai et al. (2019). This optimality theory, which is based on a highly idealized setting, has been placed in Section A of the Supplementary Material due to page constraints. The theory provides valuable insights for practitioners regarding how to train score functions to construct informative R-values. We highlight two essential messages.

Firstly, our analysis shows that the choice of optimal score function indicates that, during the training stage, we should emulate the full covariate classifier (10). This classifier learns the score

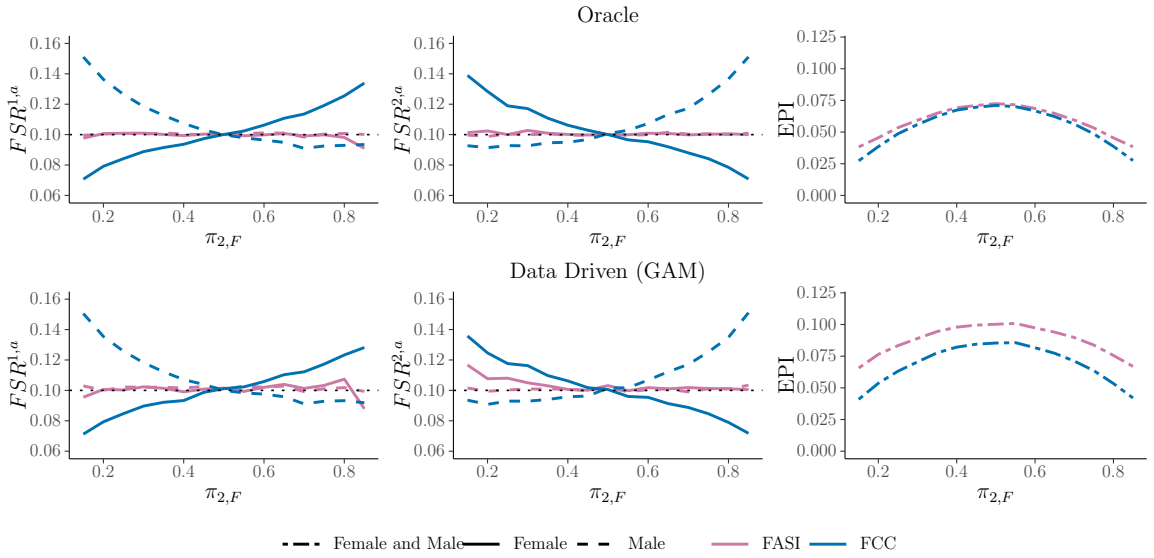

Figure 3: Simulation 1. Top row: The oracle procedure. Bottom row: A data-driven GAM fitting procedure. Left and middle column: $FSR^{1,a}$ and $FSR^{2,a}$ for both females and males. Right column: The expected proportion of indecision's (EPI).

functions using *all features*, including the sensitive attribute A, to best capture individual level information. Scores trained without the sensitive attribute [e.g., (11)] are usually suboptimal. The fairness adjustments for decisions should not be made during the training stage but in the calibration stage, where the fully informative scores can be converted to R-values to adjust the disparity in error rates across groups. This strategy shares the same spirit with the *learn then test* framework recently advocated by Angelopoulos et al. (2022).

Secondly, we find that the optimal selection rule equalizes the group-wise error rates. The intuition is that in order to maximize the EPI, the pre-specified mFSRs must be *exhausted in all separate groups*; hence the group-wise mFSRs are all equal to the nominal level (thereby automatically equalized). This implies that handling the fairness issue through a constrained optimization problem (8) leads to the asymptotic equality of our FASI algorithm. Our numerical studies corroborate this claim, though a full analysis is complicated due to the dependence between the scores. We leave this for future research.

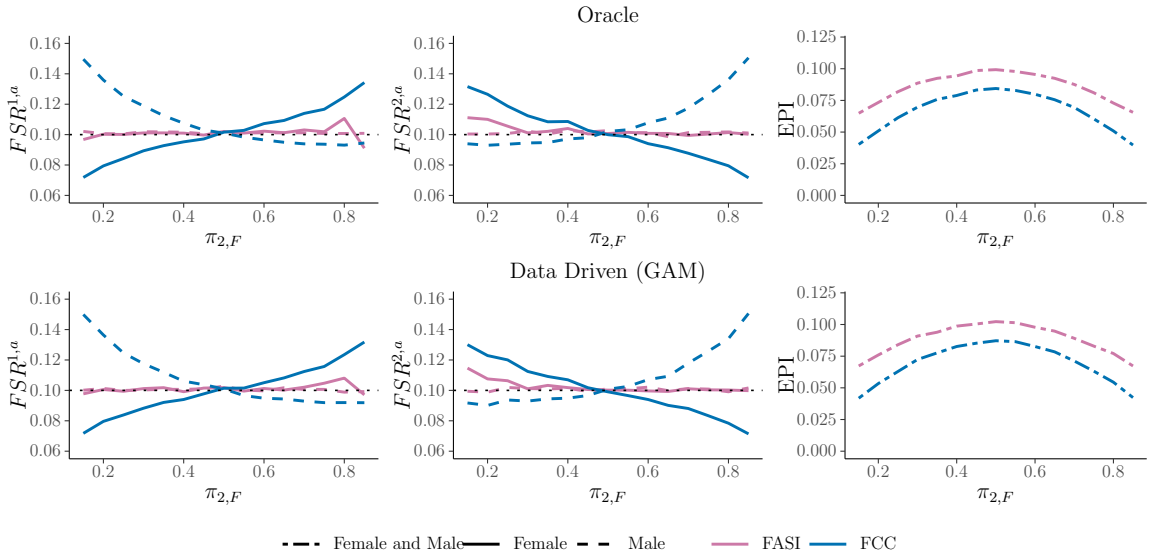

Figure 4: Simulation 2. Comparable setup to Simulation 1 except that the female and male distributions now differ from each other.

4 Simulation Results

This section presents the results from two simulation scenarios comparing FASI to the Full Covariate Classifier (FCC). The Restricted Covariate Classifier (RCC) is not included since, in our simulations, it has systematically larger deviations from the target group-wise FSR levels compared to FCC. We demonstrate that 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 posterior probabilities for S_i^c, defined in Equation 10, while the data-driven versions estimate S_i^c via a GAM classifier (Hastie et al. 2009, James et al. 2023). In our experience, similar patterns are observed when other classifiers are used to construct the base scores.

All simulations are run with sample sizes of $|\mathcal{D}| = 2{,}500$ and $|\mathcal{D}^{test}| = 1{,}000$. We generate $\mathcal{D}^{train}$ and $\mathcal{D}^{cal}$ using a random split of $\mathcal{D}$, with $|\mathcal{D}^{train}| = 1{,}500$ and $|\mathcal{D}^{cal}| = 1{,}000$. We use gender as 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 Model A.1 with four components:

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

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 $\pi_M = \pi_F = 0.5$, i.e. the numbers of females and males in the data set are equal. We will consider two scenarios in our simulation study that follow this setup.

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 males and females is 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 imbalanced FSRs across the male and female groups).

We simulate 1,000 data sets and apply both the FCC and FASI [with R-values defined in (13)] at FSR level 0.1 to the simulated data sets. The FCC method ignores the protected attributes when computing the R-values, i.e.

$$\hat{R}^{c,\mathrm{FCC}}_{n+j} = \frac{\frac{1}{n_a^{cal}+1}\left\{\sum_{i \in \mathcal{D}^{cal}} \mathbb{I}\left(\hat{S}^c_i \geq \hat{S}^c_{n+j}, Y_i \neq c\right) + 1\right\}}{\frac{1}{m_a+n_a^{cal}+1}\left\{\sum_{i \in \mathcal{D}^{test} \cup D^{cal}} \mathbb{I}\left(\hat{S}^c_i \geq \hat{S}^c_{n+j}\right) + 1\right\}}.$$

The corresponding selection rule is $\hat{\boldsymbol{Y}}^{\mathrm{FCC}} = \left\{c \cdot \mathbb{I}(R^{c,\mathrm{FCC}}_{n+j} \leq \alpha_c) : 1 \leq j \leq m\right\}$.

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 expected proportion of indecisions (EPI) obtained by averaging the results from 1,000 replications. The following patterns can be observed.

- Both the FASI method and FCC control the global FSR. For simplicity, we do not include these results in the figures below.

- Shifting our focus to the group-wise FSRs, FCC fails to control the error rate. When $\pi_{2|F} = 0.5$, by construction we have $\pi_{2|F} = \pi_{2|M}$, making the Female and Male attributes indistinguishable. However, 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.

- In comparison, 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 version of FASI is able to closely mirror the behavior of the oracle method. The FSR control is in general effective except that the FSR levels are slightly elevated in the tails.

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

Our second simulation scenario considers the setting where $F_{c,M} \neq F_{c,F}$, for both $c = 1, 2$. Denoting the mean of each distribution 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 in Figure 4.

We notice very similar patterns to our first simulation setup. Both FASI and FCC are able to control the global FSR (omitted from the figure). 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 consider the EPI (7) as the power metric. By contrast, the conventional power in classification is evaluated through Power $\coloneqq \mathbb{P}\left(\hat{Y} = c | Y = c\right)$. Under our setup, the power is first computed as $\frac{1}{m}\sum_{j=1}^{m} \mathbb{I}\left(\hat{Y}_{n+j} = c | Y_{n+j} = c\right)$, and then averaged across 1,000 replications. The difference is that this power notion only concerns a particular class c whereas the EPI combines

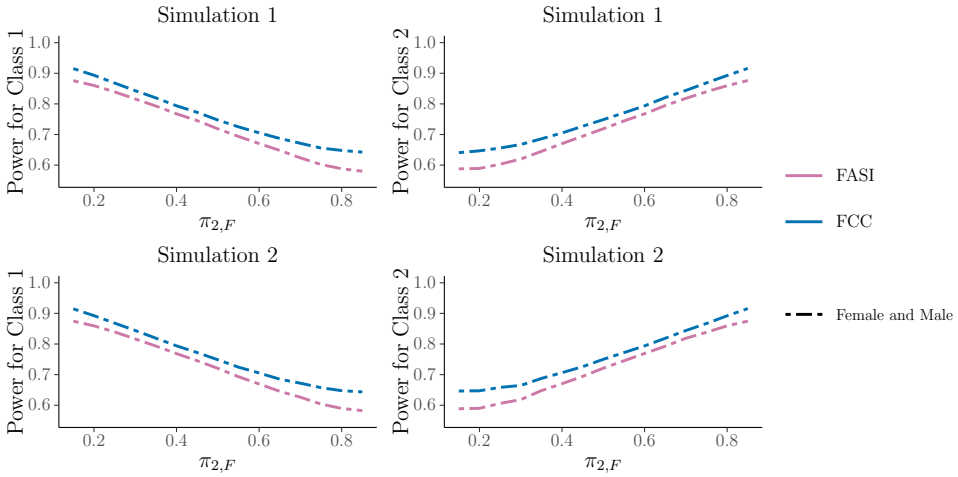

Figure 5: Power plots for Simulation 1 (top row) and Simulation 2 (bottom row). The power of both the FASI (purple) and FCC (blue) method are shown for classification 1 (left column) and classification 2 (right column).

the probabilities from all classes. Figure 5 provides the powers of FASI and FCC for both Classes 1 and 2. Similar to the EPI plots in Figures 3 and 4, we can see that (a) the FCC has higher power than FASI, which is expected because it does not need to satisfy the fairness constraint; (b) the price of fairness measured by the loss of power is relatively small.

5 Real Data Examples

In this section we demonstrate FASI's effectiveness on two real world case studies. In Section 5.1 we examine the Compas recidivism algorithm made popular by ProPublica in 2016 (Angwin et al. 2016), while in Section 5.2 we use US census data from 1994 to predict an individual's salary (Dua & Graff 2017). In both cases we compare FASI to the FCC approach described in Section 4 by randomly assigning 70% of our data set to $\mathcal{D}$ and the remaining 30% to $\mathcal{D}^{test}$. We further evenly split $\mathcal{D}$ into $\mathcal{D}^{train}$ and $\mathcal{D}^{cal}$.

Since this is a real data setting, the true posterior probabilities for S_i^c, defined in Equation 10, are unavailable to us. To estimate them, we used a GAM fitting procedure for the compas case study and an Adaboost fitting procedure for the census income prediction case study (Hastie et al. 2009, James et al. 2023).

This method can be applied to other applications through the `fasi` package available in the R language on CRAN.

5.1 COMPAS Data Analysis

In 2016, ProPublica's investigative journalists curated a data set of 6,172 individuals, 3,175 of whom were Black and the remaining 2,997 other races, that were arrested in Broward County, Florida. In this study, Black and Other are our protected attributes.

The Black and Other groups respectively had 1,773 and 1,217 individuals who actually recidivated in the 2-year time frame that the study considered. We used this two year window as a proxy for the true label of identifying recdivists.

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, we aim to use FASI to correct the possible disparity across races in terms of the *false selection rates*. Various fairness notions have been discussed in the literature (Zafar et al. 2017, Angwin et al. 2016, Dieterich et al. 2016), and we do not claim that ours is universally superior to existing ones. Our approach, while very effective in this setting, should be considered alongside many others before implementation, bearing in mind the societal trade-offs between different fairness definitions.

We randomly split the data set 1,000 times into $\mathcal{D}$ and $\mathcal{D}^{test}$, and averaged the difference between the actual and target recidivism FSR's for a range of α between 0.15 and 0.30. The first two columns of Figure 6 provide the difference between true and target FSR for the recidivist classification for FCC and FASI respectively. The last column plots the overall EPIs.

As we noted with the simulated data, while the FCC does a good job at controlling the

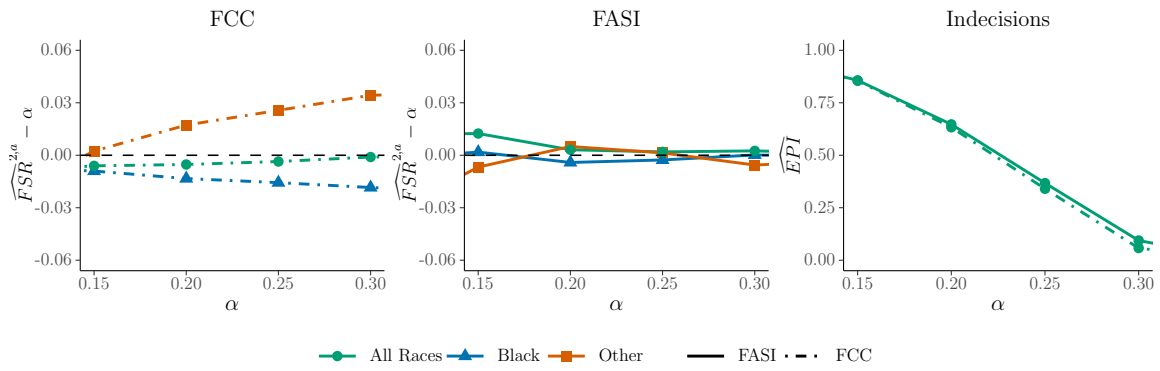

Figure 6: COMPAS data analysis for predicting recidivists. Left and Middle: False Selection Rate minus the desired control level for varying levels of α for the FCC and FASI method respectively. Right: The EPI for both the FCC and FASI method.

overall FSR (green / circle) of recidivists, it is unable to do this at the race level. In the left hand plot of Figure 6, the breakdown of the race-wise FSR control for the FCC is shown. The Black attribute (blue / triangle) systematically has FSR control lower than the desired target level. While the Other attribute (orange / square) systematically has higher FSR control than the target level. This observation holds for all values of α considered.

In comparison, the middle plot in Figure 6 shows the FASI method. For all values of α, the FSR is controlled at the desired level for both the protected attributes and for all observations. The right plot in Figure 6 also demonstrates that, in this study, FASI is able to obtain a nearly identical EPI to the FCC. This demonstrates that the price of our fairness constraint, in terms of the size of the indecision group, is nearly zero.

5.2 1994 Census Income Data Analysis

The US census is the leading body of information for producing information about the American people. Naturally, the data that they collect can directly inform future policy decisions, such as funding programs that provide economic assistance for populations in need. In particular, resources such as food, health care, job training, housing, and other economic assistance rely upon good estimates of a population's income levels. The cost of making unfair decisions when predicting ones income can be severe since the prediction helps determine how hundreds of

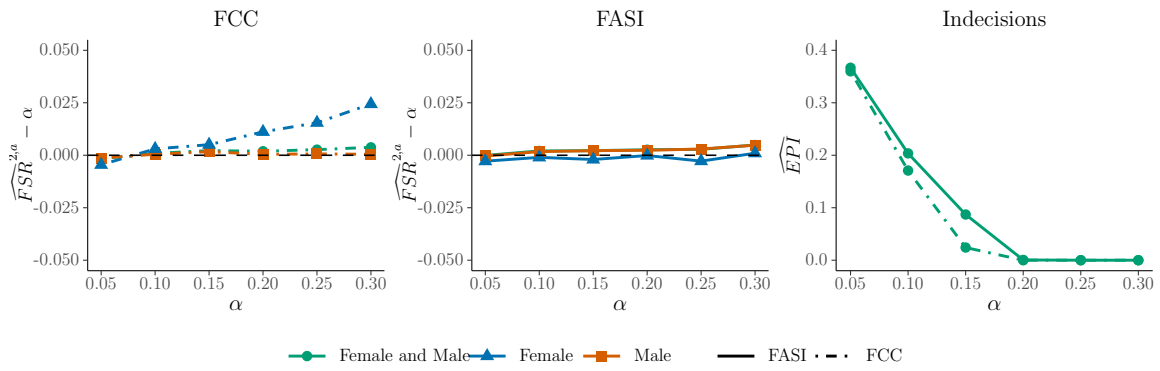

Figure 7: Census income prediction for individuals that earn more than 50K a year. Left and Middle: False Selection Rate minus the desired control level for varying levels of α for the FCC and FASI method respectively. Right: The EPI for both the FCC and FASI method.

billions of dollars in federal funding are spent for the next decade. In this case study, we use the 1994 US Census Data set from the UCI Machine Learning Repository to predict if an individual earns more than 50,000 dollars a year.

The data consist of 32,561 observations on 14, largely demographic, variables including education level, age, hours worked per week, and others. The protected attributes in this study are Female and Male. The Female group has 10,771 total observations, of which 1,179 make over \$50K a year. Similarly, the remaining 21,790 observations are from the Male attribute, of which 6,662 make over \$50K a year.

As in Section 5.1, we compare the FCC approach to the FASI method for many values of FSR control, α, ranging from 0.05 to 0.3. The left most plot of Figure 7 shows results from the FCC method, where both the overall (green / circle) and Male (orange / square) has the desired FSR control. However, the FCC is unable to maintain the desired FSR control for the Female attribute (blue / triangle) across all values of α. In comparison, the FASI method shown in the middle plot is able to maintain both the overall FSR control as well as the FSR control for Females and Males across all values of α.

The right most plot provides the estimated EPI of each approach. Unlike for the Compas data, for some values of α the FASI method returns a slightly larger indecision group on average in comparison to the FCC method.

6 Discussions

Fairness in machine learning is a complicated topic. Various works tackle representation/sampling bias, which arises when the data are collected from a population in a non-representative or non-random fashion (Mehrabi et al. 2021). This article, in contrast, addresses *algorithmic bias* – when the machine learning algorithm itself is adding bias outside of the initial input data. Our fairness criterion, described in Equation 8, is a group-wise fairness definition that assumes full knowledge of the protected groups. This is a widely adopted practice in the literature and is used in many modern applications of fairness algorithms such as medicine and criminal justice system (Manrai et al. 2016, Angwin et al. 2016); specialized software has also been developed (Bellamy et al. 2018, Saleiro et al. 2018).

The rest of this section concludes the article with a discussion of other fairness notions and related error rate concepts.

6.1 Other fairness notions

In addition to the sufficiency principle, a widely used fairness notion which, as mentioned in Section 2.1, is effectively enforced by the FASI algorithm, is the *separation principle* (Barocas et al. 2017), which requires that

$$P(Y \neq \hat{Y} | Y = c, A = a) \text{ are the same for all } a \in \mathcal{A}. \tag{20}$$

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 $\hat{Y}$, these works are concerned with the joint probabilities of $(\hat{Y}, Y)$. This fairness criterion requires that the misclassification rates are equalized across all protected groups:

$$P(Y \neq \hat{Y} | A = a) \text{ are the same for all } a \in \mathcal{A}. \tag{21}$$

Other popular fairness notions include *equalized odds* (Hardt et al. 2016, Romano, Bates & Candès 2020) and *equalized risks* (Corbett-Davies & Goel 2018). A highly controversial issue is that different fairness criteria often lead to different algorithms and decisions in practice. For example, the sufficiency and separation principles can be incompatible with each other (Kleinberg et al. 2016, Friedler et al. 2021), and classification parity / calibration can harm the very groups that the algorithms are designed to protect (Corbett-Davies & Goel 2018). We do not claim that FASI is universally superior than competitive approaches but adjusting group-wise FSRs appears to be a reasonable fairness criterion for the high-stakes applications under our consideration. Much research is still needed to fully understand the trade-offs and caveats between different approaches to fairness-adjusted inference.

Zafar et al. (2017) proposed to use cost-sensitive classifiers with group specific costs (Menon & Williamson 2018) to tackle a similar fairness issue. However their technique forces a decision to be made on all individuals, where as our work is a *selective inference* procedure that only makes "confident" decisions on a subset of individuals, returning an indecision on the rest. With human intervention, FASI can achieve better accuracy than cost-sensitive classifiers since practitioners are made aware of the cases that they should spend most of their energy on, helping them avoid mistakes with a high societal cost.

Another attractive approach is to consider individual fairness definitions that disregard protected groups, instead ensuring that similar individuals receive similar outcomes (Mukherjee et al. 2020). Due to the significant computational and theoretical challenges surrounding individual fairness algorithms, we leave this promising direction for future work.

6.2 FSR concepts in multinomial classification

The selective inference framework and FSR concepts can be extended from the binary classification setting to more general settings. 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; particularly the outlier

detection problem recently considered in Guan & Tibshirani (2021) and Bates et al. (2023) can be encompassed by 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 $\hat{\boldsymbol{Y}} = \{\hat{Y}_{n+j} : 1 \leq j \leq m\} \in \Lambda^m$. Then the FSR with respect to subset $\mathcal{C}'$ is defined as the expected fraction of erroneous selections among all selections: $\mathrm{FSR}^{\mathcal{C}'} = \mathbb{E}\left[\frac{\sum_{j=1}^m \mathbb{I}(\hat{Y}_{n+j} \in \mathcal{C}', \hat{Y}_{n+j} \neq Y_{n+j})}{\left\{\sum_{j=1}^m \mathbb{I}(\hat{Y}_{n+j} \in \mathcal{C}')\right\} \vee 1}\right]$. The group-wise FSRs taking into account the protected attribute A can be defined analogously to (8) by restricting the selections to specific groups. The EPI (7), which characterizes the power of the selection procedure, remains the same. The development of the R-values and corresponding fairness algorithms is more complicated and will be left for future research.

References

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I. & Lei, L. (2022), 'Learn then test: Calibrating predictive algorithms to achieve risk control', *arXiv preprint arXiv:2110.01052* .

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.
URL: *https://doi.org/10.1214/22-AOS2244*

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'.
URL: *https://arxiv.org/abs/1810.01943*

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.

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.
URL: *http://doi.acm.org/10.1145/2939672.2939785*

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. & Goel, S. (2018), 'The measure and mismeasure of fairness: A critical review of fair machine learning', *arXiv preprint arXiv:1808.00023* .

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'.
URL: *http://archive.ics.uci.edu/ml*

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. (2012), Fairness through awareness, *in* 'Proceedings of the 3rd innovations in theoretical computer science conference', pp. 214–226.

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.

Guan, L. & Tibshirani, R. (2021), 'Prediction and outlier detection in classification problems', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **to appear**, arXiv:1905.04396.

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.

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. (2016), 'Inherent trade-offs in the fair determination of risk scores', *arXiv preprint arXiv:1609.05807* .

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

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. (2022), 'Integrative conformal p-values for powerful out-of-distribution testing with labeled outliers', *arXiv preprint arXiv:2208.11111* .

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.
URL: *https://doi.org/10.1056/NEJMsa1507092*

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).
URL: *https://doi.org/10.1145/3457607*

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.
URL: *https://proceedings.mlr.press/v81/menon18a.html*

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', *arXiv preprint arXiv:1709.02012* .

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'.
URL: *https://arxiv.org/abs/1811.05577*

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.

Zafar, M. B., Valera, I., Gomez Rodriguez, M. & Gummadi, K. P. (2017), 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.
URL: *https://doi.org/10.1145/3038912.3052660*

Zeng, X., Dobriban, E. & Cheng, G. (2022), 'Fair bayes-optimal classifiers under predictive parity'.
URL: *https://arxiv.org/abs/2205.07182*

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

This supplement provides additional technical results (Sections A-B), proofs (Sections C-D) and additional numerical results (Section E).

A 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 theory provides practical insights for practitioners on how to train score functions to construct informative R-values.

A.1 The mixture model under an oracle setting

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 discussions are centered around a simplified random mixture model, as defined in Equation A.1 below:

$$F(x) = \sum_{a \in \mathcal{A}} \{\pi_{1,a} F_{1,a}(x) + \pi_{2,a} F_{2,a}(x)\}, \tag{A.1}$$

where $F_{1,a}(x)$ and $F_{2,a}(x)$ are the conditional CDFs of X from classes 1 and 2, respectively. Let $f_{c,a}(x)$ be the corresponding density functions of $F_{c,a}$. The probabilities $\pi_{c,a} = \mathbb{P}(Y = c, A = a)$ represent the joint probabilities of $Y = c$ and $A = a$ for $c = 1, 2$. Denote $\pi_a = P(A = a)$, $\pi_{c|a} = P(Y = c|A = a)$. For our analysis, we consider a selection rule of the form $\hat{Y}(t) = c \cdot \mathbb{I}(S^c \geq t)$,

where t is a threshold, and $\mathbb{I}(\cdot)$ is the indicator function. We assume that an oracle has knowledge of the conditional probabilities and conditional CDFs defined above.

A.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 that corresponds to every base 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).

The conversion algorithm consists of three steps. In Step 1, we find the distributional information with respect to the score S^c. Let $G^c(s) = \sum_{a\in\mathcal{A}} \pi_a G_a^c(s)$ be the CDF of S^c, where $G_a^c(s) = \pi_{1|a} G_{1,a}^c(s) + \pi_{2|a} G_{2,a}^c(s)$, with $G_{1,a}^c(s)$ and $G_{2,a}^c(s)$ denoting the conditional CDFs of S^c given $A = a$ and Y.

Suppose an oracle knows the conditional probabilities and conditional CDFs defined above. In Step 2, we compute the conditional error probability when the threshold is t:

$$Q_a^c(t) \coloneqq \mathbb{P}(Y \neq c | S^c \geq t, A = a) = \begin{cases} \frac{\pi_{2|a}\left\{1-G_{2,a}^1(t)\right\}}{1-G_a^1(t)}, & c = 1; \\ \frac{\pi_{1|a}\left\{1-G_{1,a}^2(t)\right\}}{1-G_a^2(t)}, & c = 2. \end{cases}$$

Finally, in Step 3 we compute a fair score, referred to the theoretical R-value, for an individual from group a with observed score $S^c = s$:

$$R^c(s) = \inf_{t\leq s} \left\{ Q_a^c(t) \coloneqq \mathbb{P}(Y \neq c | \hat{Y}(t) = c, A = a) \right\}, \tag{A.2}$$

where if the threshold is t, $Q_a^c(t)$ corresponds to the conditional error probability. If the base score satisfies the monotone likelihood ratio condition (MLRC, Sun & Cai 2007) then the infimum is achieved at s exactly; see Section D.1 for related discussions.

A.3 Theoretical R-value and fairness

Consider random mixture model (A.1). Suppose an oracle knows the score function $S^c(x,a) = \mathbb{P}(Y = c|X = x, A = a)$. The goal is to assign labels "0", "1" and "2" to new instances $\{(X_{n+j}, A_{n+j}) : 1 \le j \le m\}$. We assume that the instances (X_j, A_j) are independent draws from an underlying distribution $F(x,a)$. Define the marginal FSR

$$\text{mFSR}_a^c = \frac{\mathbb{E}\left\{\sum_{1\le j\le m:A_{n+j}=a} \mathbb{I}(\hat{Y}_{n+j} = c, Y_{n+j} \neq c)\right\}}{\mathbb{E}\left\{\sum_{1\le j\le m:A_{n+j}=a} \mathbb{I}(\hat{Y}_{n+j} = c)\right\}}.$$

Under the random mixture model (A.1), it can be shown that, following arguments in Storey (2003),

$$\text{mFSR}^{c,a} = \mathbb{P}(Y \neq c|\hat{Y} = c, A = a), \tag{A.3}$$

which is the conditional probability required in the sufficiency principle (A.5). A similar result was obtained in the FDR literature, as described in Storey (2003).

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

$$\text{FSR}_a^c = \text{mFSR}_a^c + o(1), \text{ when } m_a := |\{1 \le j \le m : A_{n+j} = a\}| \to \infty. \tag{A.4}$$

The connection between (A.3) and (A.4) highlights that, under the simplifying assumptions that we have made, the two criteria, namely group-wise FSR control (8), and the sufficiency principle (A.5), are closely related. However, the former is concerned with both the error rate control and fairness, whereas the latter only addresses fairness, without providing risk control in the decisions. This limitation of the sufficiency principle makes it unsuitable for high-consequence decision-making scenarios.

The next proposition, which follows directly from (A.2), shows that thresholding the theoretical R-value leads to a fair selective inference procedure.

Proposition 1. *Consider a classifier that claims* $\hat{Y} = c$ *if* $R^c \leq \alpha$. *Then*

$$\mathbb{P}(Y \neq c | \hat{Y} = c, A = a) \leq \alpha \text{ for all } a \in \mathcal{A}. \tag{A.5}$$

We would like to make two important remarks. Firstly, the theoretical R-value is the counterpart of the data-driven R-value defined in Equation (12). Essentially, it represents the minimum conditional probability required to ensure that an individual with score $S^c = s$ 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 R-values, we not only satisfy the sufficiency principle but also control the error rates, which is crucial in practical situations where decision risk must be managed.

A.4 A sketch of the optimality theory

We present and prove an intuitive result that shows $S^c(x, a) = \mathbb{P}(Y = c | X = x, A = a)$ is the optimal choice of score function 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 (A.4).

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

$$\text{Minimize the EPI, subject to } \text{mFSR}_a^c \leq \alpha_c,\ c = 1, 2 \text{ for all } a \in \mathcal{A}. \tag{A.6}$$

Here, we denote $S_{n+j}^c = \mathbb{P}(Y_{n+j} = c | X_{n+j} = x_{n+j}, A_{n+j} = a_{n+j})$, and the scores can be transformed to theoretical R-values, denoted R_{n+j}^1 and R_{n+j}^2. The process of conversion follows the general strategy outlined in Section A.2, and is described in more detail in the proof of Theorem 2 below.

Define the oracle procedure

$$\boldsymbol{\delta}_{OR} = \{\delta^j_{OR} : 1 \le j \le m\}, \text{ where } \delta^j_{OR} = \mathbb{I}(R^1_{n+j} \le \alpha_1) + 2\mathbb{I}(R^2_{n+j} \le \alpha_2). \tag{A.7}$$

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

Theorem 2. *Consider random mixture model (A.1). Assume that α_1 and α_2 have been properly chosen such that (A.7) does not have overlapping selections. Let $\mathcal{D}_{\alpha_1,\alpha_2}$ denote the collection of selection rules that satisfy $mFSR^c_a \le \alpha_c$ for $c = 1, 2$ and all $a \in \mathcal{A}$. Let $EPI_{\boldsymbol{\delta}}$ denote the EPI of an arbitrary decision rule $\boldsymbol{\delta}$. Then the oracle procedure (A.7) is optimal in the sense that $EPI_{\boldsymbol{\delta}_{OR}} \le EPI_{\boldsymbol{\delta}}$ for any $\boldsymbol{\delta} \in \mathcal{D}_{\alpha_1,\alpha_2}$.*

A.5 Connections to Storey's q-value

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 $\{H_j : 1 \le j \le m\}$ with associated p-values $\{p_j : 1 \le j \le m\}$, 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

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

which roughly measures the fraction of false discoveries when H_j is rejected. The q-value and R-value algorithms operate similarly, where an FDR/FSR analysis at level α involves obtaining the q-value/R-value for hypothesis/individual j and rejecting/selecting it if its q-value/r-value is less than α.

B R-Value, Q-Value and Conformal P-Value

In this section, we adopt a multiple testing perspective to gain further insights into the R-value. Despite its distinct motivation, we demonstrate that the R-value can be derived as the (BH) q-value of the conformal p-values (Bates et al. 2023) under the one-class classification scenario. For comparability purposes, we exclude the sensitive attribute A and concentrate on the unadjusted R-value as defined in (17).

B.1 A brief review of conformal p-values

The one-class classification problem can be formulated under the selective inference framework. Suppose we observe data from two classes and divide the observed data into subsets of inliers (labeled as "1") and outliers (labeled as "2"), respectively:

$$\mathcal{D} = \{(X_i, Y_i) : 1 \leq i \leq n\} = \mathcal{D}^1 \cup \mathcal{D}^2,$$

with $\mathcal{D}^c = \{X_i : \text{subject } i \text{ is observed with label } Y = c\}$, $c = 1, 2$.

The conformal p-value (Bates et al. 2023) is originally developed for nonparametric outlier detection. The objective is to identify outliers in $\mathcal{D}^{test}$, which is a mixture of inliers and outliers. Imagine that we simply discard labeled outliers $\mathcal{D}^2$. Viewing individuals in class "1" as the null cases, we can formulate an equivalent multiple testing problem:

$$H_{j0} : Y_{n+j} = 1 \quad vs. \quad H_{j1} : Y_{n+j} \neq 1 \ (i.e.\ Y_{n+j} = 2), \quad j = 1, \cdots, m.$$

The construction of conformal p-values involves a sample splitting step, which divides $\mathcal{D}^1$ into two parts: $\mathcal{D}^{train}$ for training a score function $\hat{\phi}^c$ and $\mathcal{D}^{cal}$ for calibrating a significance index. We view $\hat{S}^c(X)$ as a *conformity score*. The conformal p-value for testing H_{j0}, under our notational

system, corresponds to

$$\hat{\mu}(t) = \frac{\sum_{i \in D^{cal}} \mathbb{I}\{\hat{S}_i^c \geq t\} + 1}{n^{cal} + 1}, \tag{B.8}$$

where c is taken to be 2 and t is the observed score $\hat{S}_{n+j}^c = t$.

Remark 4. *We mention a minor point to avoid confusions. Under our setup, a larger score indicates a greater likelihood of being an outlier. This interpretation makes sense in our problem but is in the opposite direction compared to that in Bates et al. (2023). To make the two definitions equivalent, the expression "$S \leq t$" in the conformal p-value definition in Bates et al. (2023) has been swapped to "$S \geq t$" in our equation (B.8).*

B.2 R-value is the BH q-value of conformal p-values

To see the connection of our R-value to the conformal p-value (B.8), recall the definition of Storey's q-value

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

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 $\{\hat{S}_i : i \in \mathcal{D}^{test}\}$:

$$\hat{G}(t) = \frac{1}{m} \sum_{i \in \mathcal{D}^{test}} \mathbb{I}\{\hat{\mu}(S_i) \leq \hat{\mu}(t)\} = \frac{1}{m} \sum_{i \in \mathcal{D}^{test}} \mathbb{I}\left(\hat{S}_i^c \geq t\right), \tag{B.9}$$

where the last equality holds because, by (B.8), 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:

$$\tilde{q}_{n+j}^{BH} = \frac{\hat{\mu}(\hat{S}_{n+j}^c)}{\hat{G}(\hat{S}_{n+j}^c)}. \tag{B.10}$$

We also need to apply a monotonicity adjustment to ensure that the q-value function is non-decreasing in the conformity score, by following the steps in (14) of Section 3.1 in the main text. Specifically, let

$$\hat{q}^{BH}_{n+j} = \min_{k \in \mathcal{D}^{test}: \hat{S}^c_{n+k} < \hat{S}^c_{n+j}} \tilde{q}^{BH}_{n+k}, \quad \text{for } j \in \mathcal{D}^{test}. \tag{B.11}$$

The superscript "BH" is used because the thresholding rule

$$\hat{\boldsymbol{Y}} = \left(\mathbb{I}\left\{\hat{q}^{BH}_{n+1} \leq \alpha\right\}, \cdots, \mathbb{I}\left\{\hat{q}^{BH}_{n+m} \leq \alpha\right\}\right)$$

is equivalent to applying the Benjamini-Hochberg procedure (Benjamini & Hochberg 1995) to the conformal p-values $\hat{\mu}_{n+j}, j \in \mathcal{D}^{test}$.

Combining (B.8), (B.9) and (B.10), we can precisely recover the R-value defined in (17). Concretely, we have

$$\hat{q}^{BH}(t) = \frac{m}{n^{cal}+1} \cdot \frac{\sum_{i \in \mathcal{D}^{cal}} \mathbb{I}\left(\hat{S}^c_i \geq t\right) + 1}{\sum_{i \in \mathcal{D}^{test}} \mathbb{I}\left(\hat{S}^c_i \geq t\right)} \tag{B.12}$$

$$= \frac{m}{n^{cal}+1} \cdot \frac{\sum_{i \in \mathcal{D}^{cal}} \mathbb{I}\left(\hat{S}^c_i \geq t, Y_i \neq c\right) + 1}{\sum_{i \in \mathcal{D}^{test}} \mathbb{I}\left(\hat{S}^c_i \geq t\right)}. \tag{B.13}$$

The last equality (B.13) holds because under the one-class classification setup, $\mathcal{D}^{cal}$ is a "pure" training set in which all observations satisfy $Y_i \neq c$ trivially. We conclude that our unadjusted R-values (17), which can be called the *conformal q-value* under the one-class classification setup, is the BH q-value of conformal p-values.

B.3 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 (B.12) will be different from the R-value (17) 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 (B.13) 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.

C Proof of Theorem 1

C.1 Proof of Part (a)

C.1.1 An empirical process description of the FASI algorithm

Suppose we select subjects into class c if the base score S^c is great than t. The estimated false discovery proportion (FSP), as a function of t, in group a is given by:

$$\hat{Q}_c(t) = \frac{\frac{1}{n_a^{cal}+1}\left\{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(\hat{S}_i^c \geq t, Y_i \neq c, A_i = a) + 1\right\}}{\frac{1}{m_a}\left\{\sum_{(n+j)\in\mathcal{D}^{test}} \mathbb{I}(\hat{S}_{n+j}^c \geq t, A_j = a)\right\} \vee 1}. \tag{C.1}$$

We choose the smallest t such that the estimated FSP is less than α. Define

$$\tau = \hat{Q}_c^{-1}(\alpha) = \inf\left\{t : \hat{Q}_c(t) \leq \alpha\right\}. \tag{C.2}$$

Consider the R-value defined in (14). For $(n+j)^{th}$ observation in $\mathcal{D}^{test}$, it is easy to see that $\hat{R}^c_{n+j} = \inf_{t \le \hat{s}} \left\{ \hat{Q}_c(t) \right\}$, where $\hat{s} := \hat{S}^c(X_{n+j} = x, A_{n+j} = a)$. The FASI algorithm can be represented in two equivalent ways:

$$\mathbb{I}(\hat{R}^c_{n+j} \le \alpha) \quad \Longleftrightarrow \quad \mathbb{I}(\hat{S}^c_{n+j} \le \tau). \tag{C.3}$$

Next we turn to the description of the true FSP process of the FASI algorithm (C.3) via the representation of $\hat{S}^c$. Let

$$\begin{aligned} V^{test}(t) &= \sum_{j \in \mathcal{D}^{test}} \mathbb{I}(\hat{S}^c_j \ge t, Y_j \ne c, A_j = a), \quad \text{and} \\ R^{test}(t) &= \sum_{j \in \mathcal{D}^{test}} \mathbb{I}(\hat{S}^c_j \ge t, A_j = a) \end{aligned}$$

respectively be the count of false selections and the count of total selections in $\mathcal{D}^{test}$ when the threshold is t. We have dropped the sensitive attribute "a" to simplify the notation. Furthermore, denote

$$V^{cal}(t) = \textstyle\sum_{i \in \mathcal{D}^{cal}} \mathbb{I}(\hat{S}^c_i \ge t, Y_i \ne c, A_i = a) \text{ and } R^{cal}(t) = \sum_{i \in \mathcal{D}^{cal}} \mathbb{I}(\hat{S}^c_i \ge t, A_i = a)$$

the corresponding counts in $\mathcal{D}^{cal}$. The FSP of the proposed FASI algorithm is given by

$$\text{FSP}^{\{c\}}_a(\tau) = \frac{V^{test}(\tau)}{R^{test}(\tau) \vee 1}.$$

The operation of the FASI algorithm implies that

$$\begin{aligned} \text{FSP}^{\{c\}}_a(\tau) &= \frac{V^{test}(\tau)}{V^{cal}(\tau)+1} \cdot \frac{V^{cal}(\tau)+1}{R^{test}(\tau) \vee 1} \\ &= \hat{Q}^c(\tau) \cdot \frac{n^{cal}_a + 1}{m_a} \cdot \frac{V^{test}(\tau)}{V^{cal}(\tau)+1} \\ &\le \alpha \cdot \frac{n^{cal}_a + 1}{m_a} \cdot \frac{V^{test}(\tau)}{V^{cal}(\tau)+1}, \end{aligned}$$

where the last two steps utilize definitions (C.1) and (C.2), respectively.

C.1.2 Martingale arguments

A key step to establish the FSR control, i.e. $\mathbb{E}\left\{\text{FSP}_a^{\{c\}}(\tau)\right\} \leq \alpha$, is to show that the ratio

$$\frac{V_{test}(t)}{V^{cal}(t)+1} \tag{C.4}$$

is a martingale. Suppose that both the calibration and test data (without labels) have been given. It is natural to consider the following filtration that involves two parallel processes: $\mathcal{F}_t = \sigma\left\{V^{test}(s), V^{cal}(s) : t_l \leq s \leq t\right\}$, where t_l is lower limit of the threshold. If t_l is used, then all subjects are classified to class c.

In our proof, we focus on the following discrete-time filtration that informs the misclassification process:

$$\mathcal{F}_k = \sigma\left\{V^{test}(s_j), V^{cal}(s_j) : j = m^*, m^*-1, \ldots, k\right\}, \tag{C.5}$$

where s_k corresponds to the threshold (time) when exactly k subjects, combining the subjects in both $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$, are mistakenly classified as $Y = c$, and m^* is the total number of misclassifications in both $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$ when the threshold is t_l. $\mathcal{F}_k$ is a backward-running filtration in the sense that for $k_1 < k_2$, $\mathcal{F}_{k_2} \subset \mathcal{F}_{k_1}$.

Note that at time s_k, only one of the two following events are possible

$$\begin{aligned} A_1 &= \{V^{test}(s_{k-1}) = V^{test}(s_k), \text{ and } V^{cal}(s_{k-1}) = V^{cal}(s_k) - 1\}, \\ A_2 &= \{V^{test}(s_{k-1}) = V^{test}(s_k) - 1, \text{ and } V^{cal}(s_{k-1}) = V^{cal}(s_k)\}. \end{aligned}$$

According to Assumption 1 which claims that the data points in $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$ are exchange-

able, and the fact that FASI uses same fitted model to compute the scores, we have

$$\mathbb{P}(A_1|\mathcal{F}_k) = \frac{V^{cal}(s_k)}{V^{test}(s_k) + V^{cal}(s_k)}; \quad \mathbb{P}(A_2|\mathcal{F}_k) = \frac{V^{test}(s_k)}{V^{test}(s_k) + V^{cal}(s_k)}.$$

To see why the ratio defined in (C.4) is a discrete-time martingale with respect to the filtration $\mathcal{F}_k$, note that

$$\begin{aligned} & \mathbb{E}\left\{\frac{V^{test}(s_{k-1})}{V^{cal}(s_{k-1})+1}|\mathcal{F}_k\right\} \\ = & \frac{V^{test}(s_k)}{V^{cal}(s_k)} \cdot \frac{V^{cal}(s_k)}{V^{test}(s_k)+V^{cal}(s_k)} + \frac{V^{test}(s_k)-1}{V^{cal}(s_k)+1} \cdot \frac{V^{test}(s_k)}{V^{test}(s_k)+V^{cal}(s_k)} \\ = & \frac{V^{test}(s_k)}{V^{cal}(s_k)+1}, \end{aligned}$$

establishing the desired result.

C.1.3 FSR Control

The threshold τ defined by (C.2) is a stopping time with respect to the filtration $\mathcal{F}_k$ since $\{\tau \leq s_k\} \in \mathcal{F}_k$. In other words, the event whether the kth misclassification occurs completely depends on the information prior to time s_k (including s_k).

Let $\mathcal{D}^{test,0}/\mathcal{D}^{cal,0}$ be the index sets for subjects in the testing/calibration data that do not belong to class c. In the final step of our proof, we shall apply the optional stopping theorem to the filtration $\{\mathcal{F}_k\}$. The group-wise FSR of the FASI algorithm is

$$\begin{aligned} \text{FSR}_a^{\{c\}} &= \mathbb{E}\{\text{FSP}_a^{\{c\}}(\tau)\} \\ &\leq \alpha \cdot \mathbb{E}\left[\frac{|\mathcal{D}^{cal}|+1}{|\mathcal{D}^{test}|} \cdot \mathbb{E}\left\{\frac{V^{test}(\tau)}{V^{cal}(\tau)+1}|\mathcal{D}^{cal}, \mathcal{D}^{test}\right\}\right] \\ &= \alpha \cdot \mathbb{E}\left[\frac{|\mathcal{D}^{cal}|+1}{|\mathcal{D}^{test}|} \cdot \mathbb{E}\left\{\frac{V^{test}(t_l)}{V^{cal}(t_l)+1}|\mathcal{D}^{cal}, \mathcal{D}^{test}\right\}\right] \\ &= \alpha \cdot \mathbb{E}\left\{\frac{|\mathcal{D}^{cal}|+1}{|\mathcal{D}^{test}|} \cdot \frac{|\mathcal{D}^{test,0}|}{|\mathcal{D}^{cal,0}|+1}\right\} \\ &\leq \alpha \cdot \mathbb{E}\left\{\frac{p_c^{test,0}}{p_c^{cal,0}}\right\} := \gamma_c \alpha, \end{aligned} \tag{C.6}$$

where $p_c^{test,0}$ and $p_c^{cal,0}$ are the proportions of individuals that do not belong to class c in the test and calibration data, respectively. To get Equation (C.6) we have used the fact that when t_l is used then all subjects are classified to class c. This completes the proof.

C.2 Proof of Part (b)

The proof is more complicated as the arguments involve constructing two martingales. We follow the same organization of the proof for Part (a). Details are provided for new arguments and omitted for repeated arguments similar to those in Part (a).

C.2.1 The empirical process description

The estimated FSP in group a for a given threshold t is:

$$\hat{Q}_c(t) = \frac{\frac{1}{n_a^{cal}+1}\left\{\sum_{i\in\mathcal{D}^{cal}} \mathbb{I}(\hat{S}_i^c \geq t, Y_i \neq c, A_i = a) + 1\right\}}{\frac{1}{m_a+n_a^{cal}+1}\left\{\sum_{j\in\mathcal{D}^{test}\cup\mathcal{D}^{cal}} \mathbb{I}(\hat{S}_j^c \geq t, A_j = a) + 1\right\}}.$$

Similar to (a) define $\tau = \hat{Q}_c^{-1}(\alpha) = \inf\left\{t : \hat{Q}_c(t) \leq \alpha\right\}$. Then our data-driven algorithm is given by $\mathbb{I}(\hat{S}_{n+j}^c \leq \tau)$. Define $V^{test}(t)$, $R^{test}(t)$, $V^{cal}(t)$ and $R^{cal}(t)$ as before. Following similar arguments as in (a), we have

$$\begin{aligned}
\text{FSP}_a^{\{c,*\}}(\tau) &= \frac{V^{test}(\tau)}{V^{cal}(\tau)+1} \cdot \frac{V^{cal}(\tau)+1}{R^{cal}(\tau)+R^{test}(\tau)+1} \cdot \frac{R^{cal}(\tau)+R^{test}(\tau)+1}{R^{test}(\tau)+1} \\
&\leq \alpha \cdot \frac{n_a^{cal}+1}{n_a^{cal}+m_a+1} \cdot \frac{V^{test}(\tau)}{V^{cal}(\tau)+1} \cdot \frac{R^{cal}(\tau)+R^{test}(\tau)+1}{R^{test}(\tau)+1}.
\end{aligned}$$

Next we shall show that the last two terms in the above product are both martingales.

C.2.2 Martingale arguments

In Part (a), we have shown that $V^{test}(t)/\{V^{cal}(t)+1\}$ is a discrete-time martingale with respect to the filtration $\mathcal{F}_k$ defined by (C.5), which is defined on the misclassification process. Next we show that $\{R^{cal}(t)+R^{test}(t)+1\}/\{R^{test}(t)+1\}$ is also a discrete-time martingale.

Consider the filtration that describes the *selection process*:

$$\mathcal{F}_k^* = \sigma\left\{R^{cal}(s_j^*), R^{test}(s_j^*) : j = \tilde{m}, \tilde{m}-1, \ldots, k\right\},$$

where s_j^* corresponds to the time when exactly j subjects are selected and $\tilde{m} = |D^{cal}| + |D^{test}|$. At time s_k^*, only one of the following two events are possible:

$$\begin{aligned} A_1^* &= \left\{R^{cal}(s_{k-1}^*) = R^{cal}(s_k^*), R^{test}(s_{k-1}^*) = R^{test}(s_k^*) - 1\right\}; \\ A_2^* &= \left\{R^{cal}(s_{k-1}^*) = R^{cal}(s_k^*) - 1, R^{test}(s_{k-1}^*) = R^{test}(s_k^*)\right\}. \end{aligned}$$

On this backward running filtration, we have

$$P(A_1^*) = \frac{R^{test}(s_k^*)}{R^{cal}(s_k^*)+R^{test}(s_k^*)}, \quad P(A_2^*) = \frac{R^{cal}(s_k^*)}{R^{cal}(s_k^*)+R^{test}(s_k^*)}.$$

It follows that $\{R^{cal}(t)+R^{test}(t)+1\}/\{R^{test}(t)+1\}$ is a martingale since

$$\begin{aligned} & \mathbb{E}\left\{\frac{R^{cal}(s_{k-1}^*)+R^{test}(s_{k-1}^*)+1}{R^{test}(s_{k-1}^*)+1}|\mathcal{F}_k^*\right\} \\ = & \frac{R^{cal}(s_k^*)+R^{test}(s_k^*)}{R^{test}(s_k^*)} \cdot \frac{R^{test}(s_k^*)}{R^{cal}(s_k^*)+R^{test}(s_k^*)} + \frac{R^{cal}(s_k^*)+R^{test}(s_k^*)}{R^{test}(s_k^*)+1} \cdot \frac{R^{cal}(s_k^*)}{R^{cal}(s_k^*)+R^{test}(s_k^*)} \\ = & \frac{R^{cal}(s_k^*)+R^{test}(s_k^*)+1}{R^{test}(s_k^*)+1}. \end{aligned}$$

C.2.3 FSR Control

Note that the threshold τ is a stopping time with respect to the filtration $\mathcal{F}_k^*$. Let $\mathcal{D}^{test,0}/\mathcal{D}^{cal,0}$ be the index sets for subjects in the testing/calibration data that do not belong to class c.

$$\begin{aligned}
\text{FSR}_a^{c,*} &= \mathbb{E}\left\{\text{FSP}_a^{c,*}(\tau)\right\} \\
&\leq \alpha\mathbb{E}\left[\frac{|D^{cal}|+1}{|D^{cal}|+|D^{test}|+1}\cdot\frac{R^{cal}(\tau)+R^{test}(\tau)+1}{R^{test}(\tau)+1}\cdot\mathbb{E}\left\{\frac{V^{test}(\tau)}{V^{cal}(\tau)+1}|\mathcal{D}^{cal},\mathcal{D}^{test}\right\}\right].
\end{aligned}$$

The term $\{R^{cal}(\tau)+R^{test}(\tau)+1\}/\{R^{test}(\tau)+1\}$ can be factored out because $R^{cal}(\tau)$ and $R^{test}(\tau)$ are constant when $\mathcal{D}^{cal}$ and $\mathcal{D}^{test}$ are given. According to Part (a), $\{V^{test}(t)\}/\{V^{cal}(t)+1\}$ is a backward martingale on $\mathcal{F}_k$. When t_l is used then all subjects are classified to class c. According to the optional stopping theorem we have

$$\mathbb{E}\left\{\frac{V^{test}(\tau)}{V^{cal}(\tau)+1}|\mathcal{D}^{cal},\mathcal{D}^{test}\right\}=\frac{V^{test}(t_l)}{V^{cal}(t)+1}=\frac{|\mathcal{D}^{test,0}|}{|\mathcal{D}^{cal,0}|+1}.$$

Next, conditional on the filtration defined on the selection process, we have

$$\mathbb{E}\left\{\frac{R^{cal}(\tau)+R^{test}(\tau)+1}{R^{test}(\tau)+1}\right\}=\mathbb{E}\left\{\frac{R^{cal}(t_l)+R^{test}(t_l)+1}{R^{test}(t_l)+1}\right\}=\mathbb{E}\left\{\frac{|D^{cal}|+|D^{test}|+1}{|D^{test}|+1}\right\}.$$

Combining the above results, we have

$$\begin{aligned}
\text{FSR}_a^{c,*} &\leq \alpha\mathbb{E}\left[\frac{|D^{cal}|+1}{|D^{cal}|+|D^{test}|+1}\cdot\frac{|D^{cal}|+|D^{test}|+1}{|D^{test}|+1}\cdot\frac{|\mathcal{D}^{test,0}|}{|\mathcal{D}^{cal,0}|+1}\right] \\
&= \alpha\cdot\mathbb{E}\left\{\frac{|\mathcal{D}^{cal}|+1}{|\mathcal{D}^{cal,0}|+1}\cdot\frac{|\mathcal{D}^{test,0}|}{|\mathcal{D}^{test}|}\right\} \leq \gamma_c\alpha,
\end{aligned}$$

where γ_c is defined at the end of Section C.1.3. The proof is compete.

D Proof of Theorem 2

The theorem implies that the optimal base score for constructing R-values should be $S^c(x,a) = \mathbb{P}(Y = c|X = x, A = a)$. 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 D.1, then prove the optimality theory in Section D.2.

D.1 A monotonicity property

Suppose we use $S^c_{n+j}(x,a) = \mathbb{P}(Y_{n+j} = c|X_{n+j} = x, A_{n+j} = a)$ as the base score. The corresponding theoretical R-values can be obtained via the conversion algorithm in Appendix A.5. Under Model A.1, the mFSR level with threshold t is $\mathrm{mFSR}^{\{c\}}_a(t) = \mathbb{P}(Y \neq c|S^c \geq t, A = a)$. The theoretical R-values is defined as $R^c(s^c) = \inf_{t \leq s^c} \left\{ \mathrm{mFSR}^{\{c\}}_a(t) \right\}$. Let $Q^c_a(t)$ be the mFSR level when the threshold is t. The next proposition characterizes the monotonic relationship between $Q^c_a(t)$ and t.

Proposition 2. *$Q^c_a(t)$ is monotonically decreasing in t.*

The proposition is essential for expressing the oracle procedure as a thresholding rule based on S^c. Specifically, denote $Q^{c,-1}_a(\cdot)$ the inverse of $Q^c_a(\cdot)$. The monotonicity of $Q^c_a(t)$ and the definition of the theoretical R-value together imply that $S^c_j(x,a) = Q^{c,-1}_a(R^c_j)$ for $a \in \mathcal{A}$. For notational convenience, let $T_{n+j}(x,a) = \mathbb{P}(Y_{n+j} = 2|X_{n+j} = x, A_{n+j} = a)$. Then $S^1_{n+j} = 1 - T_{n+j}$ and $S^2_{n+j} = T_{n+j}$. Therefore the oracle rule

$$\delta^{n+j}_{OR} = \mathbb{I}(R^1_{n+j} \leq \alpha_1) + 2\mathbb{I}(R^2_{n+j} \leq \alpha_2).$$

can be equivalently written as

$$\begin{aligned}\delta_{OR}^{n+j} &= \mathbb{I}\left\{S_{n+j}^1 \geq Q_{1,a}^{-1}(\alpha_1)\right\} + 2\mathbb{I}\left\{S_{n+j}^2 \geq Q_{2,a}^{-1}(\alpha_2)\right\} \\ &= \mathbb{I}\left\{T_{n+j} \leq 1 - Q_{1,a}^{-1}(\alpha_1)\right\} + 2\mathbb{I}\left\{T_{n+j} \geq Q_{2,a}^{-1}(\alpha_2)\right\}. \end{aligned} \tag{D.1}$$

for $1 \leq j \leq m$. This provides a key technical tool in Section D.2.

Proof of Proposition 2.

Define $\tilde{Q}_a^c(t) = 1 - Q_a^c(t)$. We only need to show that $Q_a^c(t)$ is monotonically increasing in t. Let $\mathcal{M}_a = \{n+1 \leq j \leq n+m : A_j = a\}$. According to the definition of the mFSR and the definition of S_j^c, we have

$$\mathbb{E}\left\{\sum_{j\in\mathcal{M}_a}\left\{S_j^c - \tilde{Q}_a^c(t)\right\}\mathbb{I}(S_j^c > t)\right\} = 0, \tag{D.2}$$

where the expectation is taken over both $\mathcal{D}^{test}$. 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 (D.2) that $\tilde{Q}_a^c(t) > t$ otherwise the summation on the LHS must be positive, leading to a contradiction.

Next we show that $t_1 < t_2$ implies $\tilde{Q}_a^c(t_1) \leq \tilde{Q}_a^c(t_2)$. Assume instead that $\tilde{Q}_a^c(t_1) > \tilde{Q}_a^c(t_2)$. We focus on group a, then

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

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 $\tilde{Q}_a^c(t) > t$. For term III, we have $\mathbb{E}(III) < 0$ since we assume $\tilde{Q}_a^c(t_1) > \tilde{Q}_a^c(t_2)$. It follows that the assumption $\tilde{Q}_a^c(t_1) > \tilde{Q}_a^c(t_2)$ cannot be true, and the proposition is proved.

D.2 Proof of the theorem

Define the expected number of true selections ETS $= \sum_{j=1}^{m} \mathbb{I}(Y_{n+j} = c, \hat{Y}_{n+j} = c)$. 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 2, the oracle rule can be written as

$$\delta_{OR}^{n+j} = \mathbb{I}\left\{T_{n+j} \leq 1 - Q_{1,a}^{-1}(\alpha_1)\right\} + 2\mathbb{I}\left\{T_{n+j} \geq Q_{2,a}^{-1}(\alpha_2)\right\}.$$

The mFSR constraints for the oracle rule imply that

$$\mathbb{E}\left\{\sum_{j\in\mathcal{M}_a}(T_j - \alpha_1)\mathbb{I}(\delta_{OR}^j = 1)\right\} = 0,\ \mathbb{E}\left\{\sum_{j\in\mathcal{M}_a}(1 - T_j - \alpha_2)\mathbb{I}(\delta_{OR}^j = 2)\right\} = 0. \tag{D.3}$$

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

$$\mathbb{E}\left\{\sum_{j\in\mathcal{M}_a}(T_j - \alpha_1)\mathbb{I}(\delta_j = 1)\right\} \leq 0,\ \mathbb{E}\left\{\sum_{j\in\mathcal{M}_a}(1 - T_j - \alpha_2)\mathbb{I}(\delta_j = 2)\right\} \leq 0. \tag{D.4}$$

The ETS of $\boldsymbol{\delta} = (\delta_j : n+1 \leq j \leq n+m)$ is given by

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

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 (D.3) and (D.4), we have

$$\mathbb{E}\left[\sum_{j\in\mathcal{M}_a}(T_j-\alpha_1)\left\{\mathbb{I}(\delta^j_{OR}=1)-\mathbb{I}(\delta_j=1)\right\}\right]\geq 0. \tag{D.5}$$

Let $\lambda_{1,a}=(1-Q^{-1}_{1,a}(\alpha_1)-\alpha_1)/Q^{-1}_{1,a}(\alpha_1)$. It can be shown that $\lambda_{1,a}>0$. For $j\in\mathcal{M}_a$, 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

$$\mathbb{E}\left[\sum_{j\in\mathcal{M}_a}\left\{\mathbb{I}(\delta^j_{OR}=1)-\mathbb{I}(\delta_j=1)\right\}\left\{T_j-\alpha_1-\lambda_{1,a}(1-T_j)\right\}\right]\leq 0. \tag{D.6}$$

According to (D.5) and (D.6), we have

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

Note that $\lambda_{OR}>0$, the desired result follows.

E Additional Numerical Results

E.1 Comparing the R and R^+-value

In this section, we demonstrate through simulation that the R^+-value (13) is more stable than the R-value (12) when $|\mathcal{D}^{test}|$ is small. To do this, we will look at the distributions of R-value and R^+-value for a fixed base 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 base scores are constructed as the oracle class probabilities $P(Y = c|X, A)$.

In Figure 9, we compute $1,000$ R-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 9 show the histograms the R-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-value has much more variability than the R^+-value. This is because the denominator of the R-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 9, the advantage of the R^+-value slowly disappears as $|\mathcal{D}^{test}|$ increases. This causes the variability of both R-value and R^+-value to become almost identical. We conclude from this simulation that the R^+-value 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-value has more variability for small $|\mathcal{D}^{test}|$, this disadvantage can be quickly overcome through the introduction of a reasonably sized test set.

E.2 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

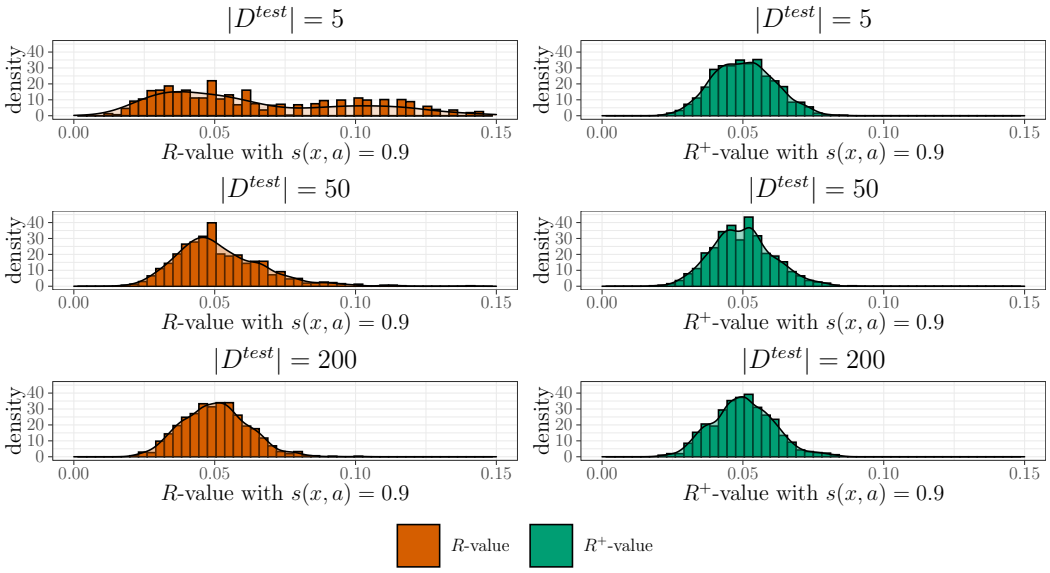

Figure 8: The comparison between the R-value and R^+-value for varying sizes of the test data set. The left column shows the histograms of the R-value (orange) and the right column shows the histograms of the R^+-value (green). The R-values and R^+-values are computed for a fixed base score of $s(x,a)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

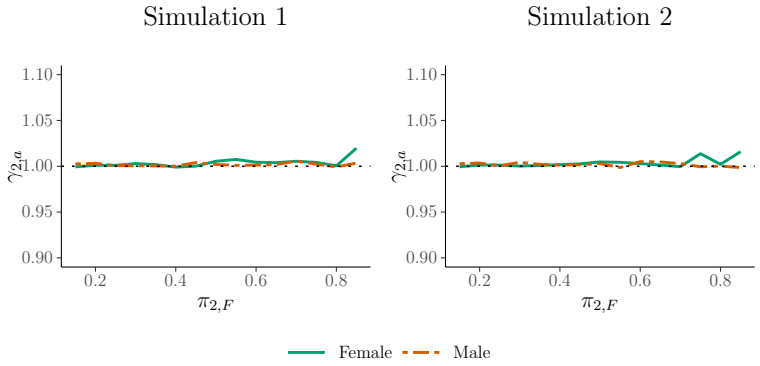

Figure 9: Estimates of $\gamma_{1,a}$ from the simulations in Section 4. The solid (green) line represents the estimate of $\gamma_{1,\mathrm{F}}$ for the Female protected group and similarly the orange (long-dashed) line for the Male protected group.

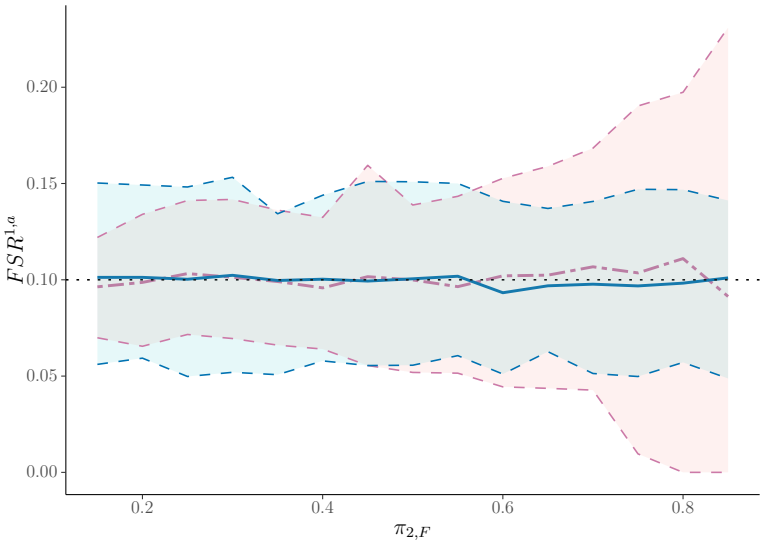

Figure 10: 90% quantiles for FSR control. Red: The female protected group. Blue: The male protected group. The x-axis varies over the true proportion of signal from the female group with the proportion of signal from the male group fixed at 50%. Goal is to control FDR at 10%.

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.

E.3 Quantiles of FSR control

Section 4 demonstrates the FASI algorithm can control the FSR in a simulation setting. The setup of this simulation is described in the same section, where the protected groups are Female and Male. 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 $\pi_{2|F} = \pi_{2|M} = 0.8$, $\boldsymbol{\mu}_1 = (0, 1, 6)^\top$ and $\boldsymbol{\mu}_2 = (2, 3, 7)^\top$. The base scores are constructed as the oracle class probabilities $P(Y = c|X, A)$. In this section, we expand on this setting and add the 90% quantiles for the FSR of classification group 1, against a varying level of the true proportion of class 2 from the female protected group.

Figure 10 demonstrates this in a data driven setting where the base scores are estimated using a GAM model (in the same way as Figure 3). The group-wise FSR represented by the solid (blue) and dot-dashed (red) lines are controlled at the desired 10% level. The quantiles are represented by the blue or red shaded regions representing the male and female protected groups respectively. For the Male protected group, whose true proportion of signal from class 2 never changes, has 90% quantiles that range between 5% and 15%. The Female protected group's quantiles are similar, except when their true proportion of signal form class 2 gets close to 1. This means that there are very little true observations for the Female protected group that truly belong to class 1, which is represented with larger quantiles for their FSR control.

E.4 Using Multiple ML 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 base scores. In this section, Figure 11 replicates the results of Simulation 1 in Section 4, for a varity of machine learning models where the data has two protected groups, Female and Male. In this section we use, logistic regression, GAM, Nonparametric Naive Bayes, and XG Boost (James et al. 2023, Hastie et al. 2009, Silverman 1986, Chen & Guestrin 2016) to estimate the base scores that will be converted to the R-values for our FASI framework.

The left column of Figure 11 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, XG Boost has a slightly higher EPI that gets

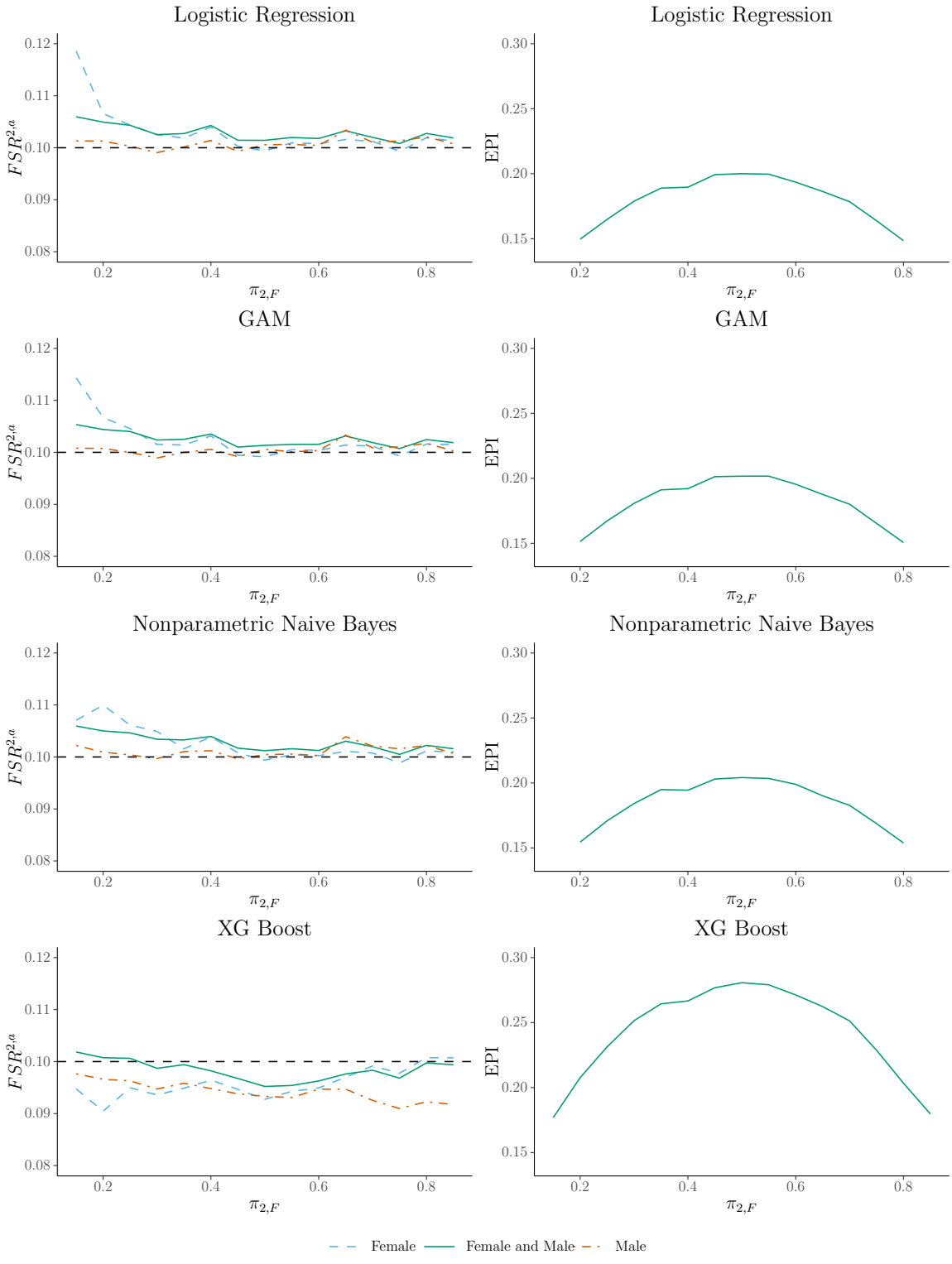

Figure 11: FSR control for the high risk classification. Left column: The resulting FSR from multiple different ML models that are used to estimate the base scores used to calculate the R-value. Right column: The corresponding EPI from different base scores. The overall FSR (green / solid) as well as 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%.

closer to 30% in the worst case. This is a consequence of the accuracy that each ML model has when estimating the true posterior 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.