In Pursuit of Machine Learning Fairness: Flexibly Detecting Subgroup and Intersectional Subgroup Biases

Julie E. Cestaro

NYU Gallatin School of Individualized Study

Master's Thesis

Advised by Dr. Daniel B. Neill

April 2026

Contents

1 Introduction

As machine learning becomes increasingly embedded in our daily lives, the impact it has on the lives and livelihoods of individuals grows as well. We are ever more aware of the presence of machine learning on social media and its ability to inform our thoughts and ideas (Tolentino, 2019), but it is also present in the job market (Dastin, 2018), in healthcare (Zbontar et al., 2018), and in the criminal justice system (Angwin et al., 2016).

This is not inherently bad or dangerous. Under conscientious oversight, machine learning can help us allocate resources more effectively, identify and reduce disparities, and improve information access and sharing for populations that would otherwise be left in the dark. The scalability of machine learning makes these benefits all the more accessible, but also means that failures can be catastrophic and widespread. In the worst cases, machine learning violates privacy, perpetuates and exacerbates societal biases, and directly harms the mental and physical well-being of people (Slattery et al., 2024).

The pursuit of ethical machine learning must be an active process. It requires careful design, formulation, and evaluation at every stage of the machine learning pipeline and throughout the lifespan of the system (Black et al., 2023). This research offers a flexible method for auditing machine learning predictions for disproportionate errors, which can be applied to any classification model for which group fairness criteria must be met.

2 Preface

2.1 The Promise of AI

AI has a bad reputation in popular culture. Some claim it is going to become the Terminator, or the machines that enslave humans in *The Matrix*. Others point to the very real problems of copyright infringement (Samuelson, 2023) and cognitive over-reliance (Zhai et al., 2024).

But many of us actually use AI every day, often without realizing it or consciously seeking

it out. When your watch automatically asks if you would like to record a walk when you start up the block or your phone suggests that you route yourself home at the end of the workday, we are engaging with predictions made using AI. We even benefit from more explicit applications of AI. It can complete tasks we humans do not want to (or cannot) do, and it can optimize computations in a way that makes overwhelmingly large tasks much more solvable. Consider self-driving cars for people whose disabilities make driving impossible or analysing large amounts of medical data quickly to assist in diagnosis (Schaich Borg et al., 2024, pp. 8–9). Even something as simple as autocorrect is a small but useful application space for AI. Used ethically, AI can measurably improve our world and make our lives easier.

2.2 Machine Learning

Machine learning is the process that gives AI its so-called "intelligence." At the most basic level, machine learning is just an algorithm in the same way that adding two numbers together is an algorithm. Algorithms exist on a spectrum of complexity, from simple arithmetic to complicated Excel formulas to something like machine learning. These algorithms all fundamentally work very similarly in that they take inputs, manipulate them, and produce some output.

Unlike a standard algorithm, where the treatment of the inputs is explicitly encoded and the outputs are entirely deterministic,[1] a machine learning algorithm learns patterns from the inputs without explicit direction from the programmer. The result is a model that has more or less independently determined how each new data point should be treated, which is both the beauty and the danger of machine learning (Blackman, 2022).

In general, machine learning will take one of a few forms. The focus area of this research is in supervised learning, but we will briefly cover the other processes for context.

When we already have some existing information about how the model should use the input data, we give the model *labelled data* so it can learn an association between the data and the label or outcome. In practice this looks like training a classifier to identify dogs by giving it a dataset of

[1]This use of the term "deterministic" in this context refers to the mathematical definition: a system in which no randomness is involved in the development of future states of the system.

images labelled "dog" and "not dog."[2] This process is called supervised learning.

Sometimes, though, we do not have enough information about the data or the prediction task to provide labels to the model. Instead, we give the model an unlabelled dataset and allow it to find commonalities between the data points. This is called unsupervised learning, and it is most similar to the kind of learning that humans do on a daily basis as we observe the world.

In both supervised and unsupervised learning, the quality of the model is extremely dependent on the quality of the input data and (when applicable) the associated labels. If the data does not accurately represent the setting in which the model is going to be deployed, the model will not have learned any patterns that are useful to its actual application space.

The third method for training a model is called reinforcement learning, which is most similar to the way that babies learn how to interact with the world. When they learn to speak, for example, they babble nonsensically until they say something that resembles a word, and then their parent or caregiver reacts positively. That positive reaction is effectively a reward and encourages the baby to repeat the same noise that had gotten them a positive response. This is reinforcement learning in humans, and it is not so different from reinforcement learning in machine learning. The algorithm makes many attempts to interact with the world and continues to repeat the actions that give it the highest reward. It then encodes the actions that give a high reward and "learns" that task.

2.3 Algorithmic Decision Makers in Practice

Even before machine learning, people were inclined to apply algorithms to settings where a fair and unbiased decision is absolutely essential, operating on the assumption that numbers are neutral. People can be so subjective, but an algorithm? An algorithm has no feelings, no opinions. It will see the inputs objectively and surely offer an objective output as a result.

A company called Northpointe was formed in 1989 to apply that assumption for a fairer criminal justice system. In 1998, its founders developed the Correctional Offender Management

[2]These labels may seem strange, but machine learning works better with binary distinctions.

Profiling for Alternative Sanctions (COMPAS) system to attempt to remove human biases from the criminal justice system. COMPAS would use machine learning to predict whether or not a defendant was likely to commit another crime, and then the judge would use that prediction to inform their sentencing decision (Christian, 2020, p. 56).

In 2001, New York State began using COMPAS to inform probation decisions based on predicted recidivism rates for a given individual. Within ten years, in 2011, New York State law was amended to require the use of systems like COMPAS when determining an individual's parole (Christian, 2020, p. 57). No one here was a bad faith actor. The *New York Times* editorial board even proposed that the adoption of COMPAS would break the New York parole board of its "subjective, often unreviewable judgements" (Editorial Board, 2015). By 2015, use of COMPAS had expanded to California, Wisconsin, and Florida, totalling approximately two hundred different jurisdictions.

The following year, in 2016, Julia Angwin and Jeff Larson with ProPublica published an alarming discovery which would become a foundational study for fairness in machine learning. The COMPAS system worked by taking some input data about a defendant and outputting a "risk score" between 1 and 10, which represented the likelihood that the individual might reoffend (1 being not likely at all, 10 being highly likely). Angwin and Larson discovered that while Black defendants were approximately equally likely to receive any of the scores from 1 to 10, White defendants were vastly more likely to be scored as a 1 and vastly less likely to be scored as a 10. Feeding recidivism data into a computer did not automatically cleanse it of bias. In fact, Angwin et al. (2016) clearly demonstrate that Black defendants are twice as likely to be mis-classified as high risk for recidivism by COMPAS compared to White defendants.

2.4 Why Should We Care About Fairness?

In a world where some groups are privileged and others are systematically disadvantaged, machine learning will learn to benefit the privileged group and harm the disadvantaged group, because that is what our society has already done. We see these historical biases most obviously in

redlining (Prince and Schwarcz, 2020) and over-policing (Lum and Isaac, 2016). Machine learning is in some ways a mirror: it learns patterns from historical data and applies those patterns to make future predictions (Schaich Borg et al., 2024, p. 57).

But what makes machine learning particularly insidious is in the ways in which it is more than a mirror. When a model trained on historically biased data is used to make decisions about who receives a loan, who should be released on parole, or who is recommended for a job interview, it does not simply reproduce existing disparities. It legitimizes them, lending the appearance of objectivity to outcomes that are anything but. Over time, the predictions of a biased model can reshape the reality it was trained to represent: if a predictive policing algorithm directs officers to over-police certain neighborhoods, the resulting arrest data will confirm the algorithm's predictions, creating a feedback loop that is difficult to detect and even harder to break (Lum and Isaac, 2016).

Even in circumstances where there are less obvious systemic biases, the systems that have led to the creation of foundational datasets in machine learning have resulted in the proliferation of biased views. The people in power, those who have financial and social advantages, are the people with access to and agency over what the data represent and, further, what the model learns from it. Especially in supervised learning, the data is the source of truth. It is the point of comparison that determines whether or not the model is accurate. The model's goal then is not to represent the objective reality of the world, but rather reality as it is determined by the data, which is compiled and labeled by a privileged group of people (Crawford, 2021).

This problem is compounded by the fact that the harms of biased machine learning are rarely distributed evenly. The communities most likely to be subject to automated decision-making in high-stakes domains are also the communities with the least political and institutional power to challenge those decisions (Eubanks, 2018). A middle-class professional denied a streaming recommendation suffers a non-trivially different harm than a low-income family whose benefits are incorrectly terminated by an automated eligibility system. The stakes are not uniform, and neither is the capacity to seek recourse.

The inherent nature of these issues means we must actively pursue and design for fairness in machine learning. But humans are fallible creatures. We are generally not very good at making decisions that maximize social good when they require sacrificing individual financial or social gain, and designing fairer models often takes additional time and resources (Schaich Borg et al., 2024, p. 110). The AI industry has become a multi-billion dollar industry, with corporate investment in the technology in the hundreds of billions as of 2024 (Stanford Institute for Human-Centered Artificial Intelligence, 2025), and many are willing to make sacrifices to be among the first to claim a piece of that investment. The incentive structure, in other words, does not naturally reward the kind of careful, deliberate work that fairness requires.

Furthermore, developing AI is genuinely fun. It presents a complex and nuanced technical challenge that can be both broadly applicable and deeply satisfying when executed successfully. This makes the field highly appealing to both experienced software engineers and casual hobbyists, many of whom begin building without fully understanding how their systems will perform in the context of our society. But enthusiasm and technical skill are not substitutes for an acute awareness of the potential risks and implementation of necessary mitigation strategies.

Because we currently lack strong regulatory frameworks or institutional oversight capable of evolving at the speed of the technology, the responsibility falls to those who develop and distribute AI to ensure that it does not cause harm. That responsibility is not optional, and it does not end at deployment. We have an obligation to ourselves and to our society to take an active, ongoing role in the fairness and, more broadly, the ethical integrity of the systems we build and maintain (Schaich Borg et al., 2024, pp. 110–111).

3 Fairness in Machine Learning Systems

3.1 Fairness Through Awareness

A natural idea when trying to create a more fair model is to simply remove the protected characteristics from the dataset and train a model that is blind to features like race, gender, or age.

This solution is based on the assumption that if the model does not have access to these protected characteristics, then it cannot discriminate based on them.

But no characteristic of a person exists in isolation. Race, gender, and age all impact a person's annual income, for example. If a prediction task is dependent on one of these demographic characteristics, the model will likely find a proxy variable for the removed characteristic and continue to exhibit bias. If all of the demographic characteristics and their proxies are removed, then the remaining features are not likely to provide enough information to train a useful classifier (Dwork et al., 2012).

Additionally, if all of the protected characteristics are removed from the dataset, it becomes much more difficult to measure and correct the bias of the classifier because we have lost all sense of these protected characteristics (Andrus et al., 2021).

3.2 Formal Definitions of Fairness

Given that class-blind fairness is not a viable option for achieving actual fairness, we have to consider a definition for fairness in terms of protected class membership.

The machine learning fairness literature formulates fairness in terms of mathematical expressions. Setting a mathematical basis for these definitions gives us concrete properties of each criterion and helps us understand the relationships among them. While these definitions may not perfectly align with philosophical bases for what it means to be fair, equitable, or just (Binns, 2018), statistical fairness criteria are helpful given the statistical nature of machine learning. There are two primary classes of statistical fairness criteria in the literature: individual fairness and group fairness. This research focuses on meeting group fairness criteria, but we will briefly cover individual fairness for additional context.

Individual fairness is rooted in the idea that similar individuals should be treated similarly. Methods for individual fairness involve measuring the similarity between two individuals and determining if the predictions they received are comparably similar (Dwork et al., 2012). While this method seems straightforward and easy to understand, its success is highly dependent on the

quality of the similarity metric used. Additionally, comparing individuals within a population can quickly become a very computationally expensive[3] process.

Group fairness allows us to conceptualize fairness in terms of the demographics of the population, which is helpful when the population is large. Additionally, fairness in terms of groups mirrors the wording of legal bases for non-discrimination. There are three formal criteria under group fairness, all of which are fundamentally mathematically incompatible except in degenerate cases. In other words, it is not possible to meet all three group fairness criteria simultaneously unless the classifier is a Bayes-optimal classifier[4] or in the case where the base rate[5] for each group is the same. Therefore, it is left to the machine learning practitioner or auditor to determine which fairness criterion is most appropriate for a given setting.

The first criterion is independence, also referred to as statistical or demographic parity. Independence requires that the sensitive characteristic is statistically independent of the prediction or score assigned by the predictor (Barocas et al., 2023, p. 55). In other words, the prediction cannot depend on an individual's protected class membership status in order to satisfy independence. This requires a strong assumption that all groups should be equally likely to receive any of the available outcomes, all else being equal. This may not necessarily hold in situations where all else is not equal and some groups have better access to resources than others (Barocas et al., 2023, p. 56).

The additional group fairness criteria, separation and sufficiency, attempt to account for the reality that not all groups are equally likely to have any of the available outcomes.

When trying to achieve separation, we want to ensure that error rates are consistent across groups. Most machine learning models are going to make errors; it is in fact highly unlikely that a model will naturally be a perfect predictor.[6] Separation asks, which group in the population bears

[3]Computational "cost" is most often measured in terms of time taken to complete computations or in terms of computational resources needed.

[4]A Bayes-optimal classifier is essentially the best possible classifier that is only achievable if we understand the true underlying probability distributions of the data. Typically the model learns an approximation because we rarely know the true underlying probability distributions of the data.

[5]A base rate is the proportion of a given real outcome for a given population. For example, if 1% of the population were medical professionals, and remaining 99% were not medical professionals, then the base rate of medical professionals is 1%.

[6]Models can only learn from patterns in the data and attempt to approximate the true probability distribution. Real-world phenomena are noisy and complex, which makes them very hard to model in a way that still generalizes to

the burden of the model's errors? Mathematically this means that in order to achieve separation, protected class membership can have no impact on the prediction when conditioned on the real, labelled outcome. In other words, we use the target variable as a sense of merit, and require that those of equal merit have an equal likelihood of receiving a positive outcome (Barocas et al., 2023, p. 57).

The third and final group fairness criterion is sufficiency. Sufficiency asks that the prediction accurately reflects an individual's merit, or real outcome, without being swayed by protected class membership (Barocas et al., 2023, p. 61). In other words, once you know what the model predicted, knowing a person's race, gender, or other protected attribute should give you no additional information about their true outcome. Essentially, the prediction is equally informative across groups — a score of 7 means the same thing regardless of the individual to whom it was assigned.

3.3 End to End Fairness

AI systems are so much more than just a machine learning algorithm making a prediction in isolation. They are the result of a series of decisions mostly made by people: from formulating a problem, collecting and processing data, choosing and tuning the model itself, to actually using the model's predictions (Black et al., 2023). Every choice that is made is a new opportunity to encode societal and individual biases.

3.4 Fairness Interventions

Given a mathematical basis for fairness, we can construct fairness interventions to be added to the machine learning pipeline (Figure 1). Most fairness interventions can be categorized as one of three types: pre-processing, in-processing, or post-processing.

Pre-processing methods, such as those described by Kamiran and Calders (2012) and Weerts et al. (2023), focus on the elements of the machine learning pipeline that occur before

unseen data.

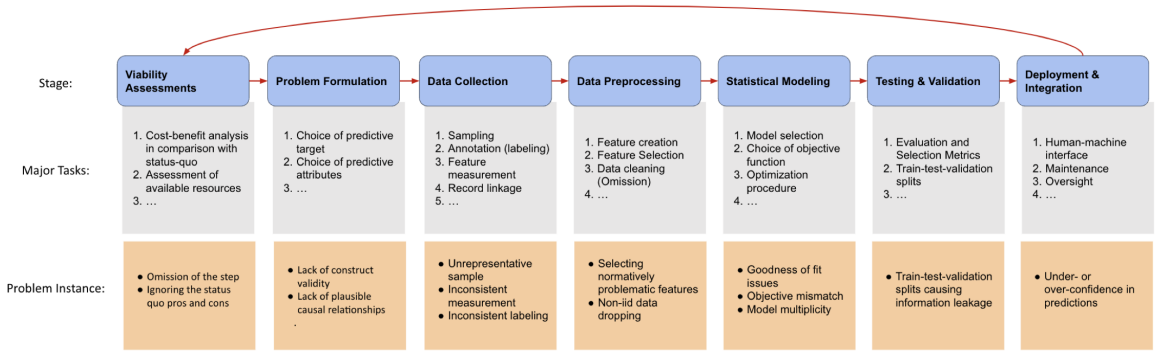

Figure 1: Reprinted from "Toward Operationalizing Pipeline-aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools" (Fig. 1), by E. Black, R. Naidu, R. Ghani, K. T. Rodolfa, D. E. Ho, & H. Heidari, 2023, *arXiv*. https://arxiv.org/abs/2309.17337

the model is chosen and trained. These methods typically focus on ensuring that the training data for the model is not contributing to bias. Often this involves transforming or modifying the data in some way so that the correlation between the sensitive attributes and the target variable (the attribute we are trying to predict) is lessened or removed. Of course, this is a delicate balance because the data should not be so different that it no longer represents the prediction setting accurately.

In-processing attempts to encode fairness criteria in the model training process. Instead of allowing the model to focus its training on simply representing the input data, in-processing methods force the model to represent the data both accurately and fairly. This is most commonly implemented by constraining or penalizing the model for drawing "unfair" connections between protected class membership and outcome during the training process (Kamishima et al., 2012). The goal is to prevent the model from learning relationships that violate fairness criteria in the first place. However, if the in-processing method is too constraining for the model, it may struggle to learn any meaningful relationships in the data and ultimately become an ineffective predictor.

Post-processing methods are applied to the model's predictions after the model has been trained and the predictions have been made. To apply post-processing is to acknowledge the biases of your predictor and to correct them before actually using the predictions to make any decisions.

Understanding how to correct the model in either pre- or post-processing methods requires

an understanding of the biases present in either the input data or the resulting predictions. This requires some kind of evaluation method to assess the placement, type, and extent of bias. It is in this particular niche of bias evaluation that this research falls.

3.5 Beyond the Code

While this research focuses primarily on the technical functionality of machine learning and the impact of its predictions, it is also worth mentioning other, non-technical aspects to the fairness of these systems.

Training a machine learning model typically requires large amounts of two resources: data and the processing power to learn from it. With data collection and use comes questions about consent and privacy. Datasets are most often created by scraping the contents of the internet for anything that might be useful, operating on the assumption that the mere existence of the information on the internet negates the need for the agreement or consent of the subject (Crawford, 2021). Furthermore, high-dimensional datasets collected from the internet, even when they do not explicitly encode personal data, can often be de-anonymized, putting individual privacy at risk (Narayanan and Shmatikov, 2008).

But beyond nearly universally exploiting people for their data is the exploitation of laborers who perform all of the repetitive digital tasks that support AI systems. Individual humans must label datasets for supervised learning tasks, review suspicious and potentially harmful content, and even manually validate reinforcement learning (Gray and Suri, 2017). These workers are often highly educated and yet are paid less than their local minimum wage for their work with no social protections like health insurance or retirement plans (Berg et al., 2018).

After collecting and processing all of this data, it needs to be stored somewhere for use. It is sent to data centers filled with servers that all run on massive amounts of electricity (Luccioni et al., 2024; Strubell et al., 2019), and whose components are made of rare earth metals that need to be mined. All of our technology, in fact, relies on electricity and rare earth elements, and the processing required to train machine learning models demands even more (Crawford, 2021).

All of this contributes to the continued destruction of our environment, and this burden is not shared equally. The worst effects of climate change fall to vulnerable populations and countries that have the fewest resources to protect themselves (World Health Organization, 2023). Even if all technical machine learning fairness criteria were always met in terms of the model's predictions, the environmental impacts of the exponential growth of this technology would still render it utterly unfair.

4 Methodology

In an effort to increase the ease with which biases in automated decisions are detected, this research proposes a post-hoc auditing method which flexibly detects biases against a subgroup defined by a single attribute (marginal subgroups) or subgroups defined by the intersection of multiple attributes (intersectional subgroups). The flexibility of this method allows auditors and machine learning practitioners to control the extent to which marginal subgroup biases can overcome intersectional subgroup biases in the detection process, or the other way around, depending on the circumstances of each particular decision system and audit. In other words, if a specific prediction context is already known to have an overwhelming marginal bias, this method will negate its effect so that intersectional biases can also be detected.

We begin by defining some dataset $\mathcal{D} = (\mathbf{X}, Y, \hat{Y}) = \{(\mathbf{X}, Y, \hat{Y})\}_{i=1}^{N}$ for N individuals indexed as $i = 1...N$. Let $\mathbf{X} \in \mathbb{R}^{N \times M}$ denote a matrix of covariates describing N individuals with M attributes. Let $\hat{Y} \in \{0, 1\}^N$ represent the binary predictions of a system under audit and let $Y \in \{0, 1\}^N$ denote the true outcomes the system is attempting to predict. We assume that all covariates in $\mathbf{X}$ are discrete-valued; continuous covariates can be discretized as a preprocessing step.

Very generally, our goal is to identify subsets S of the data, defined by a non-empty subset of values for each covariate $\mathbf{X}_1...\mathbf{X}_M$, that suffer from a disproportionate error rate (i.e., where $\hat{Y}$ has a systematic bias as compared to the true outcomes Y). Our framework is sufficiently flexible to detect violations of multiple different *group fairness criteria*, which can each be framed as

Fairness Definition	**Conditional Independence Criterion**
False Positive Rate (FPR) Balance	$\hat{y}_i \perp a_i \mid y_i = 0$
False Negative Rate (FNR) Balance	$\hat{y}_i \perp a_i \mid y_i = 1$
False Discovery Rate (FDR) Balance	$y_i \perp a_i \mid \hat{y}_i = 1$
False Omission Rate (FOR) Balance	$y_i \perp a_i \mid \hat{y}_i = 0$
Error Rate Balance	$(y_i \neq \hat{y}_i) \perp a_i$
Statistical Parity	$\hat{y}_i \perp a_i$
Outcome Parity	$y_i \perp a_i$

Table 1: Group fairness definitions. Our generalized bias scan can be applied to detect violations of any of these criteria.

conditional independence relationships between the true outcomes y_i, predicted outcomes $\hat{y}_i$, and protected class membership $a_i = \mathbf{1}\{i \in S\}$:

4.1 Scanning Method

Instead of testing every possible combination of attributes, which becomes computationally unreasonable, we utilize a multidimensional subset scanning method called Bias Scan to identify the biased subgroup, which we can do in linear rather than exponential time (Neill and Zhang, 2016). The scan aims to find a subgroup S in dataset $\mathcal{D}$ that maximizes the score function $F(S)$, which is a likelihood ratio score:

$$F(S) = \log\left(\frac{\Pr(\mathcal{D} \mid H_1(S))}{\Pr(\mathcal{D} \mid H_0)}\right) \tag{1}$$

where the null hypothesis H_0 is that the given prediction's odds are correct for all subgroups in $\mathcal{D}$ and alternative hypothesis $H_1(S)$ assumes some constant multiplicative bias in the odds for some given subgroup S:

$$H_0 : odds(y_i) = \frac{\hat{p}_i}{1-\hat{p}_i}\ \forall i \in \mathcal{D}, \text{ where } \hat{p}_i \text{ is the model's estimated probability that } y_i = 1. \tag{2}$$

$$H_1(S) : odds(y_i) = q\frac{\hat{p}_i}{1-\hat{p}_i}, \text{ where } q > 1\ \forall i \in S \text{ and } q = 1\ \forall i \notin S. \tag{3}$$

The scan iterates over each attribute until it arrives at a local maximum of $F(S)$ and uses multiple random restarts to approach the global maximum.

The original implementation of Bias Scan looks for miscalibration bias in predictions by comparing the binary outcomes y_i to $\hat{p}_i$, which is the model's estimate that $y_i = 1$, and identifying intersectional subgroups S where the probabilities are systematically biased upward or downward (Neill and Zhang, 2016). We instead focus on cases where the model's binarized predictions ($\hat{y}_i \in \{0, 1\}$) rather than its probabilistic predictions ($\hat{p}_i \in [0, 1]$) are observed. For example, for some probabilistic classification methods, $\hat{y}_i$ might be computed by comparing $\hat{p}_i$ to a threshold, e.g., $\hat{y}_i = \mathbf{1}\{\hat{p}_i > 0.5\}$, while for other binary classifiers, the most likely class $\hat{y}_i$ might be computed directly without first computing the probabilistic estimate $\hat{p}_i$.

4.2 Generalized Error Scanning

We can extend Bias Scan to evaluate error rate imbalances, detecting systematic differences between predicted outcomes $\hat{Y}$ and observed outcomes Y, by computing the observed error E and expected error rate $\hat{E}$ respectively, meaning that the null and alternative hypotheses are also updated as follows:

$$H_0 : odds(e_i) = \frac{\hat{e_i}}{1 - \hat{e_i}} \; \forall i \in \mathcal{D}, \text{ where } \hat{e_i} \text{ is an estimate of the probability that } e_i = 1. \quad (4)$$

$$H_1(S) : odds(e_i) = q\frac{\hat{e_i}}{1 - \hat{e_i}}, \text{ where } q > 1 \; \forall i \in S \text{ and } q = 1 \; \forall i \notin S. \quad (5)$$

In the simplest case of this scan, when $m = 0$, we define $\hat{e}_i$ as the global error rate, i.e., the average of all e_i values. More complex definitions of $\hat{e}_i$ and the conditions under which they are used are described in detail in subsection 4.3 and subsection 4.4. This generalized error formulation allows us to extend the scan to all of the binary error metrics outlined in Table 1.

The scanning algorithm itself is largely unchanged when scanning for error rates, save for a few additional pre-processing steps before the scan. For each error metric of interest, we restrict the dataset as described in Table 2 and compute error e_i for each row, also outlined in Table 2.

Scan Type	**Restriction of the Data**	**Formulation of e_i**
False Positive Rate (FPR)	$y_i = 0$	$\hat{y}_i$
False Negative Rate (FNR)	$y_i = 1$	$1 - \hat{y}_i$
False Discovery Rate (FDR)	$\hat{y}_i = 1$	$1 - y_i$
False Omission Rate (FOR)	$\hat{y}_i = 0$	y_i
Error Rate	none	$\mathbb{1}\{y_i \neq \hat{y}_i\}$
Statistical Parity	none	$\hat{y}_i$
Outcome Parity	none	y_i

Table 2: Scan types and relevant metrics for different group fairness definitions

Then we compute the expected error rate $\hat{e_i}$ for each row. Finally, we perform the scan to find the subgroup S that maximizes the scoring function $F(S)$, where $F(S)$ is computed using H_0 and $H_1(S)$ as defined in Equations 4 and 5 above.

4.3 Baseline Error Probability Estimation

As noted above, a simple assumption would be that, under the null hypothesis, all data records i would be expected to have equal error rates $\hat{e_i} = \hat{\Pr}(e_i = 1)$. In this case, $\hat{e_i}$ could be set equal to the global mean error rate (average of all e_i values) for all i. However, such a model might detect either marginal biases (e.g., error rates are high for Black males because they are high for both Black and male individuals) or *super-additive* intersectional biases (e.g., error rates are higher than one might expect for Black males, even controlling for race and gender separately). We now consider a new baseline model which can flexibly down-weight or ignore marginal biases in order to focus attention on super-additive intersectional biases (and therefore, increase the method's ability to detect such subgroups with super-additive intersectional biases, even when confounding marginal biases are present).

The purpose of the baseline model is to capture systematic patterns in the error rate that can be explained by marginal relationships with independent covariates. To estimate the expected error probability for each instance, we fit a logistic regression model predicting the error indicator

$\hat{e}_i$ from the covariates $\mathbf{x_i}$, which takes the form

$$\hat{e}_i = \Pr(e_i = 1 \mid \mathbf{x}_i) = \sigma\left(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}\right), \tag{6}$$

for each row i in $\mathcal{D}$, Here $\sigma(\cdot)$ denotes the logistic sigmoid function, $\sigma(x) = \frac{\exp(x)}{1+\exp(x)}$, which is used to convert the predicted log-odds of $e_i = 1$ to the predicted probability of $e_i = 1$.

Because logistic regression by nature assumes that each covariate value contributes independently to the prediction, using logistic regression to predict the probability of an error necessarily assumes that marginal group membership alone contributes to the probability of an error. When we use this as the baseline of comparison to detect subgroups for which the observed error rate significantly deviates from the expected error rate, any deviation from the expected error rate can be attributed to intersectional interaction between the covariates, which will remain unmodeled by the logistic regression, and thus can be detected by the scan as a difference between true and estimated probabilities of $e_i = 1$.

4.4 Marginal Bias Control via Multiplier

When auditing models for fairness, we may wish to control the extent to which we attribute errors to marginal effects of individual attributes and disparities arising from interactions among multiple attributes, especially when an overwhelming marginal bias is present in the predictions. To achieve this, we introduce a multiplier parameter $m \in [0, 1]$ that acts as an ℓ_1 regularization on the logistic regression used to predict $\hat{e}_i$. Effectively, m regulates the extent to which marginal bias patterns are incorporated into the baseline probability model. The relationship between m and the regularization is described in detail in subsubsection 4.4.1. When $m \approx 0$, the baseline error model is close to the global mean, meaning that little of the observed error variation predicted by the model is explained by covariates. In this setting, the scan procedure is sensitive to disparities that are the result of independent contributions of covariates.

As m increases, the logistic regression model is allowed to explain a larger portion of

the marginal variation in error rates across attributes. Consequently, disparities attributable to individual covariates are absorbed into the baseline model and no longer produce high anomaly scores. The scanning procedure therefore becomes increasingly focused on identifying subgroups whose error rates cannot be explained by marginal effects alone, and can detect biases as a result of intersections of multiple attributes. In the extreme case $m = 1$, marginal effects are ignored and only super-additive intersectional biases are detected.

4.4.1 Using m to Choose ℓ_1 Regularization

C is the inverse regularization strength for the logistic regression model and is chosen as a function of multiplier m. The goal is to regularize such that when $m = 0$ the predicted error $\hat{E}$ is equal to the global error rate and therefore does not account for marginal group membership at all, and when $m > 0$ we allow the predicted error to account for marginal group membership by a factor of m.

Prior to fitting any classifier, we standardize covariates $\mathbf{X}$ using z-score normalization with the column-wise means and standard deviations of $\mathbf{X}$. We then fit a baseline logistic regression model on the standardized features $\tilde{\mathbf{X}}$ and target variable $E \in \{0, 1\}$ with regularization strength $C_{\text{baseline}} = 1$. This represents the case where $m = 1$ and the full strength of the marginal groups can be used to predict the expected error $\hat{E}$. We then take the coefficient vector $\beta_{\text{baseline},1}...\beta_{\text{baseline},M}$ from the baseline model and use them to define the target sum T_m of the coefficients $\beta_{m,1}...\beta_{m,M}$ under multiplier m such that

$$T_m = \sum_{k=1}^{M} \beta_{m,k} = m \cdot \sum_{k=1}^{M} \beta_{\text{baseline},k}, \tag{7}$$

where M is the number of features in $\mathbf{X}$.

Because the ℓ_1 norm of the coefficient vector is a monotonic function of C, we can use binary search to find the regularization term C_m (outlined in Algorithm 1) which results in coefficients $\beta_{m,1}...\beta_{m,M}$ that satisfy Equation 7. We then use those coefficients to define logistic

regression function $f_m = \sigma\big(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta_m}\big)$, which we use to compute expected error $\hat{e}_i \in [0, 1]$ for each row $i \in \mathcal{D}$.

Algorithm 1: DYNAMICALLYFITCLASSIFIER($\tilde{\mathbf{X}}$, $\mathbf{e}$, m)

Input: standardized feature matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{N \times M}$,

binary error indicator $\mathbf{e} \in \{0, 1\}^N$,

multiplier $m \in \mathbb{R}$, $0 < m \leq 1$.

Output: Fitted logistic regression classifier $\hat{f}_{C_m} = (\beta_0, \beta_1, \ldots, \beta_M)$.

$(\beta_0, \boldsymbol{\beta}_{\text{baseline}}) \leftarrow$ LogisticRegression($\tilde{\mathbf{X}}$, $\mathbf{e}$)

$T \leftarrow m \cdot \|\boldsymbol{\beta}_{\text{baseline}}\|_1$

$C_{\text{low}} \leftarrow 0.0001$

$C_{\text{high}} \leftarrow 10.0$

while true do

  $C_{\text{mid}} \leftarrow \dfrac{C_{\text{low}} + C_{\text{high}}}{2}$

  $(\beta_0, \beta_1, \ldots, \beta_M) \leftarrow \ell_1$-LogisticRegression($\tilde{\mathbf{X}}$, $\mathbf{e}$, $C = C_{\text{mid}}$)

  $s_{\text{mid}} \leftarrow \sum_{k=1}^{M} \beta_k$

  $\delta \leftarrow |T - s_{\text{mid}}|$

  if $s_{\text{mid}} < T$ **and** $\delta > 0.001$ **then**

    $C_{\text{low}} \leftarrow C_{\text{mid}}$

  else if $s_{\text{mid}} > T$ **and** $\delta > 0.001$ **then**

    $C_{\text{high}} \leftarrow C_{\text{mid}}$

  else

    return $(\beta_0, \beta_1, \ldots, \beta_M)$

5 Evaluation

In subsection 2.3, we introduced the COMPAS system, which is used as a decision support tool by judges in the U.S. judicial system when deciding when an arrested individual should be

released prior to their trial (Angwin et al., 2016). After the initial investigation that ProPublica conducted in 2016, the COMPAS system has become a benchmark for fairness research in the machine learning fairness literature. We follow many of the pre-processing decisions made in the initial ProPublica analysis including removing traffic offenses and defining recidivism as a new arrest within two years of the initial arrest (Larson et al., 2016). After pre-processing, our dataset includes 7,214 individuals, their gender, race, age (Under 25 or 25+), charge degree (Misdemeanor or Felony), prior offenses (None, 1 to 5, or Over 5), predicted recidivism risk score s_i, where $s_i \in \{1, 2, ..., 10\}$, and whether they were re-arrested within two years of the initial arrest, where $y_i = 1$ if individual i was re-arrested and $y_i = 0$ otherwise. Given that COMPAS only provides risk scores and not predicted probabilities of reoffending, we define each defendant i's predicted probability of reoffending using maximum likelihood estimation:

$$P_{n_j} = \frac{\sum_{j=1}^{n} 1\{\mathbf{y}_i = 1 \wedge s_j = s_i\}}{\sum_{j=1}^{n} 1\{s_j = s_i\}} \tag{8}$$

Defendants with COMPAS risk scores of 5 or more $(s_i \geq 5)$ are considered "high risk" since the COMPAS documentation stipulates careful consideration by supervision agencies for these defendants (Larson et al., 2016). Therefore we define the COMPAS recommendations as $p_i = 1\{s_i \geq 5\}$.

To evaluate the ability of our generalized bias scan method to detect systematic prediction disparities, we perform 500 trials of each experiment over a semi-synthetic dataset that we generate for each trial, for each scan type. The semi-synthetic datasets are constructed based on the COMPAS data, for which we compute new outcomes (y_i) and new predictions ($\hat{y}_i$) for each entry i, and randomly select a biased subgroup (S_{bias}) for the scanning method to detect. While our experiments focus on standard separation (FPR balance, FNR balance, statistical parity) and sufficiency (FOR balance, FDR balance, outcome parity) metrics for group fairness, our generalized framework for error detection can extend to other binary error metrics by appropriate restriction of the dataset and definition of e_i.

5.1 Defining the Biased Subgroup

In our experiments, the biased subgroup S_{bias} is defined by selecting a random subset of attributes and then selecting values of those attributes that characterize membership in the subgroup. Specifically, we pick S_{bias} by choosing attributes uniformly at random, where the number of attributes is controlled by n_{bias}. We then independently include or exclude each value of those attributes with probability p_{bias} of being included in S_{bias}.

Consider a small dataset as an example, where the attributes in the dataset are age over 25 $\in \{0, 1\}$, race $\in \{\text{Black}, \text{White}\}$, and gender $\in \{\text{male}, \text{female}\}$. If $n_{bias} = 2$, then we choose two of those attributes at random, such as age over 25 and gender. Then for each value of each chosen attribute, we add that value to S_{bias} with a probability of p_{bias}. If $p_{bias} = 0.5$, we essentially flip a coin for each of the values in age over 25 and gender, and if the coin lands on heads, that value becomes part of the definition for S_{bias}. One example S_{bias} that could be created by this process is $\{\text{age over 25} : [0], \text{gender} : [\text{female}]\}$ and another example is $\{\text{age over 25} : [0, 1], \text{gender} : [\text{male}]\}$.

5.2 Generating Synthetic Predictions and Outcomes

The process for generating synthetic predictions and outcomes with injected biases to detect varies slightly depending on the type of bias that the method is looking for. Sufficiency-based scans (false omission, false discovery, outcome parity) will detect intersectional biases that appear as a result of differences in the true outcome y_i conditional on the predicted outcome $\hat{y}_i$. Separation-based scans (false positive, false negative, statistical parity) will detect intersectional biases that appear as a result of differences in the prediction $\hat{y}_i$ conditional on the true outcome y_i. We want to ensure that our scan is not searching for biases that are not actually present in the dataset, and thus we consider two different simulated injections: one that creates violations of the sufficiency-based fairness definitions (false omission rate balance, false discovery rate balance, and outcome parity), and one that creates violations of the separation-based fairness definitions (false positive rate balance, false negative rate balance, and statistical parity) respectively.

Regardless of scan type, the initial steps of the dataset generation process are the same. For each row in the dataset, we generate the following set of variables which are ultimately combined to create the predictions and outcomes in the semi-synthetic dataset.

1. For each attribute of the covariates, we draw a weight w_k from a Gaussian distribution, $w_k \sim \mathcal{N}(0, 0.2)$, which determines the relationship between features and the model's predicted outcome. These coefficients will be the basis for computing the true and predicted log odds for each row in the dataset.
2. For each attribute of the covariates, we draw an additional weight $w_{k,bias}$ from a Gaussian distribution, $w_{k,bias} \sim \mathcal{N}(0, \sigma_{marg})$, which represents the marginal bias that will act as a confounder for the intersectional bias.
3. We draw $\epsilon_{i,pred}$ from $\mathcal{N}(0, \sigma_{pred})$ and $\epsilon_{i,true}$ from $\mathcal{N}(0, \sigma_{true})$ for each individual i, to represent additional variation between individuals' true and predicted outcome values beyond what is captured in the predictor variables x_k.
4. We finally define some intersection term, which will be used to add the intersectional bias that we expect the method to be able to detect.

The process by which these variables are combined depends on the type of bias that the scan is expected to detect, as described in the subsections below.

5.2.1 For Sufficiency-based Scans

For sufficiency-based scans, we add the effects of the biases to the true log odds for each row in the dataset. To do this, we first combine the covariate weights (item 1) with the marginal bias we created (item 2) so that each covariate weight has an additional marginal bias incorporated into it. We then linearly combine those biased weights with each of the feature values to get the baseline true log odds for the row and add $\epsilon_{i,true}$ for noise (item 3). If a given row is a member of the biased subgroup S_{bias} defined as outlined in subsection 5.1 we add an additional intersectional bias ρ (item 4):

$$l_i^{true} = \sum_{k=1}^{M} (w_k + w_{k,bias}) x_{i,k} + \epsilon_{i,true} + \rho \mathbf{1}\{i \in S_{bias}\}.$$

Next, we generate the predicted log odds. We use the unbiased covariate weights (item 1) and linearly combine them with each of the feature values. This is the baseline predicted log odds for the row. We then add both ϵ_{true} and ϵ_{pred} for noise (item 3):

$$l_i^{pred} = \sum_{k=1}^{M} w_k x_{i,k} + \epsilon_{i,true} + \epsilon_{i,pred}.$$

5.2.2 For Separation-based Scans

For separation-based scans, we generate the true log odds without marginal or intersectional biases. We use the unbiased covariate weights (item 1) and linearly combine them with each of the feature values. This is the baseline true log odds for the row. We then add $\epsilon_{i,true}$ for noise (item 3):

$$l_i^{true} = \sum_{k=1}^{M} w_k x_{i,k} + \epsilon_{i,true}.$$

We then add the effects of the biases to the predicted log odds for each row in the dataset. To do this, we first combine the covariate weights (item 1) with the marginal bias we created (item 2) so that each covariate weight has an additional marginal bias incorporated into it. We then linearly combine those biased weights with each of the feature values to get the baseline predicted log odds for the row and add ϵ_{true} and ϵ_{pred} for noise (item 3). If a given row is a member of the biased subgroup S_{bias} defined in subsection 5.1 we add an additional intersectional bias (item 4):

$$l_i^{pred} = \sum_{k=1}^{M} (w_k + w_{k,bias}) x_{i,k} + \epsilon_{i,true} + \epsilon_{i,pred} + \rho \mathbf{1}\{i \in S_{bias}\}.$$

5.3 For Both Scan Types

For either simulation, once we have obtained the true log odds l_i^{true} and the predicted log odds l_i^{pred} for each individual i, we convert both of these to probabilities: $p_i^{true} = \frac{\exp(l_i^{true})}{\exp(1+l_i^{true})}$, and $p_i^{pred} = \frac{\exp(l_i^{pred})}{\exp(1+l_i^{pred})}$. We create then create binary outcomes y_i by sampling from a Bernoulli distribution, $y_i \sim \text{Bernoulli}(p_i^{true})$. Finally, we convert the probabilistic predictions p_i^{pred} to binary

predictions $\hat{y}_i$ by thresholding at 0.5:

$$\hat{y}_i = \begin{cases} 1 & \text{if } p_i^{\text{pred}} > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

5.4 Experiments

The experiments for this method are designed to demonstrate how different values of the multiplier m (defined in subsection 4.4) impact the scan's ability to detect the biased subgroup under different conditions. Because large values of m decrease the impact of marginal biases on the scoring function (outlined in subsection 4.1), we hypothesize that the method's ability to detect intersectional biases will improve as m increases, especially when large marginal biases are present in the dataset. We consider six variants of the generalized bias scan, respectively detecting (1) increased FPR; (2) decreased FNR; (3) increased $\hat{y}_i$ (violation of statistical parity); (4) increased FOR; (5) decreased FDR; and (6) increased y_i (violation of outcome parity). The first three variants correspond to disparities induced by the simulation for separation-based scans, and the last three variants correspond to disparities induced by the simulation for sufficiency-based scans. For each scan type and each value of $m \in \{0, 0.25, 0.5, 0.75, 1\}$, we tested the scan on six sets of experiments, each varying a distinct parameter from its default values of ($\rho = 2$, $\sigma_{marg} = 0.5$, $\sigma_{true} = 0.6$, $\sigma_{pred} = 0.6$, $n_{bias} = 2$, $p_{bias} = 0.5$). For each such experiment, we averaged results over 500 randomly generated synthetic signals (super-additive intersectional biases).

5.4.1 Evaluation Metrics

To evaluate our experiment results we collect precision, recall, and overlap. We define precision and recall given that true positive identifications are rows that are in S_{bias} and positively identified by the scan, false positives are rows that are positively identified but are not in S_{bias}, and false negatives are rows that are in S_{bias} but not identified by the scan. Letting S^* represent the

subset of records detected by the scan, we have:

$$\text{Precision} = \frac{|S_{bias} \cap S^*|}{|S^*|}$$

$$\text{Recall} = \frac{|S_{bias} \cap S^*|}{|S_{bias}|}$$

Overlap quantifies the similarity between the actual and found subgroups by computing the ratio between the size of their intersection and the size of their union, and is our primary metric of interest in determining successful detection of the biased subgroup by the method:

$$\text{Overlap} = \frac{|S_{bias} \cap S^*|}{|S_{bias} \cup S^*|}$$

5.5 Experimental Results

We focus here on the performance of generalized bias scan (defined as overlap coefficient between the injected and detected subsets of records) for different values of m, for experiments with (1) varying amounts of marginal bias σ_{marg} (confounder) for a fixed amount of intersectional bias $\rho = 2$ (signal); (2) varying amounts of intersectional bias ρ for fixed $\sigma_{marg} = 0.5$; and (3) varying number of affected attributes n_{bias} for fixed $\rho - 2$ and $\sigma_{marg} = 0.5$.

Figure 2 shows that as we increase the strength of the marginal bias's confounding impact on the error rate (with intersectional bias signal strength held constant), the scans using higher values of m are able to detect more of the biased subgroup. All methods' ability to detect the affected intersectional subgroup decreases as the amount of marginal bias increases, but higher m values are more robust to marginal bias, and thus the best-performing methods change from lower m values (0, 0.25, 0.5) to higher m values (0.5, 0.75) as σ_{marg} increases. We see relatively consistent results across the six scan types, with minor changes in the relative ordering of m values.

Similarly, Figure 3 shows that as we increase the intersectional bias signal (with the amount of marginal bias held constant at a moderate value of $\sigma_{marg} = 0.5$), the scans using moderate

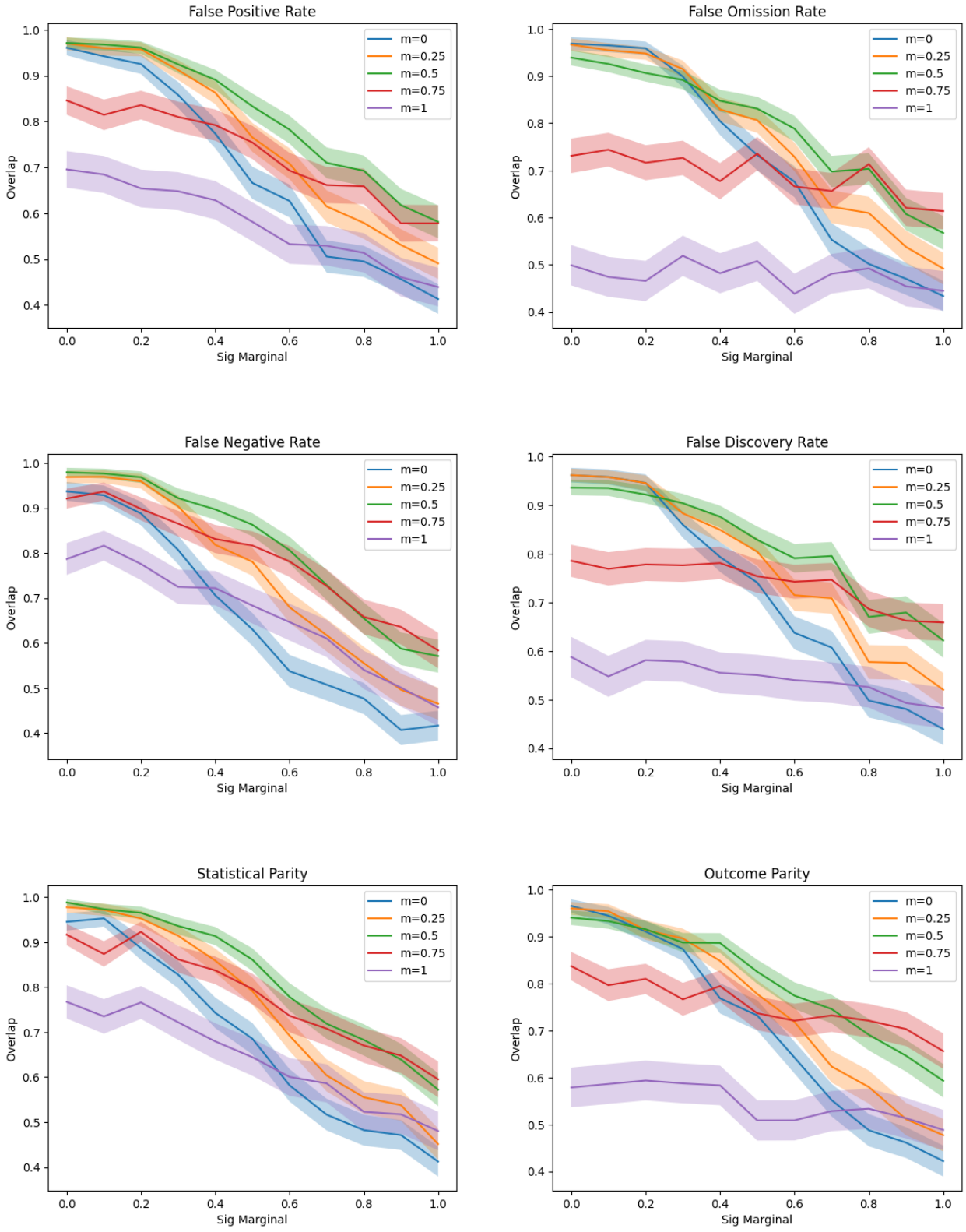

Figure 2: Detection of Each Error Type as Overlap for Increasing Marginal Bias Confounder Strength σ_{marg}

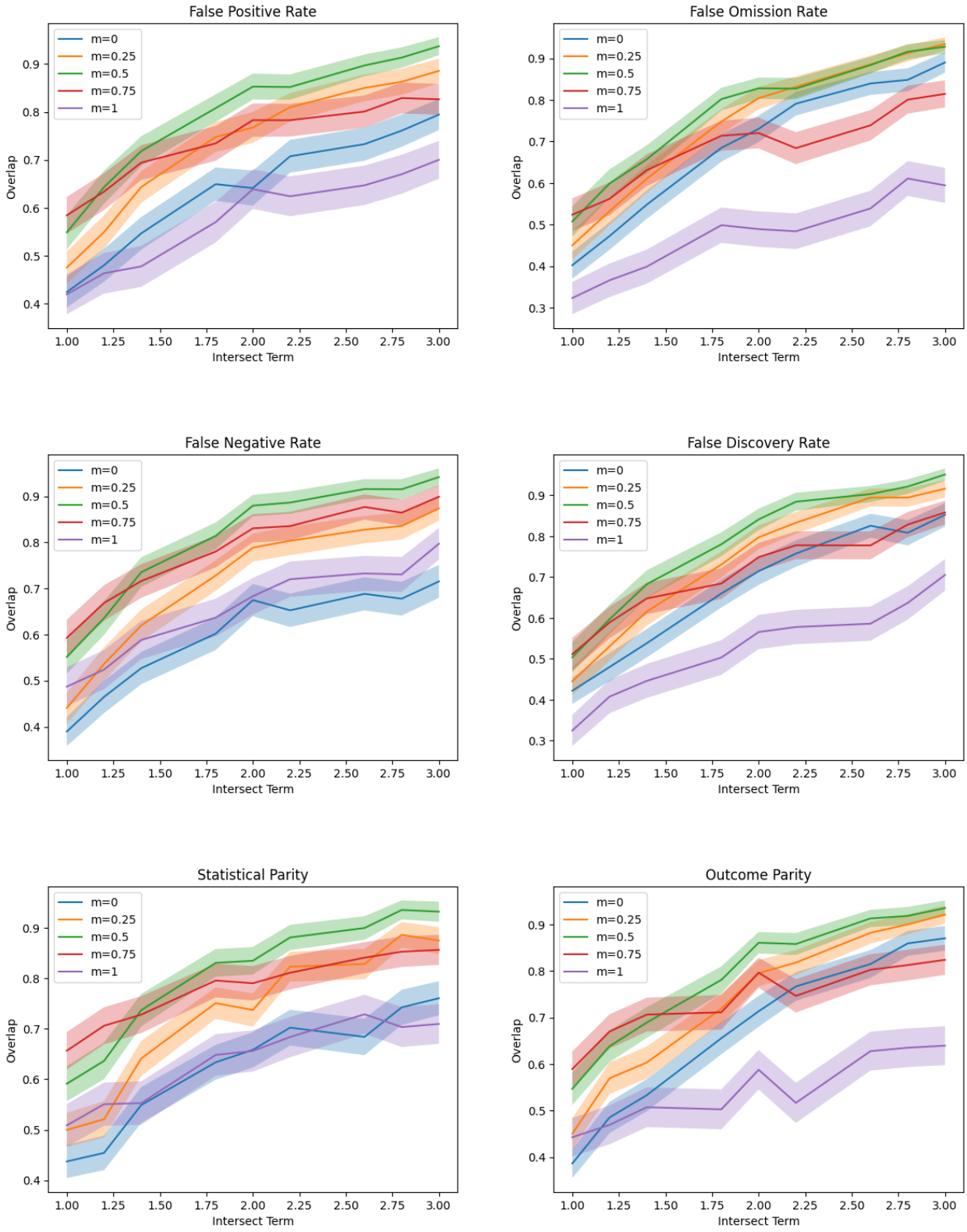

Figure 3: Detection of Each Error Type as Overlap for Increasing Intersectional Bias Signal Strength ρ

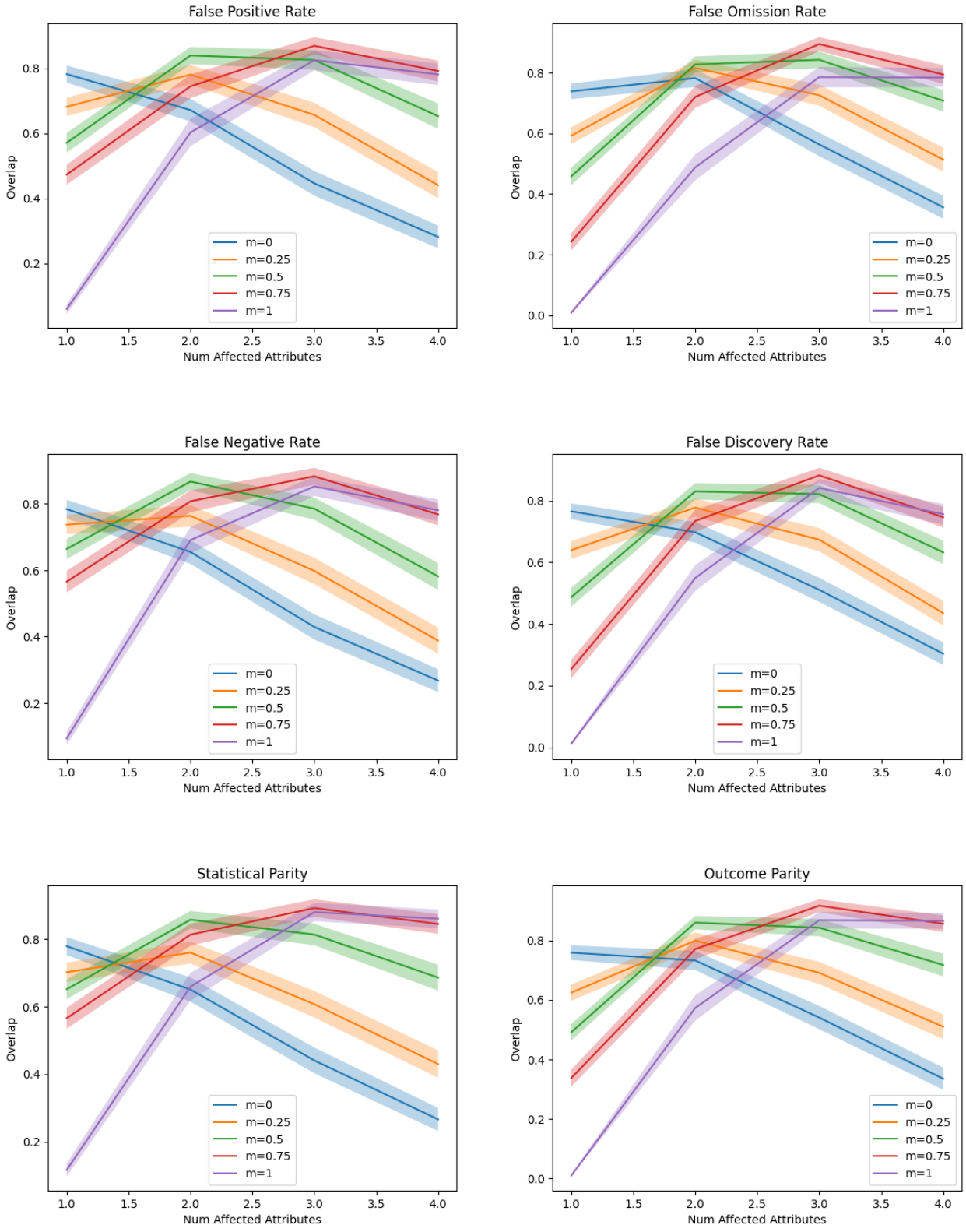

Figure 4: Detection of Each Error Type as Overlap for Increasing Number of Attributes n_{bias} Defining the Biased Subgroup

values of $m = 0.5$ are able to more accurately detect the biased subgroup. Both of these results are consistent with the intuitive understanding of how m impacts the types of biases that the scan can detect.

Finally, Figure 4 shows that larger values of m outperform smaller values of m when the number of data dimensions n_{bias} that define the affected subgroup increases to 3 or 4; while smaller values of m have the highest performance (as measured by overlap coefficient) for lower-dimensional signals. In particular, we note that $n_{bias} = 1$ corresponds to a marginal rather than intersectional bias, and thus $m = 1$ has no power to detect this bias, and as one might expect, $m = 0$ consistently does best. However, the detection performance of $m = 0$ decreases monotonically with increasing dimension n_{bias}, while detection performance for $m > 0$ tends to peak at an intermediate value, $n_{bias} = 2$ or $n_{bias} = 3$.

An interesting element of these results is the relatively poor performance of $m = 1$, as compared to intermediate values of m, even in conditions in which we imagine $m = 1$ would be the optimal setting for the scan. There are two factors at play that contribute to this diminished detection power. First, when we entirely account for marginal biases (as we do when $m = 1$) the scores computed as shown in Equation 1 are generally lower, which makes it harder for Bias Scan to find the global maximum, which often results in the scan returning no result at all. Secondly, when we use logistic regression to estimate the expected error (shown in Equation 6), the expectation is that the logistic regression is able to correctly estimate the error for all possible subgroups except the biased subgroup S_{bias}. Instead, the model accounts for the biased subgroup by overestimating the error for some subgroups and underestimating the error for others. The scan then detects the underestimated subgroup with the highest score F (computed as shown in Equation 1), which is not necessarily the subgroup of interest in the experiment.

6 COMPAS Case Study

In 2016, Julia Angwin and Jeff Larson demonstrated that Black individuals are disproportionately disadvantaged by the COMPAS system, while White individuals are disproportionately

advantaged (Angwin et al., 2016). There is a clear marginal bias in the results of this system, which we expect our method to detect when m is small in our scanning method. However, we also expect to detect different intersectional biases as we increase the value of m. We compute each defendant's predicted probability of reoffending by mapping their COMPAS risk score to the proportion of reoffending among defendants who were assigned that same risk score. Defendants with a risk score of 5 or more are considered to be "high risk", and are therefore assigned $\hat{y}_i = 1$.

For false positive, false discovery, and statistical parity scans, we scan for positive deviations (that is, a higher than expected result), which would indicate that the detected subgroup S^* is subject to more false positives or false discoveries, or that they tend to have a higher predicted score for the statistical parity scan. For false negative and false omission scans, we scan for negative deviations (that is, a lower than expected result), which would indicate that the detected subgroup S^* is subject to fewer false negatives or false omissions. All of these directions correspond to overestimation of risk, i.e., defendants who are disadvantaged by the use of COMPAS.

Figure 5 shows that different mutliplier values detect different subgroups of the COMPAS dataset for a given error scan. For values of $m = \{0, 0.25\}$, we detect some unsurprising subgroups that are subject to bias. Like Angwin and Larson, we show that Black defendants are subject to both a higher false positive rate and a lower false negative rate. (We note that for false positive scans, $m = 0.25$ or higher is needed to identify this racial bias, as the FPR scan with $m = 0$ instead picks up the subgroup of defendants with over 5 priors.) Interestingly, as we increase m, a new subgroup comes to light. Across all five fairness definitions, we see that young White women (under the age of 25) are disadvantaged compared to what one might expect considering their age, race, and gender independently: they are subject to more false positives and false discoveries, fewer false negatives and false omissions, and higher predicted scores overall (as shown by statistical parity).

Racial biases in the United States have been well documented, which makes the subgroups detected with lower values of m a predictable outcome, and all the more disappointing for its predictability. But, the results we find with $m > 0.5$ are fairly unexpected given what we already know about COMPAS. Bias against women in general is not an altogether shocking result given

an anecdotal understanding of gender bias in a patriarchal society, but the legal literature is well aligned that the opposite is true. Aside from Jennifer L. Peresie (2005), who shows that male judges exhibit bias against women in sexual harassment and sex discrimination cases, Spohn and Spears (1997) conclude that judges across the board are more lenient in their decisions for White women who have committed violent crimes when compared to Black men, Black women, and White men who have committed comparably violent crimes. This conclusion is further affirmed more recently by Geppert (2022) and Pozzulo et al. (2010). It could be that this bias is particular to COMPAS, or it could be that there really is an undocumented bias against young white women in the court system. Understanding the reasons for this result is outside of the scope of this work, but seems to merit further investigation.

7 Limitations

The method proposed here is designed to audit a classifier's predictions for biases with respect to subgroups defined by combinations of covariate attributes. It does not provide a built-in mechanism for correcting the biases it detects. This is a deliberate scoping decision rather than an oversight, but it is worth addressing directly. Combining auditing with correction and training presents a subtle risk: it can reinforce the assumption that the appropriate response to a detected subgroup bias is to retrain or post-process the model. In practice, fairness is context-specific, stakeholders may hold different and legitimate conceptions of what fairness requires, and upstream biases in data collection may make it impossible to design an optimally fair model within the existing pipeline (Black et al., 2023). The limitations of a post-hoc auditing tool should therefore be understood as an invitation to consider broader interventions, including policy-level responses, rather than simply as a call to adjust model parameters.

The multiplier m is the central mechanism by which this method controls the balance between marginal and intersectional bias detection, but its value must currently be set by the auditor prior to running the scan. There is no automated or data-driven procedure for selecting an optimal m for a given dataset or prediction context. Future work should investigate principled methods for

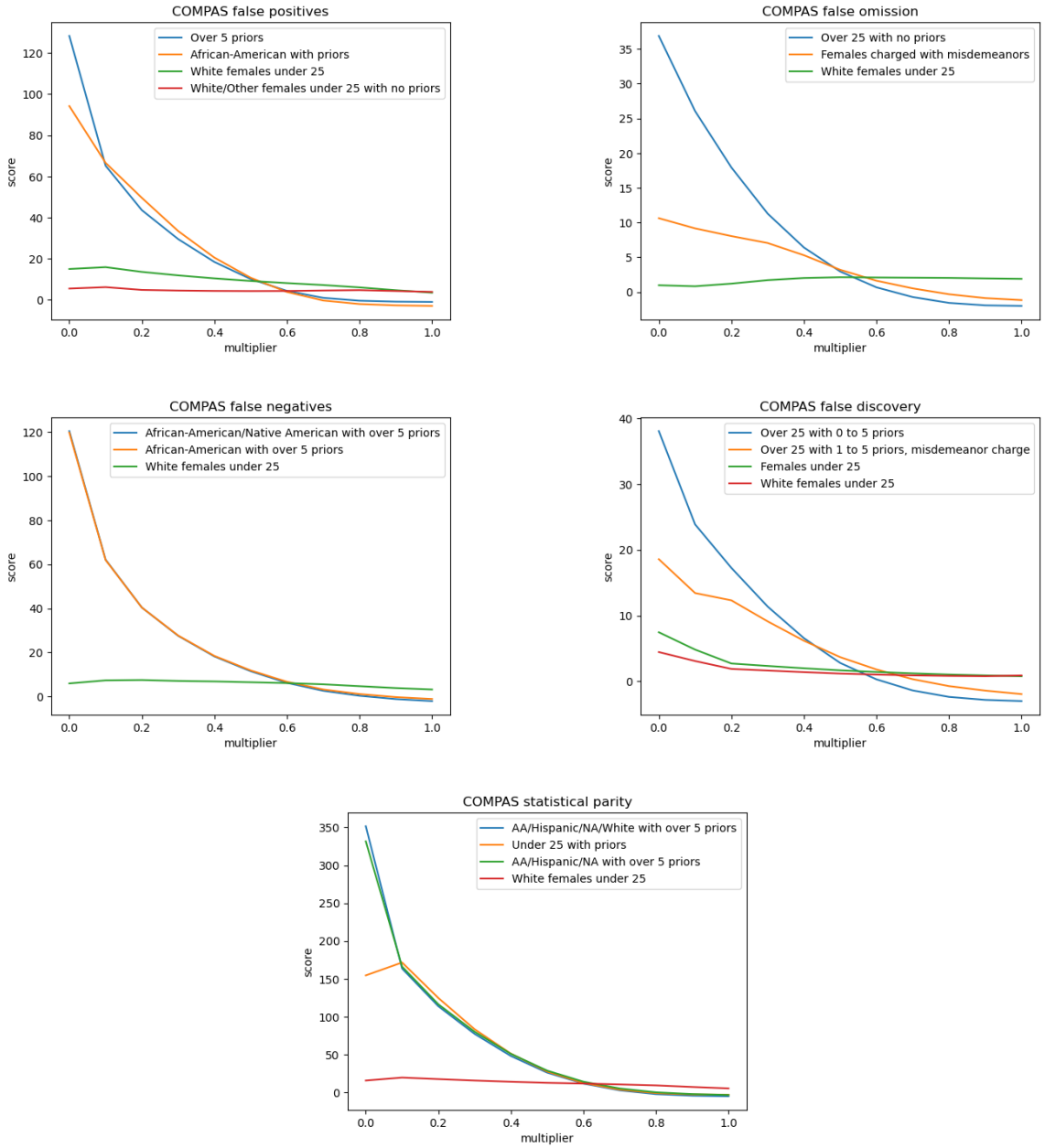

Figure 5: Subgroups Detected by Each Scan and How Their Score $F(S)$ Changes For Different Multiplier Values m

selecting m, for example through cross-validation on held-out subgroups or by deriving theoretical bounds on detection power as a function of m.

The scan procedure itself relies on a greedy coordinate ascent algorithm with multiple random restarts to identify the highest-scoring subgroup S_{bias}. As in the original bias scan (Neill and Zhang, 2016) and CBS (Boxer et al., 2023), this approach is computationally efficient due to the Additive Linear-Time Subset Scanning property of the score functions, but it does not guarantee that the globally optimal subgroup is found. In high-dimensional covariate spaces, a more severely biased subgroup may exist but go undetected. The use of permutation testing to assess statistical significance partially accounts for this, but detection power may be reduced when the true biased subgroup is small relative to the overall dataset, or when the number of attributes defining the subgroup is large (Boxer et al., 2023).

8 Conclusion

The story of COMPAS is, in many ways, the story of machine learning fairness in miniature. It was a tool built with genuine intentions – to reduce the influence of human subjectivity in a high-stakes decision-making process – that replicated the very biases it was designed to displace, and did so at scale, across hundreds of jurisdictions, for over a decade before anyone looked closely enough to notice (Angwin et al., 2016). The lesson is not that algorithmic decision-making is inherently doomed, but that good intentions are not a substitute for rigorous, ongoing evaluation. Fairness does not emerge from a model by default. It must be designed for, tested for, and continuously monitored.

This research takes up that challenge in a specific and technically grounded way. By extending the Bias Scan methodology of Neill and Zhang (2016) to support flexible detection of both marginal and intersectional subgroup biases, as well as generalizing the scan to a wide class of binary error metrics, this work addresses a meaningful gap in the existing auditing literature. The Conditional Bias Scan proposed by Boxer et al. (2023) demonstrated that intersectional bias detection was possible, but required the auditor to specify a subgroup of interest in advance; a re-

quirement that presupposes knowledge of where bias might exist before the audit has even begun. The method proposed here removes that constraint. Through the introduction of the multiplier parameter m, auditors can continuously tune the sensitivity of the scan between two poles: one that surfaces disparities attributable to individual attributes, and one that isolates the subtler, compounding harms that arise only at the intersection of multiple protected characteristics. This flexibility is not merely a technical convenience: it reflects a more honest understanding of the intersectional nature of discrimination in the real world (Buolamwini and Gebru, 2018).

That said, it would be a mistake to treat any technical auditing method as a complete solution to the problem of machine learning fairness. As the broader discussion in this thesis makes clear, bias is not a bug that appears at the prediction stage and can be cleanly patched. It accumulates across the entire machine learning pipeline (Black et al., 2023): in the historical data that encodes decades of systemic inequality, in the labeling decisions made by underpaid workers with no institutional recourse (Gray and Suri, 2017; Berg et al., 2018), in the choice of which fairness criterion to optimize for and whose interests that choice ultimately serves (Barocas et al., 2023). A post-hoc auditing method, however flexible and technically rigorous, operates at the end of that pipeline. It can identify where a deployed model is causing disproportionate harm, but it cannot undo the conditions that produced that harm in the first place.

There is also the question of what lies beyond the model entirely. The environmental cost of training and deploying machine learning systems at scale falls disproportionately on the communities least equipped to absorb it (Luccioni et al., 2024; World Health Organization, 2023). The data that fuels these systems is extracted, more often than not, without the meaningful consent of the people it represents (Crawford, 2021). These are fairness problems too, and they resist the kind of mathematical formalization that makes group fairness criteria so tractable. Acknowledging them is not a digression from the technical work, but a reminder that the technical work exists within a broader social context.

Looking forward, there are several promising directions for extending this research. The multiplier framework introduced here could be integrated into existing open-source fairness audit-

ing libraries, lowering the barrier to adoption for practitioners who may not have the background to implement it from scratch. The method could also be evaluated across a wider range of datasets and prediction domains (such as healthcare, lending, and hiring) where the structure of intersectional bias may differ substantially from the criminal justice context in which bias scan was originally developed (Neill and Zhang, 2016). More theoretically, future work might investigate whether the multiplier parameter can be chosen in a principled, data-driven way (rather than requiring the auditor to set it manually), and what guarantees can be offered about the statistical power of the scan under different choices of m.

Ultimately, the pursuit of fairness in machine learning is not a problem that will be solved once and declared finished. It is an ongoing practice that requires technical tools sharp enough to find what we are looking for, and enough intellectual honesty to keep asking whether we are looking in the right places. This research is one contribution to a collective effort that is, by necessity, far larger than any single method.

Acknowledgment of Funding

This work was partially supported by the NSF Program on Fairness in Artificial Intelligence in Collaboration with Amazon, grant IIS-2040898. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or Amazon.

References

Andrus, M., Spitzer, E., Brown, J., and Xiang, A. (2021). What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 249–260. Association for Computing Machinery.

Angwin, J., Kirchner, L., Larson, J., and Mattu, S. (2016). Machine bias. *ProPublica*.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Berg, J., Furrer, M., Harmon, E., Rani, U., and Silberman, M. S. (2018). Digital labour platforms and the future of work: Towards decent work in the online world.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 149–159.

Black, E., Naidu, R., Ghani, R., Rodolfa, K. T., Ho, D. E., and Heidari, H. (2023). Toward operationalizing pipeline-aware ml fairness: A research agenda for developing practical guidelines and tools.

Blackman, R. (2022). *Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI*. Harvard Business Review Press.

Boxer, A. et al. (2023). Auditing predictive models for intersectional biases. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91.

Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company.

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Dastin, J. (2018). Insight—amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*, pages 214–226. Association for Computing Machinery.

Editorial Board (2015). A chance to fix parole in new york.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Geppert, K. (2022). Explaining the gender gap in the criminal justice system: How family-based gender roles shape perceptions of defendants in criminal court. *Inquiries Journal*, 14(02).

Gray, M. L. and Suri, S. (2017). The humans working behind the AI curtain. *Harvard Business Review*.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2012*, volume 7524 of *Lecture Notes in Computer Science*, pages 35–50. Springer.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. ProPublica. Accessed: 2026-04-06.

Luccioni, A. S., Jernite, Y., and Strubell, E. (2024). Power hungry processing: Watts driving the

cost of AI deployment. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, pages 85–99. Association for Computing Machinery.

Lum, K. and Isaac, W. (2016). To predict and serve? *Significance*, 13(5):14–19.

Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. pages 111–125.

Neill, D. B. and Zhang, Z. (2016). Identifying significant predictive bias in classifiers. NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems.

Peresie, J. L. (2005). Female judges matter: Gender and collegial decisionmaking in the federal appellate courts. *Yale Law Journal*, 114(7):1759–1790.

Pozzulo, J. D., Dempsey, J., Maeder, E., and Allen, L. (2010). The effects of victim gender, defendant gender, and defendant age on juror decision making. *Criminal Justice and Behavior*, 37(1).

Prince, A. E. R. and Schwarcz, D. (2020). Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review*, 105:1257–1318.

Samuelson, P. (2023). Generative AI meets copyright. *Science*, 381(6654):158–161.

Schaich Borg, J., Sinnott-Armstrong, W., and Conitzer, V. (2024). *Moral AI: And How We Get There*. Pelican Books.

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., and Thompson, N. (2024). The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence.

Spohn, C. C. and Spears, J. W. (1997). Gender and case processing decisions: A comparison of case outcomes for male and female defendants charged with violent felonies. *Women & Criminal Justice*, 8(3):29–59.

Stanford Institute for Human-Centered Artificial Intelligence (2025). Economy (Chapter 4) in the 2025 AI index report.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650. Association for Computational Linguistics.

Tolentino, J. (2019). *Trick Mirror: Reflections on Self-Delusion*. Random House.

Weerts, H. et al. (2023). Fairlearn: Assessing and improving fairness of AI systems.

World Health Organization (2023). Climate change and health (fact sheet).

Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M. J., Defazio, A., Stern, R., Johnson, P., Bruno, M., Parente, M., Geras, K. J., Katsnelson, J., Chandarana, H., Zhang, Z., Drozdzal, M., Romero, A., Rabbat, M., Vincent, P., Yakubova, N., Pinkerton, J., Wang, D., Owens, E., Zitnick, C. L., Recht, M. P., Sodickson, D. K., and Lui, Y. W. (2018). fastMRI: An open dataset and benchmarks for accelerated MRI.

Zhai, C., Wibowo, S., and Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, 11(1):28.

Appendix

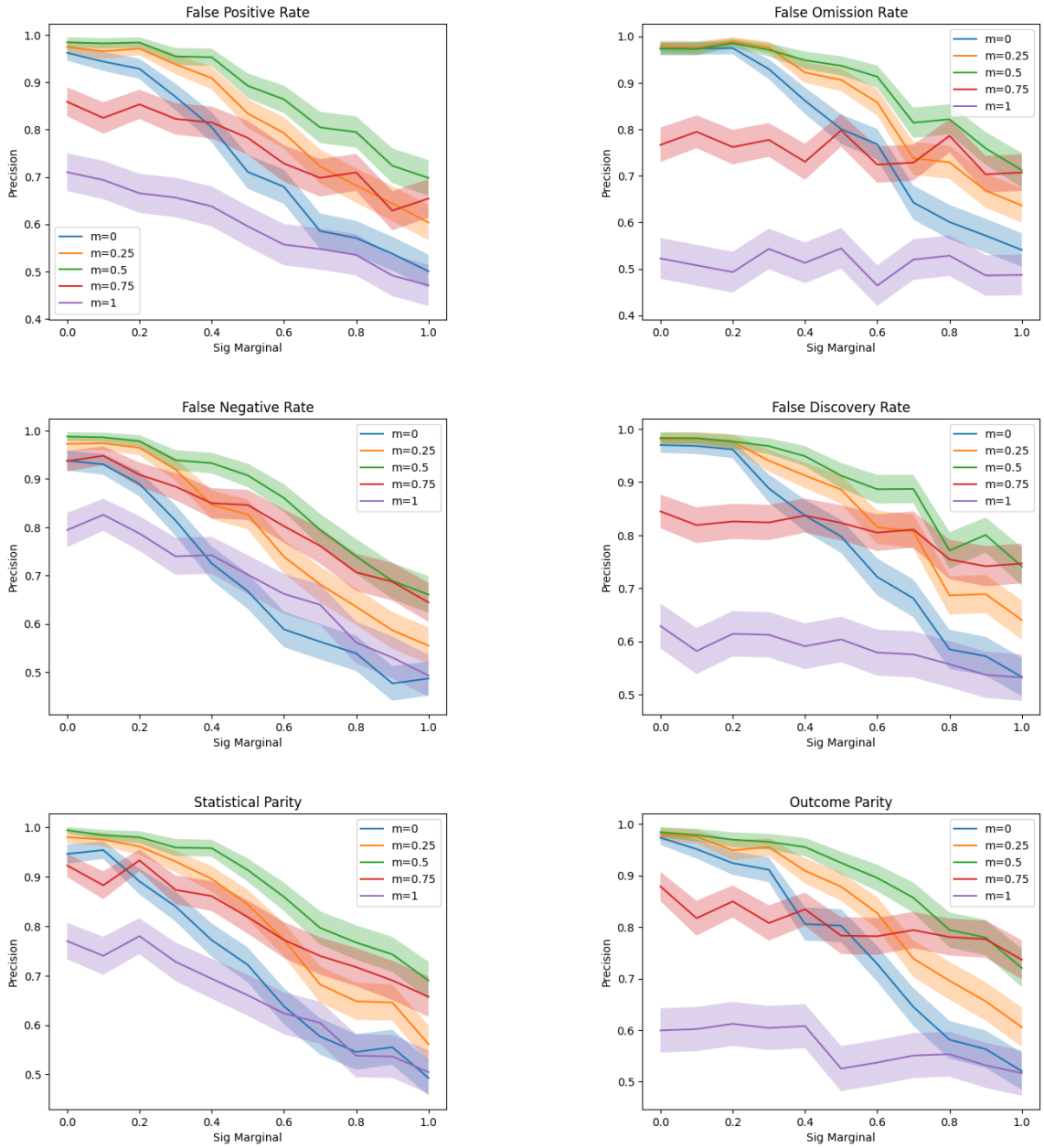

Figure 6: Detection of Each Error Type as Precision for Increasing Marginal Bias Confounder Strength σ_{marg}

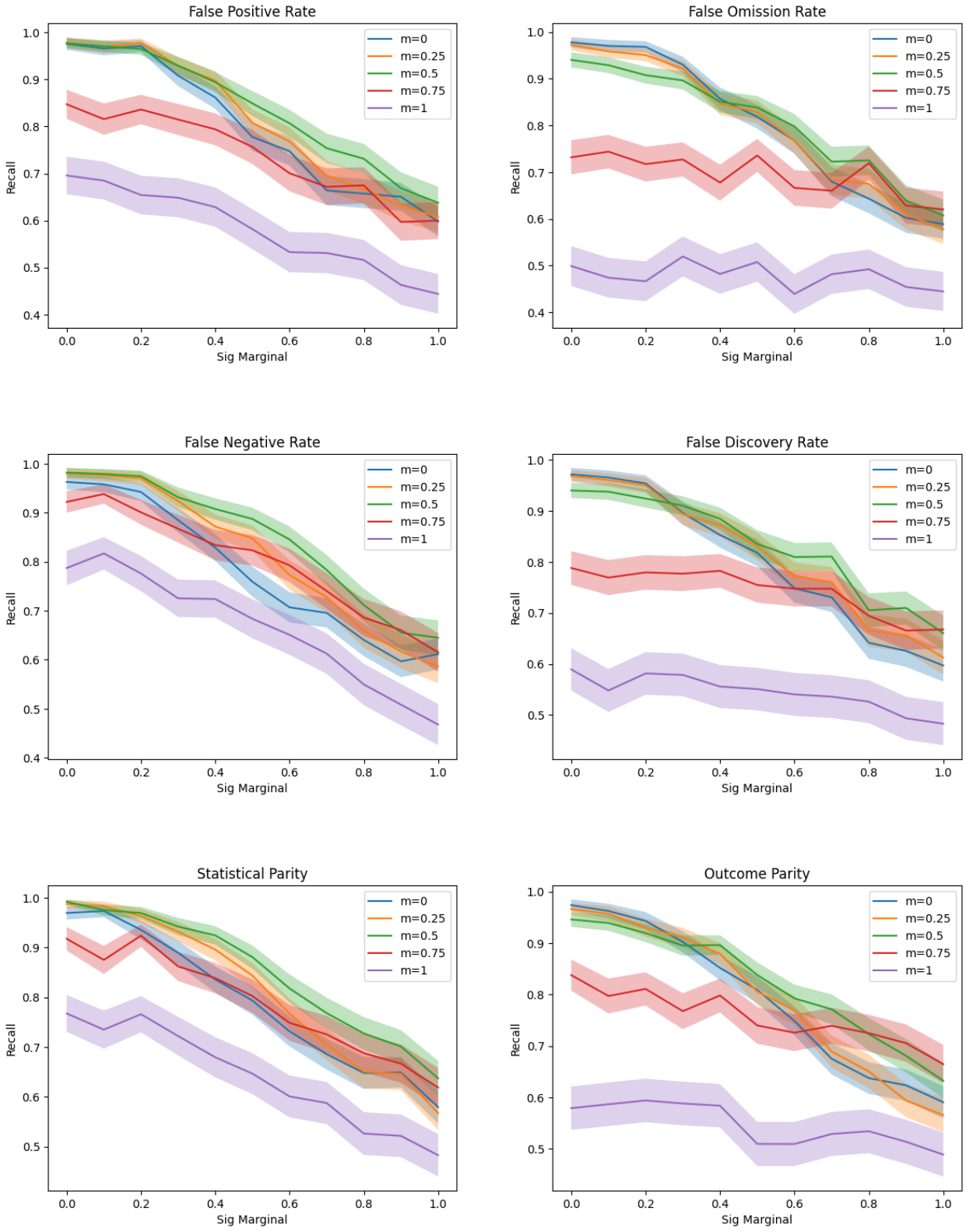

Figure 7: Detection of Each Error Type as Recall for Increasing Marginal Bias Confounder Strength σ_{marg}

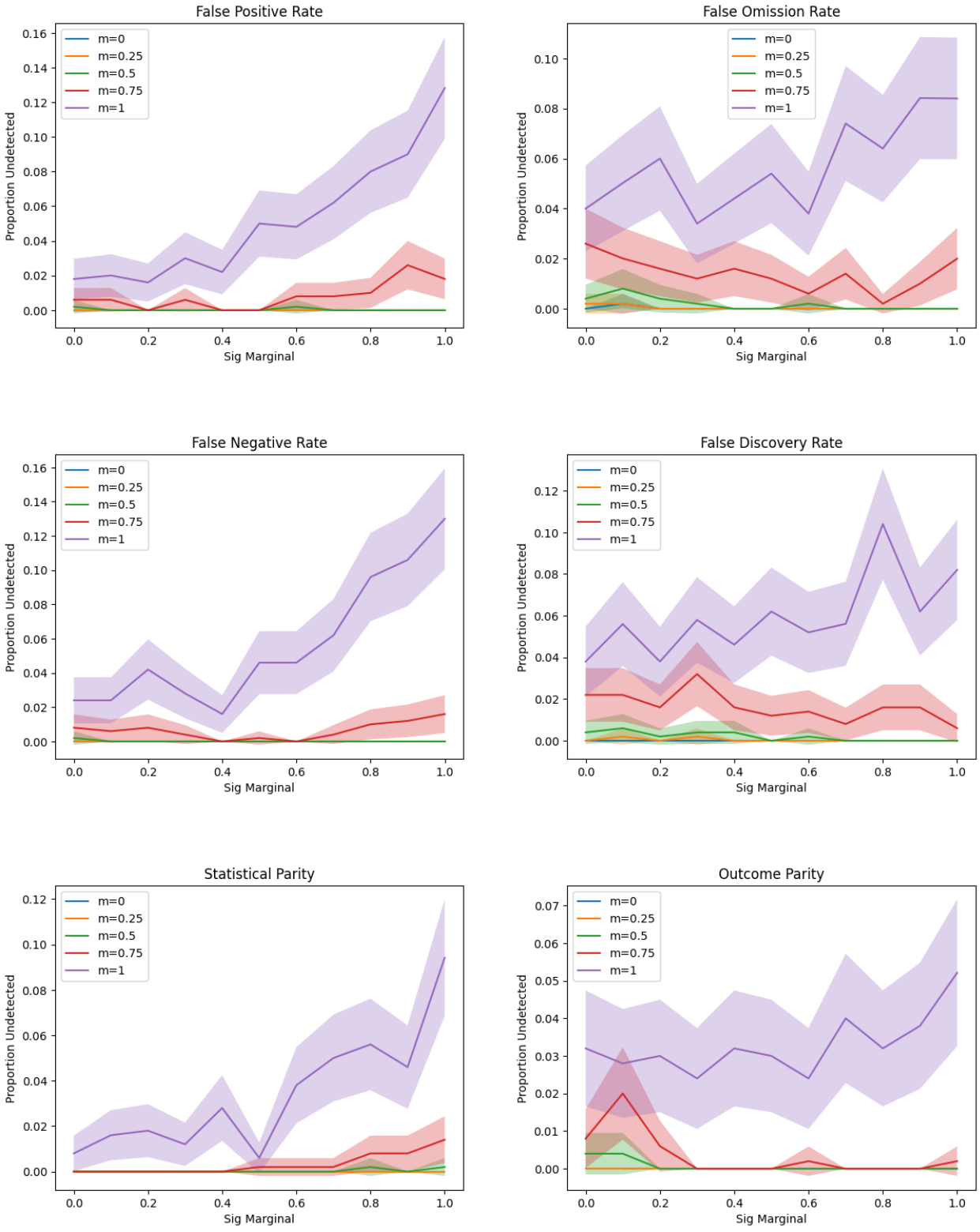

Figure 8: Proportion of Biased Subgroups Not Detected for Increasing Marginal Bias Confounder Strength σ_{marg}

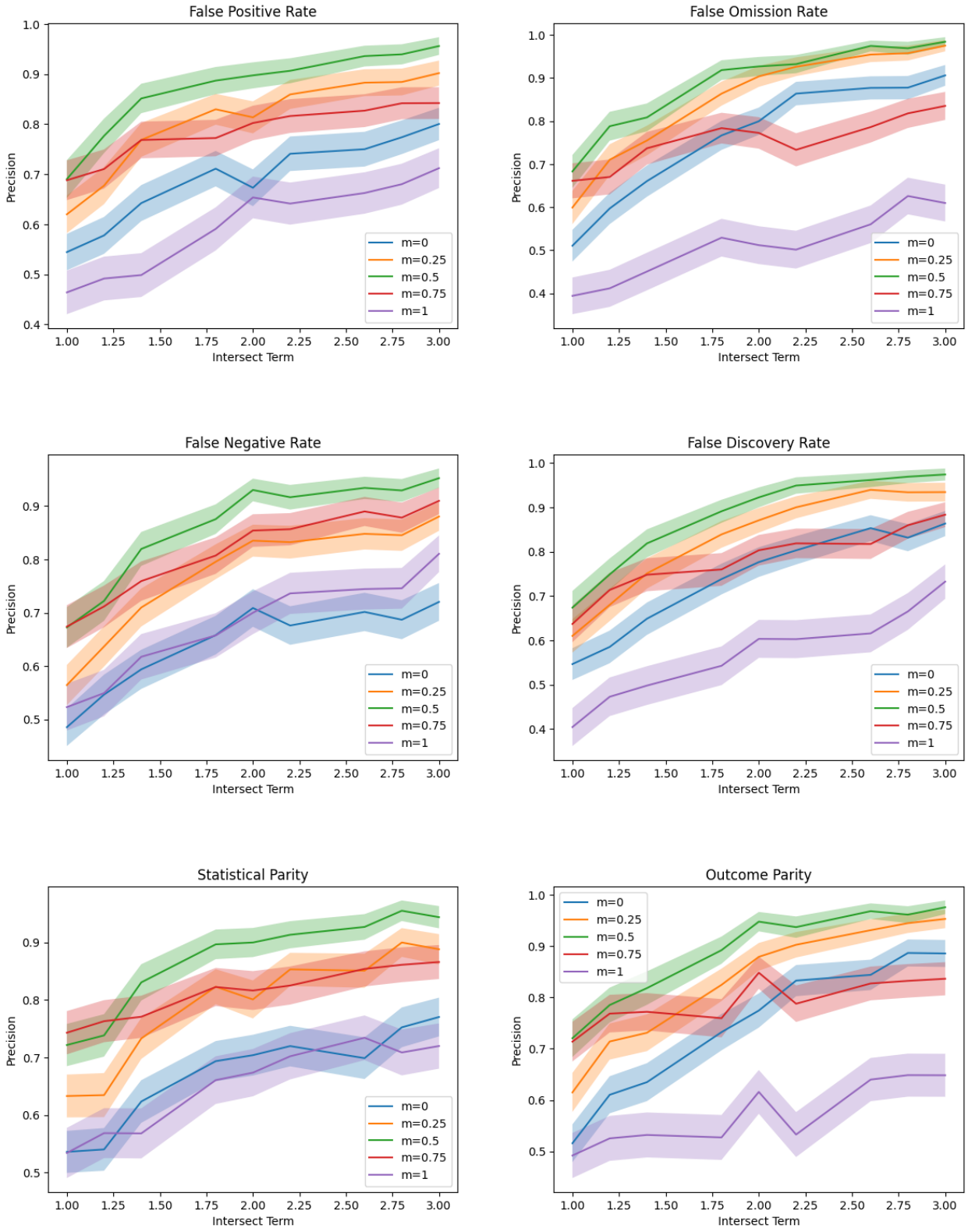

Figure 9: Detection of Each Error Type as Precision for Increasing Intersectional Bias Signal Strength ρ

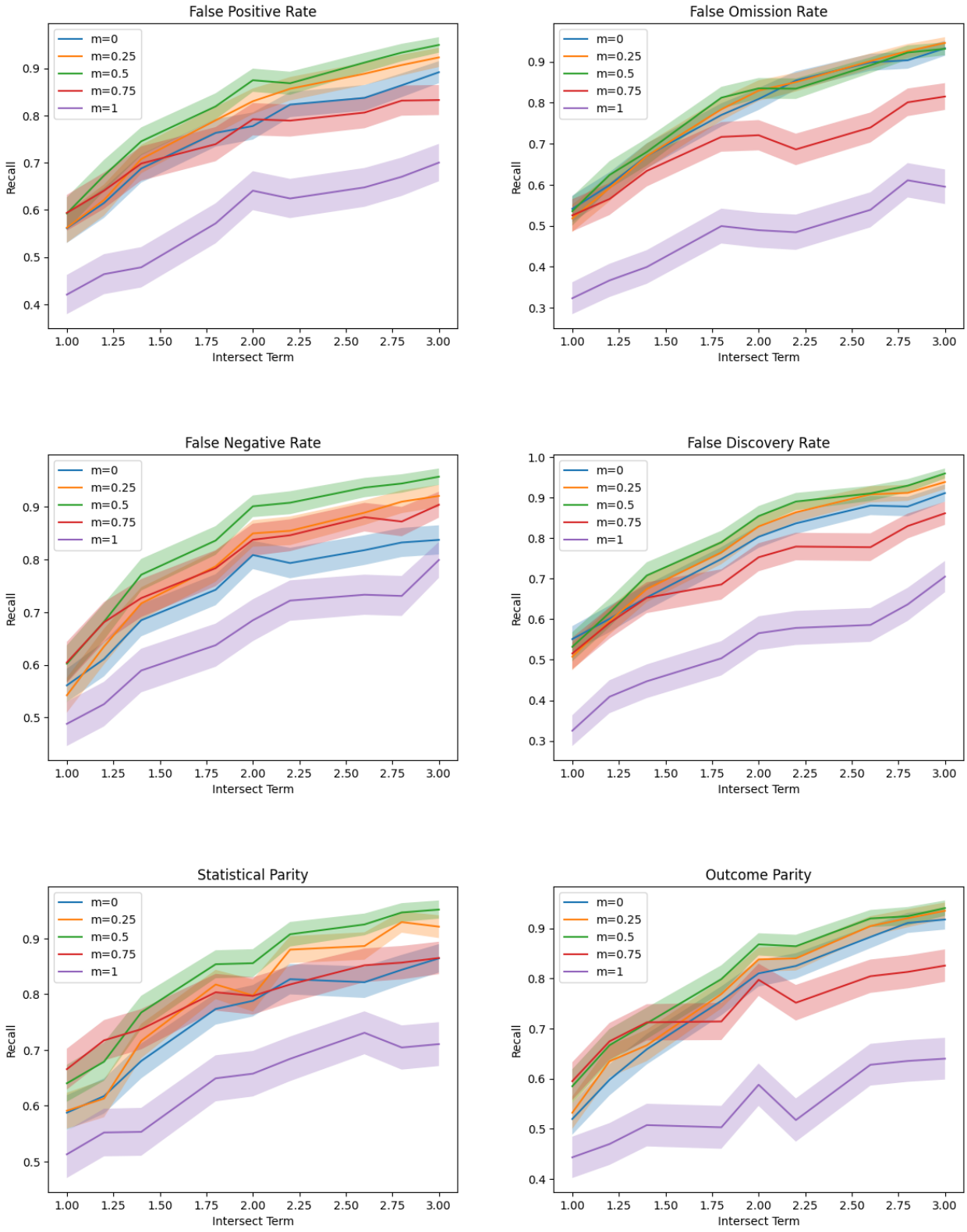

Figure 10: Detection of Each Error Type as Recall for Increasing Intersectional Bias Signal Strength ρ

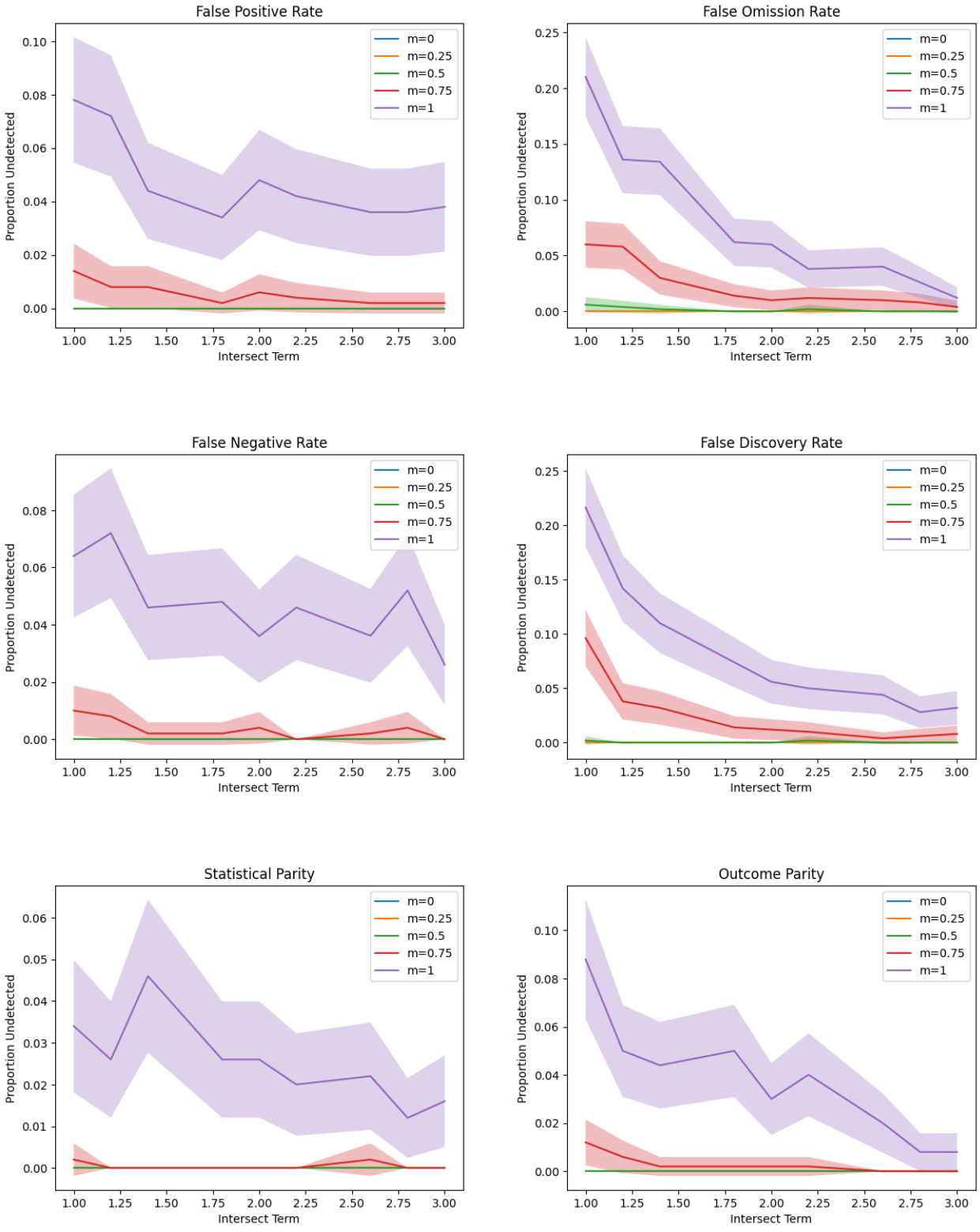

Figure 11: Proportion of Biased Subgroups Not Detected for Increasing Intersectional Bias Signal Strength ρ

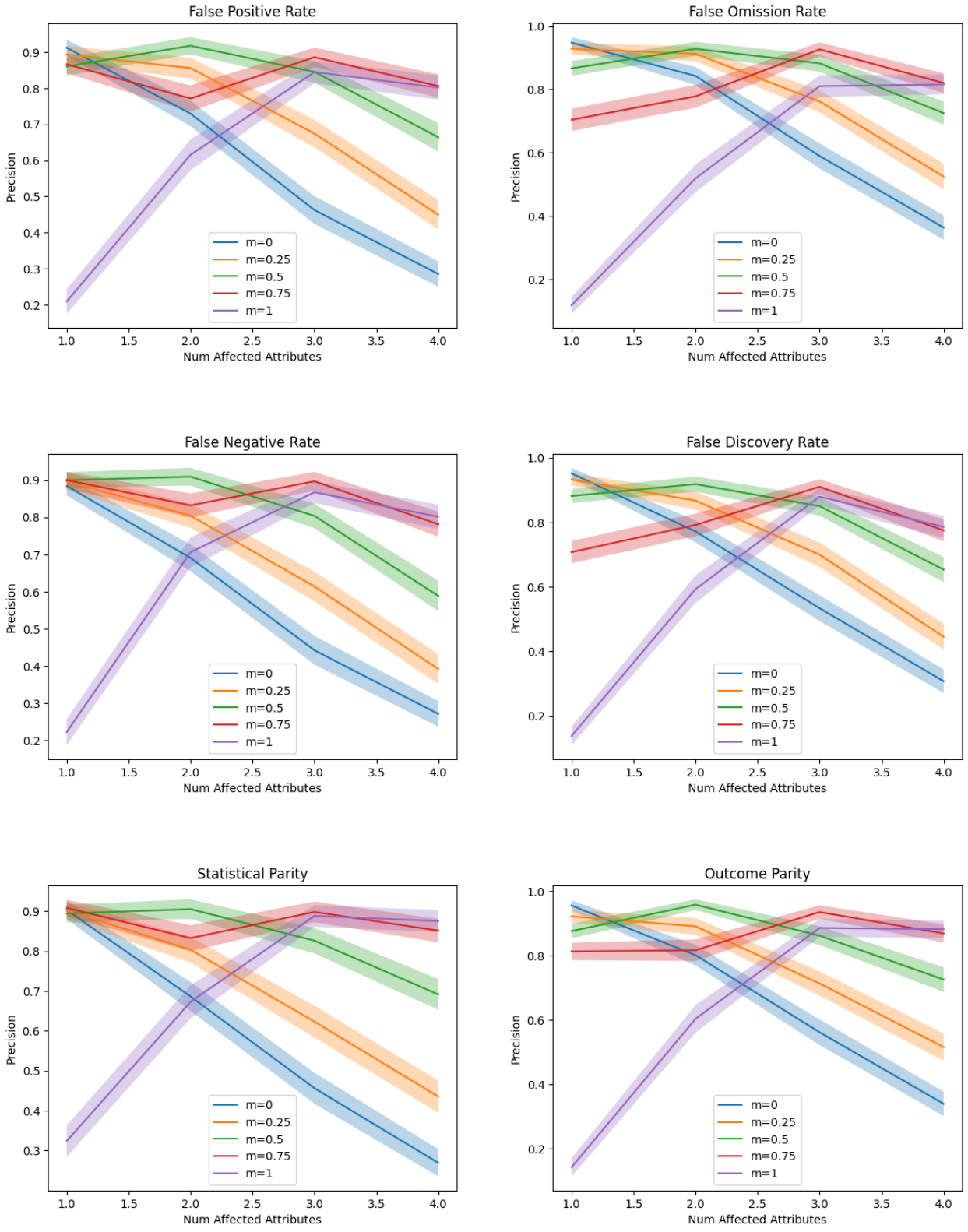

Figure 12: Detection of Each Error Type as Precision for Increasing Number of Attributes n_{bias} Defining the Biased Subgroup

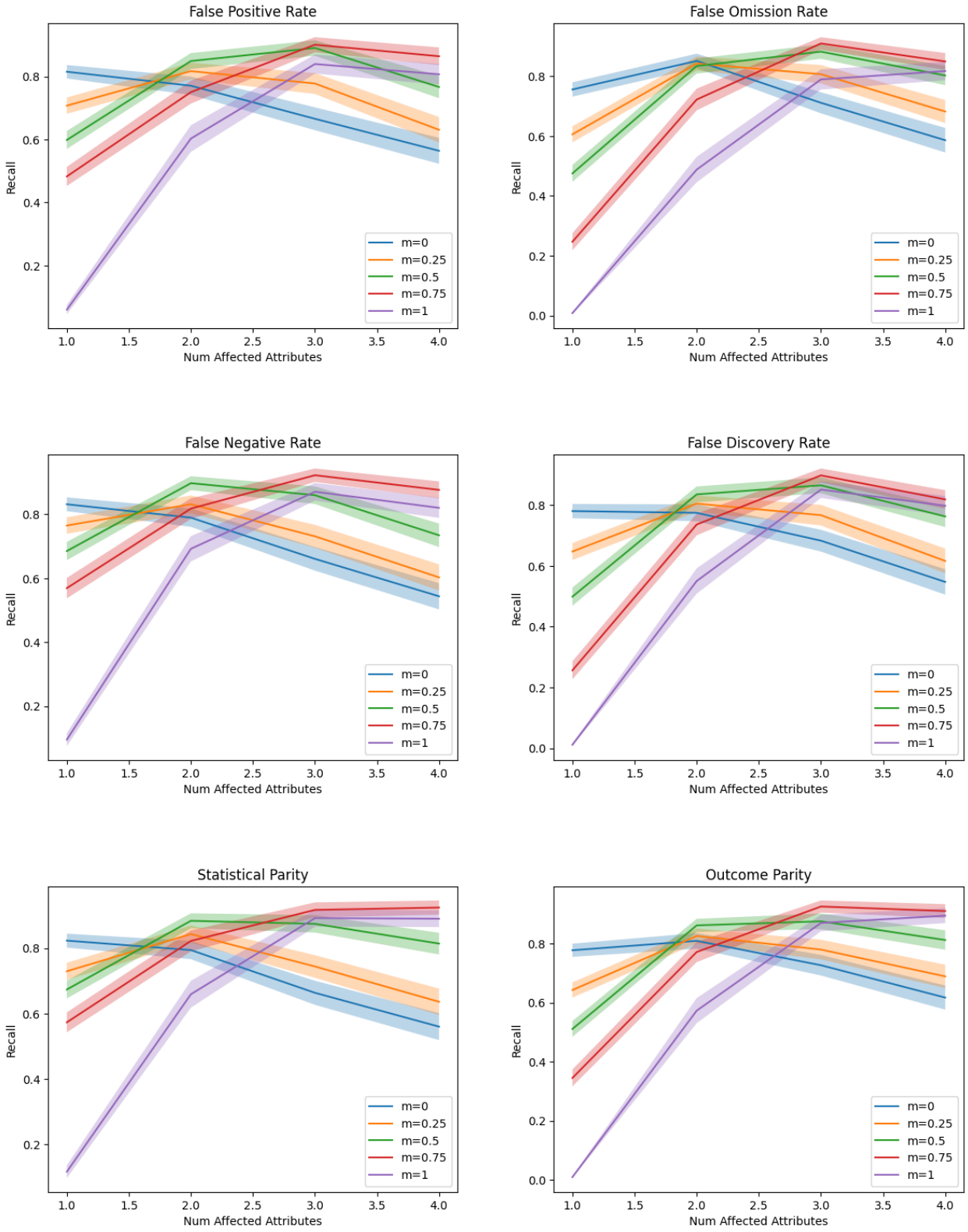

Figure 13: Detection of Each Error Type as Recall for Increasing Number of Attributes n_{bias} Defining the Biased Subgroup

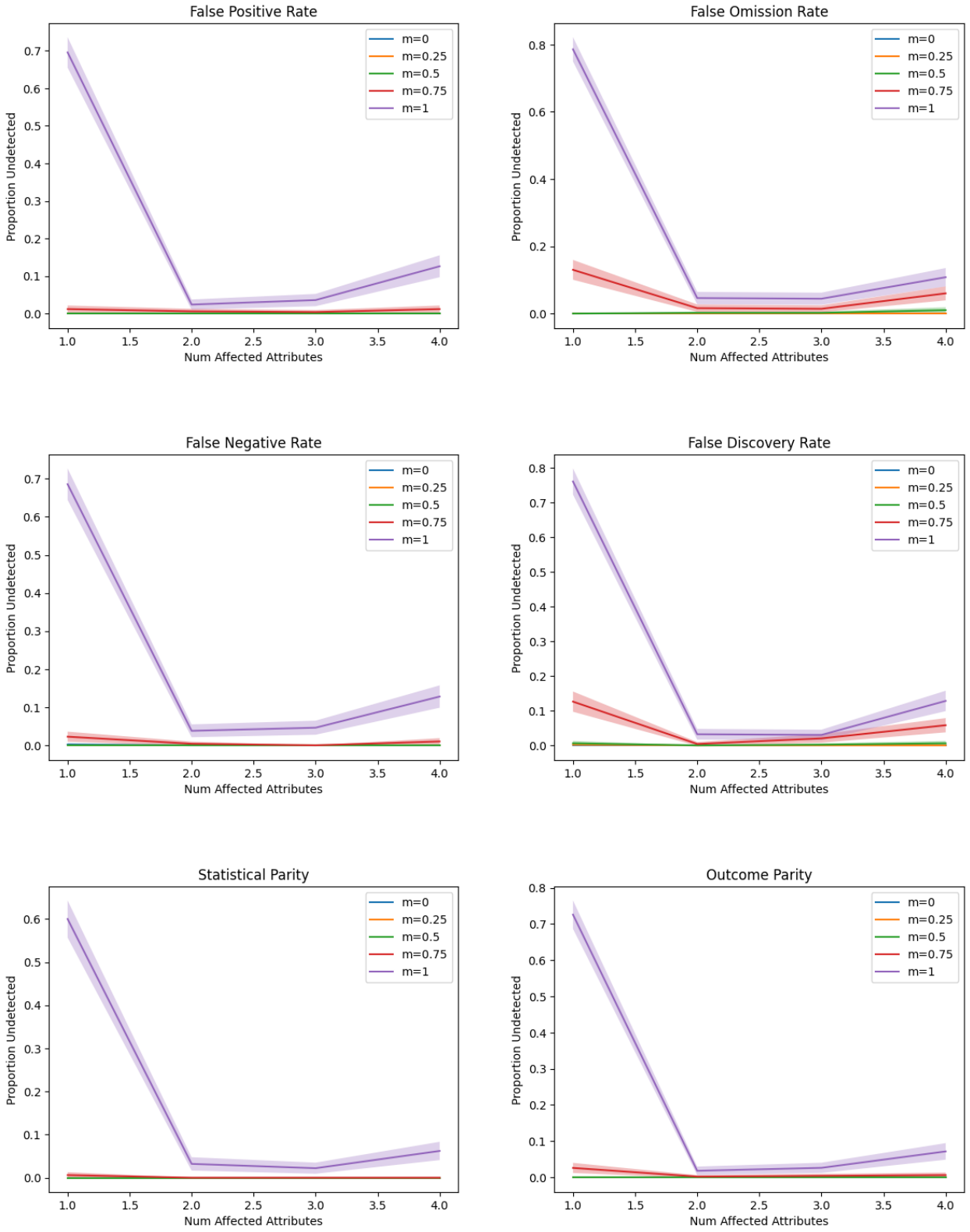

Figure 14: Proportion of Biased Subgroups Not Detected for Increasing Number of Attributes n_{bias} Defining the Biased Subgroup

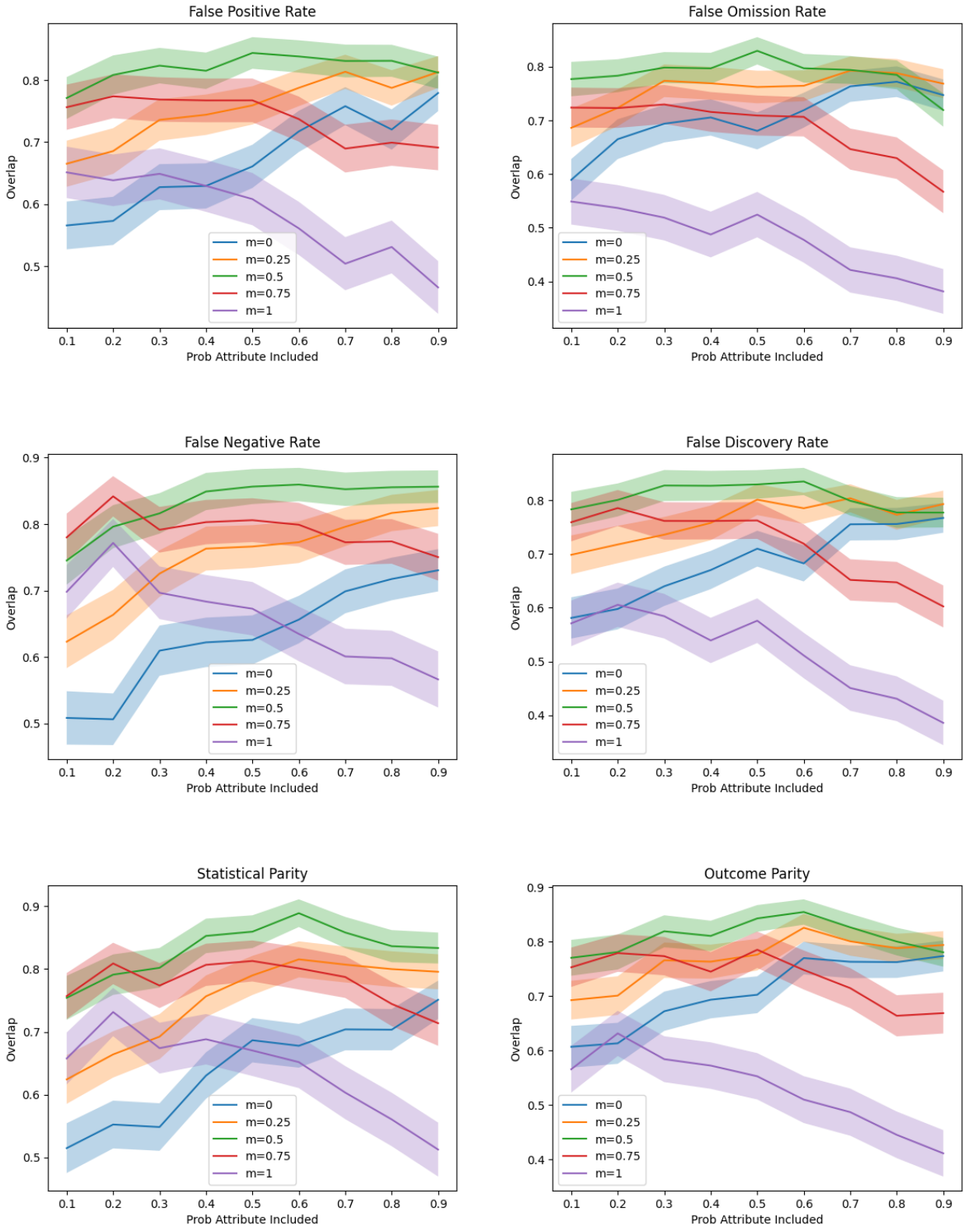

Figure 15: Detection of Each Error Type as Overlap for Increasing Probability of an Attribute Value Being Included in the Biased Subgroup p_{bias}

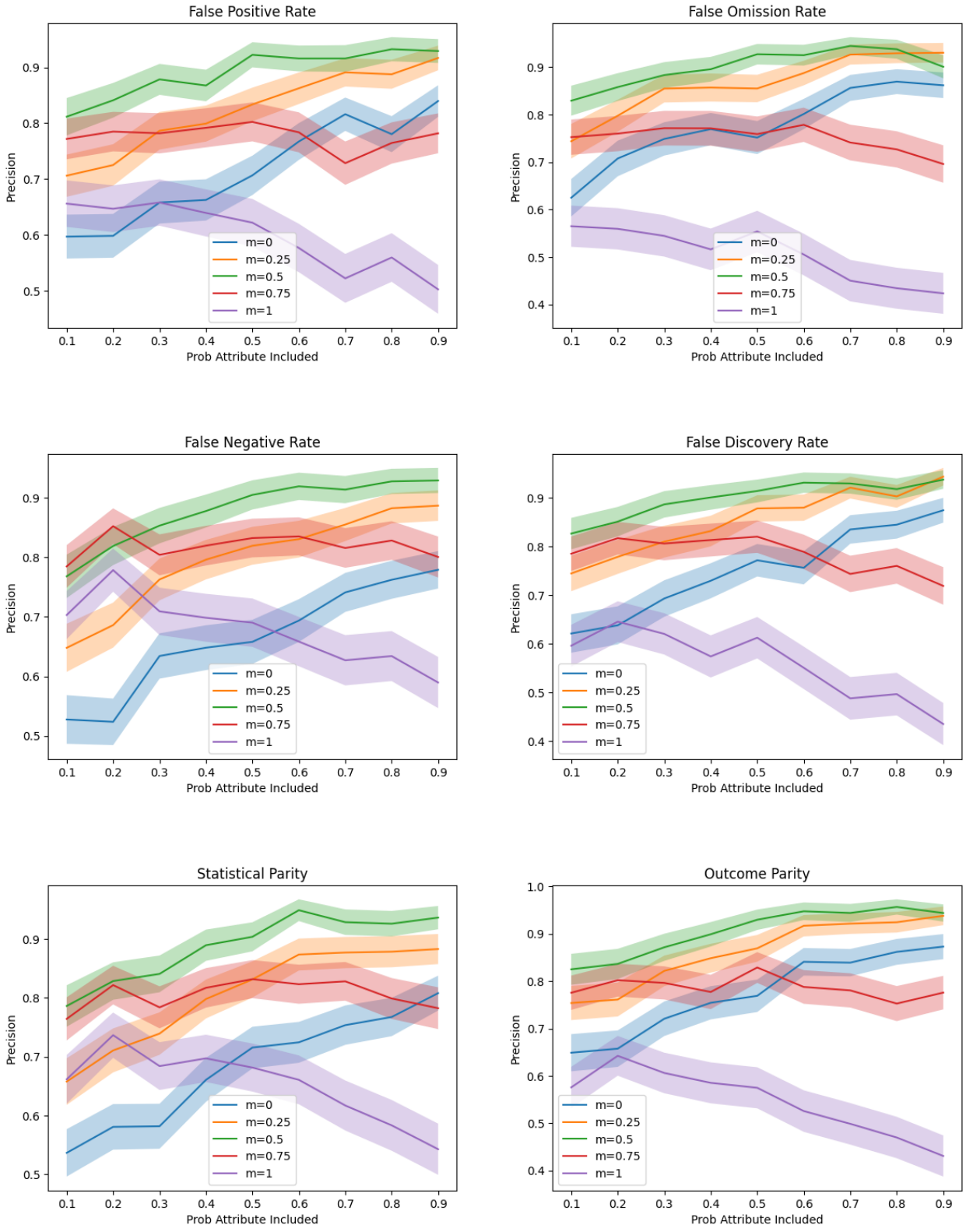

Figure 16: Detection of Each Error Type as Precision for Increasing Probability of an Attribute Value Being Included in the Biased Subgroup p_{bias}

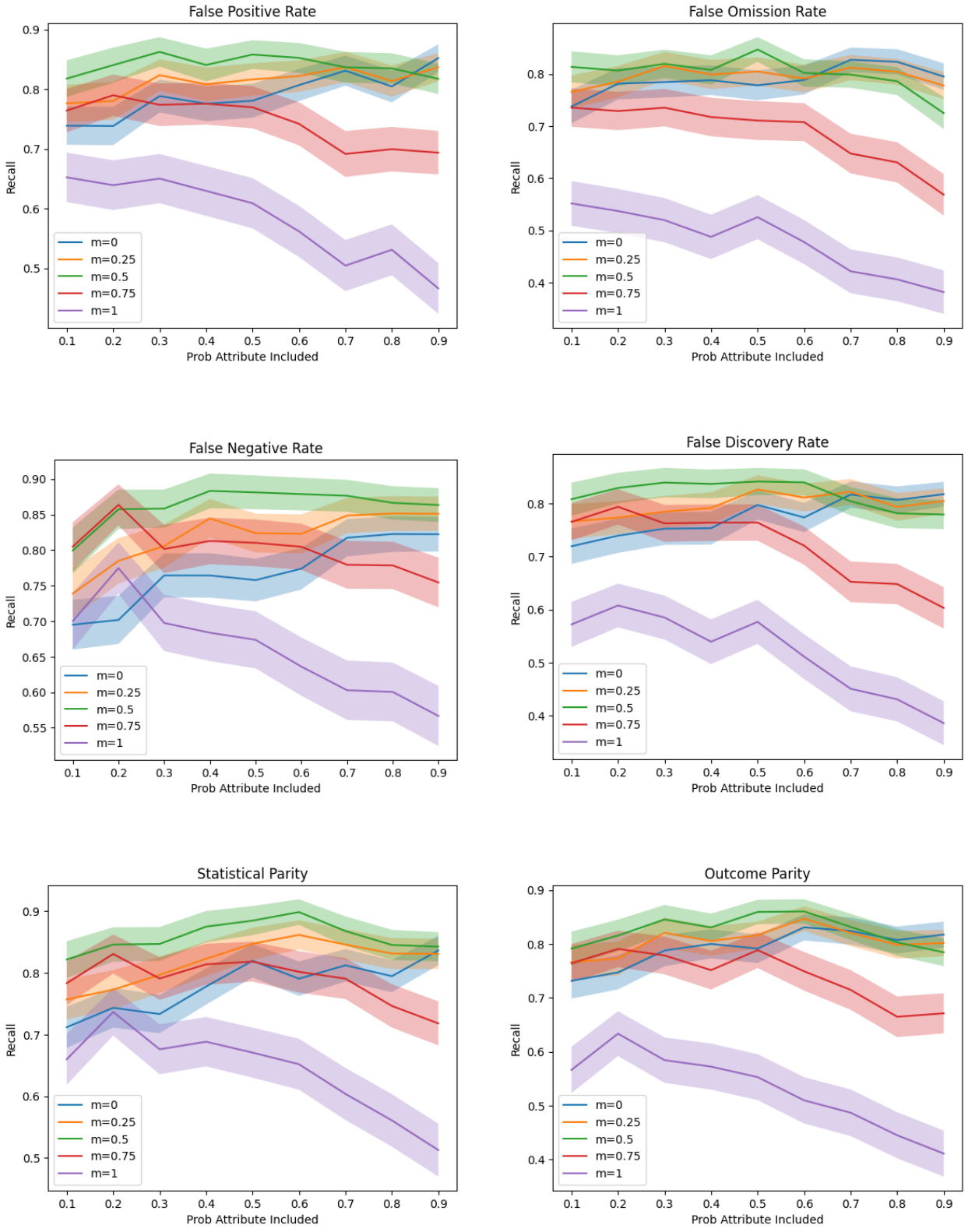

Figure 17: Detection of Each Error Type as Recall for Increasing Probability of an Attribute Value Being Included in the Biased Subgroup p_{bias}

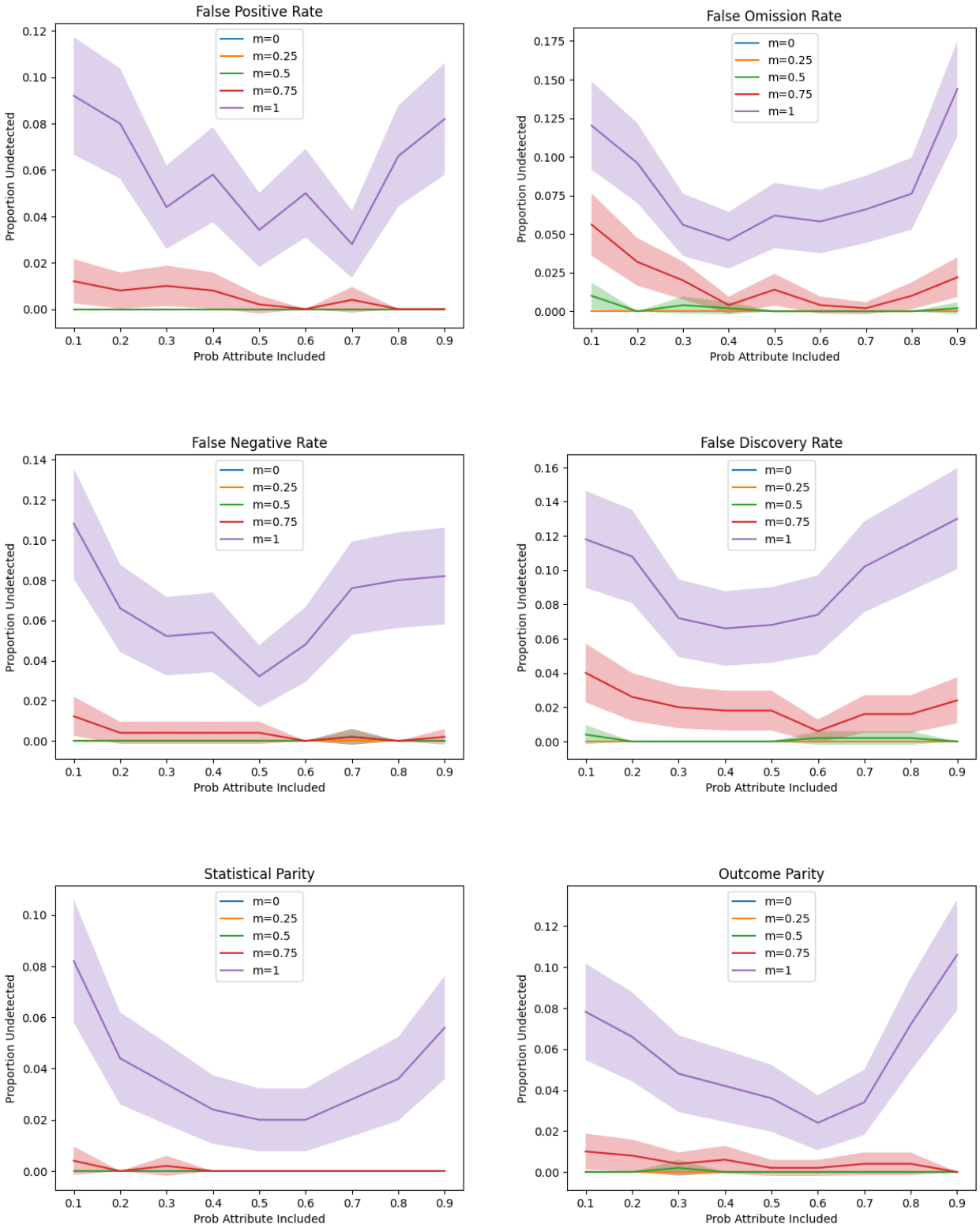

Figure 18: Proportion of Biased Subgroups Not Detected for Increasing Probability of an Attribute Value Being Included in the Biased Subgroup p_{bias}

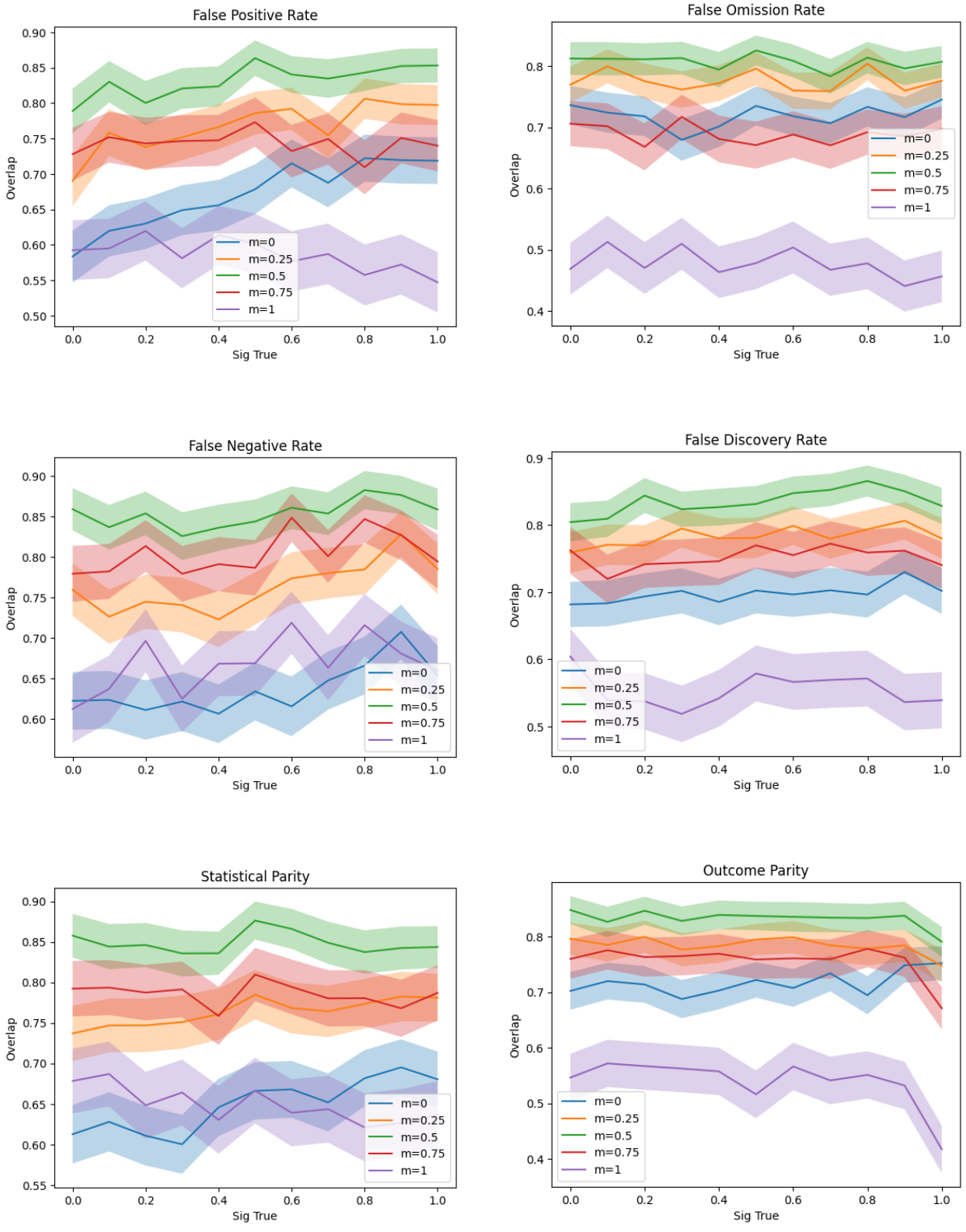

Figure 19: Detection of Each Error Type as Overlap for Increasing Individual-Level Variability in True Outcome Probability σ_{true}

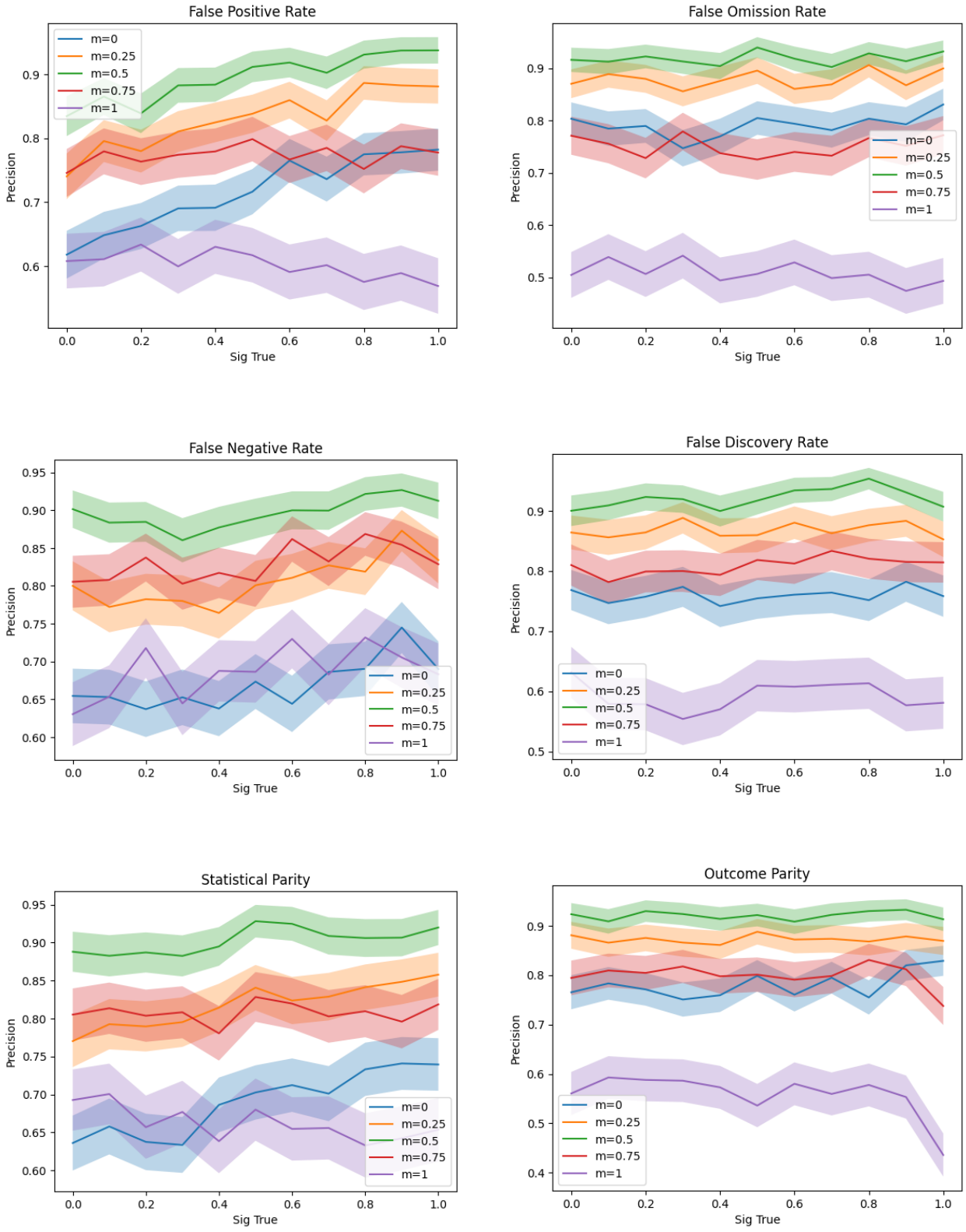

Figure 20: Detection of Each Error Type as Precision for Increasing Individual-Level Variability in True Outcome Probability σ_{true}

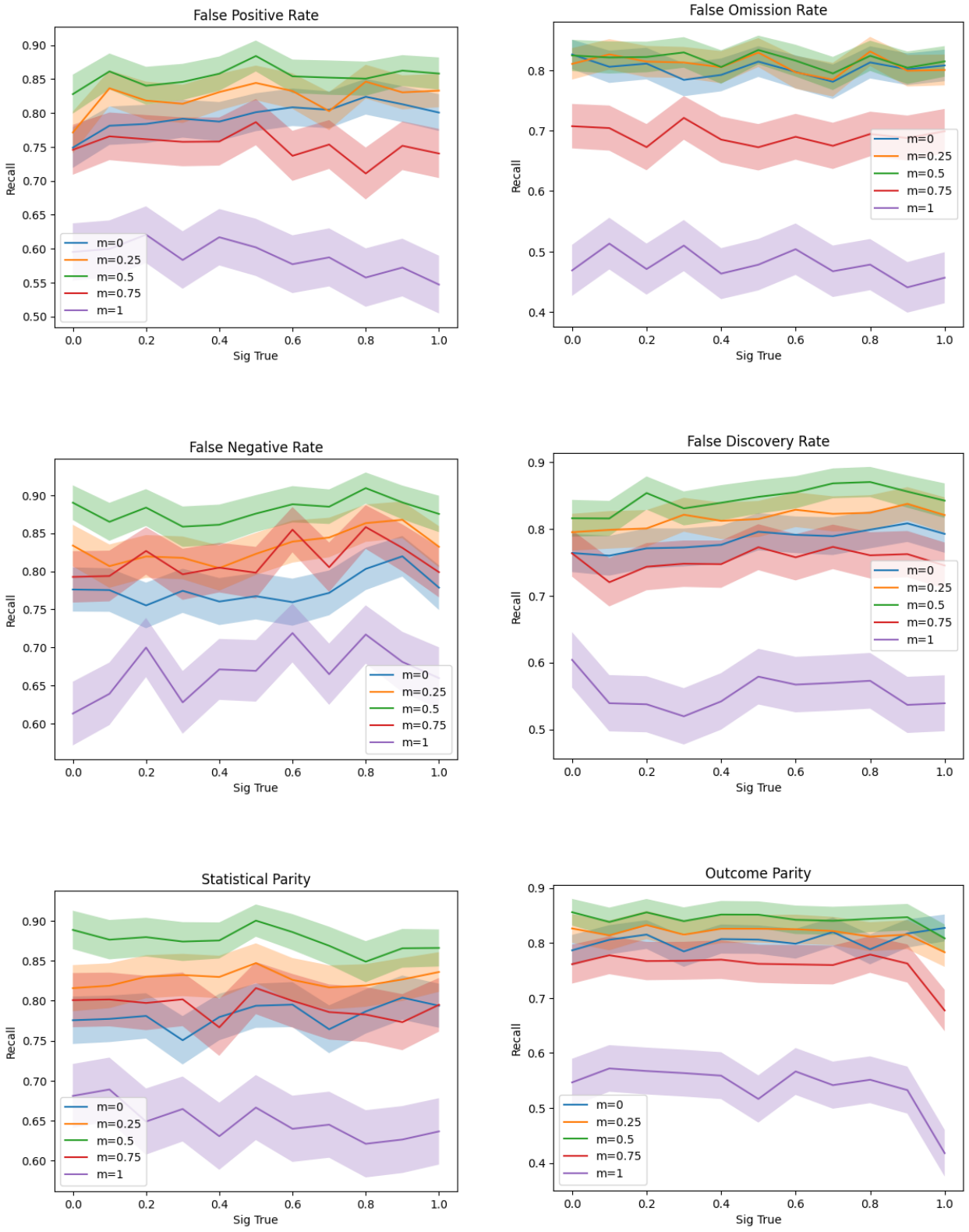

Figure 21: Detection of Each Error Type as Recall for Increasing Individual-Level Variability in True Outcome Probability σ_{true}

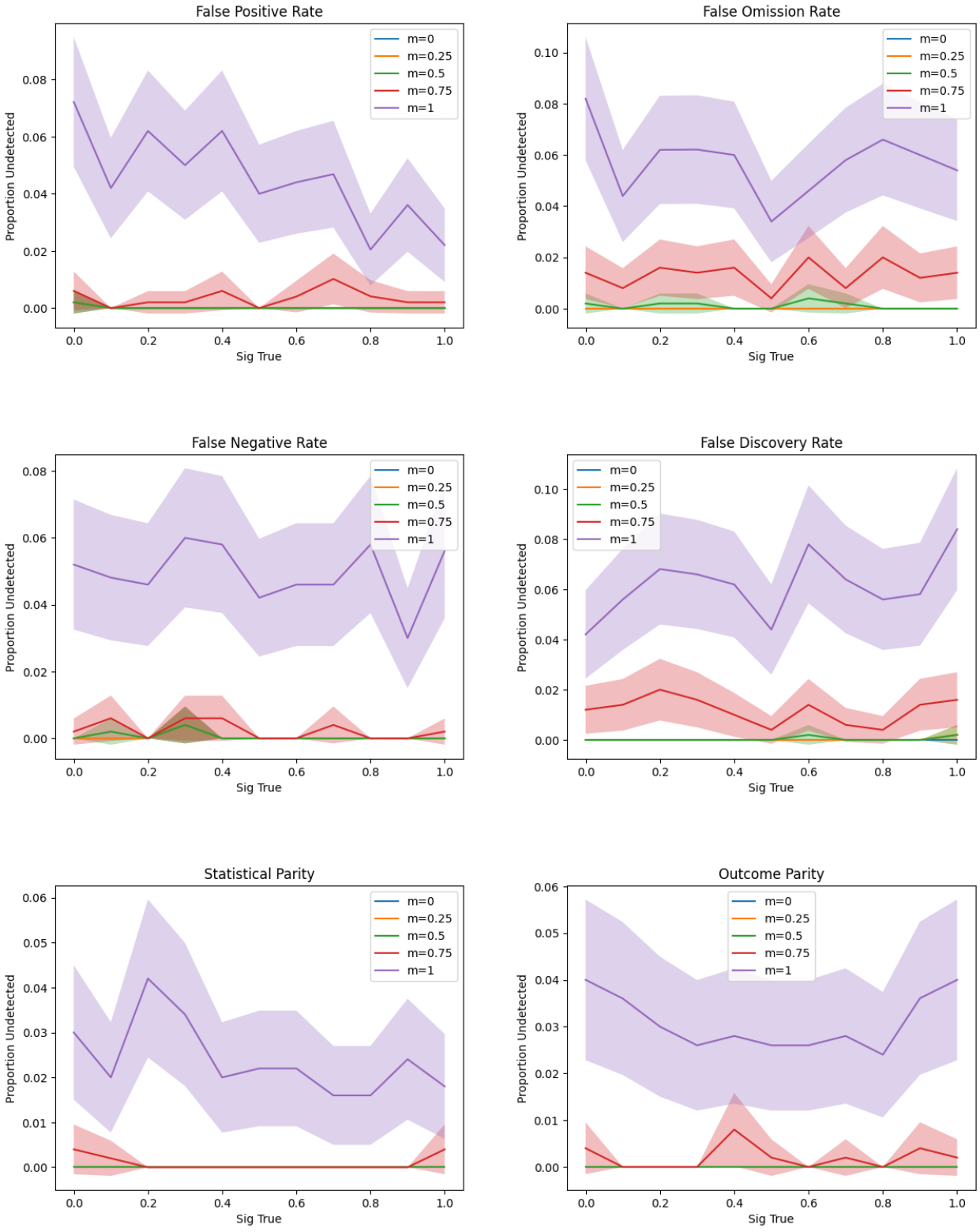

Figure 22: Proportion of Biased Subgroups Not Detected for Increasing Individual-Level Variability in True Outcome Probability σ_{true}

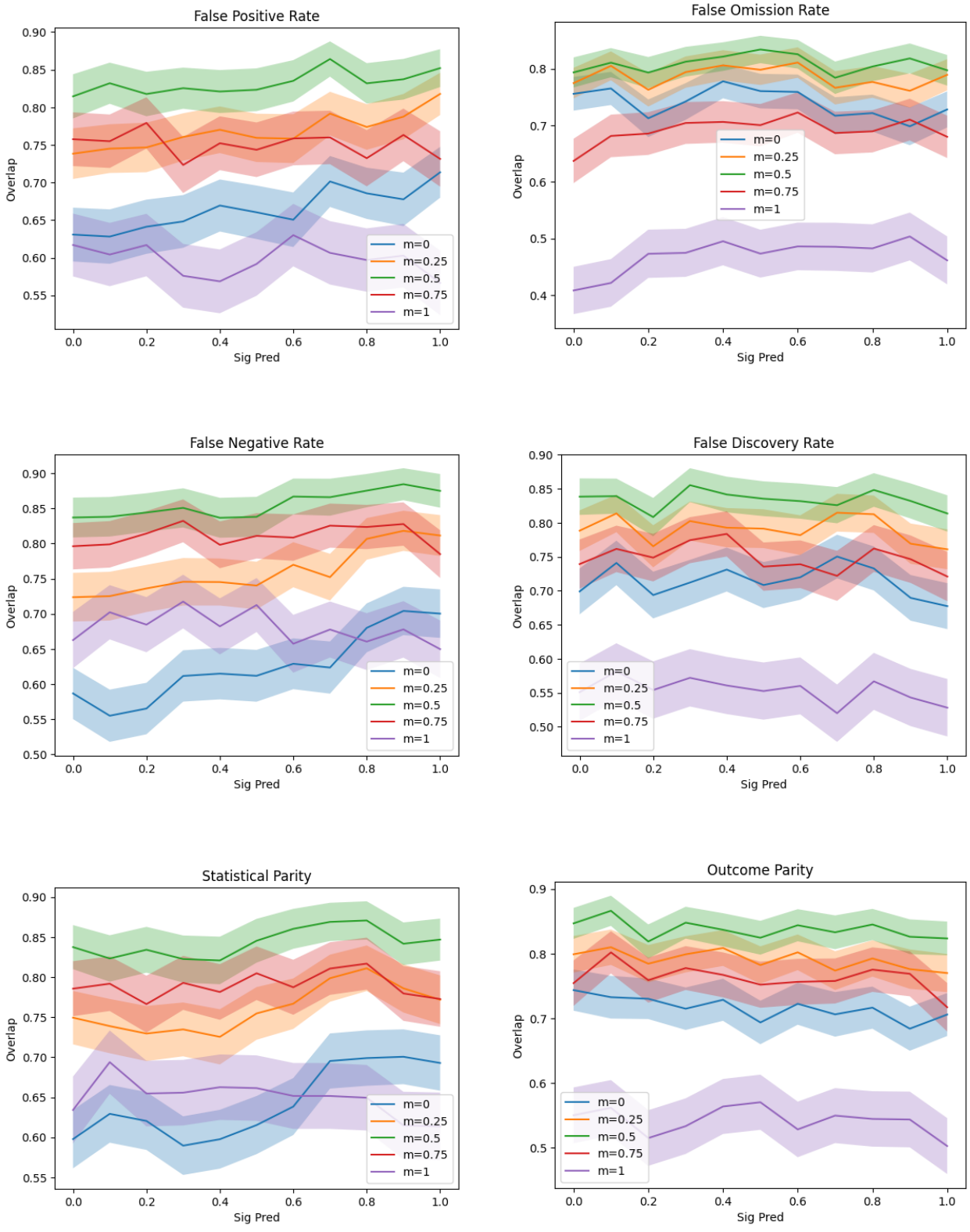

Figure 23: Detection of Each Error Type as Overlap for Increasing Individual-Level Noise in Predicted Outcome Probability σ_{pred}

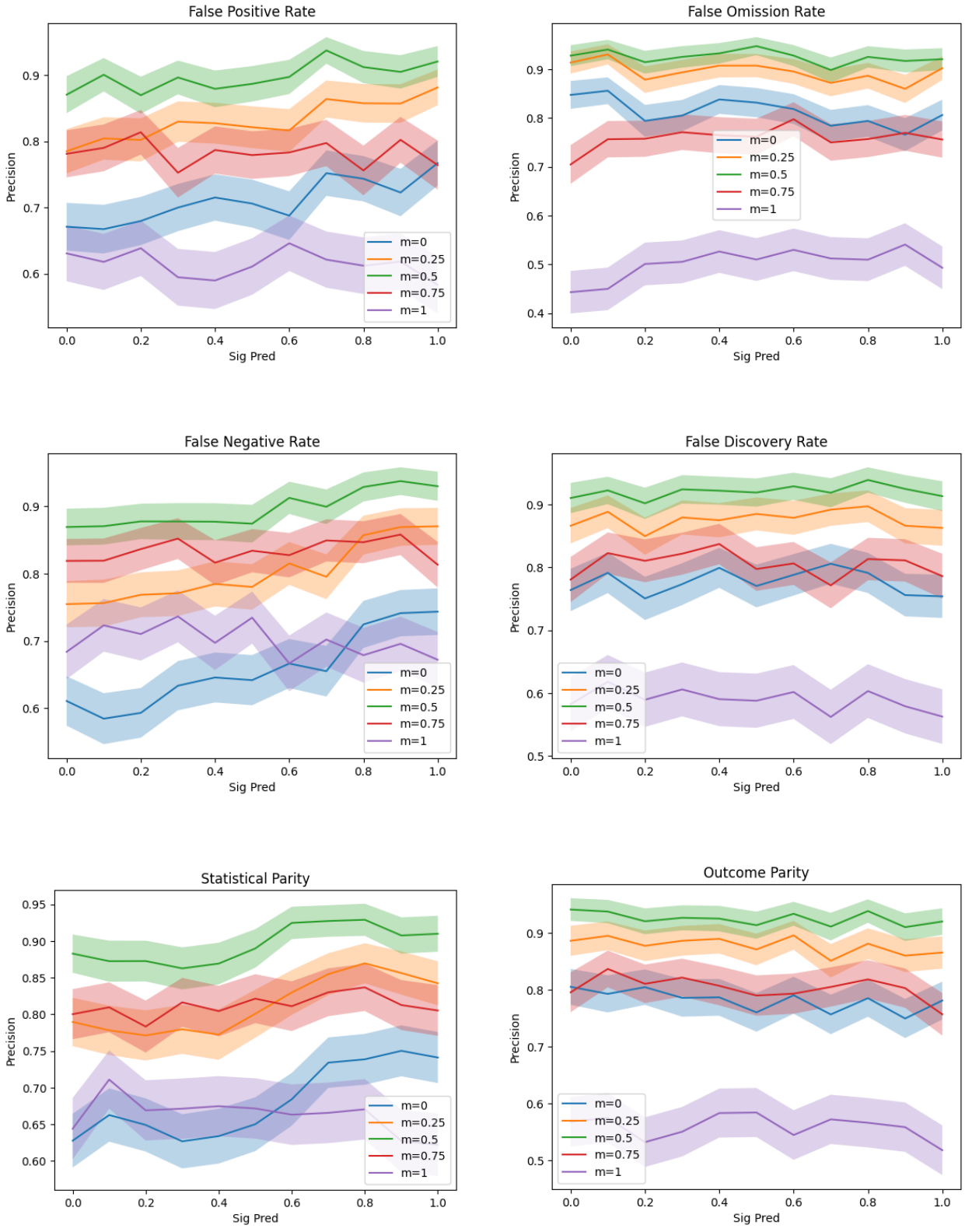

Figure 24: Detection of Each Error Type as Precision for Increasing Individual-Level Noise in Predicted Outcome Probability σ_{pred}

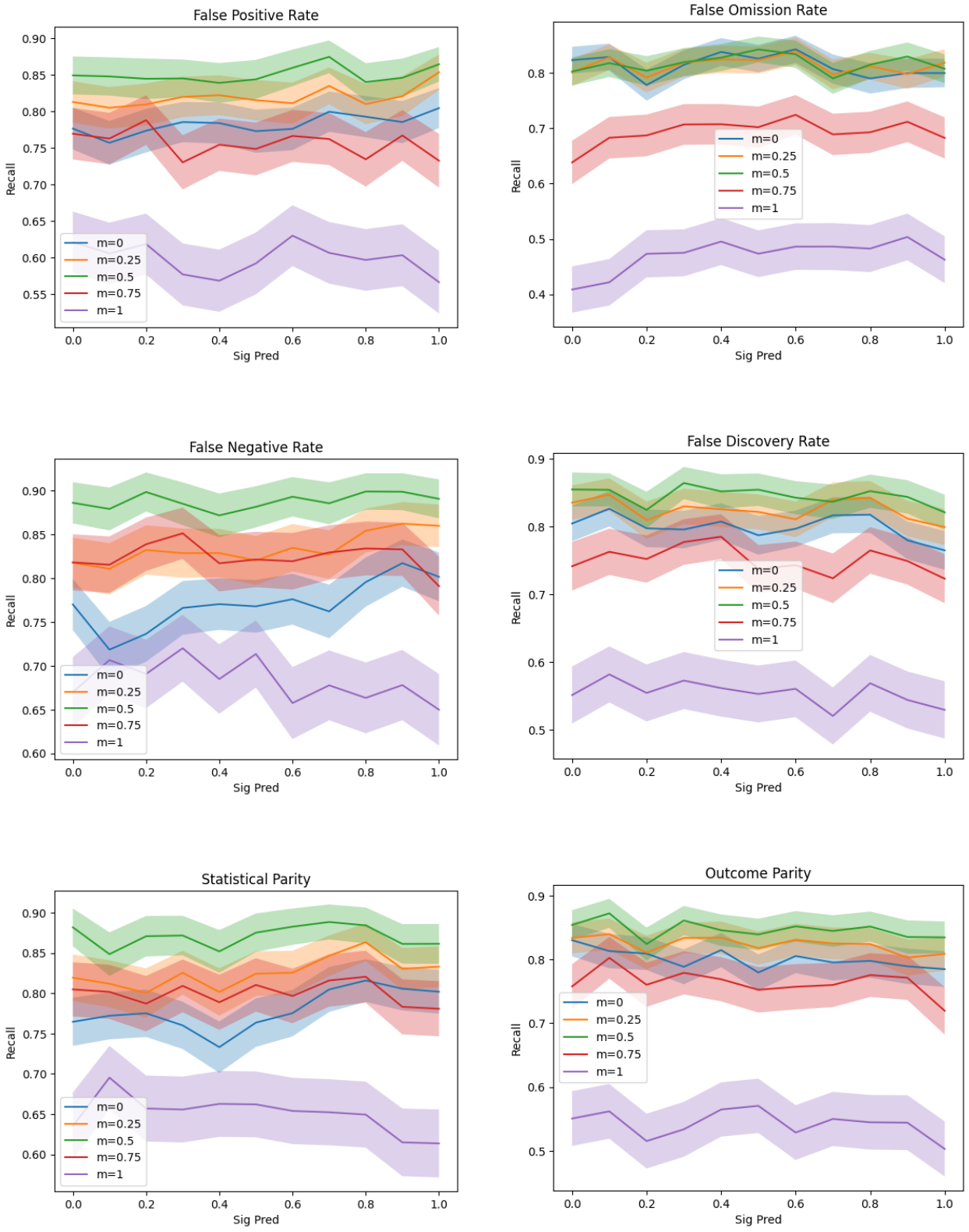

Figure 25: Detection of Each Error Type as Recall for Increasing Individual-Level Noise in Predicted Outcome Probability σ_{pred}

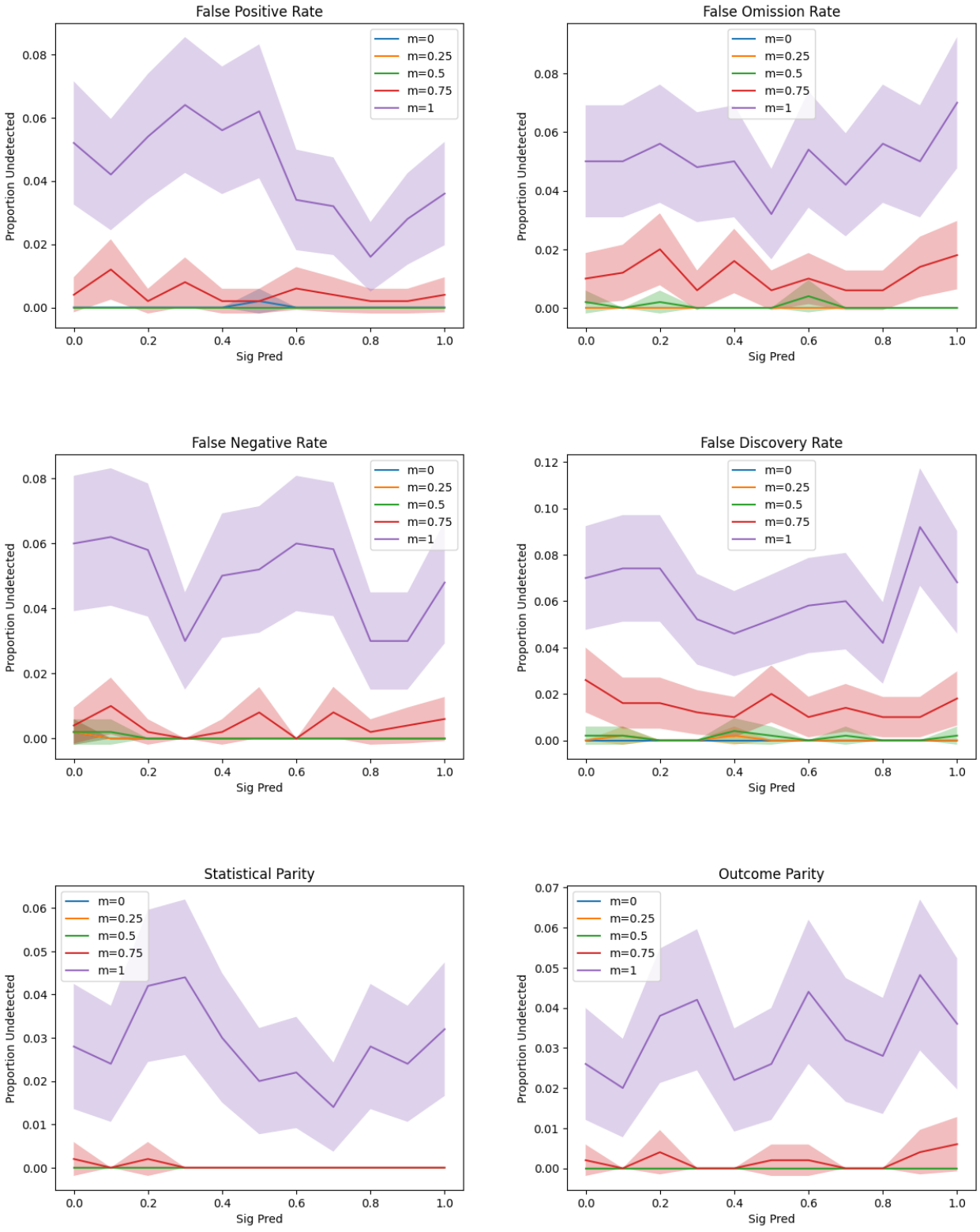

Figure 26: Proportion of Biased Subgroups Not Detected for Increasing Individual-Level Noise in Predicted Outcome Probability σ_{pred}