# A Disparity Metric Definitions

## A.1 Observational Metrics

**False Positive Rate Parity** Definition: $\hat{Y} \perp A \mid Y = 0$

Measured as: $P(\hat{Y} = 1 \mid A = 0, Y = 0) - P(\hat{Y} = 1 \mid A = 1, Y = 0)$

**False Negative Rate Parity** Definition: $\hat{Y} \perp A \mid Y = 1$

Measured as: $P(\hat{Y} = 1 \mid A = 0, Y = 1) - P(\hat{Y} = 1 \mid A = 1, Y = 1)$

**Positive Predictive Parity** Definition: $Y \perp A \mid \hat{Y} = 1$

Measured as: $P(Y = 1 \mid A = 0, \hat{Y} = 1) - P(Y = 1 \mid A = 1, \hat{Y} = 1)$

**Negative Predictive Parity** Definition: $Y \perp A \mid \hat{Y} = 0$

Measured as: $P(Y = 1 \mid A = 0, \hat{Y} = 1) - P(Y = 1 \mid A = 1, \hat{Y} = 0)$

**Equalized Odds** Definition: $Y \perp A \mid \hat{Y}$

Measured as: $\max\left\{\mathrm{FPR}(Y, A, \hat{Y}), \mathrm{FNR}(Y, A, \hat{Y})\right\}$ for false positive rate (FPR) and false negative rate (FNR) given above.

## A.2 ECP Parity Metric Definitions

**Counterfactual False Positive Rate Parity** Definition: $\hat{Y} \perp A \mid Y(D = 1) = 0$

Measured as: $P_{\mathcal{C}}(\hat{Y} = 1 \mid A = 0, Y(D = 1) = 0) - P(\hat{Y} = 1 \mid A = 1, Y(D = 1) = 0)$

**Counterfactual False Negative Rate Parity** Definition: $\hat{Y} \perp A \mid Y = 1$

Measured as: $P(\hat{Y} = 1 \mid A = 0, Y(D = 1) = 1) - P(\hat{Y} = 1 \mid A = 1, Y(D = 1) = 1)$

**Counterfactual Positive Predictive Parity** Definition: $Y(D = 1) \perp A \mid \hat{Y} = 1$

Measured as: $P(Y(D = 1) = 1 \mid A = 0, \hat{Y} = 1) - P(Y(D = 1) = 1 \mid A = 1, \hat{Y} = 1)$

**Counterfactual Negative Predictive Parity** Definition: $Y(D = 1) \perp A \mid \hat{Y} = 0$

Measured as: $P(Y(D = 1) = 1 \mid A = 0, \hat{Y} = 1) - P(Y(D = 1) = 1 \mid A = 1, \hat{Y} = 0)$

**Counterfactual Equalised Odds** Definition: $Y(D = 1) \perp A \mid \hat{Y}$

Measured as: $\max\left\{\mathrm{CF_FPR}(Y, A, \hat{Y}), \mathrm{CF_FNR}(Y, A, \hat{Y})\right\}$ for counterfactual false positive rate ($\mathrm{CF_FPR}$) and counterfactual false negative rate ($\mathrm{CF_FNR}$) given above.

# B Technical Description

## B.1 Margnalisation in DAGs

**Marginalisation Operation** Suppose $\mathbf{V}$ can be split as $\mathbf{V} = \tilde{\mathbf{V}} \cup \tilde{\mathbf{U}}$ where we are interested in the causal structure over $\tilde{\mathbf{V}}$ and do not observe the variables $\tilde{\mathbf{U}}$. We start from a causal graph $\mathcal{G}$, with unobserved $\tilde{\mathbf{U}}$ we marginalise to get to a graph $\mathcal{G}'$ which is of the form of Definition 1 by doing the following:

1. For all $U \in \tilde{\mathbf{U}}$, add an edge $Z \to \tilde{Z}$ if the current graph contains $Z \to U \to \tilde{Z}$ and then delete any edges $Z \to U$,
2. After completing the first step for all variables in $\tilde{\mathbf{U}}$, delete any $U$ if there exists another $\tilde{U} \in \tilde{\mathbf{U}}$ that influences all of the variables $U$ influences.

Evans [22] showed that there is a structural causal model over the resulting graph which preserves the causal structure over the variables $\tilde{\mathbf{V}}$. Importantly, due to the deletion step, this model has bounded number of unobserved variables, regardless of how large the set $\tilde{U}$ is.

**Graphical examples**

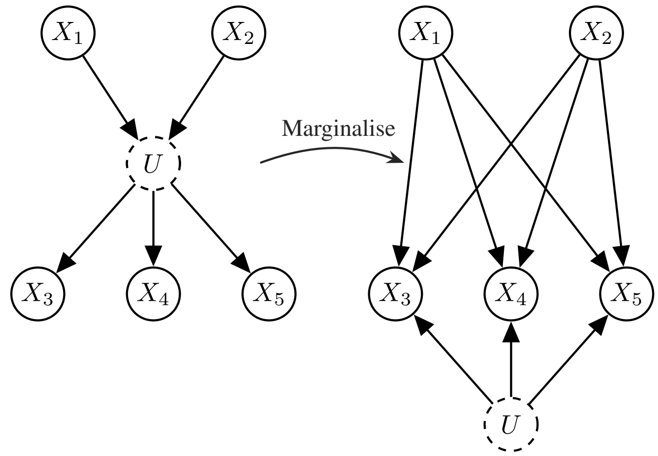



Figure 5: Example of step one in the marginalisation, taken from Evans [22].

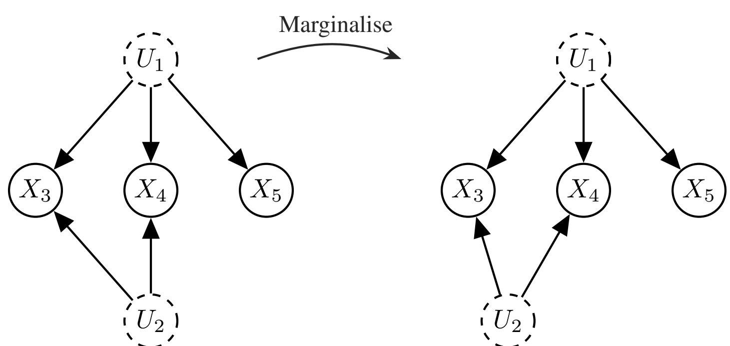



Figure 6: Example of step two in the marginalisation.

### B.2 Alternative Causal Graphs for Proxy Bias

Here we provide the following result demonstrating that a wide variety of Graphs can give the same outcome under proxy bias:

**Proposition 1.** *So long as any additional unobserved variables $U'$ satisfy the following:*

1. *$U'$ does not cause $A$.*
2. *There is no direct arrow from $U'$ to $\hat{Y}$.*

*Then marginalising over $U'$ will lead to the same graph as Figure 1a.*

*Proof.* To show this we need to demonstrate that once we have performed the marginalisation operations, no additional edges or nodes will be added to the graph. We do this step by step:

1. This step will add edges if we have two vertices $V, V'$ such that $V \to U' \to V'$. However, if neither of $V, V'$ are $\hat{Y}$ then these vertices will already be adjacent in the graph. As the graph is acyclic that means we cannot be adding any edges.
2. After removing all edges in step one, we will be left so that $U'$ has no parents and effects a subset of vertices in the graph. However, as $U'$ does not cause $A$, this must be a subset of $\{\hat{Y}, Y_P, Y\}$. As these are the vertices caused by $U$ this will lead to the deletion of $U'$.

$\square$

| Dataset | Task | Proxy Bias | Selection Bias | ECP Bias |
|---|---|---|---|---|
| Adult | Synthetic | | | |
| KDD Census-Income | Synthetic | | | |
| German credit | Credit risk | | ✓ | ✓ |
| Dutch census | Synthetic | | | |
| Bank marketing | Client | | ✓ | |
| Credit card clients | Default Risk | | ✓ | ✓ |
| COMPAS recid. | Risk prediction | ✓ | ✓ | ✓ |
| COMPAS viol. recid. | Risk prediction | ✓ | ✓ | ✓ |
| Communities&Crime | Neighborhood risk | ✓ | ✓ | ✓ |
| Diabetes | Re-admission risk | ✓ | | ✓ |
| Ricci | Promotion Prediction | ✓ | ✓ | ✓ |
| Student-Mathematics | Admissions | ✓ | ✓ | ✓ |
| Student-Portuguese | Admissions | ✓ | ✓ | ✓ |
| OULAD | Admissions | ✓ | ✓ | ✓ |
| Law School | Admissions | ✓ | ✓ | ✓ |

Table 2: Analysis of the datasets from Le Quy et al. [43], split by task. The explanation for the biases are given in Appendix E.

# C Cross Dataset Analysis

In this section we analyse the datasets presented in Le Quy et al. [43] for the three biases we present in Section 3. We describe each dataset, give the task which most closely relates to the use of this dataset, and relative to this task we decide if each of the three measurement biases are present or not. For each bias we provide a justification of our decision.

**Synthetic tasks** The synthetic tasks are hard to discuss since the biases are contextual and these tasks are purely theoretical. Given a downstream task they might or might not have the biases we discuss. Therefore we drop them from the analysis.

**Bank marketing Dataset** The goal here is to target current clients for the bank to open more accounts. Since the outcome in this case is exactly what the bank seeks to maximise, this dataset does not exhibit proxy or ECP bias. However, contacts we made via phone, so there is selection bias in whether people answered the phone.

**German credit and Credit card clients** For both of these datasets goal is to predict whether customers face default risk. The aim is to use this to decide if applying customers present a risk to the bank or not. As a result of this, there will be selection bias due to the fact that since defaults are only observed for the firms' previous customers. Finally as with the example in the main text, this exhibits extra-classificatory policy bias since the firm sets the credit limit which impacts the likelihood of default.

**COMPAS recid. and COMPAS viol. recid. and Communities and Crime** Datasets build off COMPAS have been well documented to exhibit all these biases and more [5]. These issues are not unique to COMPAS and are exhibited in all other recidivism and crime prediction datasets, as such they will also apply to communities and crime, where the aim is to predict number of historical crimes per hundred thousand population for a number of states. Moreover, a large degree of missing values in this dataset show the issues due to selection bias.

**Diabetes** For this dataset, the goal is to predict if a patient will be readmitted in the next 30 days. The aim is to use this to decide how much of a health risk a given patient is upon leaving the hospital, to decide if they should be kept there. The population is a sample of the patient pool, and so there should not be selection bias. Readmissions are different from the underling recurring illness, so this does represent a proxy, albeit it a fairly reasonable one. In this case ECP bias is a cause for concern due to the differences in quality of care by demographic group [21].

**Ricci** The Ricci dataset is an employment dataset, where the goal is to predict the likelihood of a promotion based off of a selection of available covariates. A model trained on this data would then be used to predict the potential of applying candidates in order to decide if they are invited to interview or do additional tests. This application would fall risk of all the biases we have presented and as such strong justification would be required as to the usefulness of the model. Going through one by one, proxy bias is exhibited in a similar way to the example presented in the main text, selection bias is present as the model is evaluated on a different population to the one it is trained on, and finally the firms policies will have an impact on who succeeds and is promoted at the company.

**Admissions datasets** The final datasets can all be grouped under admissions to academic institutions. Similarly to the employment example, these will exhibit all the biases we have outlined. This is because of the challenges of having a perfectly objective measure of performance, models being used on applying populations but fit on accepted populations, and the universities policies affecting the success of students. Therefore, when using these predictors arguments should be made about why using such a measure would not induce demographic skew.

# D Additional Results

## D.1 Proxy Label Results

### D.1.1 Plots from Fogliato et al. [26] under varying assumptions

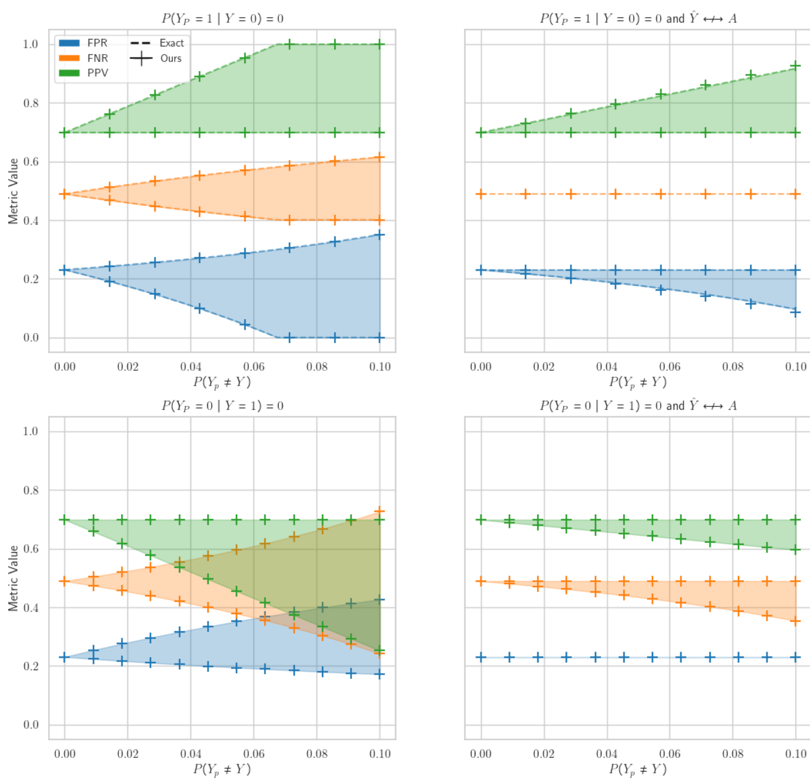



Figure 7: In this plot, we recreate the results from Fogliato et al. [26], where we are interested in the false positive rate (FPR), false negative rate (FNR), and positive predictive value (PPV) for a classifier trained on the COMPAS dataset. In this plot we consider varying for which $j$ we have $P(Y_P = 1 - j \mid Y = j)$, and we can see that doing so greatly changes the shape of the sensitivity set. Moreover, when we pair these assumptions dropping of the red dashed edge in Figure 1a we see we can identify some of the metrics of interest under any degree of bias. For $j = 1$ we identify the FNR and for $j = 0$ we identify the FPR. We prove these identification results in Appendix D.1.2
.

### D.1.2 Proxy Identification Results

In the set up in Fogliato et al. [26], the aim is the false positive/negative rate in a group $A = a$, where it is assumed that $P(Y = 1, Y_P = 0) = 0$. Now declaring the following parameters:

$$\begin{aligned}
p_{ij} &= P(Y_P = i, \hat{Y} = j \mid A = a) \\
\alpha_j &= P(Y = 1, Y_P = 0, \hat{Y} = j \mid A = a) \\
\alpha &= \alpha_0 + \alpha_1
\end{aligned}$$

Under these assumptions $\alpha_0, \alpha_1$ are sufficient to parameterise the distribution, $P(Y, Y_P, \hat{Y} \mid A = a)$. Now, following [26] we have that:

$$\begin{aligned}
\mathrm{FPR}_Y &= \frac{p_{01} - \alpha_1}{p_{00} + p_{01} - \alpha} \\
\mathrm{FNR}_Y &= \frac{p_{10} + \alpha_0}{p_{10} + p_{11} + \alpha} \\
\mathrm{PPV}_Y &= \frac{p_{11} + \alpha_1}{p_{01} + p_{11}}
\end{aligned}$$

Now, with the absence of the dashed edge, the DAG in Figure 1a implies the independence $\hat{Y} \perp Y_P \mid Y, A$. Therefore we get the following:

$$\begin{aligned}
\alpha_j &= P(Y = 1, Y_P = 0, \hat{Y} = j \mid A = a) \\
&= \frac{P(Y = 1, Y_P = 0 \mid A = a) P(Y = 1, \hat{Y} = j \mid A = a)}{P(Y = 1 \mid A = a)} \\
&= \frac{\alpha(p_{1j} + \alpha_j)}{p_{10} + p_{11} + \alpha}
\end{aligned}$$

Solving for $\alpha_j$, we get $\alpha_j = \alpha \left(\frac{p_{1j}}{p_{10}+p_{11}}\right)$. Now, inputting this for $\alpha_0$ in the expression for $\mathrm{FNR}_Y$ we get:

$$\begin{aligned}
\mathrm{FNR}_Y &= p_{10} \left( \frac{1 + \frac{\alpha}{p_{10}+p_{11}}}{p_{10} + p_{11} + \alpha} \right) \\
&= \frac{p_{10}}{p_{10} + p_{11}} \\
&= \mathrm{FNR}_{Y_P}
\end{aligned}$$

Therefore, under the assumptions given, the true false negative rate is identified and equal to the observed false negative rate on the proxy labels. Inputting the value for $\alpha_1$ into $\mathrm{FPR}_Y$ we instead get:

$$\mathrm{FPR}_Y = \frac{p_{01} - \alpha \left(\frac{p_{10}}{p_{10}+p_{11}}\right)}{(p_{00} + p_{01} - \alpha)}$$

As this is a decreasing function of $\alpha$ we can see that for $\alpha \leq \alpha_0$, $\mathrm{FPR}_Y$ is bounded as:

$$\frac{p_{01}}{(p_{00} + p_{01})} \leq \mathrm{FPR}_Y \leq \frac{p_{01} - \alpha \left(\frac{p_{10}}{p_{10}+p_{11}}\right)}{(p_{00} + p_{01} - \alpha)}$$

For PPV, we again input $\alpha_1$ to give:

$$\mathrm{PPV}_Y = \frac{p_{11} + \alpha \left(\frac{p_{11}}{p_{10}+p_{11}}\right)}{p_{01} + p_{11}}$$

Leading to the bounds:

$$\mathrm{PPV}_{Y_P} \leq \mathrm{PPV}_Y \leq \frac{p_{11} + \alpha \left(\frac{p_{11}}{p_{10}+p_{11}}\right)}{p_{01} + p_{11}}$$

The statements for the identification of the false positive rate and false negative rate are as follows:

**Proposition 2.** *Suppose we have $P(Y_P = 1 \mid Y = 0) = 0$. Then under the conditional independence statement $\hat{Y} \perp Y_P \mid Y, A$, for all level of proxy bias $P(Y_P \neq Y)$:*

$$\mathrm{FNR}_{Y|A=a} = \mathrm{FNR}_{Y_P|A=a}$$

*Where $\mathrm{FNR}_{Y|A=a}$ is the true false negative rate for the group $A = a$ and $\mathrm{FNR}_{Y_P|A=a}$ is the proxied false negative rate.*

*Proof.* Follows from the above derivations. □

Now the equivalent statement for the false positive ratio:

**Proposition 3.** *Suppose we have $P(Y_P = 0 \mid Y = 1) = 0$. Then under the conditional independence statement $\hat{Y} \perp Y_P \mid Y, A$, for all level of proxy bias $P(Y_P \neq Y)$:*

$$\mathrm{FPR}_{Y|A=a} = \mathrm{FPR}_{Y_P|A=a}$$

*Where $\mathrm{FPR}_{Y|A=a}$ is the true false negative rate for the group $A = a$ and $\mathrm{FPR}_{Y_P|A=a}$ is the proxied false negative rate.*

*Proof.* This follows from considering the distribution where $Y, Y_P$ and $\hat{Y}$ are all flipped as any statement about the false positive rate in the original distribution translates to a statement about the false negative rate in the flipped distribution. The assumption $P(Y_P = 0 \mid Y = 1)$ in the original distribution translates to $P(Y_P = 1 \mid Y = 0)$ in the flipped distribution, whereas all other assumptions are symmetric to the flipping operation. Therefore we can apply proposition 2 to see that the flipped FNR is constant under any degree of proxy noise. This leads us to conclude that under these assumptions the FPR in the original distribution must also be constant under any degree of proxy noise. □

## D.2 Selection Results

### D.2.1 Selective labels under MNAR

Here we include an experiment applying the framework to selective labels under the missing not a random assumption (MNAR)[56]. This supposes that we only see the outcome on a subset of the full dataset, with the outcome on the rest of the dataset free to vary arbitrarily.We work with the Dutch census dataset Van der Laan [60], first fitting an unconstrained logistic regression, then forming the selected population as those who have a predicted probability higher $0.3$.

Once we have formed the selected subset we then train four classifiers, each to satisfy a different parity metric. We train to false negative rate parity, false positive rate parity, positive predictive parity and negative predictive parity. False negative/positive rate parity are trained using the reductions approach [2], whereas for positive predictive parity and negative predictive parity we train 100 predictors, each weighting different parts of the distribution, taking the one with the lowest parity score above a given accuracy threshold. The plots are shown in Figure 8.

### D.2.2 Selection and Proxy Plots

In this we demonstrate the effect of selection and proxy bias jointly on the adult dataset. We include the results in Figure 9, which show that both the occurrence of multiple biases acts differently for different parity metrics.

## D.3 ECP bias results

### D.3.1 ECP experimental set up

For this experiment, we focus on finding the possible ranges for the counterfactual parity metrics from the given observational statistics. We use the sensitivity parameter of $P(Y(1) \neq Y(0))$, adding additional causal assumptions such as monotonicity ( $Y(1) \geq Y(0)$ ) and if the policy is observed or not. When simulating the policy we draw ECP $\sim \mathrm{Ber}(\frac{1}{2} + c * A)$ for $c = 0.2$ in order to skew the policy in one direction. Results show in Figure 10

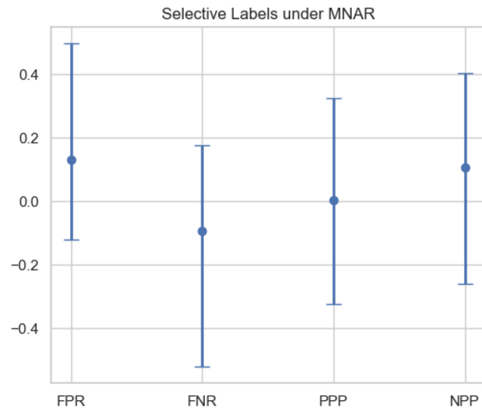



Figure 8: This plot demonstrates a sensitivity analysis for selective labels on the Dutch dataset under the missing not at random assumption.

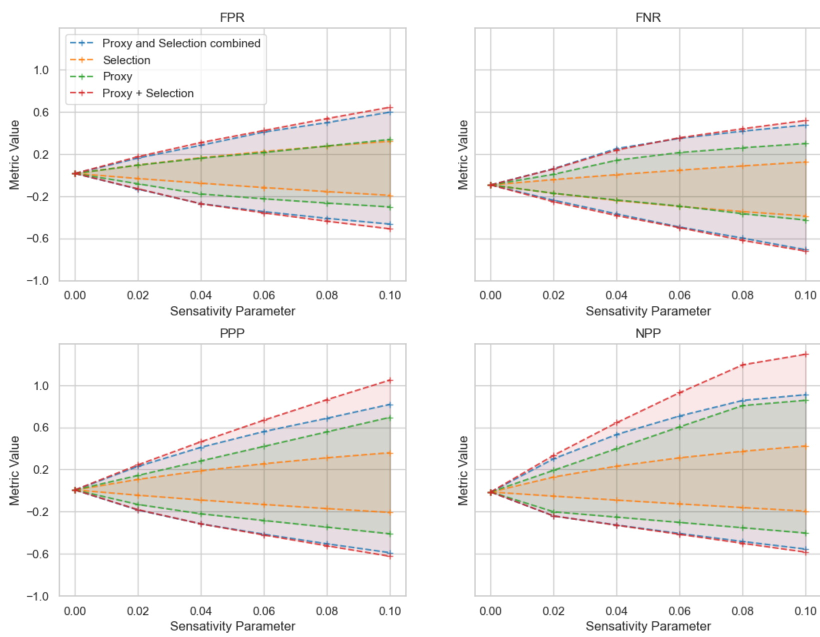



Figure 9: In these plots, we can see the effect of doing a sensitivity analysis jointly for selection and proxy bias. We can see that for the false positive rate parity (FPR) and false negative rate parity (FNR) the combined bias behaves roughly as the sum of both biases, however for positive predictive parity (PPP) and negative predictive parity (NPP) the combination behaves differently with the combined bias amounting a smaller possible range for the metrics than the sum of the range of both biases individually.

.

### D.4 Causal Fairness Experiments

In this section we give some results on applying our sensitivity analysis framework to causal fairness metrics of the variety detailed in [49]. Before doing so, we add some technical comments on these types of interventions in FairML, and some nuances of measurement bias in the context of causal inference.

Firstly, we would like to comment that our framework can still be applied to perform sensitivity analysis for measurement bias for other fairness metrics *without* having to consider counterfactuals relative protected characteristics such as race or gender, thereby avoiding difficulties with intervention

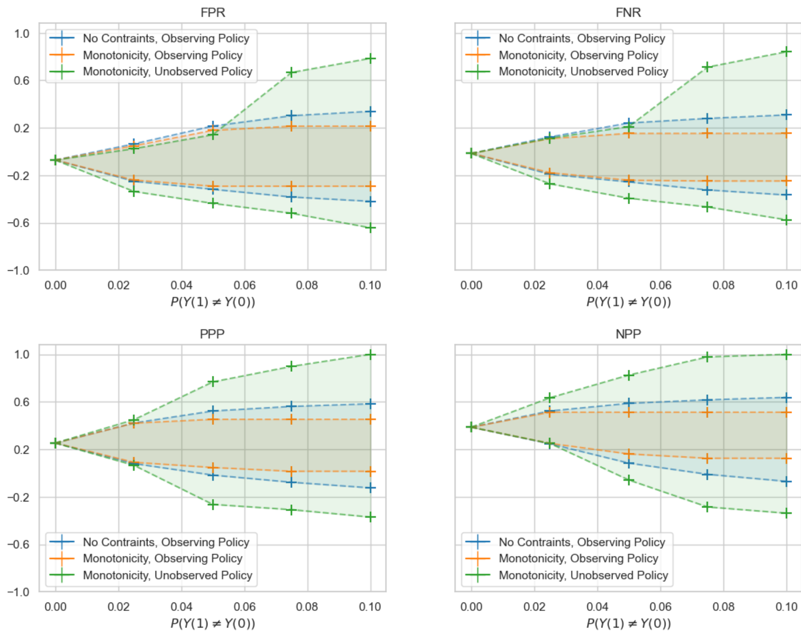



Figure 10: In these plots, we perform a sensitivity analysis for the value of the counterfactual parity metrics given in Appendix A.2. In each case we work under 3 differing levels of assumption and information

on such traits [39, 32, 36]. In this case $A$ could be seen as denoting membership to a group and indexing different graphs for each group as in Bright et al. [10]. In this case the arrows leading from $A$ would only express conditional independence relationships as opposed to causal ones. Notably in the graphs we suggest they are unconstrained.

Secondly, measurement biases and specifically selection bias in causal fairness comes with additional problems. This is because almost always, membership of such a dataset is causally downstream of the protected attribute, meaning that when conditioning on individuals being in a dataset we are introducing selection bias in some form. As Fawkes et al. [25] argue, this means DAG models will be unable to correctly capture the causal structure in most datasets we come across in FairML. Failing to account for such effects can lead to erroneous causal conclusions.

Having said this, we will proceed with applying the causal graphs in Figure 1 to do causal fairness analysis for the following metrics:

**Counterfactual Fairness (CF) [41]** We measure this as $P_{\mathcal{C}}(\hat{Y}(A=1) \neq \hat{Y}(A=0))$ which is equal to 0 exactly when $\hat{Y}$ is counterfactually fair [24].

**Total Effect (TE) [49]** Measured as $P_{\mathcal{C}}(\hat{Y}(A=1)) - P_{\mathcal{C}}(\hat{Y}(A=0))$.

**Spurious Effect (SE) [49]** Measured as $P_{\mathcal{C}}(\hat{Y}(A=a)) - P_{\mathcal{C}}(\hat{Y} \mid A=a)$.

Results are show in Figure 11, where we have assumed that counterfactual fairness is identified at a particular value. We can see that all causal fairness metrics recover a linear relationship under selection in this context.

## E Details of cross dataset bias analysis

In this section we analyse the datasets presented in Le Quy et al. [43] for the three biases we present in Section 3. We describe each dataset, give the task which most closely relates to the use of this dataset, and relative to this task we decide if each of the three measurement biases are present or not. For each bias we provide a justification of our decision.

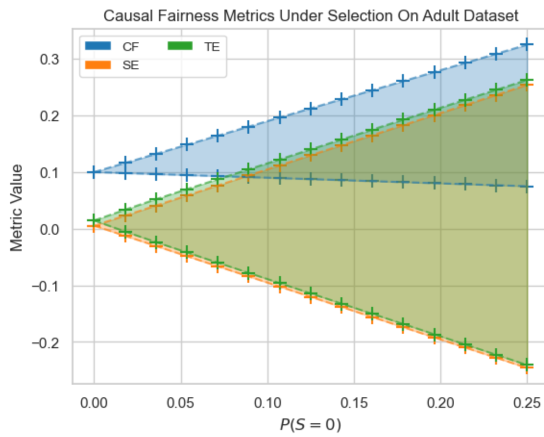



Figure 11: Causal fairness metrics under selection. We show plots for Counterfactual Fairness (CF),Total Effect (TE), and Spurious Effect (SE) using graph 1b where we have additionally assumed that counterfactual fairness is point identified.

| Dataset | Task | Proxy Bias | Selection Bias | ECP Bias |
|---|---|---|---|---|
| Adult | Synthetic | | | |
| KDD Census-Income | Synthetic | | | |
| German credit | Credit risk | | ✓ | ✓ |
| Dutch census | Synthetic | | | |
| Bank marketing | Client | | ✓ | |
| Credit card clients | Default Risk | | ✓ | ✓ |
| COMPAS recid. | Risk prediction | ✓ | ✓ | ✓ |
| COMPAS viol. recid. | Risk prediction | ✓ | ✓ | ✓ |
| Communities&Crime | Neighborhood risk | ✓ | ✓ | ✓ |
| Diabetes | Re-admission risk | ✓ | | ✓ |
| Ricci | Promotion Prediction | ✓ | ✓ | ✓ |
| Student-Mathematics | Admissions | ✓ | ✓ | ✓ |
| Student-Portuguese | Admissions | ✓ | ✓ | ✓ |
| OULAD | Admissions | ✓ | ✓ | ✓ |
| Law School | Admissions | ✓ | ✓ | ✓ |

Table 3: Analysis of the datasets from Le Quy et al. [43], split by task. The explanation for the biases are given in Appendix E.

**Synthetic tasks** The synthetic tasks are hard to discuss since the biases are contextual and these tasks are purely theoretical. Given a downstream task they might or might not have the biases we discuss. Therefore we drop them from the analysis.

**Bank marketing Dataset** The goal here is to target current clients for the bank to open more accounts. Since the outcome in this case is exactly what the bank seeks to maximise, this dataset does not exhibit proxy or ECP bias. However, contacts we made via phone, so there is selection bias in whether people answered the phone.

**German credit and Credit card clients** For both of these datasets goal is to predict whether customers face default risk. The aim is to use this to decide if applying customers present a risk to the bank or not. As a result of this, there will be selection bias due to the fact that since defaults are only observed for the firms' previous customers. Finally as with the example in the main text, this exhibits extra-classificatory policy bias since the firm sets the credit limit which impacts the likelihood of default.

**COMPAS recid. and COMPAS viol. recid. and Communities and Crime** Datasets build off COMPAS have been well documented to exhibit all these biases and more [5]. These issues are not unique to COMPAS and are exhibited in all other recidivism and crime prediction datasets, as such they will also apply to communities and crime, where the aim is to predict number of historical crimes per hundred thousand population for a number of states. Moreover, a large degree of missing values in this dataset show the issues due to selection bias.

**Diabetes** For this dataset, the goal is to predict if a patient will be readmitted in the next 30 days. The aim is to use this to decide how much of a health risk a given patient is upon leaving the hospital, to decide if they should be kept there. The population is a sample of the patient pool, and so there should not be selection bias. Readmissions are different from the underling recurring illness, so this does represent a proxy, albeit it a fairly reasonable one. In this case ECP bias is a cause for concern due to the differences in quality of care by demographic group [21].

**Ricci** The Ricci dataset is an employment dataset, where the goal is to predict the likelihood of a promotion based off of a selection of available covariates. A model trained on this data would then be used to predict the potential of applying candidates in order to decide if they are invited to interview or do additional tests. This application would fall risk of all the biases we have presented and as such strong justification would be required as to the usefulness of the model. Going through one by one, proxy bias is exhibited in a similar way to the example presented in the main text, selection bias is present as the model is evaluated on a different population to the one it is trained on, and finally the firms policies will have an impact on who succeeds and is promoted at the company.

**Admissions datasets** The final datasets can all be grouped under admissions to academic institutions. Similarly to the employment example, these will exhibit all the biases we have outlined. This is because of the challenges of having a perfectly objective measure of performance, models being used on applying populations but fit on accepted populations, and the universities policies affecting the success of students. Therefore, when using these predictors arguments should be made about why using such a measure would not induce demographic skew.

# F Details of cross dataset experiment

For this experiment, we train numerous predictors across a variety of common fairness benchmarking datasets [43] to satisfy parity constraints. For each dataset we train 18 classifiers total, where the model ML is one of logistic regression, naïve Bayes and a decision tree and the parity constraint is false negative rate parity, false positive rate parity, positive predictive parity and negative predictive parity, demographic parity and equalized odds. With the exception of positive/negative predictive parity we train all classifier to satisfy these constraints using the reductions approach [2]. For positive/negative predictive parity, we train 100 predictors, each weighting different parts of the distribution, taking the one with the lowest parity score above a given accuracy threshold. We vary the sensitivity parameter over a range of realistic values for many real-world settings, computing the sensitivity bounds for each level of the parameter. We find that, except for demographic parity, all parity measures we evaluate exhibit significant sensitivity over these parameter ranges. This makes it hard to understand what satisfying, e.x. equalised odds means on a given dataset. The caveat is that equalised odds is only satisfied as long as there are no significant measurement biases in the underlying data, which is almost never the case in FairML audits.

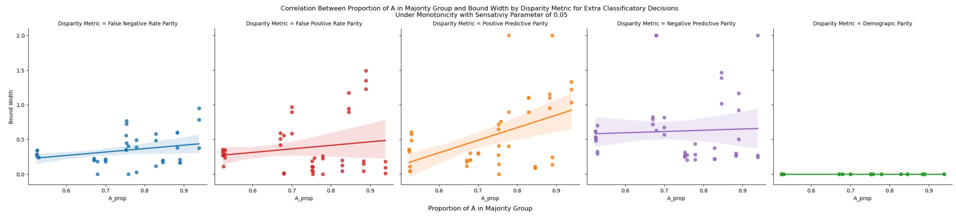
## F.1 Analysis of Results

### F.1.1 Correlational Plots

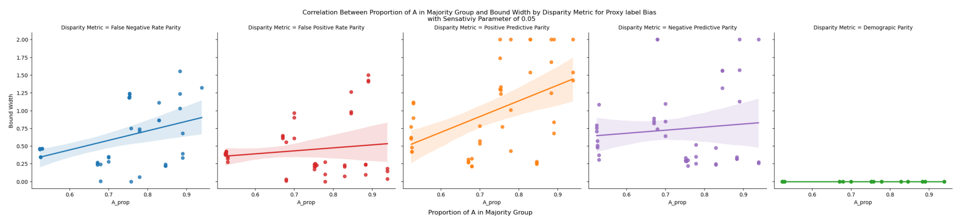
Here we explore how sensitivity varies according to class imbalance in either $A$ or $Y$. We find that for some metrics (Negative Predictive Parity) class imbalance seems to make little difference to the sensitivity of metrics, with next to no correlation observed between imbalance and sensitivity. This lies in contrast to other metrics (Positive Predictive Parity) where we can see a much more clear, positive, correlation between class imbalance and sensitivity.

### F.1.2 Cross Dataset Analysis

These results unpack some heterogeneity across datasets. We find that there are not hard and fast rules here, each dataset and each bias requires its own analysis. At the same time this is broadly consistent with our central contention that complexity goes hand in hand with fragility. In particular,



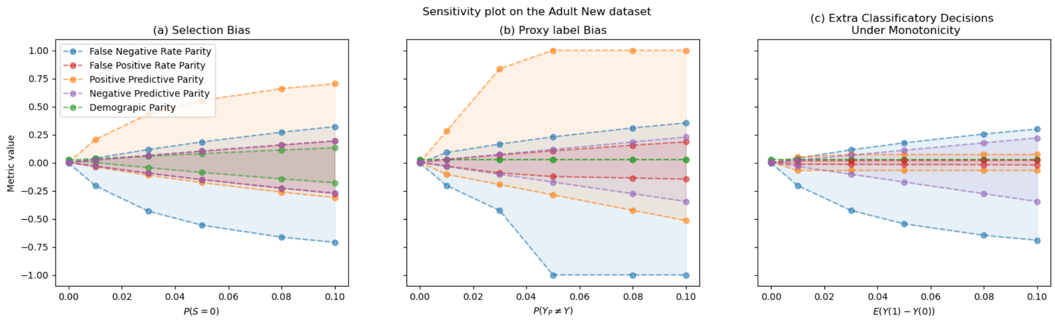


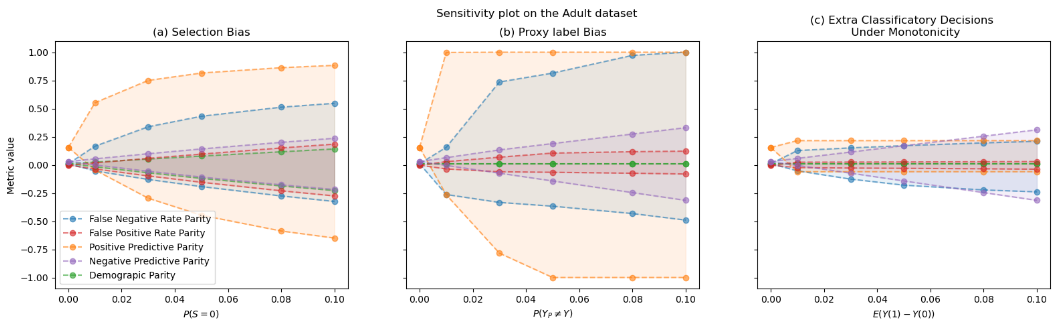


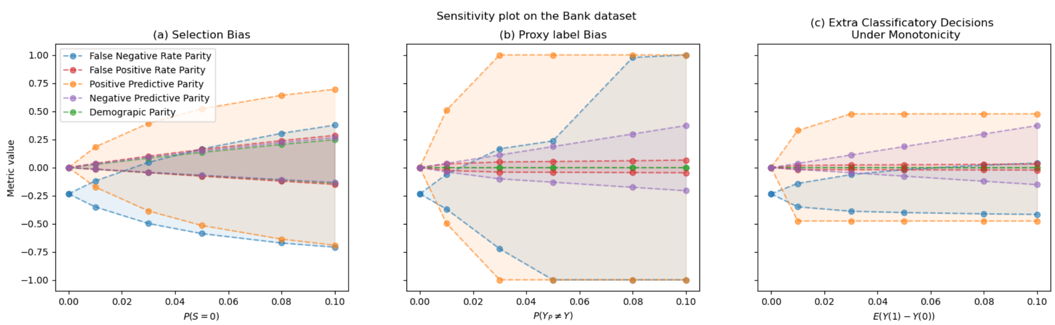


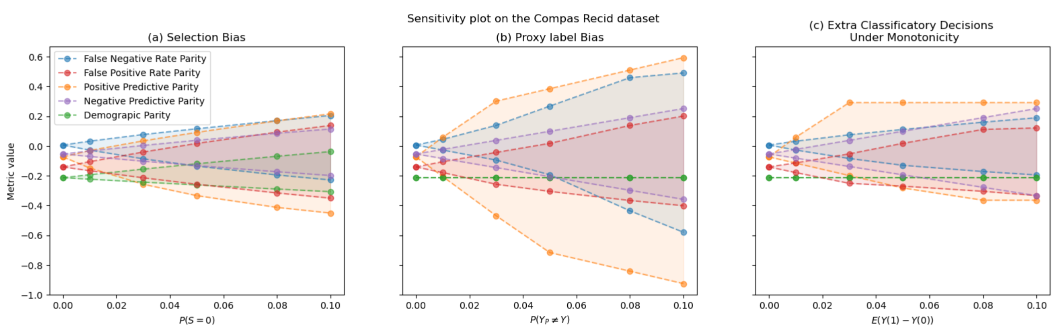


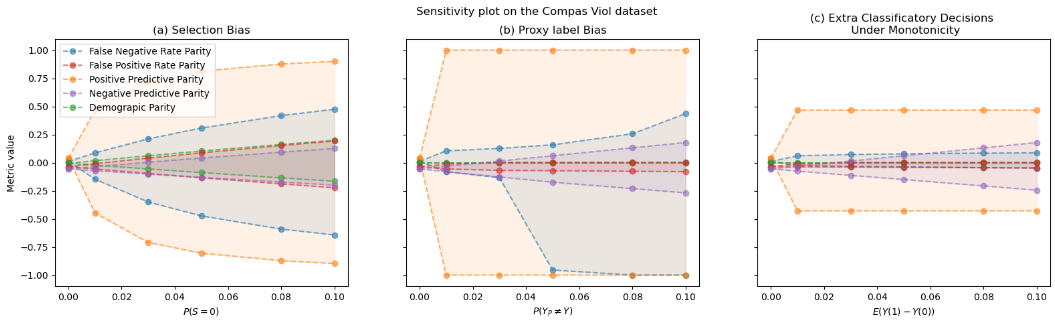
on the Adult New, Adult, Bank, Compas Recid/Viol, Credit, Dutch we see positive predictive parity is a standout fragile method across biases, followed by false negative rate pairt and negative predictive parity. On the German Credit, Law, and Ricci, Student mat/por we see NPP and FNRP tend to be the worst, followed by FPRP and PPP.

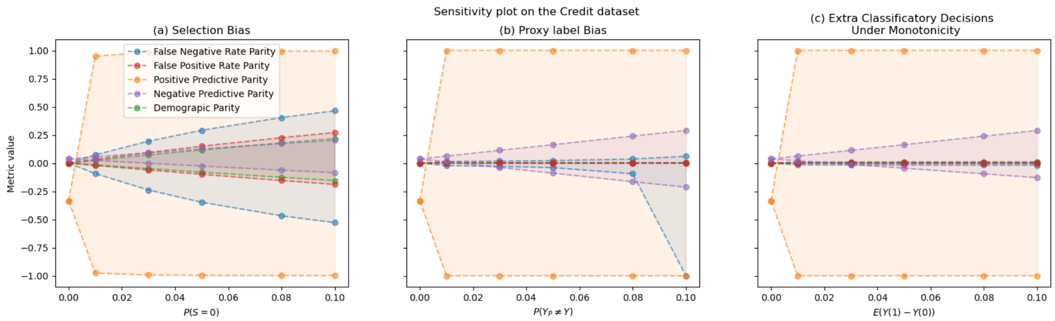

Sensitivity plot on the Adult New dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions
Under Monotonicity
False Negative Rate Parity
False Positive Rate Parity
Positive Predictive Parity
Negative Predictive Parity
Demograpic Parity
Metric value
$P(S=0)$
$P(Y_P \neq Y)$
$E(Y(1)-Y(0))$

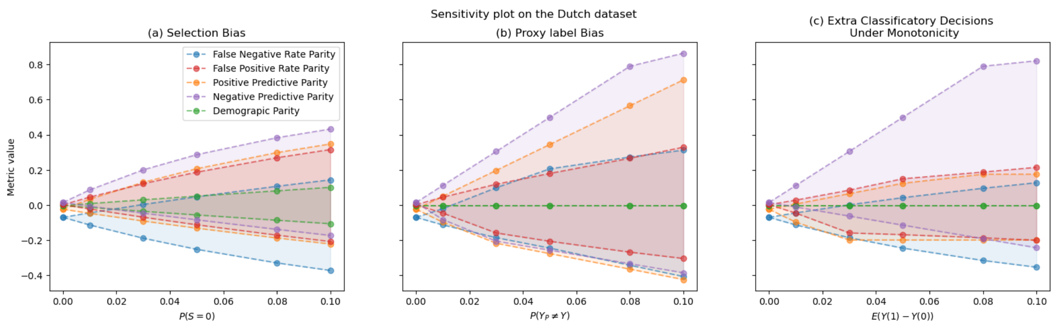

Sensitivity plot on the Adult dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions
Under Monotonicity
False Negative Rate Parity
False Positive Rate Parity
Positive Predictive Parity
Negative Predictive Parity
Demograpic Parity
Metric value
$P(S=0)$
$P(Y_P \neq Y)$
$E(Y(1)-Y(0))$

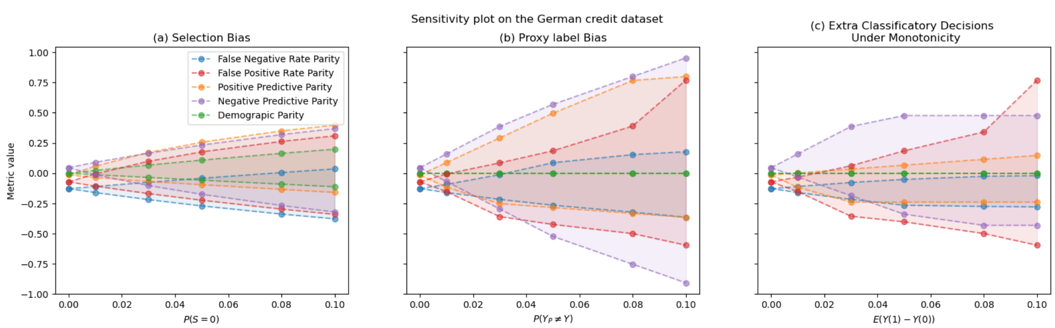

Sensitivity plot on the Bank dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions
Under Monotonicity
False Negative Rate Parity
False Positive Rate Parity
Positive Predictive Parity
Negative Predictive Parity
Demograpic Parity
Metric value
$P(S=0)$
$P(Y_P \neq Y)$
$E(Y(1)-Y(0))$

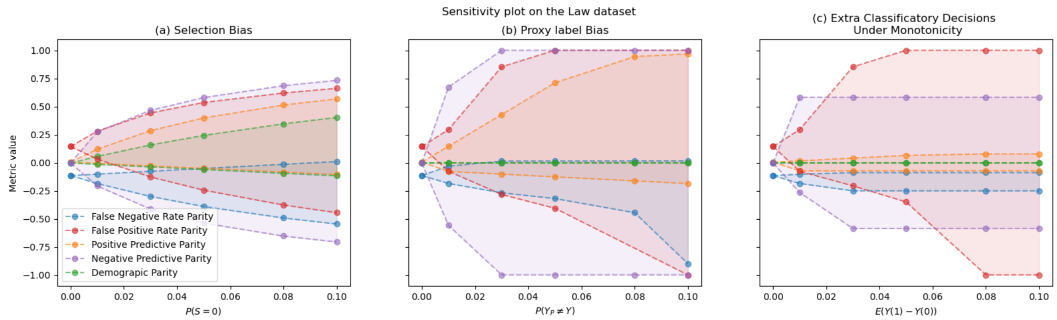

Sensitivity plot on the Compas Recid dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions
Under Monotonicity
False Negative Rate Parity
False Positive Rate Parity
Positive Predictive Parity
Negative Predictive Parity
Demograpic Parity
Metric value
$P(S=0)$
$P(Y_P \neq Y)$
$E(Y(1)-Y(0))$

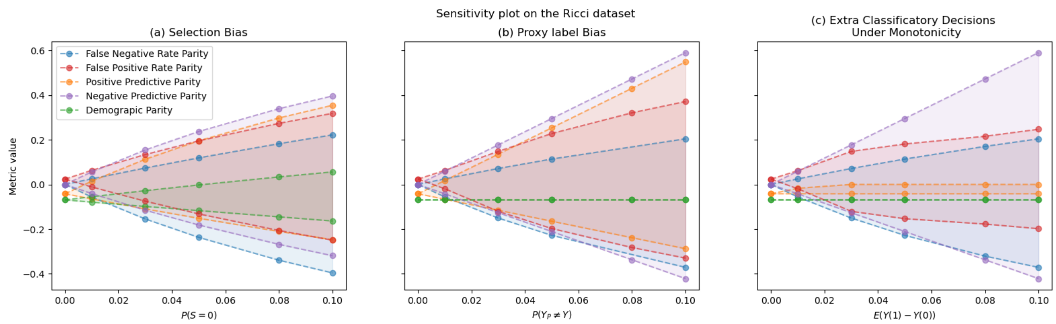

Sensitivity plot on the Compas Viol dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions
Under Monotonicity
False Negative Rate Parity
False Positive Rate Parity
Positive Predictive Parity
Negative Predictive Parity
Demograpic Parity
Metric value
$P(S=0)$
$P(Y_P \neq Y)$
$E(Y(1)-Y(0))$

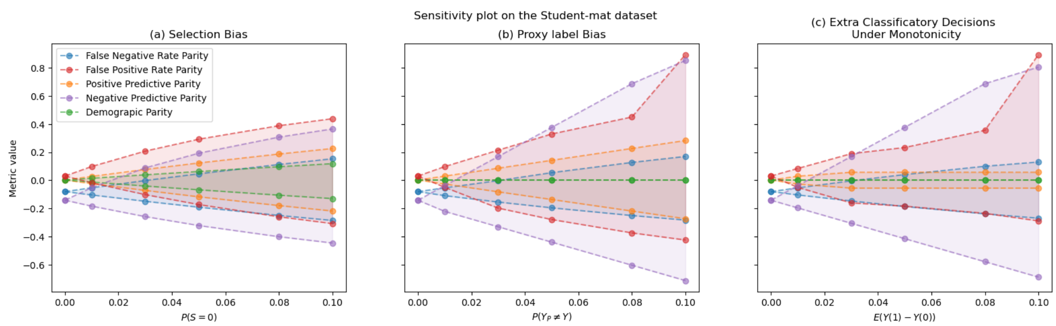

Sensitivity plot on the Credit dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions
Under Monotonicity
False Negative Rate Parity
False Positive Rate Parity
Positive Predictive Parity
Negative Predictive Parity
Demograpic Parity
Metric value
$P(S=0)$
$P(Y_P \neq Y)$
$E(Y(1)-Y(0))$

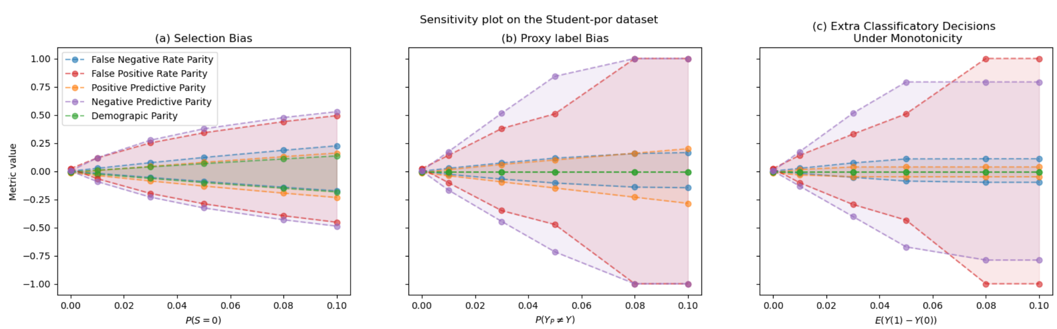

Sensitivity plot on the Dutch dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions
Under Monotonicity
False Negative Rate Parity
False Positive Rate Parity
Positive Predictive Parity
Negative Predictive Parity
Demograpic Parity
Metric value
$P(S=0)$
$P(Y_P \neq Y)$
$E(Y(1)-Y(0))$


Sensitivity plot on the German credit dataset
(a) Selection Bias
(b) Proxy label Bias
(c) Extra Classificatory Decisions
Under Monotonicity
False Negative Rate Parity
False Positive Rate Parity
Positive Predictive Parity
Negative Predictive Parity
Demograpic Parity
Metric value
$P(S=0)$
$P(Y_P \neq Y)$
$E(Y(1)-Y(0))$




## F.2 Impact Statement

This works aims to broaden the discussion of measurement biases in FairML and provide practical tools for practitioners in the area to use. Our hope is that any potential societal consequences of the work will be positive, corresponding to more equitable algorithmic decision making.