Equalized Odds Implies Partially Equalized Outcomes Under Realistic Assumptions

Daniel McNamara
Australian National University and CSIRO Data61
Canberra, ACT, Australia

Abstract

Equalized odds – where the true positive rates and false positive rates are equal across groups (e.g. racial groups) – is a common quantitative measure of fairness. Equalized outcomes – where the difference in predicted outcomes between groups is less than the difference observed in the training data – is more contentious, because it is incompatible with perfectly accurate predictions. We formalize and quantify the relationship between these two important but seemingly distinct notions of fairness. We show that under realistic assumptions, equalized odds implies partially equalized outcomes. We prove a comparable result for approximately equalized odds. In addition, we generalize a well-known previous result about the incompatibility of equalized odds and another definition of fairness known as calibration, by showing that partially equalized outcomes implies non-calibration. Our results highlight the risks of using trends observed across groups to make predictions about individuals.

1 Introduction

Definitions of fairness – and conflicts between them – are an important topic in recent quantitative fairness literature (Barocas, Hardt, and Narayanan 2018). Such definitions often involve avoiding discrimination on the basis of a particular kind of group membership, such as race or gender. In a particular situation, different definitions may be invoked by different stakeholders (Nayaranan 2018).

The controversy associated with the COMPAS recidivism prediction system showed this in practice. The news organization ProPublica claimed that the algorithm was unfair because among non-reoffenders, African-Americans were more likely to be marked high risk than whites, and among reoffenders, whites were more likely to be marked low risk than African-Americans (Angwin et al. 2016) – i.e. the algorithm violated *equalized odds*. The COMPAS response was that the algorithm was not unfair because among those marked high risk, African-Americans were not less likely to reoffend than whites (Dieterich, Mendoza, and Brennan 2016) – i.e. the algorithm satisfied *test-fairness*.

Subsequently it was shown that no algorithm can simultaneously satisfy both equalized odds and test-fairness under realistic assumptions (Chouldechova 2017). A similar result was shown in the more general setting of continuous rather than binary risk scores (Kleinberg, Mullainathan, and Raghavan 2017), replacing test-fairness with a related concept known as *calibration*. Our paper generalizes this latter result by exploring the relationship between *equalized outcomes* and equalized odds, as summarized in Table 1.

Table 1: Summary of main definitions and results.

Definitions
Equalized Odds
True positive rates same for each group
False positive rates same for each group
Partially Equalized Outcomes
Predicted difference between groups less than observed difference between groups
Calibration
Predicted probability equals observed probability for each group and each probability value

Results
Existing
Equalized Odds $\implies$ Not Calibration (Kleinberg, Mullainathan, and Raghavan 2017)
New
Equalized Odds $\implies$ Partially Equalized Outcomes
Partially Equalized Outcomes $\implies$ Not Calibration

1.1 Motivation for Equalized Odds

We motivate *equalized odds*, using recidivism prediction as a running example (see Appendix B for a more extended discussion). Among observed non-reoffenders, we may want to ensure that those from one group are not marked higher risk on average than those from another group, i.e. our false positive rates for both groups are equal. This has been dubbed *equality of opportunity* (Hardt, Price, and Srebro 2016). If we also ensure that among observed reoffenders, those from one group are not marked higher risk on average than those from another group, we have *equalized odds* (Hardt, Price, and Srebro 2016), i.e. our true positive rates for both groups are also equal.[1] It has been observed that in order to make

[1] Equality of true positive rates between groups is equivalent to equality of false negative rates between groups. Equalized odds has

the relative utility of different groups more equal, absolute utility may be reduced (Corbett-Davies and Goel 2018; Corbett-Davies et al. 2017; Menon and Williamson 2018). However, equalized odds has some intuitive appeal as a fairness measure since it ensures that mistakes do not disproportionately impact any group.

1.2 The Debate about Equalized Outcomes

Equality of outcomes between groups is a well-known fairness criterion. It has been mathematically formalized in the quantitative fairness literature through the concepts of *statistical parity* (Calders and Verwer 2010; Dwork et al. 2012), avoiding *disparate impact* (United States Equal Opportunity Employment Commission 1978; Feldman et al. 2015) and achieving *independence* between outcomes and group membership (Barocas, Hardt, and Narayanan 2018). These technical definitions have prompted debate about whether they are suitable measures of fairness.

A critique of equalized outcomes is that if the observed rates (e.g. of recidivism) are different across the two groups in the training data, then an algorithm that reflects this difference is not 'unfair' but is rather a reflection of real underlying differences (Hardt, Price, and Srebro 2016; Zafar et al. 2017). The argument goes: surely we would not want to label 'unfair' a prediction algorithm which is perfectly accurate! The job of the algorithm is to predict the world as it is; changing the world is out of scope.

However, *not* equalizing outcomes across groups creates the risk of discrimination in situations where the data collection process systematically disadvantages one group (Barocas and Selbst 2016; Zafar et al. 2017; O'Neil 2017). For example, profiling of particular populations based on pre-existing risk assessments can distort trends in reoffending. Equalized outcomes may help algorithms to avoid perpetuating this structural inequality. More generally, the question of whether redistribution should be used to reduce inequality is at the core of the left-right political divide (Jæger 2008). As such, the debate on equalized outcomes is unlikely to be definitively won or lost by either side.

1.3 Our Contribution

Our core contribution is to formalize and quantify the relationship between equalized odds and equalized outcomes, two important but seemingly distinct notions of fairness. We quantify the extent to which outcomes are equalized in an intuitive way, via a comparison between the predicted and observed differences between groups (Section 2). We prove that if we want to satisfy equalized odds, we must partially equalize outcomes – even if we only want approximately equalized odds (Section 3). In addition, we generalize a well-known existing result about the incompatibility of equalized odds and a different fairness measure known as *calibration* (Kleinberg, Mullainathan, and Raghavan 2017), using a simpler proof technique (Section 4). Our conclusion (Section 5) highlights why we should consider the reality that algorithmic decisions are imperfect when defining measures of fairness.

also been referred to as avoiding *disparate mistreatment* (Zafar et al. 2017).

2 Problem Formalization

We mathematically formalize the setting we have informally described above. In our problem setup we have input variable $X \in \mathcal{X}$ (e.g. a person's criminal record expressed as a real-valued vector), sensitive variable $S \in \{0,1\}$ encoding group membership (e.g. race coded as 1 for African-American or 0 for non African-American), target variable $Y \in \{0,1\}$ (e.g. ground truth of whether the person reoffended), and decision variable $\hat{Y} \in \{0,1\}$ (e.g. prediction of whether the person will reoffend). Let $h : \mathcal{X} \times \{0,1\} \to [0,1]$ be a stochastic hypothesis, which can also be interpreted as a scoring function.[2] Let $\hat{Y}$ be constructed such that $p(\hat{Y} = 1 | X = x, S = s) := h(x, s)$. While setting S, Y and $\hat{Y}$ to be binary variables is an assumption, this allows us to cover many cases of interest – such as the recidivism prediction example – and facilitates our analysis and interpretation. The choice of input space $\mathcal{X}$ is arbitrary.

Drawing X, S and Y and making decision $\hat{Y}$, we have a joint distribution μ over $\mathcal{X} \times \{0,1\} \times \{0,1\} \times \{0,1\}$. We may also derive marginal distributions over one or more variables, such as the marginal distribution over Y:

$$\mu_Y(Y = y) := \int_{x \in \mathcal{X}} \sum_{s \in \{0,1\}} \sum_{\hat{y} \in \{0,1\}} \mu(X = x, S = s, Y = y, \hat{Y} = \hat{y}) dx.$$

Similarly, we may derive conditional distributions, such as the marginal distribution over $\hat{Y}$ conditioned on $Y = 1$:

$$\mu_{\hat{Y}|Y=1}(\hat{Y} = \hat{y}) := \frac{\mu_{Y,\hat{Y}}(Y = 1, \hat{Y} = \hat{y})}{\mu_Y(Y = 1)}.$$

We use notation of the form $p(Y = y) := \mu_Y(Y = y)$ for marginals and $p(\hat{Y} = \hat{y} | Y = 1) := \mu_{\hat{Y}|Y=1}(\hat{Y} = \hat{y})$ for conditionals. For example, $p(\hat{Y} = 1 | Y = 1)$ is known as the true positive rate (e.g. predicted reoffence rate for reoffenders) and $p(\hat{Y} = 1 | Y = 0)$ is known as the false positive rate (e.g. predicted reoffence rate for non-reoffenders). We use the symbol $\perp$ to denote probabilistic independence between variables.

2.1 Impossibility Results

An impossibility result states several candidate properties of a joint distribution, and shows that *no* distribution can simultaneously satisfy all of these properties. A well-known impossibility result (Theorem 1.1 of Kleinberg, Mullainathan, and Raghavan 2017) considered the relationship between calibration – which requires that for both groups, each risk score accurately reflects the true risk associated with individuals assigned that score – and equalized odds. The result

[2] This underpins our comparisons with (Kleinberg, Mullainathan, and Raghavan 2017), which analyzes risk scores. Interpreting such scores as decision probabilities facilitates our analysis.

showed that it is impossible to simultaneously satisfy both fairness criteria and other realistic assumptions.

Variants exist involving approximate versions of equalized odds (Theorem 1 of Pleiss et al. 2017), calibration or both (Theorem 1.2 of Kleinberg, Mullainathan, and Raghavan 2017). We mentioned earlier the incompatibility of equalized odds and *test-fairness* – where the risk scores are binary and the true risk of individuals with a given score must be the same for both groups (Chouldechova 2017). Simple rules of conditional probability may be used to show that $\hat{Y} \perp S|Y$ – corresponding to equalized odds – and $Y \perp S|\hat{Y}$ – which is closely related to calibration – cannot both simultaneously hold under realistic assumptions (Barocas, Hardt, and Narayanan 2018). The incompatibility of equalized odds and statistical parity (i.e. the independence relationship $\hat{Y} \perp S$) has also been shown (Kleinberg, Mullainathan, and Raghavan 2017; Barocas, Hardt, and Narayanan 2018).

In our work we derive impossibility results involving equalized outcomes and equalized odds, which are of interest given the debates about these fairness criteria described in Section 1. As we shall see in Section 4, our analysis also allows us to generalize Theorem 1.1 of Kleinberg, Mullainathan, and Raghavan 2017, exploiting the relationship between equalized outcomes and calibration.

2.2 Fairness Definitions

We now formalize the definition of equalized odds.

Definition 1 (Equalized odds (Hardt, Price, and Srebro 2016; Zafar et al. 2017))**.** *Equalized odds is satisfied if both of the following hold:*

$$p(\hat{Y} = 1|S = 1, Y = 1) = p(\hat{Y} = 1|S = 0, Y = 1) \quad (1)$$

i.e. the true positive rate is the same for both groups, and

$$p(\hat{Y} = 1|S = 1, Y = 0) = p(\hat{Y} = 1|S = 0, Y = 0) \quad (2)$$

i.e. the false positive rate is the same for both groups.

We now present a novel formalization of equalized outcomes.

Definition 2 (Equalized outcomes)**.** *Let*

$$\begin{aligned} &p(\hat{Y} = 1|S = 1) - p(\hat{Y} = 1|S = 0) \\ &= \alpha(p(Y = 1|S = 1) - p(Y = 1|S = 0)) \end{aligned} \quad (3)$$

where α is a constant we refer to as the equalized outcomes coefficient. If (3) *holds for $\alpha = 0$ we have* fully equalized outcomes*. If* (3) *holds for some $\alpha \in (0, 1)$ we have* partially equalized outcomes*. If* (3) *holds for $\alpha = 1$ we have* non-equalized outcomes*.*

Fully equalized outcomes corresponds to the well-known definition of *statistical parity* (Calders and Verwer 2010; Dwork et al. 2012), or equivalently the *independence* $\hat{Y} \perp S$ (Barocas, Hardt, and Narayanan 2018). The value of introducing the parameter α is that we quantify the extent to which outcomes are equalized in an intuitive way, via a comparison with the observed difference between groups. Under partially equalized outcomes, the predicted difference between groups is smaller than the observed difference between groups. Non-equalized outcomes means that predicted outcomes are *faithful* to the observed difference in outcomes between groups. If $\alpha > 1$ the predicted difference amplifies the observed difference, while if $\alpha < 0$ the predicted difference flips the sign of the observed difference. These options do not appear advantageous in terms of either fairness or accuracy, and we do not focus on them.

2.3 Realistic Assumptions

We now introduce three realistic assumptions which we use in some parts of our analysis (we flag when each assumption is being used). The first assumption is that the observed rates (e.g. of recidivism) are different across groups, which is true for most cases of interest.

Assumption 1 (Different observed rates)**.**

$$p(Y = 1|S = 1) \neq p(Y = 1|S = 0) \quad (4)$$

The other two assumptions are that our decisions are *imperfect* (i.e. they are not always accurate) and *non-vacuous* (i.e. they have some predictive power). This covers the bulk of realistic situations in which algorithmic decisions are used. We observe that the imperfect decisions assumption will hold if Y cannot be expressed as a deterministic function of X and S. In this case, changing $\hat{Y}$ will not help. This is typically the case when we are making predictions about the future actions of individuals.

Assumption 2 (Imperfect decisions)**.** *At least one of the following holds:*

$$p(\hat{Y} = 1|Y = 0) > 0 \quad (5)$$

i.e. some negative examples are misclassified, or

$$p(\hat{Y} = 1|Y = 1) < 1 \quad (6)$$

i.e. some positive examples are misclassified.

Assumption 3 (Non-vacuous decisions)**.**

$$p(\hat{Y} = 1|Y = 1) > p(\hat{Y} = 1|Y = 0) \quad (7)$$

i.e. the decision is more likely to be positive for positive examples than for negative examples.

3 The Relationship Between Equalized Odds and Equalized Outcomes

Assuming equalized odds is satisfied, we show there is a quantifiable trade-off between accuracy and the extent to which outcomes are equalized. As a corollary, we show that equalized odds implies partially equalized outcomes. We consider the cases where equalized odds either exactly or approximately holds.[3]

[3]While the exact version is a special case of the approximate version, we consider the exact case first as it makes the presentation of the results more intuitive.

3.1 Perfectly Equalized Odds

We show in Theorem 1 that given perfectly equalized odds, the extent to which we equalize outcomes is given by the difference α between the true positive rate and false positive rate. This novel result is of interest because it precisely quantifies the relationship between the well-known but seemingly distinct notions of equalized odds and equalized outcomes.

Theorem 1 (Equalized outcomes given equalized odds)**.** *Let*

$$\alpha := p(\hat{Y} = 1|Y = 1) - p(\hat{Y} = 1|Y = 0).$$

Suppose (1) *and* (2) *hold, i.e. equalized odds is satisfied. Then* (3) *is satisfied, i.e.* α *is the equalized outcomes coefficient satisfying*

$$\begin{aligned} &p(\hat{Y} = 1|S = 1) - p(\hat{Y} = 1|S = 0) \\ &= \alpha(p(Y = 1|S = 1) - p(Y = 1|S = 0)). \end{aligned}$$

Proof. We have

$$\begin{aligned} &p(\hat{Y} = 1|S = 1) \\ &\quad = p(Y = 1|S = 1)p(\hat{Y} = 1|S = 1, Y = 1) \\ &\quad\quad + p(Y = 0|S = 1)p(\hat{Y} = 1|S = 1, Y = 0) \quad (8) \end{aligned}$$

and

$$\begin{aligned} &p(\hat{Y} = 1|S = 0) \\ &\quad = p(Y = 1|S = 0)p(\hat{Y} = 1|S = 0, Y = 1) \\ &\quad\quad + p(Y = 0|S = 0)p(\hat{Y} = 1|S = 0, Y = 0) \quad (9) \end{aligned}$$

by the law of total probability.
Applying (1) and (2) to (8) yields

$$\begin{aligned} p(\hat{Y} = 1|S = 1) = p(Y = 1|S = 1)p(\hat{Y} = 1|Y = 1) \\ + p(Y = 0|S = 1)p(\hat{Y} = 1|Y = 0) \quad (10) \end{aligned}$$

and similarly, applying (1) and (2) to (9) yields

$$\begin{aligned} p(\hat{Y} = 1|S = 0) = p(Y = 1|S = 0)p(\hat{Y} = 1|Y = 1) \\ + p(Y = 0|S = 0)p(\hat{Y} = 1|Y = 0). \quad (11) \end{aligned}$$

Subtracting (11) from (10) and using the definition of α, we have

$$\begin{aligned} &p(\hat{Y} = 1|S = 1) - p(\hat{Y} = 1|S = 0) \\ &\quad = \alpha(p(Y = 1|S = 1) - p(Y = 1|S = 0)). \quad (12) \end{aligned}$$

□

As a consequence of Theorem 1 and our realistic assumptions, if we have perfectly equalized odds then we have partially equalized outcomes, as shown in Corollary 1. While as we mentioned above the incompatibility of equalized odds and fully equalized outcomes (i.e. $\alpha = 0$, *statistical parity* or *independence*) was already known, we are the first to show that equalized odds is also incompatible with non-equalized outcomes ($\alpha = 1$) or indeed any value of α outside the interval $(0, 1)$ under our realistic assumptions.

Corollary 1 (Equalized odds implies partially equalized outcomes under realistic assumptions)**.** *Suppose* (1) *and* (2) *hold, i.e. we have equalized odds. Suppose also that Assumptions 1, 2 and 3 hold. Then satisfying* (3) *requires* $\alpha \in (0, 1)$*, i.e. we have partially equalized outcomes.*

Proof. By Theorem 1 we know that given equalized odds, (3) is satisfied for $\alpha = p(\hat{Y} = 1|Y = 1) - p(\hat{Y} = 1|Y = 0)$. Applying Assumption 1 (different observed rates), this is the *only* value of α satisfying (3). Applying Assumption 2 (imperfect decisions) we have $\alpha < 1$. Applying Assumption 3 (non-vacuous decisions) we have $\alpha > 0$. The result follows. □

3.2 Approximately Equalized Odds

We consider a relaxation of the equalized odds condition, allowing the false positive rates to slightly differ across groups and the false negative rates to likewise slightly differ across groups. The parameter δ quantifies the degree of this relaxation, with $\delta = 0$ corresponding to perfectly equalized odds.

Definition 3 (Approximately equalized odds)**.** *For some constant* $\delta \geq 0$ *we have*

$$\begin{aligned} &p(\hat{Y} = 1|S = 1, Y = 1), p(\hat{Y} = 1|S = 0, Y = 1) \in \\ &[(1 - \delta)p(\hat{Y} = 1|Y = 1), (1 + \delta)p(\hat{Y} = 1|Y = 1)] \quad (13) \end{aligned}$$

i.e. the true positive rate is approximately the same for both groups.
We also have

$$\begin{aligned} &p(\hat{Y} = 1|S = 1, Y = 0), p(\hat{Y} = 1|S = 0, Y = 0) \in \\ &[(1 - \delta)p(\hat{Y} = 1|Y = 0), (1 + \delta)p(\hat{Y} = 1|Y = 0)] \quad (14) \end{aligned}$$

i.e. the false positive rate is approximately the same for both groups.

In Theorem 2 we show that if δ-approximately equalized odds is satisfied, then the extent to which we equalize outcomes is given by an interval. This midpoint of the interval is determined by the difference α between the true positive rate and false positive rate. The size of the interval is determined by δ and a distribution-dependent parameter β.

Theorem 2 (Equalized outcomes given approximately equalized odds)**.** *Let*

$$\alpha := p(\hat{Y} = 1|Y = 1) - p(\hat{Y} = 1|Y = 0),$$

$$\epsilon := p(Y = 1|S = 1) + p(Y = 1|S = 0)$$

and

$$\beta := \epsilon p(\hat{Y} = 1|Y = 1) + (2 - \epsilon)p(\hat{Y} = 1|Y = 0).$$

Observe that $\beta \geq 0$*. Suppose* (13) *and* (14) *hold, i.e.* δ*-approximately equalized odds is satisfied. Then*

$$\begin{aligned} &p(\hat{Y} = 1|S = 1) - p(\hat{Y} = 1|S = 0) \in \\ &[\alpha(p(Y = 1|S = 1) - p(Y = 1|S = 0)) - \delta\beta, \\ &\alpha(p(Y = 1|S = 1) - p(Y = 1|S = 0)) + \delta\beta]. \end{aligned}$$

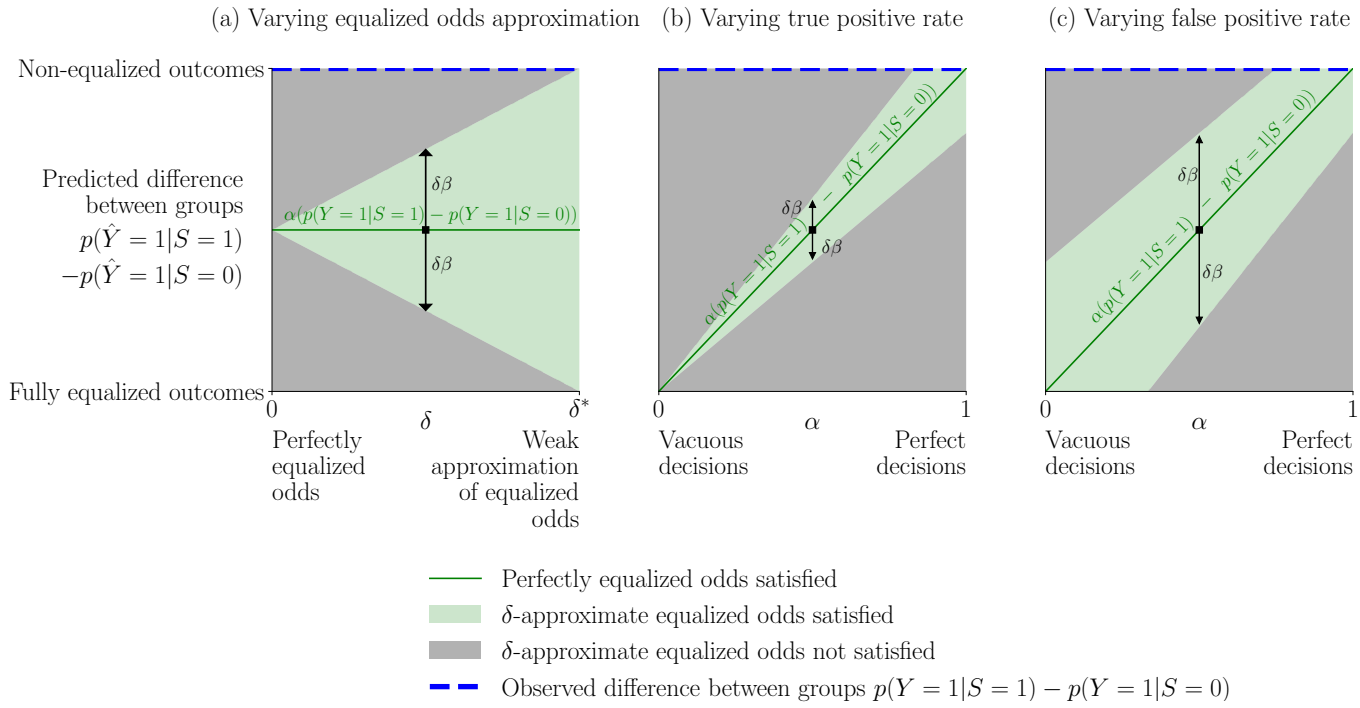

Figure 1: Visualization of our key results. Certain combinations of equalized outcomes, equalized odds and accuracy are possible (light green regions), while other combinations are impossible (dark gray regions). In (a) we vary equalized odds approximation parameter δ, fixing accuracy parameter $\alpha := p(\hat{Y}=1|Y=1) - p(\hat{Y}=1|Y=0)$. In (b) we vary the $p(\hat{Y}=1|Y=1)$ term in α and in (c) we vary the $p(\hat{Y}=1|Y=0)$ term in α, fixing δ. β is a distribution-dependent parameter (see Theorem 2).

Proof idea. As in Theorem 1, use the law of total probability to express $p(\hat{Y}=1|S=1)$ and $p(\hat{Y}=1|S=0)$. Then apply the δ-approximately equalized odds assumption to upper and lower bound their difference. See Appendix A for complete proof. □

3.3 Interpretation

We visualize our results in Figure 1. In each plot the y-axis shows the predicted difference between groups, i.e. the extent to which outcomes are equalized, on a scale from zero (bottom) to the observed difference between groups (top). We vary other parameters along the x-axes of the plots.

If perfectly equalized odds is satisfied there is an exact relationship between equalized outcomes and α (see Theorem 1, green line on plots). If δ-approximate equalized odds is satisfied there is a region of permissible combinations of equalized outcomes and α values (see Theorem 2, light green region on plots). Combinations outside this region violate δ-approximate equalized odds (dark gray region on plots).

In Figure 1(a), we see that if we relax the constraint on equalized odds by increasing the parameter δ (see Definition 3), we have a larger region of possible combinations.[4] The size of this region is quantified by the slack term $\delta\beta$. The region is an interval centred on the product of α and the observed difference between groups. We see visually why for $\alpha \in (0,1)$, i.e. for decisions that are imperfect and non-vacuous, we have partially equalized outcomes.

Figures 1(b) and 1(c) show that if we have equalized odds, then increasing accuracy (measured by α) moves towards non-equalized outcomes.[5] We may increase α by increasing the true positive rate, as in Figure 1(b), where we assume no false positives. We may also increase α by decreasing the false positive rate, as in Figure 1(c), where we assume no false negatives. Under perfectly equalized odds the effect on equalized outcomes is the same, while under approximately equalized odds the permissible regions differ because β depends on the false positive rate and the true positive rate.

4 Generalization of Calibration-Equalized Odds Impossibility Result

The relationship between equalized odds and equalized outcomes, in addition to its intrinsic interest, allows us to generalize a well-known result about the impossibility of simultaneously satisfying calibration and equalized odds (Theorem 1.1 of Kleinberg, Mullainathan, and Raghavan 2017). We use a proof technique involving elementary probabilities, which also provides a simpler proof of the previous result.

[4] The value of δ for which the edge of this region intersects the x-axis is given by $\delta^* := \frac{\alpha}{\beta}(p(Y=1|S=1) - p(Y=1|S=0))$. Figure 1(a) uses the fixed parameters $\alpha := 0.5$, $\beta := 1$ and $\epsilon := 1$.

[5] Figures 1(b) and 1(c) use the fixed parameters $\epsilon := 1$ and $\delta := 0.2(p(Y=1|S=1) - p(Y=1|S=0))$.

4.1 Review of Existing Result

We first introduce the definition of group-conditional calibration proposed in previous work (Kleinberg, Mullainathan, and Raghavan 2017; Pleiss et al. 2017). This means that for both groups, each risk score equals the observed risk associated with individuals assigned that score.

Definition 4 (Group-conditional calibration (Kleinberg, Mullainathan, and Raghavan 2017; Pleiss et al. 2017))**.** *Both of the following statements hold* $\forall c \in [0, 1]$*:*

$$p(Y = 1|h(x, s) = c, S = 1) = c \tag{15}$$

$$p(Y = 1|h(x, s) = c, S = 0) = c \tag{16}$$

We now state the well-known calibration-equalized odds impossibility result (Theorem 1.1 of Kleinberg, Mullainathan, and Raghavan 2017, restated to align with our definitions).

Theorem 3 (Calibration-equalized odds impossibility result (Kleinberg, Mullainathan, and Raghavan 2017))**.** *Suppose* (1)*,* (2)*,* (15) *and* (16) *hold, i.e. equalized odds and group-conditional calibration are both satisfied. Then at least one of Assumption 1 or Assumption 2 is violated, i.e. the observed rates are the same for both groups and/or the decision is perfect.*

In other words, equalized odds implies not calibration under realistic assumptions, as stated in Table 1.

4.2 Group-Conditional Calibration Implies Non-Equalized Outcomes

In preparation for generalizing Theorem 3, we show in Lemma 1 that group-conditional calibration implies non-equalized outcomes but not vice versa. Using the contrapositive of the fact that group-conditional calibration implies non-equalized outcomes, partially equalized outcomes implies not calibration as stated in Table 1. We observe that in contrast to group-conditional calibration, test-fairness as proposed in Chouldechova 2017 does not in general imply non-equalized outcomes.

Lemma 1 (Group-conditional calibration implies non-equalized outcomes but not vice versa)**.** *If* (15) *and* (16) *hold, then* (3) *holds for* $\alpha = 1$*, i.e. group-conditional calibration implies non-equalized outcomes.*

However, if (3) *holds for* $\alpha = 1$*, then it is not the case that* (15) *and* (16) *must hold, i.e. non-equalized outcomes does not imply group-conditional calibration.*

Proof idea. Use laws of probability to show that group-conditional calibration implies non-equalized outcomes. Then construct a single example to show that non-equalized outcomes does not imply group-conditional calibration. See Appendix A for complete proof. □

4.3 The Generalized Result

The existing result stated in Theorem 3 shows that if group-conditional calibration and equalized odds hold, realistic assumptions are violated. Our new result in Theorem 4 shows that if non-equalized outcomes and equalized odds hold, the same realistic assumptions are violated. As we just showed in Lemma 1, group-conditional calibration implies non-equalized outcomes but not vice versa, i.e. non-equalized outcomes is a weaker condition than group-conditional calibration. Therefore Theorem 4 is more general than Theorem 3, since with a weaker condition we arrive at the same conclusion. It is straightforward to see that Lemma 1 and Theorem 4 together imply Theorem 3. We observe that our proof technique appears simpler, since it relies only on elementary manipulation of probabilities.

Theorem 4 (Generalization of calibration-equalized odds impossibility result)**.** *Suppose* (1) *and* (2) *hold, and* (3) *holds for* $\alpha = 1$*, i.e. equalized odds and non-equalized outcomes are both satisfied. Then at least one of Assumption 1 or Assumption 2 is violated, i.e. the observed rates are the same for both groups and/or the decision is perfect.*

Proof. Suppose (3) holds for $\alpha = 1$, i.e. non-equalized outcomes is satisfied.

Suppose (1) and (2) hold, i.e. equalized odds is satisfied. Applying Theorem 1,

$$\begin{aligned} &p(\hat{Y} = 1|S = 1) - p(\hat{Y} = 1|S = 0) \\ &= (p(\hat{Y} = 1|Y = 1) - p(\hat{Y} = 1|Y = 0)) \\ &\times (p(Y = 1|S = 1) - p(Y = 1|S = 0)). \end{aligned} \tag{17}$$

Combining (3) and (17), we have

$$\begin{aligned} &p(Y = 1|S = 1) - p(Y = 1|S = 0) \\ &= (p(\hat{Y} = 1|Y = 1) - p(\hat{Y} = 1|Y = 0)) \\ &\times (p(Y = 1|S = 1) - p(Y = 1|S = 0)). \end{aligned} \tag{18}$$

We conclude from (18) that at least one of the following holds:

$$p(Y = 1|S = 1) = p(Y = 1|S = 0) \tag{19}$$

$$p(\hat{Y} = 1|Y = 1) - p(\hat{Y} = 1|Y = 0) = 1 \tag{20}$$

If (19) holds then Assumption 1 is violated, i.e. the observed rates are the same for both groups. If (20) holds, then $p(\hat{Y} = 1|Y = 1) = 1$ and $p(\hat{Y} = 1|Y = 0) = 0$. Therefore Assumption 2 is violated, i.e. the decision is perfect. □

5 Conclusion

When algorithms make predictions of the future actions of individuals, a certain degree of inaccuracy seems inevitable. In this context, naively using trends observed across groups to make predictions about individuals – a problem known as group-to-individual inference (Fisher, Medaglia, and Jeronimus 2018) – creates the risk of unfairness, in the legal system and beyond. We have formalized the intuition that when algorithms conduct group-to-individual inference – or in other words, *stereotype* – they tend to be unfair to individuals who are 'atypical' (e.g. non-reoffenders from a group with higher reoffence rates). In particular, we have seen that an imperfect algorithm for which the predicted and observed differences between groups are equal will violate equalized odds. Avoiding this requires partially equalized outcomes, which can be seen as an instantiation of 'algorithmic affirmative action' (Chander 2016).

Acknowledgments I would like to thank the reviewers for their useful feedback. I would also like to thank Bob Williamson and Cheng Soon Ong for discussions during the development of this work. The research was supported by an Australian Government Research Training Program Scholarship and a CSIRO Data61 Top-Up Scholarship.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. ProPublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, S., and Selbst, A. D. 2016. Big Data's Disparate Impact. *California Law Review* 104:671.

Barocas, S.; Hardt, M.; and Narayanan, A. 2018. *Fairness and Machine Learning*. fairmlbook.org.

Calders, T., and Verwer, S. 2010. Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Mining and Knowledge Discovery* 21(2).

Chander, A. 2016. The Racist Algorithm. *Michigan Law Review* 115.

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5(2).

Corbett-Davies, S., and Goel, S. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv*.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Dieterich, W.; Mendoza, C.; and Brennan, T. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. *Northpointe Inc.*

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*.

Edwards, H., and Storkey, A. 2016. Censoring Representations with an Adversary. *International Conference on Learning Representations*.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Fisher, A. J.; Medaglia, J. D.; and Jeronimus, B. F. 2018. Lack of Group-to-Individual Generalizability is a Threat to Human Subjects Research. *Proceedings of the National Academy of Sciences*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*.

Jæger, M. M. 2008. Does Left–Right Orientation have a Causal Effect on Support for Redistribution? Causal Analysis with Cross-Sectional Data using Instrumental Variables. *International Journal of Public Opinion Research* 20(3).

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. *Proceedings of Innovations in Theoretical Computer Science*.

Menon, A. K., and Williamson, R. C. 2018. The Cost of Fairness in Binary Classification. In *Conference on Fairness, Accountability and Transparency*.

Nayaranan, A. 2018. Tutorial: 21 Fairness Definitions and their Politics. Conference on Fairness, Accountability and Transparency, https://www.youtube.com/watch?v=jIXIuYdnyyk.

O'Neil, C. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*.

United States Equal Opportunity Employment Commission. 1978. Uniform Guidelines on Employee Selection Procedures.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*.

A Supplementary Proofs

We present the proofs of Theorem 2 and Lemma 1.

A.1 Proof of Theorem 2 (Equalized Outcomes Given Approximately Equalized Odds)

Proof. We have

$$p(\hat{Y}=1|S=1) = p(Y=1|S=1)p(\hat{Y}=1|S=1,Y=1)+ p(Y=0|S=1)p(\hat{Y}=1|S=1,Y=0)$$

and

$$p(\hat{Y}=1|S=0) = p(Y=1|S=0)p(\hat{Y}=1|S=0,Y=1)+ p(Y=0|S=0)p(\hat{Y}=1|S=0,Y=0)$$

by the law of total probability.

Assuming δ-approximately equalized odds, we have

$$\begin{aligned}
&p(\hat{Y}=1|S=1) - p(\hat{Y}=1|S=0)\\
&\leq p(Y=1|S=1)(1+\delta)p(\hat{Y}=1|Y=1)\\
&\quad + p(Y=0|S=1)(1+\delta)p(\hat{Y}=1|Y=0)\\
&\quad - p(Y=1|S=0)(1-\delta)p(\hat{Y}=1|Y=1)\\
&\quad - p(Y=0|S=0)(1-\delta)p(\hat{Y}=1|Y=0)\\
&= \alpha(p(Y=1|S=1) - p(Y=1|S=0)) + \delta\beta.
\end{aligned}$$

The equality follows by rearranging the terms and using the definitions of α and β.

Similarly,

$$\begin{aligned}
&p(\hat{Y}=1|S=1) - p(\hat{Y}=1|S=0)\\
&\geq p(Y=1|S=1)(1-\delta)p(\hat{Y}=1|Y=1)\\
&\quad + p(Y=0|S=1)(1-\delta)p(\hat{Y}=1|Y=0)\\
&\quad - p(Y=1|S=0)(1+\delta)p(\hat{Y}=1|Y=1)\\
&\quad - p(Y=0|S=0)(1+\delta)p(\hat{Y}=1|Y=0)\\
&= \alpha(p(Y=1|S=1) - p(Y=1|S=0)) - \delta\beta.
\end{aligned}$$

□

A.2 Proof of Lemma 1 (Group-Conditional Calibration Implies Non-Equalized Outcomes but Not Vice Versa)

Proof. Suppose (15) and (16) hold, i.e. group-conditional calibration is satisfied. Then

$$\begin{aligned}
&p(Y=1|S=1) - p(Y=1|S=0)\\
&= \int_0^1 p(h(x,s)=c|S=1)p(Y=1|h(x,s)=c,S=1)\,dc\\
&\quad - \int_0^1 p(h(x,s)=c|S=0)p(Y=1|h(x,s)=c,S=0)\,dc
\end{aligned}$$

by the law of total probability

$$\begin{aligned}
&= \int_0^1 p(h(x,s)=c|S=1)c\,dc\\
&\quad - \int_0^1 p(h(x,s)=c|S=0)c\,dc
\end{aligned}$$

by group-conditional calibration, substituting in (15) and (16)

$$\begin{aligned}
&= \int_0^1 p(h(x,s)=c|S=1)p(\hat{Y}=1|h(x,s)=c,S=1)\,dc\\
&\quad - \int_0^1 p(h(x,s)=c|S=0)p(\hat{Y}=1|h(x,s)=c,S=0)\,dc
\end{aligned}$$

by the definition $p(\hat{Y}=1|X=x,S=s) := h(x,s)$

$$= p(\hat{Y}=1|S=1) - p(\hat{Y}=1|S=0)$$

by the law of total probability. Hence (3) holds for $\alpha = 1$ and we have shown that group-conditional calibration implies non-equalized outcomes.

However, we may have non-equalized outcomes without group-conditional calibration. For example, suppose

$$h(x,s) := p(Y=1|S=s) + \eta$$

where η is generated by random noise with range $[-p(Y=1|S=s), 1-p(Y=1|S=s)]$ and mean zero.

Therefore

$$p(\hat{Y}=1|S=1) = p(Y=1|S=1)$$

and

$$p(\hat{Y}=1|S=0) = p(Y=1|S=0).$$

Hence (3) holds for $\alpha = 1$, i.e. we have non-equalized outcomes.

We also have $\forall c \in [0,1]$

$$p(Y=1|h(x,s)=c,S=1) = p(Y=1|S=1)$$

and

$$p(Y=1|h(x,s)=c,S=0) = p(Y=1|S=0).$$

Hence (15) and (16) do not in general hold and we have shown that non-equalized outcomes does not imply group-conditional calibration. □

B Motivating Examples of Equalized Odds and its Relationship to Equalized Outcomes

We present a motivating example for equalized odds using recidivism prediction. We also present an example which motivates the relationship between equalized odds and equalized outcomes.

Table 2: Test set results for a recidivism prediction model on the ProPublica dataset. The example motivates equalized odds.

Metric	African-American	Non African-American	Overall
Observed reoffence rate	54.1%	39.9%	47.2%
Predicted reoffence rate	55.2%	40.7%	48.1%
Predicted reoffence rate among non-reoffenders	47.8%	36.2%	41.4%
Predicted reoffence rate among reoffenders	61.5%	47.4%	55.7%

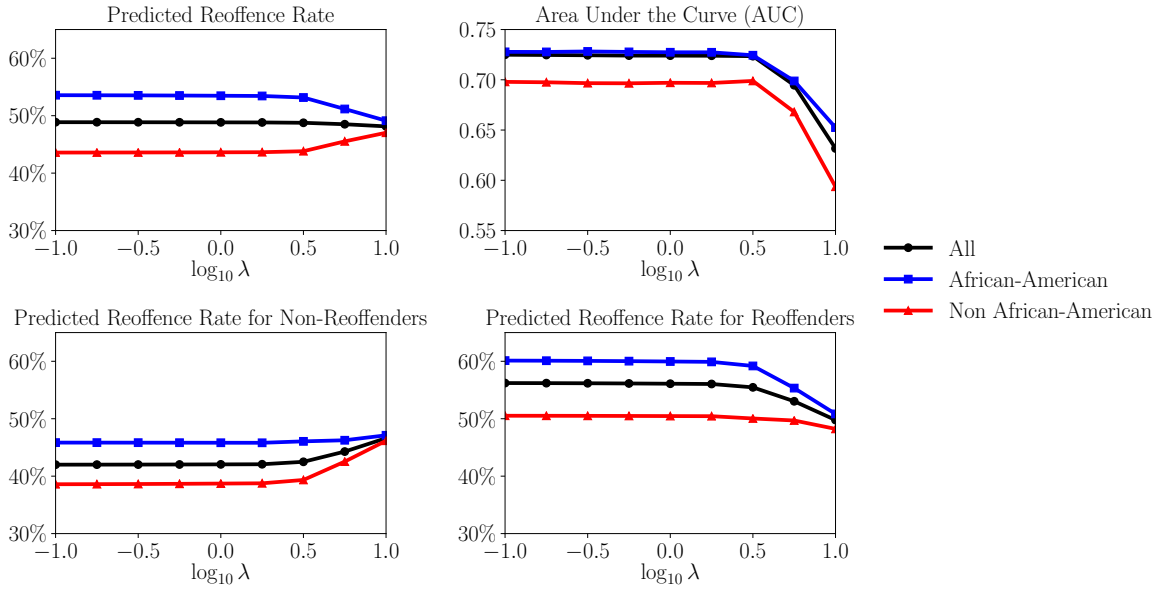

Figure 2: Motivating example: equalized odds appears related to equalized outcomes. The x-axis shows the parameter λ used in pre-processing (see text) on a log scale, while the y-axis shows several performance measures of interest.

B.1 Equalized Odds

We explore the definition of and rationale behind *equalized odds*, using recidivism prediction with the ProPublica dataset, which contains information about 7214 criminal offences committed in Broward County, Florida. We used the individual's age, gender, race and criminal history to predict whether they would reoffend within two years.[6] We applied a 70/30 training/test split of the data, trained a logistic regression model[7] on the training set, and used this model to predict the probability that each individual in the test set would reoffend.

The model achieved an area under the curve (AUC) of 0.72 on the test set, indicating that the model is far from perfect but a lot better than a random guess.[8] The results are shown in Table 2. We note there is a difference in the observed reoffence rates between African-American and non African-American individuals in the data. The predicted reoffence rates are close to the observed reoffence rates for both groups, and thus show a difference of a similar magnitude. The model rates African-American individuals as higher risk on average, but one could justify this by arguing that the model simply reflects trends in the data.

However, looking separately at those individuals who were observed as non-reoffenders, and those who were observed as reoffenders, tells a different story. Looking at the non-reoffenders, for African-Americans the predicted reoffence rate is 47.8% while for non African-Americans it is 36.2%. In other words, the *false positive rate* is much higher for African-Americans than for non African-Americans. Now looking only at the reoffenders, we notice a difference in the *true positive rate* across racial groups – for African-Americans the predicted reoffence rate is 61.5% while for non African-Americans it is 47.4%. Equivalently, the *false negative rate* for non African-Americans (52.6%) is much higher than for African-Americans (38.5%).

Among non-reoffenders, non African-Americans are better off since they are less likely to be incorrectly classified as high risk. Among reoffenders, non African-Americans are

[6]The dataset is available at `https://github.com/propublica/compas-analysis/blob/master/compas-scores.csv`. We predicted the column `is_recid` using `sex`, `age_cat`, `juv_fel_count`, `juv_misd_count`, `juv_other_count` `priors_count` and `c_charge_degree`, representing categorical variables as a one-hot encoding.

[7]Implemented in Python using the `sklearn` package.

[8]AUC can be interpreted as the probability that a randomly selected positive example will receive a higher score than a randomly selected negative example. A perfect classifier achieves an AUC of 1, while a random classifier achieves an AUC of 0.5.

also better off since they are more likely to be incorrectly classified as low risk. These two types of discrimination are precisely what ProPublica reported about the COMPAS algorithm (Angwin et al. 2016). Our example shows how easily this can occur, even if on the face of it the model seems to just reflect differences between two groups in its training data. It also shows how individuals are impacted by inferences made from past observations of others who appear similar to them – in effect they are stereotyped by the algorithm.

In summary, our example shows how a model's true positive rates and false positive rates may differ across groups, which may disadvantage a particular group. This observation motivates the definition of equalized odds – the true positive rates and false positive rates are equal across groups – which, if satisfied, prevents this form of disadvantage (Hardt, Price, and Srebro 2016).

B.2 The Relationship between Equalized Odds and Equalized Outcomes

Continuing with our ProPublica dataset example, we ask whether our findings – that our observed and predicted reoffence rates were close for both groups, and that we violated equalized odds – are quirks of this particular algorithm or dataset? As our theoretical results have shown, this is far from a coincidence – in fact, under realistic assumptions this combination is inevitable!

The core contribution of our work is to formalize the relationship between equalized odds and equalized outcomes. To provide intuition on this relationship, we pre-processed the ProPublica data to suppress information about race using a technique proposed in Edwards and Storkey 2016. The technique is governed by a parameter λ – increasing this parameter changes the data to make it harder to distinguish between the records of African-Americans and non African-Americans.[9] We then ran logistic regression (as in Section B.1) on the pre-processed data and reported results on the test set, as shown in Figure 2.

This technique yielded more *equalized outcomes* with increasing λ, i.e. the predicted reoffence rates for African-Americans and non African-Americans became closer (top left). The accuracy of the model as measured by AUC declined somewhat with increasing λ (top right). The predicted reoffence rates for non-reoffenders became closer for the two groups with increasing λ (bottom left). The predicted reoffence rates for reoffenders for the two groups also became closer (bottom right). In other words, we achieved a tighter approximation of equalized odds with increasing λ. Are these trends in equalized outcomes, accuracy and equalized odds a coincidence? Can we have equalized odds without equalizing outcomes? What about equalized odds with fully equalized outcomes?

In summary, our example shows anecdotal evidence of a relationship between equalized odds and equalized outcomes, and raises questions about whether this relationship has a mathematical foundation. The technical results in this paper address these questions.

[9] More specifically, we learn a map which is applied to each data point to transform it. In learning this map, we optimize an objective function which jointly depends on how well the transformed data approximates the original data, and how well an adversary can estimate a particular attribute (in this case race) from it. The latter is more important for larger λ. We observe that simply omitting the attribute is not sufficient, since it may be possible to infer the attribute from other columns.