the Effects of Calibration in Fairness

Kimia Kazemian KK983@CORNELL.EDU

Abstract

Algorithmic fairness is an important topic as predictive algorithms, including those based on artificial intelligence, become increasingly prevalent. There are a variety of fairness measures to choose from, but calibrated uncertainty estimates are often necessary for accurate predictions. In this study, we analyze the interplay between calibration and fairness measures and found that it is possible to achieve fairness while maintaining calibration through proper thresholding of the non-binary scores associated with classification tasks. Our findings highlight the complexity of balancing calibrated performance with fairness considerations in machine learning applications and the importance of carefully selecting the threshold in order to achieve the desired fairness metrics.

1. Related Work

Algorithmic fairness is receiving significant attention due to the increasing use of predictive algorithms, including those based on artificial intelligence. In a binary classification task, well-calibration means that individuals assigned score s must have probability s of belonging to the positive class, regardless of group membership. However, in practice, the score is often continuous-valued and must be binarized through a thresholding operation to generate a binary prediction. This raises questions about the interplay between calibration and fairness metrics, predictive parity, equalized odds, and statistical parity.

Previous research has shown that equalized odds, defined as the condition in which both groups have the same average binary score, may not be compatible with well-calibration when there are unequal base rates and imperfect prediction Kleinberg et al. (2016). One proposed solution is to weaken the definition of equalized odds in order to achieve calibration Pleiss et al. (2017), but this approach has limitations, particularly in sensitive cases such as healthcare. This approach involves post-processing algorithms that withhold labels for a portion of the data, which may not be practical or desirable in certain situations. There has been some further work on a reconciliation Reich and Vijaykumar (2020) under some conditions.

There is research that shows incorporating fairness metrics may not always improve net benefit compared to calibrated models followed by thresholding Pfohl et al. (2022). However, Garg et al. (2020) has shown that through proper thresholding, well-calibration is not necessarily incompatible with statistical parity. This issue has not been studied in greater depth.

In this paper, we propose a more concrete analysis of the interplay between calibration and fairness metrics, specifically focusing on proper thresholding of calibrated scores in a binary setting as proposed by Garg et al. (2020) while considering the complexities involved in working with continuous and binary domains. We show it is possible to use proper thresholding of calibrated scores to satisfy equalized odds and statistical parity. We highlight the nuances between calibration and predictive parity and give specific obtainable threshold to achieve predictive parity given calibrated

scores. We also discuss the need to identify the appropriate threshold on a given dataset, and the potential implications of relaxing definitions of fairness metrics.

2. Problem setting

2.1. fairness metrics

equalized odds: A predictor satisfies *equalized odds* if both the true positive rate (TPR) and (separately) the false positive rate (FPR) are the same across groups. More formally, equalized odds requires that the group-specific TPR satisfy $p(\widetilde{y} = 1|y = 1, G = 0) = p(\widetilde{y} = 1|y = 1, G = 1)$ and that the group-specific FPR satisfy $p(\widetilde{y} = 1|y = 0, G = 0) = p(\widetilde{y} = 1|y = 0, G = 1)$.

predictive parity: we consider that *predictive parity* is satisfied when the positive predictive value(PPV) is the same for both groups. PPV is defined as the probability that individuals *predicted* to belong to the positive class *actually* belong to the positive class. Mathematically, predictive parity therefore requires $p(y = 1|\widetilde{y} = 1, G = 0) = p(y = 1|\widetilde{y} = 1, G = 1)$.

We note that some authors define predictive parity in a more constrained manner, requiring not only parity for PPV, but also for its counterpart, negative predictive value (NPV), which requires additionally satisfying $p(y = 0|\widetilde{y} = 0, G = 0) = p(y = 0|\widetilde{y} = 0, G = 1)$. The terms "overall predictive parity" or "conditional use accuracy equality" have been usedtto describe a predictor with equality across groups in both PPV and NPV.

Statistical parity: sometimes referred to as *group fairness* or *demographic parity*) is achieved when members of both groups are predicted to belong to the positive class at the same rate. Mathematically, this means satisfying $p(\widetilde{y} = 1|G = 0) = p(\widetilde{y} = 1|G = 1)$. Notably, this metric gives no consideration to the outcomes y. Therefore, when the base rates $p(y|G)$ differ across the two groups, statistical parity rules out the perfect predictor.

To provide an initial framing, it is interesting to note that using the basic probability relation $p(A, B) = p(A|B)p(B) = p(B|A)p(A)$, the respective probability distributions associated with each of these three metrics can be expressed as follows:

$$p(y, \widetilde{y}|G) = \underbrace{p(y|\widetilde{y}, G)}_{\text{Predictive Parity}} \times \underbrace{p(\widetilde{y}|G)}_{\text{Statistical Parity}} = \underbrace{p(\widetilde{y}|y, G)}_{\text{Equalized Odds}} \times \underbrace{p(y|G)}_{\text{Base Rate}} \tag{1}$$

With respect to metrics such as statistical parity, equalized odds and predictive parity that evaluate fairness by comparing binary predictions with binary outcomes, pairwise combinations of metrics can be simultaneously satisfied only under very limited conditions, if at all. Garg et al. (2020).

2.2. calibration

in this paper, we are interested in whether we could threshold calibrated scores to achieve each of these metrics, we start by defining calibration in a binary setting:

calibration: An algorithm is *calibrated* if for all scores s, the individuals who have the same score have the same probability of belonging to the positive class, regardless of group membership. Mathematically, this is expressed through $p(y = 1|S = s, G = 0) = p(y = 1|S = s, G = 1)$

There is another related metric termed *well-calibration* or *calibration within groups* that imposes an additional, more stringent condition. In order for a model to be well-calibrated (or to have calibration within groups), individuals assigned score s must have probability s of belonging to the positive class.

The difference between calibration and well-calibration is simply one of mapping; the scores of a calibrated predictor can, using a suitable transformation, be converted to scores satisfying well-calibration.

2.3. Balance for positive/negative class

Kleinberg et al. (2016) have noted that when the average score s for all individuals constituting the group-specific positive class is the same for both groups of interest, it can be said that there exists *balance for the positive class*. Similarly, *balance for the negative class* is satisfied when the average score s for members of the negative class are equal, regardless of group membership. Mathematically this is expressed in terms of expected values. For the negative class, balance requires $\mathbb{E}[s|y = 0, G = 0] = \mathbb{E}[s|y = 0, G = 1]$, and for the positive class balance requires $\mathbb{E}[s|y = 1, G = 0] = \mathbb{E}[s|y = 1, G = 1]$.

2.4. calibration and predictive parity

It is possible to view calibration as a generalization of predictive parity to the non-binary setting. Of course, in general the score s is not binary. A continuous-valued score can be binarized through a thresholding operation to generate a binary prediction $\widetilde{y}$. However, it is *not* the case that thresholding a calibrated score in this manner necessarily leads to predictive parity. consider a threshold $s = t \in [0, 1]$, such that $\forall s > s = t, \widetilde{y} = 1$ and $\widetilde{y} = 0$ otherwise. Hence, the distribution relevant to predictive parity $p(y = 1|\widetilde{y} = 1, G)$ can be expressed $p(y = 1|s > s = t, G)$. Using this we can write:

$$p(y, s > s = t|G) = \int_{s=t}^{1} \underbrace{p(y|s, G)}_{\text{calibration term}} \; p(s|G)ds$$

a binary prediction $\widetilde{y}$. However, it is *not* the case that thresholding a calibrated score in this manner necessarily leads to predictive parity.

To prove this, consider a threshold $s = t \in [0, 1]$, such that $\forall s > s = t, \widetilde{y} = 1$ and $\widetilde{y} = 0$ otherwise. Hence, the distribution relevant to predictive parity $p(y = 1|\widetilde{y} = 1, G)$ can be expressed $p(y = 1|s > s = t, G)$. Using this we can write:

$$p(y, s > s = t|G) = \int_{s=t}^{1} \underbrace{p(y|s, G)}_{\text{calibration term}} \; p(s|G)ds \tag{2}$$

$$\implies \underbrace{p(y|s > s = t, G)}_{\text{predictive parity term}} = \frac{\int_{s=t}^{1} p(y|s, G)p(s|G)ds}{\int_{s=t}^{1} p(s|G)ds} \tag{3}$$

The above equation relates predictive parity to calibration, showing that even when the calibration term $p(y|s, G)$ is the same for both groups, the probability distribution of the score, expressed in equation 3 through $p(s|G)$, can vary across groups in a way that causes predictive parity not to be satisfied. To make this more intuitive, we will consider a special case where there are only two score values s_1 and s_2 above the threshold $s = t$ such that $p(s|G) \neq 0$. In other words, all individuals who receive risk scores above the threshold have the possibility of receiving one of only two scores, s_1 or s_2. Hence, $p(s > s = t|G) = p(s = s_1|G) + p(s = s_2|G)$.

Under this special case equation 3 reduces to:

$$p(y = 1|\widetilde{y} = 1, G) = \frac{p(y = 1|s = s_1, G)p(s = s_1|G) + p(y = 1|s = s_2, G)p(s = s_2|G)}{p(s = s_1|G) + p(s = s_2|G)} \tag{4}$$

Using this scenario, consider an example in which we have 100 people in each of two groups: orange and blue (this is a new example, unrelated to the example using orange and blue groups introduced earlier in the paper). Consider further an algorithm that only gives one of three possible scores (0.25, 0.5 or 0.75) to every individual's loan application. Suppose that scores are being binarized using a threshold of 0.49, such that any individual with a score above 0.49 is deemed to belong to the positive class. In this example, this would mean there are two possible scores (0.5 and 0.75) that can lead to a positive prediction. This is illustrated in Table 1 given below.

Table 1: Predictive Parity and Calibration Example

Score	Orange Group	Blue Group	Prediction after threshold with $s = t = 0.49$
0.25	40 (16)	40 (16)	Negative
0.5	20 (10)	40 (20)	Positive
0.75	40 (30)	20 (15)	Positive
Total	100(56)	100(51)	

The first column represents the score that the model assigned. In the second and third columns, the numbers outside the parentheses convey the number of people in the group assigned that score. The numbers in parentheses represent the number of people from those assigned that score who actually belong to the positive class. In this example the predictor is calibrated, since given a score, the fraction of people who actually belong to the positive class is independent of the group. Does this model satisfy predictive parity? Choosing 0.49 as the threshold gives a total of 60 positive predictions for both the orange group and the blue group. However, of the people with scores greater than 0.49, only 40 members in the orange group and 35 members of the blue group are actually in the positive class, resulting in a PPV of $40/60 = 0.66$ for the orange group and $35/60 = 0.583$ for the blue group. Thus, while the predictor is calibrated, choosing a threshold of 0.49 does not lead to a set of binary predictions that satisfy predictive parity.

It is also interesting to note that if all persons who had a score of 0.25 are instead given a score of 0.4, the model will not only be calibrated (because, as before, people with the same score have the same probability of belonging to the positive class) but also *well*-calibrated (because, due to this change in scoring, for all scores the score itself would give the probability of belonging to the positive class). However, this change in score would have no impact on the thresholding example above, illustrating that even a well-calibrated model does not, after applying a threshold to produce binary predictions, necessarily satisfy predictive parity.

2.5. calibration and statistical parity

It is possible to simultaneously achieve both statistical parity and calibration. For example, in the scenario described in table 1, applying a threshold of 0.49 to the calibrated scores resulted in binary predictions that satisfied statistical parity for both groups, even though the base rates were different. In this case, 60 individuals from each group were predicted to belong to the positive class, despite the fact that a total of 56 individuals from the orange group and 51 individuals from the blue group actually belonged to the positive class. Additionally, even if all individuals with a score of 0.25 were given a score of 0.4, the model would still be well-calibrated and satisfy statistical parity for the threshold of 0.49, despite the imperfect prediction. This demonstrates that statistical parity, as defined in section 2, is not necessarily incompatible with calibration and well-calibration, even when the base rates are different and the prediction is not perfect.

It is important to note that fairness metrics designed for use with scores s may differ from those intended for use with binary predictions. While it is easy to convert scores to binary values through thresholding, this process can mask the complexities of whether fairness metrics are still met after the thresholding. For instance, in the example above, the scores were calibrated and applying a threshold of 0.49 resulted in binary predictions that exhibited statistical parity. However, if the same distribution of scores had been subject to a threshold of 0.55, the resulting predictions would not have satisfied statistical parity. This emphasizes the need to consider the impact of threshold selection on fairness metrics.

3. results

3.1. calibration and predictive parity

One question that arises is whether there always exists a threshold at which thresholding calibrated scores satisfies predictive parity. To address this, we can consider a scenario in which predictive parity is satisfied for a given calibrated score at the threshold $s = t$. In this case, we must have:

$$\frac{\int_{s=t}^{1} p(y|s)p(s|G_1)ds}{\int_{s=t}^{1} p(s|G_1)ds} - \frac{\int_{s=t}^{1} p(y|s)p(s|G_2)ds}{\int_{s=t}^{1} p(s|G_2)ds} = 0 \tag{5}$$

Note that we abbreviated $p(y|s, G_1) = p(y|s, G_2)$ to $p(y|s)$. Here's a further analysis of these functions: Also note that at s=0,

$$\frac{\int_{0}^{1} p(y|s)p(s|G_1)ds}{\int_{0}^{1} p(s|G_1)ds} = \frac{p(y, s > 0|G)}{\int_{s=t}^{1} p(s|G_1)ds} = p(y|G)$$

so 5 reduces to the difference of the base rates, which we could assume is positive, without loss of generality. At $s = 1$ using the l'Hôpital rule we have,

$$\lim_{t\to 1} \frac{\int_{s=t}^{1} p(y|s)p(s|G_1)ds}{\int_{s=t}^{1} p(s|G_1)ds} - \frac{\int_{s=t}^{1} p(y|s)p(s|G_2)ds}{\int_{s=t}^{1} p(s|G_2)ds} = \lim_{t\to 1} p(y|t) - p(y|t) = 0 \tag{6}$$

So to show that the function has zeros inside $(0, 1)$ it's enough to show that it's not always positive,(by the intermediate value theorem). For that, it might be useful to look at the derivative.
However, an observation gives us a more immediate answer: Knowing that continuous scores are in practice discrete, given our highest score is h , a cut off right below h would satisfy predictive parity. This is because

$$\begin{aligned} &p(y = 1|\tilde{y}, G = 0) = p(y = 1|s > h - \epsilon, G = 0) = p(y = 1|s = h, G = 0) \\ &= p(y = 1|s = h, G = 1) = p(y = 1|\tilde{y}, G = 1) \end{aligned}$$

Our approach enables us to efficiently locate a threshold at which calibrated scores satisfy predictive parity. However, it is important to recognize that other thresholds may also satisfy this fairness metric. As a result, it is necessary to carefully analyze the appropriate choice of threshold among those that are allowed, taking into account the specific characteristics of the data and the question being addressed. Therefore, the appropriate choice of threshold might depend on the context.

3.2. calibration and equalized odds

One question that arises is whether there exists a threshold that can be applied to calibrated scores in order to achieve equalized odds. To address this question, we begin by making an observation:

$$\int_{s=t}^{1} p(y|s,G)p(s|G)ds = p(y|G)\int_{s=t}^{1} p(s|y,G)$$

so if the answer to the above question is yes, we must have:

$$\frac{\int_{s=t}^{1} p(y|s,G_1)p(s|G_1)ds}{p(y|G_1)} - \frac{\int_{s=t}^{1} p(y|s,G_2)p(s|G_2)ds}{p(y|G_2)} = 0 \tag{7}$$

$$\textit{or} \quad \frac{\int_{s=t}^{1} p(y|s,G_1)p(s|G_1)ds}{\int_{s=t}^{1} p(y|s,G_2)p(s|G_2)ds} - \frac{p(y|G_1)}{p(y|G_2)} = 0 \tag{8}$$

first we observe that $t = 0, 1$ are solutions. At $t = 0$,

$$\int_{0}^{1} p(y|s,G)p(s|G)ds = p(y|G)\int_{0}^{1} p(s|y,G) = p(y|G)$$

$$\textit{this implies that 7 reduces to} \quad \frac{p(y|G_1)}{p(y|G_1)} - \frac{p(y|G_2)}{p(y|G_2)} = 0$$

And at $t = 1$, using L'Hôpital rule we have 8 reduces to:

$$\lim_{t\to 1} \frac{\int_{s=t}^{1} p(y|s,G_1)p(s|G_1)ds}{\int_{s=t}^{1} p(y|s,G_2)p(s|G_2)ds} - \frac{p(y|G_1)}{p(y|G_2)} =$$

$$\frac{p(y|t)p(t|G_1)}{p(y|t)p(t|G_2)} - \frac{p(y|G_1)}{p(y|G_2)} = 0$$

To understand where the zeros of this function are, we take derivative to find critical points:

$$\frac{p(y|t,G_1)p(t|G_1)}{p(y|G_1)} - \frac{p(y|t,G_2)p(t|G_2)}{p(y|G_2)} = 0 \tag{9}$$

$$\iff p(y|t)\left[\frac{p(t|G_1)}{p(y|G_1)} - \frac{p(t|G_2)}{p(y|G_2)}\right] = 0 \tag{10}$$

$$\iff \frac{p(t|G_1)}{p(t|G_2)} - \frac{p(y|G_1)}{p(y|G_2)} = 0 \tag{11}$$

Which illustrates that there will be a critical point where the ratio of statistical parity terms for groups is equal to the base rates. Also note that

$$9 \iff \frac{p(y|G_1)p(t|y,G_1)}{p(y|G_1)} - \frac{p(y|G_2)p(t|y,G_2)}{p(y|G_2)} = 0 \tag{12}$$

$$\iff p(t|y,G_1) - p(t|y,G_2) = 0 \tag{13}$$

Which describes those critical points as points where the equalized odds term is satisfied(note that this is not the same as equalized odds being satisfied). To show that 7 has zeros outside of $(0, 1)$ We only need to check if the function 7 assumes values of different sign at these 11 ,13 points.

Based on these observations, it appears that it may be possible to find a threshold that satisfies equalized odds when working with calibrated scores. However, further analysis is needed to fully understand the relationship between calibrated scores, thresholding, and equalized odds. The appropriate choice of threshold might depend on the specific characteristics of the data and the desired outcomes.

3.3. calibration and statistical parity

Another question that arises is whether there always exists a threshold that can be applied to calibrated scores in order to achieve statistical parity. To address this question, we can derive the following equation:

$$p(s > s = t|G) = \frac{\int_{s=t}^{1} p(y|s,G)p(s|G)ds}{\int_{s=t}^{1} p(y|s,G)ds} = \frac{p(y|G)\int_{s=t}^{1} p(s|y,G)ds}{\int_{s=t}^{1} p(y|s,G)ds}$$

this will imply, assuming we satisfy both statistical parity and calibration, we have to satisfy:

$$\frac{\int_{s=t}^{1} p(y|s)p(s|G_1)ds}{\int_{s=t}^{1} p(y|s,G_1)ds} = \frac{\int_{s=t}^{1} p(y|s)p(s|G_2)ds}{\int_{s=t}^{1} p(y|s,G_2)ds}$$

with the same notation of abbreviating $p(y|s,G_1) = p(y|s,G_2)$ to $p(y|s)$.

4. conclusion and further questions

In this paper, we demonstrated that while calibrated scores may not inherently satisfy other fairness metrics such as predictive parity, equalized odds, and statistical parity, it is possible to use proper thresholding to achieve these desired outcomes in some cases. Specifically, we found that by using a certain threshold, calibrated scores can generate predictions that satisfy predictive parity. Additionally, we showed that it may also be possible to use thresholding to achieve equalized odds and statistical parity, though the choice of the threshold can play an important role in determining whether these conditions are satisfied. In our study, we also highlighted the complexities involved in working with continuous and binary domains.

Furthermore, we discussed the role of observability in selecting fairness metrics and the importance of considering the specific problem at hand. We also noted that while there is evidence that incorporating fairness metrics may not always improve net benefit compared to calibrated models followed by thresholding Pfohl et al. (2022), it is still valuable to explore the underlying reasons for thresholding and whether it automatically satisfies fairness metrics or requires careful selection of the threshold.

In addition, we suggested that it may be interesting to test our analysis for equalized odds with several popular distributions to gain further insights. Finally, we raised the question of whether it is possible to relax definitions of fairness metrics and calibration in a sensible way and how this might impact our results. Overall, our study highlights the complex nature of balancing calibrated performance with fairness considerations in machine learning applications.

References

Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3662–3666. IEEE, 2020.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Stephen Pfohl, Yizhe Xu, Agata Foryciarz, Nikolaos Ignatiadis, Julian Genkins, and Nigam Shah. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1039–1052, 2022.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

Claire Lazar Reich and Suhas Vijaykumar. A possibility in algorithmic fairness: Can calibration and equal error rates be reconciled? *arXiv preprint arXiv:2002.07676*, 2020.