With Malice Towards None: Assessing Uncertainty via Equalized Coverage

Yaniv Romano* Rina Foygel Barber† Chiara Sabatti*‡ Emmanuel J. Candès*§

March, 2020

Abstract

An important factor to guarantee a fair use of data-driven recommendation systems is that we should be able to communicate their uncertainty to decision makers. This can be accomplished by constructing prediction intervals, which provide an intuitive measure of the limits of predictive performance. To support equitable treatment, we force the construction of such intervals to be unbiased in the sense that their coverage must be equal across all protected groups of interest. We present an operational methodology that achieves this goal by offering rigorous distribution-free coverage guarantees holding in finite samples. Our methodology, *equalized coverage*, is flexible as it can be viewed as a wrapper around any predictive algorithm. We test the applicability of the proposed framework on real data, demonstrating that equalized coverage constructs unbiased prediction intervals, unlike competitive methods.

Keywords: conformal prediction, calibration, uncertainty quantification, fairness, quantile regression, unbiased predictions.

Media Summary

Machine learning algorithms are increasingly deployed in sensitive applications to inform the selection of job candidates, to inform bail and parole decisions, and to filter loan application, among many others. Such practices have become the subject of intense scrutiny, as society must be concerned about whether these algorithms reinforce discrimination and make the status quo normative. A major area of study has therefore been to propose mathematical definitions of appropriate notions of fairness or algorithmic models of fairness, and to make sure that learned models comply with such prescriptions. Because fairness is rather ill defined, it has been reported that such notions can be incompatible and/or hurt the groups they intend to protect.

In this work, we follow the prescription of Corbett-Davies and Goel, 2018 and decouple the statistical problem of risk assessment from the policy problem of taking actions or designing interventions. Rather than dictating policy, our aim will solely be the design of statistical algorithms providing the decision maker with information summarizing the knowledge that can be extracted from state-of-the-art machine learning systems in a way that is mathematically guaranteed to be unbiased regardless of a person's protected attributes, providing an operational definition of fairness. This is accomplished by constructing prediction sets, which provide an intuitive measure of the limits of predictive performance. To support equitable treatment, we force the construction of such intervals to be unbiased in the sense that their coverage must be equal across all protected groups of interest. We present an operational methodology that achieves this goal by offering rigorous distribution-free coverage guarantees holding in finite samples. Our methodology is flexible in the sense that it can be wrapped around any predictive algorithm. For instance, in a stylized application where a recommendation system for college admission predicts the GPA of candidate students after two years of undergraduate

*Department of Statistics, Stanford University
†Department of Statistics, University of Chicago
‡Department of Biomedical Data Science, Stanford University
§Department of Mathematics, Stanford University

education, our methodology would modify this system to produce, for each student, a range of values obeying two properties. First, the range contains the true outcome 90% of the time (or any other percentage). Second, this property holds regardless of the group to which the student belongs.

1 Introduction

1.1 The problem of equitable treatment

We are increasingly turning to machine learning systems to support human decisions. While decision makers may be subject to many forms of prejudice and bias, the promise and hope is that machines would be able to make more equitable decisions. Unfortunately, whether because they are fitted on already biased data or otherwise, there are concerns that some of these data driven recommendation systems treat members of different classes differently, perpetrating biases, providing different degrees of utilities, and inducing disparities. The examples that have emerged are quite varied:

1. **Criminal justice**: courts in the United States may use COMPAS—a commercially available algorithm to assess a criminal defendant's likelihood of becoming a recidivist—to help them decide who should receive parole, based on records collected through the criminal justice system. In 2016 ProPublica analyzed COMPAS and "found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk" Dieterich, Mendoza, and Brennan, 2016.[1,2]

2. **Recognition system**: a department of motor vehicles (DMV) may use facial recognition tools to detect people with false identities, by comparing driver's license or ID photos with other DMV images on file. In a related context, Buolamwini and Gebru, 2018 evaluated the performance of three commercial classification systems that employ facial images to predict individuals' gender, and reported that the overall classification accuracy on male individuals was higher than female individuals. They also found that the predictive performance on lighter-skinned individuals was higher than darker-skinned individuals.

3. **College admissions**: a college admission office may be interested in a new algorithm for predicting the college GPA of a candidate student at the end of their sophomore year, by using features such as high-school GPA, SAT scores, AP courses taken and scores, intended major, levels of physical activity, and so on. On a similar matter, the work reported in Gardner, Brooks, and Baker, 2019 studied various data-driven algorithms that aim to predict whether a student will drop out from a massive open online course (MOOC). Using a large data set available from Gardner, Brooks, Andres, and Baker, 2018, the authors found that in some cases there are noticeable differences between the models' predictive performance on male students compared to female students.

4. **Disease risk**: healthcare providers may be interested in predicting the chance that an individual develops certain disorders. Diseases with a genetic component have different frequencies in different human populations, reflecting the fact that disease—causing mutations arose at different times and in individuals residing in different areas: for example, Tay-Sachs disease is approximately 100 times more common in infants of Ashkenazi Jewish ancestry (central-eastern Europe) than in non-Jewish infants Kaback, O'Brien, and Rimoin, 1977. The genotyping of DNA polymorphisms can lead to more precise individual risk assessment than that derived from simply knowing to which ethnic group the individual belongs. However, given our still partial knowledge of the disease causing mutations and their prevalence in different populations, the precision of these estimates varies substantially across ethnic groups. For instance, the study reported in Kessler, Yerges-Armstrong, Taub, Shetty, Maloney, et al., 2016 found a preference for European genetic variants over non-European variants in two genomic databases that are widely used

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[2] For a recent analysis of COMPAS findings, see Rudin, Wang, and Coker, 2020.

by clinical geneticists (this reflects the fact that most studies have been conducted on European populations). Relying on this information only would result in predictions that are more accurate for individuals of European descent than for others.

The breadth of these examples underscores how data must be interpreted with care; the method that is advocated in this paper is useful regardless of whether the disparity is due to factors of inequality/bias, or instead due to genetic risk. Indeed, policymakers have issued a call that Executive Office of the President, 2014

> *"we must uphold our fundamental values so these systems are neither destructive nor opportunity limiting. [...] In order to ensure that growth in the use of data analytics is matched with equal innovation to protect the rights of Americans, it will be important to support research into mitigating algorithmic discrimination, building systems that support fairness and accountability, and developing strong data ethics frameworks."*

This is a broad call that covers multiple aspects of data collection, mining and interpretation; clearly, a response requires a multi-faced approach. Encouragingly, the machine learning community is beginning to respond to this challenge. A major area of study has been to propose mathematical definitions of appropriate notions of fairness Dieterich et al., 2016; Dwork, Hardt, Pitassi, Reingold, and Zemel, 2012; Hardt, Price, and Srebro, 2016; Hebert-Johnson, Kim, Reingold, and Rothblum, 2018; Kim, Ghorbani, and Zou, 2019; Zafar, Valera, Gomez Rodriguez, and Gummadi, 2017 or algorithmic models of fairness Kusner, Loftus, Russell, and Silva, 2017. In many cases, these definitions are an attempt to articulate in mathematical terms what it means not to discriminate on the basis of "protected characteristics"; U. S. law identifies these as sex, race, age, disability, color, creed, national origin, religion, and genetic information. Now, discrimination can take many forms, and it is not surprising that it might be difficult to identify one analytical property that detects it in every context. Moreover, the call above is broader than the specific domains where discrimination is forbidden by law and invites us to develop analytical frameworks that guarantee an ethical use of data.

1.2 Responses from the machine learning community

To begin to formalize the problem, it is useful to consider the task of predicting the value of Y, a binary $\{0, 1\}$ variable, with a guess $\hat{Y}$. We assume that $Y = 1$ represents a more "favorable" state, and that the value of $\hat{Y}$ will influence the decider, so that predicting $\hat{Y} = 1$ for some individuals gives them an advantage. In this context, $\mathbb{P}(\hat{Y} = 0 \mid Y = 1)$, the *false negative rate*, represents the probability with which an opportunity is denied to a "well deserving" individual. It is obvious that this is a critical error rate to control in scenarios such as deciding parole (see Example 1): freedom is a fundamental right, and nobody should be deprived of it needlessly. We then wish to require that $\mathbb{P}(\hat{Y} = 0 \mid Y = 1, A = a)$ is equal across values of the protected attribute A Hardt et al., 2016. In the case of distributions of goods (as when giving a loan), one might argue for parity of other measures such as $\mathbb{P}(\hat{Y} = 1 \mid A = a)$ which would guarantee that resources are distributed equally across the different population categories Dwork et al., 2012; Feldman, Friedler, Moeller, Scheidegger, and Venkatasubramanian, 2015; Zafar, Valera, Gomez-Rodriguez, and Gummadi, 2019. Indeed, these observations are at the basis of two notions of fairness considered in the literature.

Researchers have noted several problems with fairness measures that, as the above, ask for (approximate) parity of some statistical measure across all of these groups. Without providing a complete discussion, we list some of these problems here. (a) To begin with, it is usually unclear how to design algorithms that would actually obey these notions of fairness from finite samples, especially in situations where the outcome of interest or protected attribute is continuous. (b) Even if we could somehow "operationalize" the fairness program, these measures are usually incompatible: it is provably impossible to design an algorithm that obeys all notions of fairness simultaneously Chouldechova, 2017; Kleinberg, Mullainathan, and Raghavan, 2017. (c) This is particularly troublesome, as the appropriate measure appears to be context dependent. Consider Example 4 and suppose that $Y = 1$ corresponds to having Tay-Sachs, whose rate differs across populations. Due to the unbalanced nature of the disease, one would expect the predictive model to have a lower true positive rate for non-Jewish infants than that of Ashkenazi Jewish infants (for which the disease is much more common). Here, forcing parity of true positive rates Hardt et al., 2016 would conflict with accurate predictions for each group

Hebert-Johnson et al., 2018. (d) Finally, and perhaps more importantly, researchers have argued that enforcing frequently discussed fairness criteria "can often harm the very groups that these measures were designed to protect" Corbett-Davies and Goel, 2018.

In light of this, it has been suggested to decouple the statistical problem of risk assessment from the policy problem of taking actions and designing interventions. Quoting from Corbett-Davies and Goel, 2018, "an algorithm might (correctly) infer that a defendant has a 20% chance of committing a violent crime if released, but that fact does not, in and of itself, determine a course of action." Keeping away from policy then, how can we respond to the call in Executive Office of the President, 2014 and provide a policymaker the best information gleaned from data while supporting equitable treatment? Our belief is that multiple approaches will be needed, and with this short paper our aim is to introduce an additional tool to evaluate the performance of algorithms across different population groups.

1.3 This paper: equalized coverage

One fundamental way to support data ethics is not to overstate the power of algorithms and data-based predictions, but rather always accompany these with measures of uncertainty that are easily understandable by the user. This can be done, for example, by providing a plausible range of predicted values for the outcome of interest. For instance, consider a recommendation system for college admission (Example 3), not knowing about the accuracy of the prediction algorithm, we would like to produce for, each student, a predicted GPA interval $[\hat{Y}_{\text{lo}},\ \hat{Y}_{\text{hi}}]$ obeying the following two properties: the interval should be faithful in the sense that the true *unknown* outcome Y lies within the predicted range 90% of the time, say; second, this should be unbiased in that *the average coverage should be the same within each group*.

Such a predictive interval has the virtue of informing the decision maker about the evidence machine learning can provide while being explicit about the limits of predictive performance. If the interval is long, it just means that the predictive model can say little. Each group enjoys identical treatment, receiving *equal coverage* (e.g., 90%, or any level the decision maker wishes to achieve). Hence, the results of data analysis are unbiased to all. In particular, if the larger sample size available for one group overly influences the fit, leading to poor performance in the other groups, the prediction interval will make this immediately apparent through much wider confidence bands for the groups with fewer samples. Prediction intervals with equalized coverage, then, naturally assess and communicate the fact that an algorithm has varied levels of performance on different subgroups.

It seems impossible a priori to present information to the policymaker in such a compelling fashion without a strong model for dependence of the response Y on the features X or protected attributes A. In our college admission example, one may have trained a wide array of complicated predictive algorithms such as random forests or deep neural networks, each with its own suite of parameters; for all practical purposes, the fitting procedure may just as well be a black box. The surprise is that such a feat is possible under no assumption other than that of having samples that are drawn exchangeably—e.g., they may be drawn i.i.d.—from a population of interest. We propose a concrete procedure, which acts as a wrapper around the predictive model, to produce valid prediction intervals that provably satisfy the equalized coverage constraint for any black box algorithm, sample size and distribution. Such a procedure can be formulated by refining tools from conformal inference, a general methodology for constructing prediction intervals Lei, G'Sell, Rinaldo, Tibshirani, and Wasserman, 2018; Lei, Robins, and Wasserman, 2013; Papadopoulos, Proedrou, Vovk, and Gammerman, 2002; Romano, Patterson, and Candès, 2019; Vovk, Gammerman, and Shafer, 2005; Vovk, Gammerman, and Saunders, 1999; Vovk, Nouretdinov, and Gammerman, 2009. Our contribution extends classical conformal inference as we seek a form of *conditional rather than marginal* coverage guarantee Barber, Candès, Ramdas, and Tibshirani, 2019; Lei and Wasserman, 2014; Vovk, 2012.

The specific procedure we suggest to construct predictive intervals with equal coverage, then, supports equitable treatment in an additional dimension. Specifically, we use the same learning algorithm for all individuals, borrowing strength from the entire population, and leveraging the entire dataset, while adjusting "global" predictions to make "local" confidence statements valid for each group. Such a training strategy may also improve the statistical efficiency of the predictive model, as illustrated by our experiments in Section 3. Of course, our approach comes with limitations as well: we discuss these and possible extensions in Section 4.

2 Equalized coverage

Let $\{(X_i, A_i, Y_i)\}$, $i = 1, \ldots, n$, be some training data where the vector $X_i \in \mathbb{R}^p$ may contain the sensitive attribute $A_i \in \{0, 1, 2, \ldots\}$ as one of the features. Consider a test point with known X_{n+1} and A_{n+1} and aim to construct a prediction interval $C(X_{n+1}, A_{n+1}) \subseteq \mathbb{R}$ which contains the unknown response $Y_{n+1} \in \mathbb{R}$ with probability at least $1 - \alpha$ on average within each group; here, $0 < 1 - \alpha < 1$ is a desired coverage level. Our ideas extend to categorical responses in a fairly straightforward fashion; for brevity, we do not consider these extensions in this paper. Interested readers will find details on how to build valid conformal prediction sets in classification problems in Shafer and Vovk, 2008. Formally, we assume that the training and test samples $\{(X_i, A_i, Y_i)\}_{i=1}^{n+1}$ are drawn exchangeably from some arbitrary and unknown distribution P_{XAY}, and we wish that our prediction interval obeys the following property:

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1}, A_{n+1}) \mid A_{n+1} = a\} \geq 1 - \alpha \tag{1}$$

for all a, where the probability is taken over the n training samples and the test case. Once more, (1) must hold for any distribution P_{XAY}, sample size n, and regardless of the group identifier A_{n+1}. (While this only ensures that coverage is *at least* $1 - \alpha$ for each group—and, therefore, the groups may have unequal coverage level—we will see that under mild conditions the coverage can also be upper bounded to lie very close to the target level $1 - \alpha$.)

In this section we present a methodology to achieve (1). Our solution builds on classical conformal prediction Lei et al., 2018; Vovk et al., 2005 and the recent conformalized quantile regression (CQR) approach Romano et al., 2019 originally designed to construct **marginal** distribution-free prediction intervals (see also Kivaranovic, Johnson, and Leeb, 2019). CQR combines the rigorous coverage guarantee of conformal prediction with the statistical efficiency of quantile regression Koenker and Bassett, 1978 and has been shown to be adaptive to the local variability of the data distribution under study. Below, we present a modification of CQR obeying (1). Then in Section 2.2, we draw connections to conformal prediction Lei et al., 2018; Papadopoulos et al., 2002 and explain how classical conformal inference can also be used to construct prediction intervals with equal coverage across protected groups.[3]

Before describing the proposed method we introduce a key result in conformal prediction, adapted to our conditional setting. Variants of the following lemma appear in the literature Lei et al., 2018; Romano et al., 2019; Tibshirani, Foygel Barber, Candes, and Ramdas, 2019; Vovk, 2012; Vovk et al., 2005.

Lemma 1. *Suppose the random variables $Z_1, \ldots, Z_{m+1}$ are exchangeable conditional on $A_{m+1} = a$, and define $Q_{1-\alpha}$ to be the $(1 - \alpha)(1 + 1/m)$-th empirical quantile of $\{Z_i : 1 \leq i \leq m\}$. For any $\alpha \in (0, 1)$,*

$$\mathbb{P}\{Z_{m+1} \leq Q_{1-\alpha} \mid A_{m+1} = a\} \geq \alpha.$$

Moreover, if the random variables $Z_1, \ldots, Z_{m+1}$ are almost surely distinct, then it also holds that

$$\mathbb{P}\{Z_{m+1} \leq Q_{1-\alpha} \mid A_{m+1} = a\} \leq \alpha + 1/(m + 1).$$

2.1 Group-conditional conformalized quantile regression (CQR)

Our method starts by randomly splitting the n training points into two disjoint subsets; a *proper training set* $\{(X_i, A_i, Y_i) : i \in \mathcal{I}_1\}$ and a *calibration set* $\{(X_i, A_i, Y_i) : i \in \mathcal{I}_2\}$. Then, consider any algorithm $\mathcal{A}$ for quantile regression that estimates conditional quantile functions from observational data, such as quantile neural networks Taylor, 2000 (described in Appendix 4.3). To construct a prediction interval with $1 - \alpha$ coverage, fit two conditional quantile functions on the proper training set,

$$\{\hat{q}_{\alpha_{\text{lo}}}, \hat{q}_{\alpha_{\text{hi}}}\} \leftarrow \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_1\}), \tag{2}$$

[3] We build on the split conformal methodology Lei et al., 2018; Papadopoulos et al., 2002 rather than its transductive (or full) version Vovk et al., 2005 due to the high computational cost of the latter. We refer the reader to Lei et al., 2018; Vovk et al., 2005 for more details about transductive conformal prediction, its advantages and limitations.

Algorithm 1: Group-conditional CQR.

Input:

Data $(X_i, A_i, Y_i) \in \mathbb{R}^p \times \mathbb{N} \times \mathbb{R},\ 1 \leq i \leq n$.
Nominal coverage level $1 - \alpha \in (0, 1)$.
Quantile regression algorithm $\mathcal{A}$.
Training mode: joint/groupwise.
Test point $X_{n+1} = x$ with sensitive attribute $A_{n+1} = a$.

Process:

Randomly split $\{1, \ldots, n\}$ into two disjoint sets $\mathcal{I}_1$ and $\mathcal{I}_2$.
If *joint training*:
Fit quantile functions on the whole proper training set: $\{\hat{q}_{\alpha_{\text{lo}}}, \hat{q}_{\alpha_{\text{hi}}}\} \leftarrow \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_1\})$.
Else use *groupwise training*:
Fit quantile functions on the proper training examples from group $A_{n+1} = a$:
$\{\hat{q}_{\alpha_{\text{lo}}}, \hat{q}_{\alpha_{\text{hi}}}\} \leftarrow \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_1 \text{ and } A_i = a\})$.
Compute E_i for each $i \in \mathcal{I}_2(a)$, as in (3).
Compute $Q_{1-\alpha}(E, \mathcal{I}_2(a))$, the $(1 - \alpha)(1 + 1/|\mathcal{I}_2(a)|)$-th empirical quantile of $\{E_i : i \in \mathcal{I}_2(a)\}$.

Output:

Prediction interval $\hat{C}(x, a) = [\hat{q}_{\alpha_{\text{lo}}}(x) - Q_{1-\alpha}(E, \mathcal{I}_2(a)),\ \hat{q}_{\alpha_{\text{hi}}}(x) + Q_{1-\alpha}(E, \mathcal{I}_2(a))]$ for the unknown response Y_{n+1}.

at levels $\alpha_{\text{lo}} = \alpha/2$ and $\alpha_{\text{hi}} = 1 - \alpha/2$, say, and form a first estimate of the prediction interval $\hat{C}_{\text{init}}(x) = [\hat{q}_{\alpha_{\text{lo}}}(x),\ \hat{q}_{\alpha_{\text{hi}}}(x)]$ at $X = x$. $\hat{C}_{\text{init}}(x)$ is constructed with the goal that a new case with covariates x should have probability $1 - \alpha$ of its response lying in the interval $\hat{C}_{\text{init}}(x)$, but the interval $\hat{C}_{\text{init}}(x)$ was empirically shown to under- or over-cover the test target variable Romano et al., 2019. (Quantile regression algorithms are not supported by finite sample coverage guarantees Meinshausen, 2006; Steinwart and Christmann, 2011; Takeuchi, Le, Sears, and Smola, 2006; Zhou and Portnoy, 1996, 1998.)

This motivates the next step that borrows ideas from split conformal prediction Lei et al., 2018; Papadopoulos et al., 2002 and CQR Romano et al., 2019. Consider a group $A = a$, and compute the empirical errors (often called conformity scores) achieved by the first guess $\hat{C}_{\text{init}}(x)$. This is done by extracting the calibration points that belong to that group,

$$\mathcal{I}_2(a) = \{i : i \in \mathcal{I}_2 \text{ and } A_i = a\},$$

and evaluating

$$E_i := \max\{\hat{q}_{\alpha_{\text{lo}}}(X_i) - Y_i, Y_i - \hat{q}_{\alpha_{\text{hi}}}(X_i)\}, \quad i \in \mathcal{I}_2(a). \tag{3}$$

This step provides a family of conformity scores $\{E_i : i \in \mathcal{I}_2(a)\}$ that are restricted to the group $A = a$. Each score measures the signed distance of the target variable Y_i to the boundary of the interval $\hat{C}_{\text{init}}(x)$; if Y_i is located outside the initial interval, then $E_i > 0$ is equal to the distance to the closest interval endpoint. If Y_i lies inside the interval, then $E_i \leq 0$ and its magnitude also equals the distance to the closest endpoint. As we shall see immediately below, these scores may serve to measure the quality of the initial guess $\hat{C}_{\text{init}}(\cdot)$ and used to calibrate it as to obtain the desired distribution-free coverage. Crucially, our approach makes no assumptions on the form or the properties of $\hat{C}_{\text{init}}(\cdot)$—it may come from any model class, and is not required to meet any particular level of accuracy or coverage. Its role is to provide a "base algorithm" that effectively estimates the underlying uncertainty, around which we will build our predictive intervals.

Finally, the following crucial step builds a prediction interval for the unknown Y_{n+1} given $X_{n+1} = x$ and $A_{n+1} = a$. This is done by computing

$$Q_{1-\alpha}(E, \mathcal{I}_2(a)) := (1 - \alpha)(1 + 1/|\mathcal{I}_2(a)|)\text{-th empirical quantile of } \{E_i : i \in \mathcal{I}_2(a)\},$$

which is then used to calibrate the first interval estimate as follows:

$$\hat{C}(x,a) := \left[\hat{q}_{\alpha_{\text{lo}}}(x) - Q_{1-\alpha}(E, \mathcal{I}_2(a)),\ \hat{q}_{\alpha_{\text{hi}}}(x) + Q_{1-\alpha}(E, \mathcal{I}_2(a))\right]. \tag{4}$$

Before proving the validity of the interval in (4), we pause to present two possible training strategies for the initial quantile regression interval $\hat{C}_{\text{init}}(x)$. We refer to the first as *joint training* as it uses the whole proper training set to learn a predictive model, see (2). The second approach, which we call *groupwise training*, constructs a prediction interval for Y_{n+1} separately for each group; that is, for each value $a = 0, 1, 2, \ldots$, we fit a regression model to all training examples with $A_{n+1} = a$. These two variants of the CQR procedure are summarized in Algorithm 1. While the statistical efficiency of the two approaches can differ (as we will see in Section 3), both are guaranteed to attain valid group-conditional coverage for any data distribution and regardless of the choice or accuracy of the quantile regression estimate.

Theorem 1. *If (X_i, A_i, Y_i), $i = 1, \ldots, n+1$ are exchangeable, then the prediction interval $\hat{C}(X_{n+1}, A_{n+1})$ constructed by Algorithm 1 obeys*

$$\mathbb{P}\{Y_{n+1} \in \hat{C}(X_{n+1}, A_{n+1}) \mid A_{n+1} = a\} \geq 1 - \alpha$$

for each group $a = 0, 1, 2, \ldots$. Moreover, if the conformity scores $\{E_i :\ i \in \mathcal{I}_2(a) \cup \{n+1\}\}$ for $A_{n+1} = a$ are almost surely distinct, then the group-conditional prediction interval is nearly perfectly calibrated:

$$\mathbb{P}\{Y_{n+1} \in \hat{C}(X_{n+1}, A_{n+1}) \mid A_{n+1} = a\} \leq 1 - \alpha + \frac{1}{|\mathcal{I}_2(a)| + 1}$$

for each group $a = 0, 1, 2, \ldots$.

Proof. Fix any group a. Since our calibration samples are exchangeable, the conformity scores (3) $\{E_i :\ i \in \mathcal{I}_2(a)\}$ are also exchangeable. Exchangeability also holds when we add the test score E_{n+1} to this list. Consequently, by Lemma 1,

$$1 - \alpha \leq \mathbb{P}(E_{n+1} \leq Q_{1-\alpha}(E, \mathcal{I}_2(a)) \mid A_{n+1} = a) \leq 1 - \alpha + \frac{1}{|\mathcal{I}_2(a)| + 1}, \tag{5}$$

where the upper bound holds under the additional assumption that $\{E_i :\ i \in \mathcal{I}_2(a) \cup \{n+1\}\}$ are almost surely distinct, while the lower bound holds without this assumption.

To prove the validity of $\hat{C}(X_{n+1}, A_{n+1})$ conditional on $A_{n+1} = a$, observe that, by definition,

$$Y_{n+1} \in \hat{C}(X_{n+1}, A_{n+1}) \quad \text{if and only if} \quad E_{n+1} \leq Q_{1-\alpha}(E, \mathcal{I}_2(a)).$$

Hence, the result follows from (5).

□

Variant: asymmetric group-conditional CQR When the distribution of the conformity scores is highly skewed, the coverage error may spread asymmetrically over the left and right tails. In some applications it may be better to consider a variant of Algorithm 1 that controls the coverage of the two tails separately, leading to a stronger conditional coverage guarantee. To achieve this goal, we follow the approach from Romano et al., 2019 and evaluate two separate empirical quantile functions: one for the left tail,

$$Q_{1-\alpha_{\text{lo}}}(E_{\text{lo}}, \mathcal{I}_2(a)) := \\ (1 - \alpha_{\text{lo}})(1 + 1/|\mathcal{I}_2(a)|)\text{-th empirical quantile of } \{\hat{q}_{\alpha_{\text{lo}}}(X_i) - Y_i : i \in \mathcal{I}_2(a)\};$$

and the second for the right tail

$$Q_{1-\alpha_{\text{hi}}}(E_{\text{hi}}, \mathcal{I}_2(a)) := \\ (1 - \alpha_{\text{hi}})(1 + 1/|\mathcal{I}_2(a)|)\text{-th empirical quantile of } \{Y_i - \hat{q}_{\alpha_{\text{hi}}}(X_i) : i \in \mathcal{I}_2(a)\}.$$

Next, we set $\alpha = \alpha_{\text{lo}} + \alpha_{\text{hi}}$ and construct the interval for Y_{n+1} given $X_{n+1} = x$ and $A_{n+1} = a$:

$$\hat{C}(x, a) := [\hat{q}_{\alpha_{\text{lo}}}(x) - Q_{1-\alpha_{\text{lo}}}(E_{\text{lo}}, \mathcal{I}_2(a)),\ \hat{q}_{\alpha_{\text{hi}}}(x) + Q_{1-\alpha_{\text{hi}}}(E_{\text{hi}}, \mathcal{I}_2(a))]. \tag{6}$$

The validity of this procedure is stated below.

Theorem 2. *Suppose the samples (X_i, A_i, Y_i), $i = 1, \ldots, n+1$ are exchangeable. With the notation above, put* $\text{lower}(X_{n+1}) = \hat{q}_{\alpha_{\text{lo}}}(X_{n+1}) - Q_{1-\alpha_{\text{lo}}}(E_{\text{lo}}, \mathcal{I}_2(A_{n+1}))$ *and* $\text{upper}(X_{n+1}) = \hat{q}_{\alpha_{\text{hi}}}(X_{n+1}) + Q_{1-\alpha_{\text{hi}}}(E_{\text{hi}}, \mathcal{I}_2(A_{n+1}))$ *for short. Then*

$$1 - \alpha_{\text{lo}} \leq \mathbb{P}\{Y_{n+1} \geq \text{lower}(X_{n+1}) \mid A_{n+1} = a\} \leq 1 - \alpha_{\text{lo}} + \frac{1}{|\mathcal{I}_2(a)| + 1}$$

and

$$1 - \alpha_{\text{hi}} \leq \mathbb{P}\{Y_{n+1} \leq \text{upper}(X_{n+1}) \mid A_{n+1} = a\} \leq 1 - \alpha_{\text{hi}} + \frac{1}{|\mathcal{I}_2(a)| + 1},$$

where the lower bounds above always hold while the upper bounds hold under the additional assumption that the residuals are almost surely distinct. Under these circumstances, the interval (6) *obeys*

$$1 - (\alpha_{\text{lo}} + \alpha_{\text{hi}}) \leq \mathbb{P}\{Y_{n+1} \in \hat{C}(X_{n+1}, A_{n+1}) \mid A_{n+1} = a\} \leq 1 - (\alpha_{\text{lo}} + \alpha_{\text{hi}}) + \frac{2}{|\mathcal{I}_2(a)| + 1}.$$

Proof. As in the proof of Theorem 1, the validity of the lower and upper bounds is obtained by applying Lemma 1 twice. □

2.2 Group-conditional conformal prediction

The difference between CQR Romano et al., 2019 and split conformal prediction Papadopoulos et al., 2002 is that the former calibrates an estimated quantile regression interval $\hat{C}_{\text{init}}(X)$, while the latter builds a prediction interval around an estimate of the conditional mean $\hat{Y} = \hat{\mu}(X)$. For instance, $\hat{\mu}$ can be formulated as a *classical regression* function estimate, obtained by minimizing the mean-squared-error loss over the proper training examples. To construct predictive intervals for the group $A = a$, then simply replace both $\hat{q}_{\alpha_{\text{lo}}}$ and $\hat{q}_{\alpha_{\text{hi}}}$ with $\hat{\mu}$ in Algorithm 1 (or in its two-tailed variant). The theorems go through, and this procedure gives predictive intervals with exactly the same guarantees as before. As we will see in our empirical results, a benefit of explicitly modeling quantiles is superior statistical efficiency.

3 Case study: predicting utilization of medical services

The Medical Expenditure Panel Survey (MEPS) 2016 data set,[4] provided by the Agency for Healthcare Research and Quality, contains information on individuals and their utilization of medical services. The features used for modeling include age, marital status, race, poverty status, functional limitations, health status, health insurance type, and more. We split these features into dummy variables to encode each category separately. The goal is to predict the health care system utilization of each individual; a score that reflects the number of visits to a doctor's office, hospital visits, etc. After removing observations with missing entries, there are $n = 15656$ observations on $p = 139$ features. We set the sensitive attribute A to **race**, with $A = 0$ for non-white and $A = 1$ for white individuals, resulting in $n_0 = 9640$ samples for the first group and $n_1 = 6016$ for the second. In all experiments we transform the response variable by $Y = \log(1 + (\text{utilization score}))$ as the raw score is highly skewed.

Below, we illustrate that empirical quantiles can be used to detect prediction bias. Next, we show that usual (marginal) conformal methods do not attain equal coverage across the two groups. Finally, we compare the performance of joint vs. groupwise model fitting and show that, in this example, the former yields shorter predictive intervals.

[4] https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181

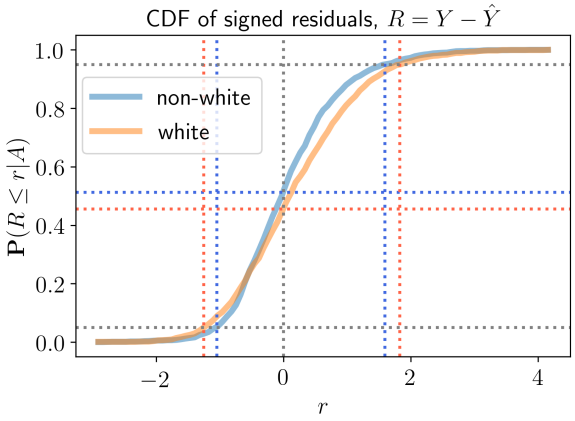

Figure 1: Empirical cumulative distribution function of the signed residuals $R = Y - \hat{Y}$ for both values of the sensitive attribute, computed on test samples. The blue dashed horizontal line is the value of $\mathbb{P}\{Y \leq \hat{Y}|A = 0\}$ equal to 0.51. Similarly, the red dashed horizontal line is $\mathbb{P}\{Y \leq \hat{Y}|A = 1\} = 0.45$. The dashed vertical colored lines present the 0.05th and 0.95th quantiles of each group, defined in (7) and (8), respectively. Left side: $r_0^{\text{lo}} = -1.04$ (blue), $r_1^{\text{lo}} = -1.25$ (red). Right side: $r_0^{\text{hi}} = 1.59$ (blue) and $r_1^{\text{hi}} = 1.83$ (red).

3.1 Bias detection

We randomly split the data into training (80%) and test (20%) sets and standardize the features to have zero mean and unit variance; the means and variances are computed using the training examples. Then we fit a neural network regression function $\hat{\mu}$ on the training set, where the network architecture, optimization, and hyper-parameters are similar to those described and implemented in Romano et al., 2019. The code for reproducing all the experiments is available online at https://github.com/yromano/cqr. Next, we compute the signed residuals of the test samples,

$$R_i = Y_i - \hat{Y}_i,$$

where $\hat{Y}_i = \hat{\mu}(X_i)$, and plot the resulting empirical cumulative distribution functions $\mathbb{P}\{R \leq r|A = 0\}$ and $\mathbb{P}\{R \leq r|A = 1\}$ in Figure 1. Observe that $\mathbb{P}\{R \leq r|A = 0\} \neq \mathbb{P}\{R \leq r|A = 1\}$. In particular, when comparing the two functions at $r = 0$, we see that $\hat{\mu}$ overestimates the response of the non-white group and underestimates the response of the white group, as

$$\mathbb{P}\{Y \leq \hat{Y}|A = 0\} = 0.51 > 0.45 = \mathbb{P}\{Y \leq \hat{Y}|A = 1\}.$$

Recall that the lower and upper quantiles of the signed residuals are used to construct valid group-conditional prediction intervals. While these must be evaluated on calibration examples (see next section), for illustrative purposes we present below the 0.05th and 0.95th quantiles of each group using the two cumulative distribution functions of test residuals. To this end, we denote by r_0^{lo} and r_1^{lo} the lower empirical quantiles of the non-white and white groups, defined to be the smallest numbers obeying the relationship

$$\mathbb{P}\{R \leq r_0^{\text{lo}}|A = 0\} \geq 0.05 \quad \text{and} \quad \mathbb{P}\{R \leq r_1^{\text{lo}}|A = 1\} \geq 0.05. \tag{7}$$

Following Figure 1, this pair is equal to

$$r_0^{\text{lo}} = -1.04 > -1.25 = r_1^{\text{lo}},$$

implying that for at least 5% of the test samples of each group, the fitted regression function $\hat{\mu}$ overestimates the utilization of medical services with larger errors for white individuals than for non-white individuals.

As for the upper empirical quantiles, we compute the smallest r_0^{hi} and r_1^{hi} obeying

$$\mathbb{P}\{R \leq r_0^{\text{hi}} | A = 0\} \geq 0.95 \quad \text{and} \quad \mathbb{P}\{R \leq r_1^{\text{hi}} | A = 1\} \geq 0.95, \tag{8}$$

and obtain

$$r_0^{\text{hi}} = 1.59 < 1.83 = r_1^{\text{hi}}.$$

Here, in order to cover the target variable for white individuals at least 95% of the time we should inflate the regression estimate by an additive factor equal to 1.83. For non-white individuals, the additive factor is *smaller* and equal to 1.59. This shows that $\hat{\mu}$ systematically predicts higher utilization of non-white individuals relative to white individuals.

3.2 Achieving equalized coverage

We now verify that our proposal constructs intervals with equal coverage across groups. Below, we set $\alpha = 0.1$. To avoid the coverage errors to be spread arbitrarily over the left and right tails, we choose to control the two tails independently by setting $\alpha_{\text{lo}} = \alpha_{\text{hi}} = \alpha/2 = 0.05$ in (6). We arbitrarily set the size of the proper training and calibration sets to be identical. (The features are standardized as discussed earlier.)

For our experiments, we test six different methods for producing conformal predictive intervals. We compare two types of constructions for the predictive interval:

- Conformal prediction (CP), where the predictive interval is built around an estimated mean $\hat{\mu}$ (as described in Section 2.2);
- Conformalized quantile regression (CQR), where the predictive interval is constructed around initial estimates $\hat{q}_{\alpha_{\text{lo}}}$ and $\hat{q}_{\alpha_{\text{hi}}}$ of the lower and upper quantiles.

In both cases, we use a neural network to construct the models; we train the models using the software provided by Romano et al., 2019, using the same neural network design and learning strategy. For both the CP and CQR constructions, we then implement three versions:

- Marginal coverage, where the intervals $\hat{C}(X)$ are constructed by pooling all the data together rather than splitting into subgroups according to the value of A;
- Conditional coverage with groupwise models, where the initial model for the mean $\hat{\mu}$ or for the quantiles $\hat{q}_{\alpha_{\text{lo}}}, \hat{q}_{\alpha_{\text{hi}}}$ is constructed separately for each group $A = 0$ and $A = 1$;
- Conditional coverage with a joint model, where the initial model for the mean $\hat{\mu}$ or for the quantiles $\hat{q}_{\alpha_{\text{lo}}}, \hat{q}_{\alpha_{\text{hi}}}$ is constructed pooling data across both groups $A = 0$ and $A = 1$.

The results are summarized in Table 1, displaying the average length and coverage of the marginal and group-conditional conformal methods. These are evaluated on *unseen test data* and averaged over 40 train-test splits, where 80% of the samples are used for training (the calibration examples are a subset of the training data) and 20% for testing. All the conditional methods perfectly achieve 90% coverage per group (this is a theorem after all). On the other hand, the marginal CP method under-covers in the white group and over-covers in the non-white group (interestingly, though, the marginal CQR method almost attains equalized coverage even though it is not designed to give such a guarantee).

Turning to the statistical efficiency of the conditional conformal methods, we see that conditional CQR outperforms conditional CP in that it constructs shorter and, hence, more informative intervals, especially for the non-white group. The table also shows that the intervals for the white group are wider than those for the non-white group across all four conditional methods, and that joint model fitting is here more effective than groupwise model fitting as the former achieves shorter prediction intervals.

Method	Group	Avg. Coverage	Avg. Length
*Marginal CP	Non-white	0.920	2.907
	White	0.871	2.907
Conditional CP (groupwise)	Non-white	0.903	2.764
	White	0.901	3.182
Conditional CP (joint)	Non-white	0.904	2.738
	White	0.902	3.150
*Marginal CQR	Non-white	0.905	2.530
	White	0.894	3.081
Conditional CQR (groupwise)	Non-white	0.904	2.567
	White	0.900	3.203
Conditional CQR (joint)	Non-white	0.902	2.527
	White	0.901	3.102

Table 1: Length and coverage of both marginal and group-conditional prediction intervals ($\alpha = 0.1$) constructed by conformal prediction (CP) and conformalized quantile regression (CQR) for MEPS dataset. The results are averaged across 40 random train-test (80%/20%) splits. *Groupwise* – two independent predictive models are used, one for non-white and another for white individuals; *joint* – the same predictive model is used for all individuals. In all cases, the model is formulated as a neural network. The methods marked by an asterisk are not supported by a group-conditional coverage guarantee.

4 Discussion

4.1 Larger intervals for a subpopulation

It is possible that the intervals constructed with our procedure have different lengths across groups. For example, our experiments show that, on average, the white group has wider intervals than the non-white group. Although one might argue that the different length distribution is in itself a type of unfairness, we want to caution the reader against assuming that a "fair" statistical procedure must necessarily produce intervals of equal length.

There are multiple aspects to consider. First, we believe that when there is a difference in performance across the protected groups, one needs to make this evident to the user and to understand the reasons behind it (we discuss below the issue with artificially forcing the two intervals to be of the same length). In some cases this difference might be reduced by improving the predictive algorithm, collecting more data for the population associated with poorer performance, introducing new features with higher predictive power, and so on. For example, in the context of studies that aim to predict disease risk on the basis of genetic features, it has become apparent that existing risk assessment tools suffer bias due to being constructed based on samples coming primarily from European populations; these tools will be much more effective if based on a larger sample that better reflects the diversity in the general population. It may also be the case that higher predictive precision in one group versus another may arise from bias, whether intentional or not, in the type of model we use, the choice of features we measure, or other aspects of our regression process—e.g., if historically more emphasis was placed on finding accurate models for a particular group a, then we may be measuring features that are highly informative for prediction within group a while another group a' would be better served by measuring a different set of variables. Crucially, we do not want to mask this differential in information, but rather make it explicit—thereby possibly motivating the decision makers to take action.

We also note that in some cases, reducing the difference in performance might not be possible while increasing information. For example, the collection of a large enough sample for a minority population might be impossible due to privacy considerations and financial burden. Or the outcome in question might have

structurally different variability across the groups. In such cases, equal length prediction intervals might be constructed only artificially, reducing the precision of the statements one can make for a given group—a choice that should be made with the participation of users and policy makers, rather than by data analysts alone.

4.2 The use of protected attribute

The debate around fairness in general, and our proposal in particular, requires the definition of "classes" of individuals across which we would like an unbiased treatment. In some cases these coincide with "protected attributes" where discrimination on their basis is prohibited by the law. The legislation sometimes does not allow the decision maker to know/use the protected attribute in reaching a conclusion, as a measure to caution against discrimination. While "no discrimination" is a goal everyone should embrace regardless whether the law mandates it or not, we shall consider the opportunity of using "protected attributes" in data-driven recommendation systems. On the one hand, ignoring protected attributes is certainly not sufficient to guarantee absence of discrimination (see, e.g., Buolamwini and Gebru, 2018; Chouldechova, 2017; Corbett-Davies and Goel, 2018; Dieterich et al., 2016; Dwork et al., 2012; Gardner et al., 2019; Hardt et al., 2016; Zafar et al., 2017). On the other hand, information on protected attributes might be necessary to guarantee equitable treatment. Our procedure relies on the knowledge of protected attributes, so we want to expand on this last point a little. In absence of knowledge of what are the causal determinants of an outcome, "protected attributes" can be an important component of a predictor. To quote from Corbett-Davies and Goel, 2018: "in the criminal justice system, for example, women are typically less likely to commit a future violent crime than men with similar criminal histories. As a result, gender-neutral risk scores can systematically overestimate a woman's recidivism risk, and can in turn encourage unnecessarily harsh judicial decisions. Recognizing this problem, some jurisdictions, like Wisconsin, have turned to gender-specific risk assessment tools to ensure that estimates are not biased against women." For disease risk assessment (Example 4 earlier) or related tasks such as diagnosis and drug prescription, race often provides relevant information and is routinely used. Presumably, once we understand the molecular basis of diseases and drug responses, and once sufficiently accurate measurements on patients are available, race may cease to be useful. Given present circumstances, however, Risch, Burchard, Ziv, and Tang, 2002 argue that "identical treatment is not equal treatment" and that "a race-neutral or color-blind approach to biomedical research is neither equitable nor advantageous, and would not lead to a reduction of disparities in disease risk or treatment efficacies between groups." In our context, the use of protected attributes allows a rigorous evaluation of the potentially biased performance for different groups.

Clearly, our current proposal can be adopted only when data on protected attributes has been collected; generalizations of the proposed methodology to situations where the group identifier is unknown are topics for further research.

4.3 Conclusion and future work

We add to the tools that support fairness in data-driven recommendation systems by developing a highly operational method that can augment any prediction rule with the best available *unbiased* uncertainty estimates across groups. This is achieved by constructing prediction intervals that attain valid coverage regardless of the value of the sensitive attribute. The method is supported by rigorous coverage guarantees, as demonstrated on real data examples. Although the focus of this paper is on continuous response variables, one can adapt tools from conformal inference Vovk et al., 2005 to construct prediction sets with equalized coverage for categorical target variables as well.

In this paper, we have not discussed other measures of fairness: we believe an appropriate comparison would require much larger space, and would benefit from the inclusion of multiple voices. In evaluating the different proposals of the growing literature on algorithmic fairness, it might be useful to keep in mind a distinction between properties that should be *required* versus properties that are merely *desirable*. As an analogy, in statistical hypothesis testing, most commonly we require a bound on the false positive rate (Type I error); under this constraint, high power (low Type II error) is then desirable.

One century of statistical reasoning has taught us the importance of quantifying uncertainty and error. No algorithm should be ever deployed without a precise and intelligible description of the errors it makes and the

statistical guarantees it offers. As practitioners know too well, it is most often not possible to guarantee that all errors are below a certain threshold. It becomes then crucial to select which statistical guarantee is most relevant for a problem, and fairness requires it to hold across different population groups. So, in the case of parole, we might think that the most crucial error to avoid is that of denying freedom to a deserving individual, and we should then enforce the probability of this error to be below the desired threshold in each population group. Or, as in the case of this paper, we might want to provide the user with a 90% predictive interval for the GPA of a student, and we then need to require that its coverage is as advertised in each population. Equality in other measures of performance which have not been identified as primary (as the length of the predictive intervals) might then be "desirable," but should not be "prescribed" and automatically pursued, without a conscious evaluation of the associated costs.

The knowledgeable reader will recognize that our approach is therefore different from the principle of equalized odds advocated in Hardt et al., 2016, which enforces that the two types of errors one can make in a binary classification problem must both be the same across the groups under study. (The cost is here that the algorithm would then need to change the predictions in at least one group to achieve the desired objective; this may be far from desirable and would not treat individuals equitably.) Returning to the distinction between a prescription and a wishlist, we make equalized coverage prescriptive. This does not mean that the data analyst cannot pay attention to other measures of fairness. For instance, she has the freedom to select predictive algorithms which score high on other metrics, e.g., by adding empirical constraints to the construction of prediction sets (or intervals). We hope to report on progress in this direction in a future publication.

Appendix: quantile neural networks

We follow Koenker and Bassett, 1978 and cast the estimation problem of the conditional quantiles of $Y|X{=}x$ as an optimization problem. Given training examples $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$, we fit a parametric model using the pinball loss Koenker and Bassett, 1978; Steinwart and Christmann, 2011, defined by

$$\rho_\alpha(y - \hat{y}) := \begin{cases} \alpha(y - \hat{y}), & \text{if } y - \hat{y} > 0, \\ (\alpha - 1)(y - \hat{y}), & \text{otherwise} \end{cases}$$

where $\hat{y}$ is the output of a regression function $\hat{q}_\alpha(x)$ formulated as a deep neural network. The network design and training algorithm are identical to those described in Romano et al., 2019 (once again, the source code is available online at https://github.com/yromano/cqr). Specifically, we use a two-hidden-layer neural network, with ReLU nonlinearities. The hidden dimension of both layers is set to 64. We use Adam optimizer Kingma and Ba, 2014, with minibatches of size 64 and a fixed learning rate of 5×10^{-4}. We employ weight decay regularization with parameter equal to 10^{-6} and also use dropout Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov, 2014 with a dropping rate of 0.1. We tune the number of epochs using cross validation (early stopping), with an upper limit of 1000 epochs.

Acknowledgements

E. C. was partially supported by the Office of Naval Research (ONR) under grant N00014-16-1-2712, by the Army Research Office (ARO) under grant W911NF-17-1-0304, by the Math + X award from the Simons Foundation and by a generous gift from TwoSigma. Y. R. was partially supported by the ARO grant and by the same Math + X award. Y. R. thanks the Zuckerman Institute, ISEF Foundation and the Viterbi Fellowship, Technion, for providing additional research support. R. F. B. was partially supported by the National Science Foundation via grant DMS-1654076 and by an Alfred P. Sloan fellowship. C. S. was partially supported by the National Science Foundation via grant DMS-1712800.

References

Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2019). The limits of distribution-free conditional predictive inference. *arXiv preprint arXiv:1903.04684*.

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

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. doi:10.1089/big.2016.0047.

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

Dieterich, W., Mendoza, C., & Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc.*

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). doi:10.1145/2090236.2090255.

Executive Office of the President. (2014). *Big data: Seizing opportunities, preserving values*. Createspace Independent Pub.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259–268). doi:10.1145/2783258.2783311.

Gardner, J., Brooks, C., Andres, J. M., & Baker, R. S. (2018). MORF: A framework for predictive modeling and replication at scale with privacy-restricted mooc data. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 3235–3244). doi:10.1109/BigData.2018.8621874.

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics and Knowledge* (pp. 225–234). doi:10.1145/3303772.3303791.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29* (pp. 3315–3323).

Hebert-Johnson, U., Kim, M., Reingold, O., & Rothblum, G. (2018). Multicalibration: Calibration for the (Computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 1939–1948).

Kaback, M. M., O'Brien, J. S., & Rimoin, D. L. (1977). Tay-Sachs disease: Screening and prevention. *Alan R. Liss (pub.)*

Kessler, M. D., Yerges-Armstrong, L., Taub, M. A., Shetty, A. C., Maloney, K., et al. (2016). Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry. *Nature Communications*, *7*(1), 12521. doi:10.1038/ncomms12521.

Kim, M. P., Ghorbani, A., & Zou, J. (2019). Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 247–254). doi:10.1145/3306618.3314287.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kivaranovic, D., Johnson, K. D., & Leeb, H. (2019). Adaptive, distribution-free prediction intervals for deep neural networks. *arXiv preprint arXiv:1905.10634*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Vol. 67, 43:1–43:23). doi:10.4230/LIPIcs.ITCS.2017.43.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems 30* (pp. 4066–4076).

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094–1111. doi:10.1080/01621459.2017.1307116.

Lei, J., Robins, J., & Wasserman, L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association*, *108*(501), 278–287. doi:10.1080/01621459.2012.751873.

Lei, J., & Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 71–96. doi:10.1111/rssb.12021.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, 983–999.

Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). Inductive confidence machines for regression. In *Machine Learning: European Conference on Machine Learning ECML 2002* (pp. 345–356).

Risch, N., Burchard, E., Ziv, E., & Tang, H. (2002). Categorization of humans in biomedical research: Genes, race and disease. *Genome Biology*, *3*(7), comment2007.1. doi:10.1186/gb-2002-3-7-comment2007.

Romano, Y., Patterson, E., & Candès, E. (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems 32* (pp. 3543–3553).

Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, *2*(1).

Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, *9*, 371–421.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Steinwart, I., & Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, *17*(1), 211–225. doi:10.3150/10-BEJ267.

Takeuchi, I., Le, Q. V., Sears, T. D., & Smola, A. J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, *7*, 1231–1264.

Taylor, J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, *19*(4), 299–311. doi:10.1002/1099-131X(200007)19:4<299::AID-FOR775>3.0.CO;2-V.

Tibshirani, R. J., Foygel Barber, R., Candes, E., & Ramdas, A. (2019). Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems 32* (pp. 2530–2540).

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning* (Vol. 25, pp. 475–490).

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. doi:10.1007/b106715.

Vovk, V., Gammerman, A., & Saunders, C. (1999). Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning* (pp. 444–453).

Vovk, V., Nouretdinov, I., & Gammerman, A. (2009). On-line predictive linear regression. *Annals of Statistics*, *37*(3), 1566–1590. doi:10.1214/08-AOS622.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & 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* (pp. 1171–1180). doi:10.1145/3038912.3052660.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, *20*(75), 1–42.

Zhou, K. Q., & Portnoy, S. L. (1996). Direct use of regression quantiles to construct confidence sets in linear models. *Annals of Statistics*, *24*(1), 287–306. doi:10.1214/aos/1033066210.

Zhou, K. Q., & Portnoy, S. L. (1998). Statistical inference on heteroscedastic models based on regression quantiles. *Journal of Nonparametric Statistics*, *9*(3), 239–260. doi:10.1080/10485259808832745.