We thank the reviewers and the editor for their helpful comments, which will improve our manuscript. We are pleased to see two positive reviews, and we believe that the most serious concerns raised by the two more critical reviewers are due to easily-addressed misunderstandings; we proceed with a point-by-point response next.

Reviewer 1 (R1) suggested that it will be valuable to see the trade-off between predictive loss and equalized odds violations for varying hyperparameter values. We report on the trade-off in Figure 1 and will include this in the manuscript.

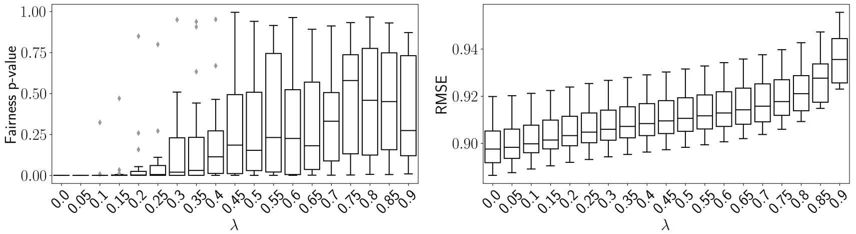

Figure 1: The effect of the regularization parameter for MEPS data set. The results are evaluated on 20 random splits of the data. We use a deep neural network as the underlying predictive model.

Reviewer 2 (R2) expresses concern that the proposed framework does not have a theoretical guarantee. While correct, we are not aware of any learning algorithm that is supported by a guarantee of conditional independence, except in the special case of binary classification. The reason is that this is an extremely difficult problem, especially for neural network algorithms that involve the minimization of non-convex loss functions. Working within these constraints, we rigorously derived our proposed loss function, emerging from the property we proved in Proposition 1. To quote Reviewer 4 (R4), our proposal "uses principled methods and thinks mostly rigorously about the correct way to approach equalized odds." Importantly, for identical levels of fairness, our simulations show the proposed procedure leads to improved accuracy over the alternative heuristic procedures. As a closing remark, because theoretical guarantees are never available, we pay close attention to our new hypothesis test for rigorously detecting violations of the equalized odds property.

R2 also expresses disappointment that our hypothesis test can be viewed as a special case of the abstract Holdout Randomization Test (HRT), as we state in Section 3. We believe this fact does not diminish its utility; the HRT is a special case of the conditional randomization test [18], but why is this a weakness? There is value in being concrete and in proposing novel useful methods: our proposal is the only tool in the literature to rigorously check whether a learned model actually obeys equalized odds. To quote R1, "The [...] general statistical test for equalized odds is also valuable." We are also puzzled about R2's question whether "the hypothesis test is better than the existing notion of equalized odds." In contrast, our test checks whether equalized odds is satisfied by a predictor. This test does not introduce a new notion of fairness. We will clarify this and further clarify that our contribution is twofold: (1) a technique for fitting models that satisfy equalized odds in regression and multi-class classification problems that is shown to be more accurate than the few alternatives and (2) the first rigorous test to detect violations of this property.

Turning to Reviewer 3 (R3), the concern most vigorously expressed is that *fairness through unawareness* has key shortcomings, which we emphatically agree with. To be explicit, *fairness through unawareness* [1] is a notion of fairness where the analyst does not use the sensitive attribute in modeling. In contrast, equalized odds is a different notion of fairness proposed in [6] to address the problems R3 describes. That prior work establishes the benefits of using equalized odds instead of *fairness through unawareness*. A primary goal of our present work is precisely to avoid the problems of *fairness through unawareness* articulated by R3 by introducing a technique to instead achieve equalized odds—a better notion of fairness—in regression and multi-class classification problems. We are therefore puzzled as to where the discussion about *fairness through awareness* is coming from.

Secondly, we must point out an omission from R3's summary that we simply use the randomly sampled sensitive attributes "instead of real sensitive features." Our use of randomization is more sophisticated than it may seem: we use the fair dummies to define a loss function that aims to drive the predictive rule $\hat{f}$ to achieve equalized odds. Certainly, other forms of randomization have been studied before, but randomizing *conditionally* on the observed Y is the crucial idea necessary to promote equalized odds in model fitting, and the technique we provide is entirely new.

Lastly, we thank R3 for asking to explain how to estimate $P(A|Y)$ when A is a continuous sensitive attribute. In short, one can sample fair dummies by fitting a conditional distribution estimator, e.g., using quantile regression, and we will add a proper discussion. In response to R3's related question, equalized odds does not exclude or discourage a perfect predictor. Due to limited space, we must refer the reader to [6] rather than discuss this point.

Finally, R4's most serious concern is about novelty: "I probably would have proposed the same tools that the authors suggest." We believe this comment strengthens the validity of the proposed method and stress that despite the widely-discussed importance of addressing fairness in machine learning, the fundamental problem of fitting models that satisfy equalized odds in multi-class classification and regression problems remained entirely open until recently, and our approach offers improved accuracy over the very few alternatives.