We thank the reviewers for their feedback.

Reviewer 1

Thank you for the encouraging words. We have tried as much as possible to provide insight on what xAUC discrepancies mean and to provide diagnostic tools to inspect them in order to inform practice. We agree further work is needed to push this in specific applications. While this is beyond the scope of this present paper, we have employed xAUC/xROC metrics to assess fairness in more recent work and will continue to push this direction in application work.

Reviewer 2

We appreciate the kind words and your expectation for xAUC to be widely adopted. And, yes, thank you: we will mention the connection to separation (separation for a score implies but is strictly stronger than $\Delta\,\mathrm{xAUC} = 0$; note also that $\Delta\,\mathrm{xAUC}$ is a metric for assessing level of unfairness rather than just a criterion). Re code dependencies: we will update the checklist; we use Python 2.7 with numpy/pandas/sklearn/matplotlib/scipy.

Reviewer 3

“fairness literature is awash with fairness metrics [...] make a compelling case for why we should prefer their metric to the existing alternatives."

- Firstly, there are no metrics specifically for **disparate impacts of continuous risk probability scores**, as acknowledged by R1. We distinguish between conditional probability estimation with observed binary outcomes, and fair regression which studies parity constraints on loss with respect to observed regression outcomes (Agarwal et al, 2019; Zink and Rose 2019).
- We specifically featured the COMPAS example, a commonly studied dataset, to make a compelling case for this metric: attention to xAUC, its probabilistic interpretation, and its decompositions, illuminates causes of unfairness. We highlighted that previous arguments (e.g., Dieterich et al. [18]) which compared within-group AUCs did not provide a meaningful notion of accuracy equity in decision-support settings, so that **relative to such naive metrics that exist, xAUC should be preferred**. We further featured other datasets to demonstrate the breadth of applicability.

Re: base rates: In Proposition 1, we provide a decomposition that illustrates the connection between base rates, xAUC disparities, and marginal AUC. This shows that base rates are just one ingredient and disparities can exist even with equal base rates.

Re: what action should be taken after xAUC is measured:

- The legitimacy of direct post-training adjustment is highly contested both in the fairness literature and in practice and is highly context-dependent, and this is in no way specific to the xAUC. Part of this ambiguity relates to the overloaded use of “predictive risk scores” in the fairness literature, which our work in part seeks to clarify. We highlight that different interpretations may be more or less directly applicable based on the problem setting. If the setting does not admit adjustment (e.g. risk assessment in healthcare to inform patient decisions), we highlight work that discusses alternative approaches for directly improving performance, such as choice of covariates. When these avenues are available, it is unclear if adjustment is advisable.
- What action should be taken is highly context-dependent, and we illustrate how different contexts lead to different recommendations. Recognizing that the issue of direct adjustment (we believe) remains contested in the fairness literature and for practitioners, due to legal and Pareto-violation concerns, we (or anyone) wouldn’t be able to provide an honest and realistic algorithmic solution to address xAUC disparities. See e.g. Hellman 2019, “Measuring Algorithmic Unfairness”, which seeks to establish general principles for preferring calibration vs. impact-based or base-rate sensitive metrics.
- Nonetheless, we relate the idea of post-processing adjustment of scores to post-processing classifiers by adjusting thresholds. Furthermore, the results we include in the supplement do provide the tools for $\Delta\,\mathrm{xAUC}$-minimizing adjustment of scores for those who would choose to do so.