We thank the reviewers for their feedback. We will release our implementation on github. We thank R1 for pointing out our calibration metric of choice, D-Calibration (D-Cal), has been published at JMLR 2020 (we cite this as [H] here).

[R3: Relation to D-Cal] Building on [H]'s D-cal, we propose X-cal to regularize a model to have low D-Cal. X-cal is a differentiable approximation of an upper-bound on D-Cal, amenable to stochastic optimization.

[R1, R2: Real world dataset with censoring; Survival benchmarks] We evaluated X-Cal on [Avati et al]'s alternative MIMIC set with 70% censored points. D-Cal goes from $2 \times 10^{-4} \to 9 \times 10^{-5}$ as λ increases from $0 \to 10^3$. We will add this to the paper. We are glad to include specific evaluations/benchmark sets that the reviewers think are relevant.

[R2: Comparison with MTLR/ approaches mentioned in [H]] We do not include methods from [H] because our work focuses on using flexible models with good likelihood but poor calibration, like S-CRPS. We use a categorical model because it is a common flexible likelihood that can approximate many continuous distributions given enough bins. This allows us to evaluate X-Cal without parametric restrictions. We did cite an approach like MTLR [Ranganath 2018]. We will cite MTLR [Yu et al. 2011] and its neural version [Fotso 2018]. We ran MTLR using PyCox library on the uncensored synthetic gamma dataset. This gave a model with D-cal 0.7486, which is higher than any model we study.

[R3,R4: Comparison to other established calibration metrics] The alternative notion of calibration for fixed time t* suggested by reviewers [Yadlowsky 2019, Royston/Altman 2013] are described in [H] as "1-Calibration". [H] proves that D-Cal (with fixed bins) and 1-Cal for time t* (with fixed bins) measure different aspects of the survival distribution: 0 D-Cal and 0 1-Cal do not imply each other. A practitioner may need calibration at several times e.g. 6 months, 1 year. Future work is to regularize models with approximations of 1-Cal. measures (e.g. Hosmer-Lemeshow statistic) using soft indicators. Our focus is to maintain a certain level of calibration based on the specific metric, D-cal.

[R2: p-value] The p-value reported by [H] is the result of a χ^2-test on the D-Cal test statistic. Thus, if models are ordered in the test statistic their corresponding p-values are ordered in the same way. While p-values help test for perfect calibration, our focus is on *improving* calibration of existing models which we demonstrate in in our experiments.

[R3: Discontinuous learned conditional survival model] As mentioned in 4.1, a discontinuous model will have a lower bound greater than 0 for D-Cal because its CDF values will not be a continuous Unif(0,1) variable. However, minimizing D-Cal will still spread out the CDF values to whichever extent possible and thus improve calibration. In the case of a categorical model, this lower bound decreases to 0 as the size of each bin goes to 0 when adding more bins.

[R3: Adjustments for right censoring / MNIST censoring] This is an important issue. In line 151 of our paper, we handle right censoring using the technique proposed in appendix B.5 in [H] and proved to result in a valid test statistic. As noted in [H] on page 47, the estimation of D-cal on a censored dataset will not equal the estimate when censored times are revealed. This is due to the fact that in the censored dataset [H]'s correction for right censoring gives a few bins the correct weight for free meaning D-cal will be lesser. However, for a given dataset, D-Cal is 0 for any bin for the true conditional $p(T \mid X)$ for any non-informative censoring process that meets a "positivity" assumption. Thus, two models evaluated on the *same* data (censored or uncensored) can be compared with D-Cal. Reweighting methods, such as Yadlowsky et al. that R3 suggests, can be used to adjust for censoring. One option is to adjust with with $p(C \mid X)$. This requires $C \perp T \mid X$ and solving a censored survival problem $p(C \mid X)$ with a high-dimensional conditioning set. Another option is to adjust with the lower dimensional conditioning set $p(C \mid risk_\theta(X))$. This requires $C \perp T \mid risk_\theta(X)$ and differentiating through the *estimation* of $p(C \mid risk_\theta(X))$ w.r.t. θ. The approach we take requires neither another censored survival problem nor differentiating through estimation.

[R2: λ and γ] Choosing λ: the user first decides on a threshold of D-Cal and then increase λ until D-Cal evaluated on a held-out validation set meets this threshold. See Table 1 for the role of γ. With low γ, soft D-cal approximates poorly and D-Cal/NLL suffer. For γ too large, gradients vanish and D-Cal/NLL suffer. We found $\gamma = 10^4$ allowed for easy optimization with soft D-Cal approximating D-Cal well.

γ	10	10^4	1.1×10^4	10^5
D-Cal	0.4	0.0005	0.0002	0.0003
NLL	4.33	1.82	1.88	2.49

Table 1: Soft D-cal as γ varies.

[R2, Tightness of upper-bound] Table 2 shows that models ordered by the upper-bound are ordered in D-cal the same way. Further, when batch size is large enough, if $\lambda_i < \lambda_j$, the bound for λ_j is less than D-Cal for λ_i.

λ	Batch size	D-Cal	Bound
10	500	0.0040	0.0059
	5000	”	0.0042
50	500	0.0006	0.0024
	5000	”	0.0008
100	500	0.0003	0.0022
	5000	”	0.0005

Table 2: Slack in the Upper Bound

[R4: Choosing D-Cal on a validation set] During training we evaluated NLL+D-cal on a validation set at every epoch and save the model. Then, we report test metrics for the model with best validation NLL + D-Cal. If we only select/optimize X-cal on a validation set, the predictive likelihood may get arbitrarily worse. This issue occurs with Platt scaling as well.

[Minor comments/ Typos] We thank the reviewers for detailed feedback about our writing. We will define Harrell's Concordance Index, change 10k to our intended 10^4, and rephrase "calibration means accurate prognosis".