# Robust Risk Prediction from Noisy Data

**Nathan Stromberg** [1] **Tyler Sypherd** [1] **Visar Berisha** [1] **Lalitha Sankar** [1]

## Abstract

Risk prediction models in healthcare are often trained in a supervised fashion by learning a mapping from a set of clinical variables to an outcome of interest. For example, self-reported symptom data is combined with disease diagnosis to train predictive models for characterizing disease risk from symptoms; however, in practice, these data are often noisy. Arguably, the most common approach to characterizing risk in clinical applications is the logistic model owing to its interpretability. We present a method for robust risk prediction using the logistic model which preserves not only the output predicted probabilities (a proxy for a composite risk score), but also the odds ratios of the model (a proxy for covariate-level risk) under unknown amounts of label noise. We demonstrate the efficacy of our method on a large COVID-19 self-reported survey dataset.

## 1. Introduction

Risk prediction models in healthcare estimate the risk of a clinical outcome based on one or more clinical variables. For example, self-reported survey data of symptoms and diagnosis of a disease of interest is commonly used to learn models for estimation of disease risk from the symptoms. Unfortunately, in practice, data frequently contain significant and unknown amounts of noise, either in the features or in the class label (Rauscher et al., 2008; Gorber et al., 2009). This noise could be present as the result of an adversarial agent, data entry errors, or benign inaccuracies. For this reason, any model learning on this data should be robust to noise. In this paper we focus on noise in the class labels, symmetric label noise (SLN).

The logistic model, widely used because of its interpretability (Menni et al., 2020; Jewell et al., 2020; Tu, 1996), is not only well-calibrated[1] but also defines a measure of covariate importance through the odds ratio. This allows for an interpretable risk score at feature-level (we refer to this as covariate-level risk) and at the model level via the posterior (we refer to this as composite risk) (Liang et al., 2020). Unfortunately, robustness to noise is not inherent to the logistic model (Pregibon, 1981). In the logistic model, label noise is mainly corrected in the literature by making modifications to the canonical log-loss including the following well-known hyperparameterized losses: focal loss (Lin et al., 2017), NCE+RCE (Ma et al., 2020), and $\alpha$-loss (Sypherd et al., 2022) among others. Both NCE+RCE and $\alpha$-loss have proven robustness to label noise.

We focus on adding robustness to the logistic model through loss functions and specifically examine $\alpha$-loss because it captures common losses such as log-loss ($\alpha = 1$) and exponential loss ($\alpha = \frac{1}{2}$) and has additional theoretical guarantees (Sypherd et al., 2021) regarding the learned posterior. We make the following contributions:

- We present a framework for incorporating tunable loss functions into the logistic model as a method to preserve both covariate-level risk and composite risk under symmetric label noise.
- We translate theoretical results from Sypherd et al. (2021) to healthcare risk prediction.
- We apply our method to a novel COVID-19 self-reported survey dataset (Salomon et al., 2021) and demonstrate the robustness of $\alpha$-loss.

## 2. Problem Setup

We assume a supervised classification setting in which the goal is to learn both composite and covariate-level risk from class label $Y \in \mathcal{Y}$ and features, or covariates, $X \in \mathcal{X}$. $X$ and $Y$ are jointly distributed according to $P(X,Y) = P(X)P(Y|X)$. We additionally assume the logistic model, that is

$$\log \frac{P(Y=1|X=x)}{P(Y=-1|X=x)} = \theta^T x \tag{1}$$

$$P(Y=1|X=x) = \sigma(\theta^T x). \tag{2}$$

We note that an equivalent definition of the logistic regression of $Y$ by $X$ is given by the minimization of log-loss

[1] Arizona State University. Correspondence to: Nathan Stromberg <nstrombe@asu.edu>.

*Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML)*. Copyright 2022 by the author(s).

[1] That is, it learns the true posterior distribution in expectation.

under sigmoid classifiers, specifically

$$\theta^* = \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathbb{E}_{X,Y}[\ell(Y, \sigma(\theta^T X)], \tag{3}$$

where $\sigma : \mathbb{R} \to (0, 1)$ is the sigmoid function and $\ell(\cdot, \cdot)$ is log-loss. We will return to this definition of logistic regression in our treatment of $\alpha$-loss in Section 3.1.

### 2.1. Measuring Risk

We examine the logistic model because it gives two soft risk measures: composite risk (the learned posterior) and covariate-level risk (odds ratios).

**Definition 2.1** (Risk (Composite Risk))**.** We define the risk that a sample $x$ is in the positive class as $\hat{P}(Y = 1|X = x)$.

Composite risk quantifies the likelihood of a negative outcome ($Y = 1$) given a set of features in an intuitive way, but many learning models are not calibrated in this risk measure (Minderer et al., 2021; Guo et al., 2017). This measure provides a view of risk at the sample level, but risk can be determined for a covariate as well by measuring the change in odds when a covariate is changed.

**Definition 2.2** (Odds)**.** The odds of a particular sample $x$ being in the positive class is given by

$$\text{Odds}(x) = \frac{\hat{P}(Y = 1|X = x)}{\hat{P}(Y = -1|X = x)}. \tag{4}$$

In light of Equation (4), Equation (2) can be seen as defining the logistic regression as a regression on the log odds.

**Definition 2.3** (Odds Ratio (Covariate-Level Risk))**.** Let the covariate of interest be $X_i$. Take $x \in \mathcal{X}$ and let $\tilde{x}$ be equal to $x$ in all components except the $i^{th}$. Let $\tilde{x}_i$ be a unit change of $x_i$, i.e. if $x_i \in \mathbb{R}$, $\tilde{x}_i = x_i + 1$. We then define the odds ratio as the ratio of the odds of $\tilde{x}$ to the odds of $x$. Alternatively, the odds ratio is the change in log odds when $x_i$ is increased by unit value. Specifically note that

$$\frac{\text{Odds}(\tilde{x})}{\text{Odds}(x)} = e^{\theta_i}. \tag{5}$$

The odds ratio of a given feature is of interest, specifically in medical research, because it allows a practitioner to examine the relative impact of a feature on the outcome odds.

### 2.2. Learning Risk Measures Under Noise

Noise in datasets can take on a variety of forms, many of them difficult to model and to eliminate. We examine a specific type of noise called symmetric label noise and analyze its impact on the learned composite and covariate-level risk.

**Definition 2.4** (Symmetric Label Noise)**.** Let $D$ be a dataset with features $X \in \mathcal{X}$ and binary class label $Y \in \{-1, 1\}$. Symmetric label noise is a corruption of $D$ which preserves features values $X$ but flips $Y$ equally from each class with total probability of flip $p$. We define the twisted posterior induced by symmetric label noise,

$$\begin{aligned}\tilde{P}(Y = 1|X = x) = (1 - p)P(Y = 1|X = x) \\ + pP(Y = -1|X = x).\end{aligned} \tag{6}$$

Well-calibrated models, i.e., those that learn the posterior of the data, will only learn this twisted posterior $\tilde{P}(Y|X)$ instead of the true posterior $P(Y|X)$ when learning on noisy samples. Such a model (e.g. log-loss) in turn would propagate the errors in the training set in its predictions, since the composite risk, as defined in Definition 2.1, would not be preserved. To preserve both risk and odds ratios, we now present a tunable framework to directly learn a correction of this twisted posterior using $\alpha$-loss.

## 3. Tunable Loss Functions

As mentioned earlier, many tunable and robust loss functions exist for classification, but we focus on $\alpha$-loss for its ability to capture several canonical losses and its strong theoretical guarantees.

### 3.1. $\alpha$-loss

**Definition 3.1** ($\alpha$-loss (Sypherd et al., 2022))**.** Let $\mathcal{P}(\mathcal{Y})$ be the set of probability distributions over $\mathcal{Y}$. Let $P(y|x) := P(Y = y|X = x)$ denote the true posterior for compactness. We define $\alpha$-loss for $\alpha \in (0, 1) \cup (1, \infty)$, $\ell^\alpha : \mathcal{Y} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}_+$ as

$$\ell^\alpha(y, P(y|x)) := \frac{\alpha}{\alpha - 1}(1 - P(y|x)^{1-1/\alpha}), \tag{7}$$

and by continuous extension,

$$\begin{aligned}\ell^1(y, P(y|x)) &:= -\log P(y|x), \\ \ell^\infty(y, P(y|x)) &:= 1 - P(y|x).\end{aligned}$$

When $P_{XY}$ is known, the minimizer of $\alpha$-loss, for every $\alpha$, is derived in Sypherd et al. (2022). We summarize the result as a way to build intuition for the ensuing results.

**Proposition 3.2** ((Sypherd et al., 2022))**.** *For each $\alpha \in (0, \infty]$,*

$$\underset{\hat{P}(Y|X)}{\arg\min}\, \mathbb{E}_{X,Y}[\ell^\alpha(Y, \hat{P}(Y|X))] = \hat{P}^*_\alpha, \tag{8}$$

*where, given $x \in \mathcal{X}$,*

$$\hat{P}^*_\alpha(y|x) = \frac{P(y|x)^\alpha}{\sum\limits_{y} P(y|x)^\alpha}, \forall y \in \mathcal{Y} \tag{9}$$

Throughout, we refer to $\hat{P}^*_\alpha$ as the $\alpha$-tilted distribution.

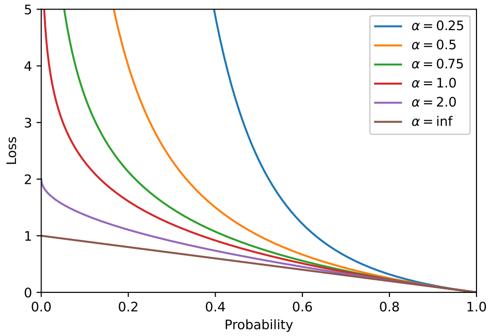


*Figure 1.* $\alpha$-loss Equation (7) as a function of the probability for several values of $\alpha$. We see that for large $\alpha$, outliers are penalized less than for small $\alpha$. This implies that large $\alpha$ is robust to outliers in the data.

### 3.2. Performance Guarantees for $\alpha$-loss

For compactness, let $\eta_c : \mathcal{X} \to [0,1]$ denote the true posterior mapping $P(Y = 1|X = x)$ and let $\eta_t : \mathcal{X} \to [0,1]$ denote the twisted posterior mapping $\tilde{P}(Y = 1|X = x)$ (Equation (6)). Further, we write $\eta_t^{(\alpha)}$ to denote the $\alpha$-tilted posterior learned on the twisted dataset.

To evaluate the effect of tuning $\alpha$ in recovering the true posterior, we use the KL divergence as a measure of closeness. Using this KL measure, Sypherd et al. (2021) proved that, for a broad noise class[2], choosing an $\alpha_0 > 1$ strictly reduces the expected KL divergence over the feature space. We restate this result in terms of SLN for our application.

**Proposition 3.3** ((Sypherd et al., 2021)). *For symmetric label noise less than* 50%*, the following ordering holds*

$$\mathbb{E}[\mathrm{KL}(\eta_c(X), \eta_t^{(1)}(X))] > \mathbb{E}[\mathrm{KL}(\eta_c(X), \eta_t^{(\alpha_0)}(X))], \tag{10}$$

*for some* $\alpha_0 > 1$.

We additionally note that the result Proposition 3.3 is entirely independent of the level of imbalance (the priors on each class). Arguments in Sypherd et al. (2021) suggest that $\alpha$ need not be tuned to exactly $\alpha_0$, but that there exists a broad range of $\alpha$ which performs similarly in terms of KL divergence from the clean model. Thus, our method performs well even on highly imbalanced data with unknown levels of noise, as exhibited in the following application.

## 4. Application to COVID-19 Risk Prediction

We now demonstrate the efficacy of our method in preserving both composite and covariate-level risk on a large self-report survey dataset.

### 4.1. COVID-19 Dataset

We use the COVID-19 Trends and Impact Survey (CTIS) dataset from Carnegie-Mellon University (Salomon et al., 2021) that was collected and collated in collaboration with Facebook. We compress the dataset to contain 42 categorical and real-valued features including symptom data, behaviors, and comorbidities. Each sample is labeled either as RT-PCR-confirmed COVID positive (1) or negative (0) based on self-reported diagnoses by study participants. Samples with spurious responses (e.g., negative number of people in household) or responses with missing features were removed for training and testing. This preprocessing resulted in a dataset of 864,154 samples with a class imbalance of $14 : 86$ of positive to negative COVID cases.

### 4.2. Experimental Setup

**Model** For better accuracy and a simpler, interpretable logistic model, we restrict the model to predict using a smaller set of 8 features; we choose these as the features with the largest odds ratio on the validation set: age, gender, anosmia, shortness of breath, aches, tiredness, cough, and fever. Hyperparameters such as learning rate were also selected on a validation set. Models were trained over a grid of possible noise values, $p$, and $\alpha$ values, $(p, \alpha) \in [0, 0.15] \times [0.6, 3]$. For each pair $(p, \alpha)$, 5 models were trained with a different random noise seed and results were averaged across these samples for every metric.

**Noise Modeling** To enable reproducibility, we introduce synthetic SLN and tune the percentage of noise. This allows our experiments to quantify the level of noise in the data and analyze the effectiveness of our method across noise levels.

**Baseline** Because the underlying true statistics are not available as a ground truth, a "clean" model is selected as a baseline comparison. We select this model to be one with no added noise ($p = 0$) and log-loss ($\alpha = 1$). Because log-loss ($\alpha = 1$) is calibrated, the "clean" model will learn the true posterior distribution in expectation without the presence of noise.

### 4.3. Risk Preservation Metrics

In order to examine both composite and covariate-level risk, we evaluate both the KL divergence between the clean posterior and the $\alpha$-tilted posterior and the mean squared error (MSE) of odds ratios (Definition 2.3).

**Composite Risk Robustness** In order to measure the robustness of our method in composite risk, we now examine the effects of tuning $\alpha$ on the KL divergence between $\eta_c$, the clean posterior, and $\eta_t^{(\alpha)}$, the posterior learned on the twisted

[2] This was formally dubbed the class of Bayes-blunting twists.

| | KL Divergence | | |
|---|---|---|---|
| Noise, $p$ | $\alpha = 1$ KL | $\alpha_0$ KL | $\alpha_0$ value |
| 0% | $\mathbf{0}$ | 4.95e−5 | 1.1 |
| 1% | $\mathbf{1.00}$e$\mathbf{-5}$ | 3.95e−5 | 1.1 |
| 2.5% | 5.80e−5 | $\mathbf{5.79}$e$\mathbf{-5}$ | 1.1 |
| 5% | 2.15e−4 | $\mathbf{1.56}$e$\mathbf{-4}$ | 1.1 |
| 7.5% | 4.69e−4 | $\mathbf{3.59}$e$\mathbf{-4}$ | 1.2 |
| 10% | 8.00e−4 | $\mathbf{6.15}$e$\mathbf{-4}$ | 1.2 |
| 12.5% | 1.19e−3 | $\mathbf{9.67}$e$\mathbf{-4}$ | 1.2 |
| 15% | 1.64e−3 | $\mathbf{1.39}$e$\mathbf{-4}$ | 1.2 |

*Table 1.* For non-zero noise levels, we see that tuning $\alpha$ greater than 1 yields a model whose estimated posterior is closer to the clean posterior than $\alpha = 1$ in terms of KL divergence. Additionally, we note that $\alpha_0$ increases with increasing levels of noise.

data with $\alpha$-loss. This is highlighted in Table 1 where we see that not only does $\alpha_0 > 1$ outperform log-loss at every non-trivial noise level, but that as noise increases in the data (i.e., there is a greater twist), $\alpha_0$ increases as well. This suggests a connection between the level of noise and the optimal value of $\alpha$ for training a logistic model, an observation that is strengthened by the results of Proposition 3.3. We note however that selecting $\alpha$ larger than optimal degrades performance, indicating that the hyperparameter needs to be carefully tuned.

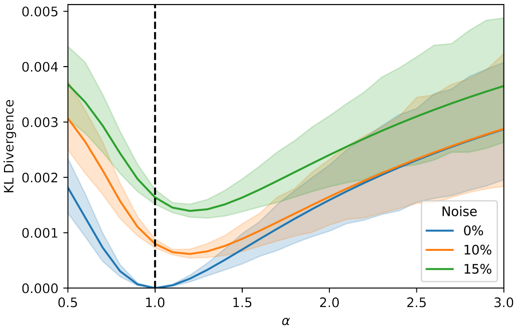


*Figure 2.* 95% confidence intervals around KL divergence vs. $\alpha$. We see that the KL divergence of $\eta_t^{(\alpha)}$ and $\eta_c$ is decreased by tuning $\alpha > 1$ but that $\alpha$ too large decreases effectiveness.

**Covariate-Level Robustness** We calculate the MSE between the clean and untwisted odds ratios as follows:

$$\mathrm{MSE}(\theta_c, \theta_t^{(\alpha)}) = \frac{\|\theta_c - \theta_t^{(\alpha)}\|^2}{d} \tag{11}$$

where $\theta_c$ is the vector of clean weights and $\theta_t^{(\alpha)}$ is the vector of weights learned by $\alpha$-loss on the twisted dataset.

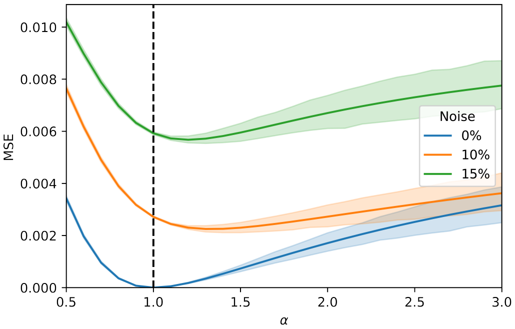


*Figure 3.* 95% confidence intervals around MSE vs. $\alpha$. We see that the MSE between the clean and twisted odds ratios is minimized for $\alpha > 1$.

We see in Figure 3 and Table 2 that the optimal $\alpha$ generally increases with the level of noise, and importantly we see that the optimal $\alpha > 1$, indicating that perhaps not only could the model output be preserved as indicated in Sypherd et al. (2021) but the model weights themselves.

### 4.4. Classification Metrics

While the main objective of the paper was to preserve risk metrics when training on noisy data, many applications may still make a hard decision. Classification metrics such as accuracy are strongly affected by class imbalance, so we additionally examine sensitivity to assess classification performance on the minority (positive) class. In both of these metrics, training with $\alpha$-loss shows some gains (see Figures 4a and 4b), suggesting that increased robustness in composite and covariate-level risk does not come at the cost of hard classification accuracy.

## 5. Conclusion

In this paper we have presented a method of combating label noise through the application of $\alpha$-loss. Our method preserves both composite and covariate-level risk inherent to the logistic model, and thus preserves the interpretability of the model when noise is present in the training data. We have examined the application of this method using the motivating example of noisy self-reported survey data, and have shown empirical results on a large COVID-19 self-reported survey dataset. These results indicate that a model learned through $\alpha$-loss outperforms log-loss in composite and covariate-level risk robustness.

## References

Gorber, S. C., Schofield-Hurwitz, S., Hardt, J., Levasseur, G., and Tremblay, M. The accuracy of self-reported smoking: A systematic review of the relationship between self-reported and cotinine-assessed smoking status. *Nicotine & Tobacco Research*, 11(1):12–24, 01 2009. ISSN 1462-2203. doi: 10.1093/ntr/ntn010. URL https://doi.org/10.1093/ntr/ntn010.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks, 2017.

Jewell, N. P., Lewnard, J. A., and Jewell, B. L. Predictive Mathematical Models of the COVID-19 Pandemic: Underlying Principles and Value of Projections. *JAMA*, 323(19):1893–1894, 05 2020. ISSN 0098-7484. doi: 10.1001/jama.2020.6585. URL https://doi.org/10.1001/jama.2020.6585.

Liang, W., Liang, H., Ou, L., Chen, B., Chen, A., Li, C., Li, Y., Guan, W., Sang, L., Lu, J., et al. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with covid-19. *JAMA internal medicine*, 180(8):1081–1089, 2020.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection, 2017. URL https://arxiv.org/abs/1708.02002.

Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., and Bailey, J. Normalized loss functions for deep learning with noisy labels. In *International Conference on Machine Learning*, pp. 6543–6553. PMLR, 2020.

Menni, C., Valdes, A. M., Freidin, M. B., Sudre, C. H., Nguyen, L. H., Drew, D. A., Ganesh, S., Varsavsky, T., Cardoso, M. J., El-Sayed Moustafa, J. S., Visconti, A., Hysi, P., Bowyer, R. C. E., Mangino, M., Falchi, M., Wolf, J., Ourselin, S., Chan, A. T., Steves, C. J., and Spector, T. D. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat Med*, 26(7):1037–1040, 07 2020.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks, 2021.

Pregibon, D. Logistic regression diagnostics. *The annals of statistics*, 9(4):705–724, 1981.

Rauscher, G. H., Johnson, T. P., Cho, Y. I., and Walk, J. A. Accuracy of Self-Reported Cancer-Screening Histories: A Meta-analysis. *Cancer Epidemiology, Biomarkers & Prevention*, 17(4):748–757, 04 2008. ISSN 1055-9965. doi: 10.1158/1055-9965.EPI-07-2629. URL https://doi.org/10.1158/1055-9965.EPI-07-2629.

Salomon, J. A., Reinhart, A., Bilinski, A., Chua, E. J., La Motte-Kerr, W., Rönn, M. M., Reitsma, M. B., Morris, K. A., LaRocca, S., Farag, T. H., Kreuter, F., Rosenfeld, R., and Tibshirani, R. J. The us covid-19 trends and impact survey: Continuous real-time measurement of covid-19 symptoms, risks, protective behaviors, testing, and vaccination. *Proceedings of the National Academy of Sciences*, 118(51), 2021. ISSN 0027-8424. doi: 10.1073/pnas.2111454118. URL https://www.pnas.org/content/118/51/e2111454118.

Sypherd, T., Nock, R., and Sankar, L. Being properly improper, 2021. URL https://arxiv.org/abs/2106.09920.

Sypherd, T., Diaz, M., Cava, J. K., Dasarathy, G., Kairouz, P., and Sankar, L. A tunable loss function for robust classification: Calibration, landscape, and generalization. *IEEE Transactions on Information Theory*, pp. 1–1, 2022. doi: 10.1109/TIT.2022.3169440.

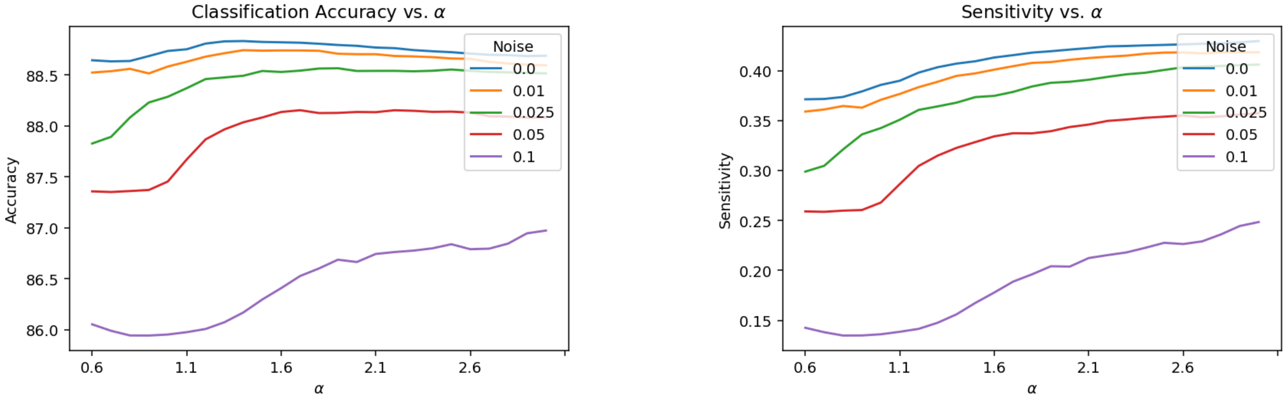
Tu, J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11):1225–1231, 1996. ISSN 0895-4356. doi: https://doi.org/10.1016/S0895-4356(96)00002-9. URL https://www.sciencedirect.com/science/article/pii/S0895435696000029.

## A. Risk Preservation Metrics

| | MSE of Model Weights | | |
|---|---|---|---|
| Noise,$p$ | $\alpha = 1$ MSE | $\alpha_0$ MSE | $\alpha_0$ value |
| 0 | $\mathbf{0}$ | 5.22e$-$5 | 1.1 |
| 1% | 3.33e$-$5 | $\mathbf{2.12}$e$-\mathbf{5}$ | 1.1 |
| 2.5% | 1.96e$-$4 | $\mathbf{1.00}$e$-\mathbf{4}$ | 1.2 |
| 5% | 7.31e$-$4 | $\mathbf{4.31}$e$-\mathbf{4}$ | 1.3 |
| 7.5% | 1.58e$-$3 | $\mathbf{1.13}$e$-\mathbf{3}$ | 1.4 |
| 10% | 2.72e$-$3 | $\mathbf{2.52}$e$-\mathbf{3}$ | 1.3 |
| 12.5% | 4.16e$-$3 | $\mathbf{3.78}$e$-\mathbf{3}$ | 1.3 |
| 15% | 5.92e$-$3 | $\mathbf{5.67}$e$-\mathbf{3}$ | 1.2 |

*Table 2.* Mean Squared Error of Model Weights. We see that for every non-zero noise level, $\alpha \neq 1$ is able to achieve a lower MSE than $\alpha = 1$.

## B. Classification Metric Figures



(a) Classification Accuracy for models trained under differing noise levels. We see that tuning $\alpha$ gives gains at every noise level.

(b) Sensitivity score for models under differing noise levels. Tuning $\alpha > 1$ gives an increase in sensitivity across noise levels.