Regretful Decisions under Label Noise

Anonymous authors
Paper under double-blind review

Abstract

Machine learning models are routinely used to support decisions that affect individuals – be it to screen a patient for a serious illness or to gauge their response to treatment. In these tasks, we are limited to learning models from datasets where the labels are subject to noise. In this work, we study the impact of learning under label noise at the instance level. We introduce a notion of *regret* for this regime, which measures the number of unforeseen mistakes when learning from noisy labels. We show that standard approaches to learn models from noisy labels can return models that perform well at a population level while subjecting individuals to a *lottery of mistakes*. We develop machinery to estimate the likelihood of mistakes at an instance level from a noisy dataset, by training models over plausible realizations of datasets without label noise. We present a comprehensive empirical study of label noise in clinical prediction tasks. Our results reveal how our failure to anticipate mistakes can compromise model reliance and adoption, and demonstrate how we can address these challenges by anticipating and abstaining from regretful decisions.

1 Introduction

Machine learning models are routinely used to support or automate decisions that affect individuals – be it to screen a patient for a mental illness [47] or estimate their risk for an adverse treatment response [2]. In such applications, we fit models from datasets with *label noise* – i.e., where the labels reflect a noisy observation of the outcome that we wish to predict. In practice, label noise may arise as a result of human annotation [e.g., due to inherent ambiguity 26] or measurement error [e.g., due to noisy readings from a wearable sensor 20]. In such cases, label noise can have detrimental effects on model performance [10].

Over the past decade, these challenges have led to extensive work on *learning from noisy datasets* [see 10, 45, for surveys]. These advances have improved our ability to mitigate label noise at a population level. In contrast, there has been little work studying the effects of label noise at an instance level. At a high level, this oversight reflects the fact that we cannot provide meaningful guarantees on individual predictions under label noise. Even in the best-case scenario, where we have perfectly specified distributional assumptions on label noise, we may learn a model that performs well on average but cannot identify the points where mistakes are made (see Fig. 1).

As shown in Fig. 1, when we learn under label noise, we build a model that predicts accurately but cannot determine where it makes its mistakes. In this regime, individuals are subject to a "lottery" of erroneous predictions. These effects handicap model reliance, as well as any downstream applications that rely on the correctness of individual predictions - e.g., model explanations [43, 44], post-hoc analyses [22, 30], clinical decision support [31].

In effect, label noise arises in many real-world applications where we use models to support or automate individual decisions [see, e.g., 52, for a recent metareview of 72 cases in medicine]. In decision support applications, our failure characterize the correctness of predictions may lead to overreliance – as physicians to rely on predictions that may be incorrect [5, 25, 29]. In applications for automation , our failure to characterize the correctness or confidence of predictions at an instance level– e.g., debugging [1, 22] or by abstention [9, 16].

In this work, we study how label noise affects these individual predictions. Our work is motivated by the fact that – even as we may be unable to resolve the effects of label noise at an instance level – we can mitigate harm and reap benefits from models through exposition and uncertainty quantification.

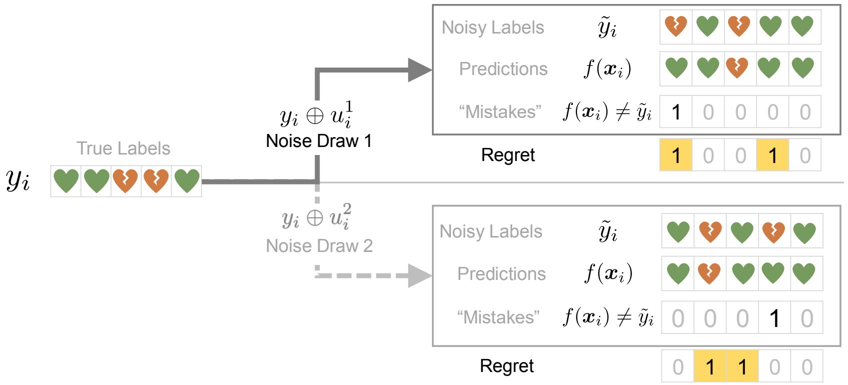

True Labels
y_i
$y_i \oplus u_i^1$
Noise Draw 1
$y_i \oplus u_i^2$
Noise Draw 2
Noisy Labels $\tilde{y}_i$
Predictions $f(\boldsymbol{x}_i)$
"Mistakes" $f(\boldsymbol{x}_i) \neq \tilde{y}_i$ 1 0 0 0 0
Regret 1 0 0 1 0
Noisy Labels $\tilde{y}_i$
Predictions $f(\boldsymbol{x}_i)$
"Mistakes" $f(\boldsymbol{x}_i) \neq \tilde{y}_i$ 0 0 0 1 0
Regret 0 1 1 0 0

Figure 1: Prediction problems with noisy labels only contain a single draw of label noise. In such tasks, we can learn a model that performs well at a population level but cannot anticipate its mistakes at an individual level. In such cases, *regret* characterizes the number of individuals who are subjected to a lottery of mistakes by measuring the difference between anticipated mistakes and actual mistakes.

Our goal is to reveal these effects and develop machinery to mitigate them. To this end, the main contributions of our work are as follows:

1. We introduce a notion of regret when learning from noisy datasets. Regret captures how uncertainty in labels affects individual predictions and can be generalized to other settings where a dataset exhibits uncertainty.
2. We show how learning under label noise leads to inevitable regret. Our analysis characterizes key limitations in a wide class of methods to learn from label noise.
3. We develop a method to flag regretful predictions by training models on plausible realizations of a clean dataset. Our method can measure the sensitivity of individual predictions under label noise and explore common noise assumptions while allowing control over plausibility.
4. We present results from a comprehensive empirical study on clinical prediction tasks. The results highlight the practical implications of label noise at the instance level, and demonstrate how our approach can support safety by flagging potential mistakes.

Related Work Our work is related to a stream of research on learning from noisy labels. We focus on applications where we cannot resolve label noise by acquiring clean labels [see e.g., 10, 45, for surveys]. Many methods learn models in this regime by hedging for uncertainty in labels [28, 36, 39]. As we show in Section 2, these approaches can mitigate loss in model performance at a population level yet assign unpredictable mistakes. In practice, the individuals who are subject to unpredictability exceeds the noise rate – meaning that many of them are subject to a lottery of mistakes. Our work highlights the limitations of this regime. In this sense, our results complement the work of Oyen et al. [38], who characterize the lack of robustness to label noise at a population level under general distributional assumptions.

We propose to mitigate these issues through a principled approach for uncertainty quantification. Our approach ties in with recent work on model multiplicity, which shows how changes in the machine learning pipeline can produce models that assign conflicting predictions [3, 6, 18, 32, 35, 48, 49] and lead to downstream effects on fairness, explanations, and recourse [4, 15, 23, 33]. With respect to the literature on label noise, our approach is similar to the work of Reed et al. [42], who propose training an ensemble of deep neural networks by sampling alternative realizations of clean labels. In contrast, our procedure samples plausible realizations of clean labels and retrains plausible models to quantify uncertainty at an individual level rather than to predict.

2 Framework

We consider a classification task where we wish to learn a model $f : \mathcal{X} \to \mathcal{Y}$ to accurately predict a label $y \in \mathcal{Y}$ from a feature vector $\boldsymbol{x} \in X \subseteq \mathbb{R}^d$.

In a standard classification task, we would be given a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where each example $(\boldsymbol{x}_i, y_i)$ is drawn from a joint distribution of random variables X and Y. Given the dataset, we would fit a model $f : \mathcal{X} \to \mathcal{Y}$ that performs well in deployment – i.e., that minimizes the *true risk* $R(f) := \mathbb{E}_{X,Y}[\mathbb{I}[f(X) \neq Y]]$.

We consider a variant of this task where we are given a noisy dataset $\tilde{\mathcal{D}} = \{(\boldsymbol{x}_i, \tilde{y}_i)\}_{i=1}^n$ where each *noisy label* $\tilde{y}_i$ represents an uncertain observation of a *true label* y_i. We represent the uncertainty through binary variable $u_i := \mathbb{I}[y_i \neq \tilde{y}_i]$, which indicates that the noisy label $\tilde{y}_i$ has been *flipped* from its true value y_i. Given u_i, we can express the noisy labels in terms of true labels and vice-versa:

$$\tilde{y}_i := y_i \oplus u_i \qquad\qquad y_i := \tilde{y}_i \oplus u_i.$$

Given a noisy dataset, we denote the flips for all n examples as a vector that we call the *noise draw*.

Definition 1. Given a binary classification task with n examples, the *noise draw* $\boldsymbol{u} = [u_1, \ldots, u_n] \subseteq \{0,1\}^n$ is a realization of n random variables $U := [U_1, \ldots, U_n] \subseteq \{0,1\}^n$,

Given an example $(\boldsymbol{x}_i, y_i)$, each u_i is drawn from Bernoulli distribution with parameters $p_{u|y_i,\boldsymbol{x}_i} := \Pr(U_i = 1 \mid X = \boldsymbol{x}_i, Y = y_i)$. Thus, each noisy label $\tilde{y}_i$ is generated by the random process:

$$U_i \sim \mathsf{Bern}(p_{u|y_i,\boldsymbol{x}_i})$$
$$\tilde{y}_i = y_i \oplus U_i$$

We assume that the parameters $p_{u|y_i,\boldsymbol{x}_i}$ are specified by a *noise model* such as those in Table 1. In what follows, we write $p_{u|y_i,\boldsymbol{x}_i}$ instead of p_u when its conditioning is clear from context.

We view the noise in a noisy dataset as the output of a single draw of label noise. We refer to this draw as the *true draw* and denote it $\boldsymbol{u}^{\text{true}} := [u_1^{\text{true}}, \ldots, u_n^{\text{true}}]$. In practice, the true draw $\boldsymbol{u}^{\text{true}}$ is fixed but unknown. From the perspective of a practitioner, $\boldsymbol{u}^{\text{true}}$ could be any realization of the random variable U. If they knew $\boldsymbol{u}^{\text{true}}$, they could trivially resolve label noise as they could recover the true labels for each point as $y_i = \tilde{y}_i \oplus u_i^{\text{true}}$.

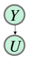
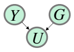
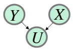
Noise Model	PGM	Parametric Representation	Inference Requirements	Sample Use Case
Uniform		$p_u = \Pr(U = 1)$	$q_u = \Pr(U = 1)$	Uniform measurement error
Class Level		$p_{u\|y} = \Pr(U = 1 \mid Y = y)$ $q_{u\|\tilde{y}} = \Pr\left(U = 1 \mid \tilde{Y} = \tilde{y}\right)$	$\pi_y = \Pr(Y = y)$	Data-driven discovery tasks where $\tilde{Y}$ is an experimental outcome confirmed by a hypothesis test with type I/II error [14]
Subgroup Level		$p_{u\|y,g} = \Pr(U = 1 \mid Y = y, G = g)$ $q_{u\|\tilde{y},g} = \Pr\left(U = 1 \mid \tilde{Y} = \tilde{y}, G = g\right)$	$\pi_{y,g} = \Pr(Y = y \mid G = g)$	Tasks where noise $\tilde{Y}$ changes based on annotator characteristics [46] or across patient subpopulations [12].
Feature Level		$p_{u\|y,\boldsymbol{x}} = \Pr(U = 1 \mid Y = y, X = \boldsymbol{x})$ $q_{u\|\tilde{y},\boldsymbol{x}} = \Pr\left(U = 1 \mid \tilde{Y} = \tilde{y}, X = \boldsymbol{x}\right)$	$\pi_{y,\boldsymbol{x}} = \Pr(Y = y, X = \boldsymbol{x})$	Chest X-ray diagnosis where label noise $\tilde{Y}$ changes based on image quality X and the disease Y [e.g., pneunomia vs COVID 13].

Table 1: Common noise models expressed in terms of the noise draw U. We represent each model as a probability distribution with parameters $p_{u|y,\boldsymbol{x}}$. Given a dataset with noisy labels, we infer noise draws using a posterior distribution with parameters $p_{u|\tilde{y},\boldsymbol{x}}$ and the prior distribution π_y. We assume that $p_{u|y,\boldsymbol{x}} < 0.5$ to ensure that there are more clean labels than noisy labels [36, 50].

On the Regret of Prediction Consider a practitioner who trains a model $f : \mathcal{X} \to \mathcal{Y}$ from a noisy dataset using an algorithm to learn from noisy labels. In such settings, they may be able to recover a model that performs well at a population level. However, they will be unable to determine where their model makes mistakes. In this regime, individuals are subject to a *lottery of mistakes*. We say that an individual are assigned a *regretful prediction* if they "win" this lottery.

Definition 2. Consider a classification task with label noise where we train a model $f : \mathcal{X} \to \mathcal{Y}$. We measure the *regret* for an example $(\boldsymbol{x}_i, \tilde{y}_i)$ as:

$$\text{Regret}(f(\boldsymbol{x}_i), \tilde{y}_i, U_i) := \mathbb{I}\left[e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i) \neq e^{\text{true}}(f(\boldsymbol{x}_i), y_i(U_i))\right] \tag{1}$$

Here, $e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i)$ denotes an *anticipated mistake*, and $e^{\text{true}}(f(\boldsymbol{x}_i), y_i(U_i)) := \mathbb{I}[f(\boldsymbol{x}_i) \neq y_i(U_i)]$ denotes an *actual mistake* with respect to the true label $y_i(U_i) = \tilde{y}_i \oplus U_i$.

In practice, $e^{\text{pred}}(\cdot)$ is determined by how we account for noise, if at all. If we fit a model via standard ERM on the noisy labels, then $e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i) = \mathbb{I}\left[f(\boldsymbol{x}_i) \neq \tilde{y}_i\right]$. If we fit a model using noise-tolerant ERM [e.g., 36, 39], then $e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i) := \tilde{\ell}_{01}(f(\boldsymbol{x}_i), \tilde{y}_i)$ where $\tilde{\ell}_{01}(\cdot)$ is a unbiased loss defined so that $\mathbb{E}_U[\tilde{\ell}_{01}(\boldsymbol{x}_i, y_i(U_i))] = \ell_{01}(f(\boldsymbol{x}_i), \tilde{y}_i)$.

One of the key benefits of studying regret in this setting is for exposition. Regret captures the irreducible error we incur due to aleatoric uncertainty in the noise draw U_i. In online learning, the concept of regret arises because we cannot foresee randomness in the *future*. In learning from noisy labels, regret arises because we cannot infer randomness from the *past*. Using regret, we can disambiguate the effects of label noise at the population level and the instance level, as shown through the following result.

Proposition 3. Consider a classification task where we learn a classifier f from a noisy dataset. Given a noisy example $(\boldsymbol{x}, \tilde{y})$ let $q_{u|\boldsymbol{x},\tilde{y}} := \Pr(U = 1 \mid \tilde{Y} = \tilde{y}, X = \boldsymbol{x})$. Then:

$$\mathbb{E}_U\left[\text{Regret}(f(\boldsymbol{x}), \tilde{y}, U)\right] = q_{u|\boldsymbol{x},\tilde{y}}.$$

Prop. 3 provides an opportunity to discuss several implications of label noise at the instance level. On the one hand, the result states that we can use the noise rate to gauge the *expected* number of anticipated mistakes. In practice, however, we cannot tell how these mistakes are distributed over instances. In this case, each instance where $q_{u|\boldsymbol{x},\tilde{y}} > 0$ is subjected to a lottery of mistakes. In a task where we have a uniform noise model with a noise rate of 5%, we would only to assign regretful predictions to 5% of instances. Even so, 100% of instances could be assigned an *unanticipated mistake* since the noise draw is always unknown.

On the Regret of Hedging Many algorithms for learning from noisy labels are designed to *hedge* against label noise [41]. Given a noisy dataset and a noise model, hedging seeks to minimize the *expected risk over all possible noise draws*. In some cases, algorithms may implement hedging explicitly via ERM with a modified loss [see e.g., 34, 36]. In others, the hedging may be implicit – e.g., by assigning sample weights to instances that are chosen to minimize expected risk over all possible draws [see e.g., 28, 39, 51]. In the best-case scenario, where we can correctly specify the noise model, we can expect algorithms that hedge to return a model that minimizes the expected excess risk with respect to all noise draws. In this case, we have $\mathbb{E}_{U|X,Y}[\Delta\text{Error}(f, \tilde{\mathcal{D}}, U)] = 0$ where:

$$\Delta\text{Error}(f, \tilde{\mathcal{D}}; U) := \underbrace{\sum_{i=1}^{n} e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i)}_{\text{Predicted Training Error}} - \underbrace{\sum_{i=1}^{n} e^{\text{true}}(f(\boldsymbol{x}_i), y_i)}_{\text{True Training Error}} \tag{2}$$

However, the resulting model f would still incur regret:

$$\text{Regret}(f, \tilde{\mathcal{D}}, U) := \sum_{i=1}^{n} \mathbb{I}\left[e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i) \neq e^{\text{true}}(f(\boldsymbol{x}_i), y_i(U_i))\right]. \tag{3}$$

We formalize this intuition in Prop. 3 where we show that despite $\Delta\text{Error}(f, \tilde{\mathcal{D}}; U) \approx 0$, $\text{Regret}(f, \tilde{\mathcal{D}}, U) > 0$ for the classical hedging algorithm of Natarajan et al. [36].

Proposition 4. Consider training a model $f : \mathcal{X} \to \mathcal{Y}$ on a noisy dataset via ERM with a modified loss function $\tilde{\ell} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ such that $\mathbb{E}_U[\tilde{\ell}(f(\boldsymbol{x}), \tilde{y})] = \ell(f(\boldsymbol{x}), y)$ for all $(\boldsymbol{x}, \tilde{y})$. In this case, the model minimizes risk for an *implicit noise draw* $\boldsymbol{u}^{\text{mle}} = [u_1^{\text{mle}} \ldots u_n^{\text{mle}}]$ where each u_i^{mle} corresponds to the value with maximal likelihood under the posterior noise model $q_{u|\tilde{y}_i,\boldsymbol{x}_i}$.

Prop. 4 implies that learning a model by hedging will incur regret — unless the noise in the dataset matches the implicit noise draw $\boldsymbol{u}^{\text{mle}} := \boldsymbol{u}^{\text{true}}$. In practice, this event is unlikely as $\Pr\left(\boldsymbol{u}^{\text{mle}} = \boldsymbol{u}^{\text{true}}\right)$ becomes vanishingly small as $n \to \infty$ (see Appendix A). In some cases, $\Pr\left(\boldsymbol{u}^{\text{mle}} = \boldsymbol{u}^{\text{true}}\right) = 0$ in a finite sample regime because the implicit noise draw is unrealizable.

3 Anticipating Mistakes with Plausible Models

In this section, we describe a principled approach to anticipate regretful predictions given a dataset of noisy labels and a noise model.

3.1 Algorithm

Seeing how regret stems from our inability to anticipate mistakes at an instance level, We want to produce information that can help us anticipate mistakes at the instance level. Specifically, we wish to estimate the likelihood of assigning a regretful prediction to each instance, We refer to this quantity as *ambiguity* and estimate it using models that we train using the procedure in Algorithm 1. Given a noisy dataset and a noise model, this procedure generates plausible realizations of a clean dataset and trains a set of plausible models to estimate ambiguity. In practice, we can use these estimates as confidence scores for a model that we learn under label noise. In this way, we can reap benefits from a wide range of techniques that use confidence scores for selective classification [11] or for active learning [8].

Sampling Plausible Draws Given a noisy dataset $\tilde{\mathcal{D}}$, noise model $p_{u|y,\boldsymbol{x}}$, and prior distribution $\pi_{y,\boldsymbol{x}} := \Pr(Y = y \mid X = \boldsymbol{x})$, we sample noise draws from the posterior distribution:

$$q_{u|\tilde{y},\boldsymbol{x}} = \frac{(1-\pi_{\tilde{y},\boldsymbol{x}})\cdot p_{u|1-\tilde{y},\boldsymbol{x}}}{p_{u|\tilde{y},\boldsymbol{x}}\cdot(1-\pi_{\tilde{y},\boldsymbol{x}}) + (1-p_{u|\tilde{y},\boldsymbol{x}})\cdot \pi_{\tilde{y},\boldsymbol{x}}} \tag{4}$$

In principle, one can sample noise draws from the posterior distribution in Eq. (4). In practice, this approach can output *atypical noise draws* – i.e., "edge case" draws that are highly unlikely under a given noise model.[1] In settings where we wish to estimate ambiguity using a fixed number of draws, atypical draws represent can severely bias our estimates and undermine their utility. Although we can moderate such effects by constructing estimates using more draws, this has practical challenges: we would need to train a large number of models. Given these challenges, we sample noise draws in a way that can control for their atypicality.

Definition 5. Given a noise draw $\boldsymbol{u} \in \{0,1\}^n$, let $q_{u|\tilde{y}} := \Pr(U = 1 \mid \tilde{Y} = \tilde{y})$ denote its true posterior noise rate, and $\hat{q}_{u|\tilde{y}} := \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left[u_i = 1 \mid \tilde{y}_i = y\right]$ denote its estimate. Given any $\epsilon \in [0,1]$, the *set of plausible draws* contains all draws whose empirical distribution is within ϵ of the true posterior noise rate:

$$\mathcal{U}_\epsilon(\tilde{\boldsymbol{y}}) := \{\boldsymbol{u} \in \{0,1\}^n \mid |\hat{q}_{u|\tilde{y}} - q_{u|\tilde{y}}| < \epsilon \cdot q_{u|\tilde{y}} \text{ for all } u \in \{0,1\}\}.$$

The set of plausible draws is a strongly typical set and its behavior follows well-known results in information theory [7]. Given a noisy dataset where n is large, for example, we can expect most draws to concentrate in $\mathcal{F}_\epsilon^{\text{plaus}}$ [see, e.g., Theorem 3.1.2 in 7]. We can control the typicality of draws by setting the atypicality parameter ϵ, which represents the relative deviation in noise rate from $q_{u|\tilde{y}}$. In practice, this parameter can be set apriori: given a uniform noise model with a noise rate of $q_{u|\tilde{y}} = 0.1$, we can set $\epsilon = 0.2$ to consider draws that flip between 8% to 12% of instances. In settings where we wish to consider a specific noise draw $\boldsymbol{u}_0$, we can set ϵ to guarantee that $\boldsymbol{u} \in \mathcal{F}_\epsilon^{\text{plaus}}$ with high probability (see Prop. 9 in Appendix A.2). By default, we set $\epsilon = 0.1$ to consider draws within 10% of what we would expect.

Training the Set of Plausible Models Given a plausible noise draw $\boldsymbol{u}^k$, we construct a *plausible* realization of a clean dataset by pairing each feature vector $\boldsymbol{x}_i$ with a *plausible* realization of the true label $\hat{y}_i^k = u^k \oplus \tilde{y}_i$.

Definition 6. The *set of ϵ-plausible models* contains all models trained using ϵ-plausible datasets:

$$\mathcal{F}_\epsilon^{\text{plaus}} := \left\{ \hat{f} \in \operatorname*{argmin}_{f\in\mathcal{F}} \hat{R}(f,\hat{\mathcal{D}}) \mid \hat{\mathcal{D}} := \{(\boldsymbol{x}_i, \hat{y}_i^k)\}_{i=1}^n, \boldsymbol{u} \in \mathcal{U}_\epsilon(\tilde{\boldsymbol{y}}) \right\}.$$

3.2 Estimation

In a idealized case where we would recover a plausible draw that matches the true draw $\boldsymbol{u}^k = u^{\text{true}}$, our procedure would return a plausible dataset $\hat{\mathcal{D}}^k$ and model $\hat{f}^k$ that perfectly flags all regretful

[1] For example, a noise draw that flips 30% of labels under a uniform noise model with a noise rate of 10%.

Algorithm 1 Generate Plausible Draws, Datasets, and Models

```
Input  noisy dataset $(\boldsymbol{x}_i, \tilde{y}_i)_{i=1}^n$, noise model $p_{u|y}$, number of models $m \geq 1$, atypicality $\epsilon \in [0,1]]$
Initialize  $\hat{\mathcal{F}}_\epsilon^{\text{plaus}} \leftarrow \{\}$
repeat
    $u_i \leftarrow U \sim \text{Bern}(p_{u|\tilde{y},\boldsymbol{x}})$ for $i \in [n]$          generate noise draw by posterior inference
    if $\boldsymbol{u} = [u_1, \ldots, u_n] \in \mathcal{U}_\epsilon$ then          check if draw is plausible (i.e., Def. 5)
        $\hat{y}_i \leftarrow \tilde{y}_i \oplus u_i$ for $i \in [n]$
        $\hat{\mathcal{D}}^k \leftarrow \{(\boldsymbol{x}_i, \hat{y}_i)\}_{i=1}^n$          construct plausible clean dataset
        $\hat{f}^k \leftarrow \text{argmin}_{f \in \mathcal{F}} \hat{R}(f; \hat{\mathcal{D}}^k)$          train plausible model
        $\hat{\mathcal{F}}_\epsilon^{\text{plaus}} \leftarrow \hat{\mathcal{F}}_\epsilon^{\text{plaus}} \cup \{\hat{f}^k\}$          update plausible models
until $|\hat{\mathcal{F}}_\epsilon^{\text{plaus}}| = m$
Output  $\hat{\mathcal{F}}_\epsilon^{\text{plaus}}$, sample of $m$ plausible models from $\mathcal{F}_\epsilon^{\text{plaus}}$
```

predictions. Seeing how $\boldsymbol{u}^{\text{true}}$ is unknown, we repeat this process m times and use the m plausible models $\mathcal{F}_\epsilon^{\text{plaus}}$ to estimate the prevalence of an anticipated mistake for each point in our dataset. We refer to this measure as *ambiguity* and define it below.

Definition 7. Given a noisy example $(\boldsymbol{x}, \tilde{y}) \in \tilde{\mathcal{D}}$, we measure its expected *ambiguity* over the set of ϵ-plausible models as:

$$\mu(\boldsymbol{x}) := \Pr\left(f(\boldsymbol{x}) \neq \hat{y} \mid f \in \mathcal{F}_\epsilon^{\text{plaus}}, \hat{y} = u \oplus \tilde{y}, u \sim q_{u|\tilde{y}}\right). \tag{5}$$

Given a set of m plausible models, we can estimate ambiguity using the sample mean:

$$\hat{\mu}(\boldsymbol{x}) := \frac{1}{m} \sum_{k \in [m]} \mathbb{I}\left[\hat{f}^k(\boldsymbol{x}) \neq \hat{y}^k\right]. \tag{6}$$

Ambiguity measures the likelihood of a mistake at the instance level. This measure incorporates information from the noise distribution (i.e., by considering multiple *plausible* realizations of the true labels), and our learning process (i.e., by training models for each set of clean labels). We formalize this intuition in Appendix B.

3.3 Discussion

The reliability of our estimates depends on the following modeling assumptions:

Typicality of the True Noise Draw: The first assumption is that the true noise draw $\boldsymbol{u}^{\text{true}}$ is a typical noise draw. Although the true draw is unknown, we can assume that most draws to be typical given results in typical set theory [7].

Noise Model: Our estimates will also depend on the specification of the noise model p_u. As we show in Fig. 2, the impact of depends on the degree of misspecification. In the worse case – e.g., if we assume the noise rate is 5% when in reality it is 20% – misspecification can lead to highly unreliable estimates as always sample edge cases. In practice, we can moderate the potential effect of misspecification. For example, when working with simple noise models – e.g., uniform or class level – we can be conservative and assume a higher noise rate or choose a higher ϵ to capture a larger set of plausible draws. In settings where we are unsure of the noise model, we can generate a data-driven estimate using the noisy dataset [see e.g., 27, 28, 39].

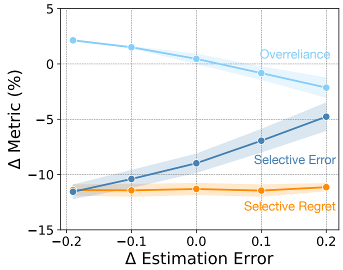


Figure 2: Impact of misspecifying the noise model for an LR model for `shock_mimic` dataset. We consider a setting where label noise is drawn from a uniform noise model with a true noise rate of 20%, but we estimate ambiguity using a misspecified noise rates from between $[1\%, 40\%]$. As shown, misspecification leads to moderate effects on overreliance and selective error but does not affect selective regret.

4 Experiments

In this section, we present results from an empirical study of label noise in clinical prediction tasks. Our goals are to highlight the effects of label noise at the instance level, and to evaluate the ability of our approach to identify and abstain from regretful predictions. We include additional details and results in Appendix C, and code to reproduce our results in an anonymized repository.

Setup We work with 5 classification datasets from real-world clinical applications where models are used to support individual medical decisions (see Table 3, and Appendix C for details). We split each dataset into a training sample (80%, to fit a model), and a test sample (20%, to measure out-of-sample performance). Starting with the true labels from each dataset, we generate noisy labels by sampling noise draws from 6 noise distributions: 2 noise models (uniform, class level) × 3 noise rates per model $[5\%, 20\%, 40\%]$. We use the noisy datasets to fit a logistic regression model (LR) and a neural network (DNN) using two training procedures: (1) Ignore, where we ignore label noise and fit a model to predict noisy labels; and (2) Hedge where we hedge against noisy labels using the method of Natarajan et al. [36]. This setup yields 24 models for each dataset: 6 noise regimes × 2 model classes × 2 training procedures. For each model, train a sample of $m = 200$ plausible models from a plausible set with $\epsilon = 10\%$ using the procedure in Section 3, estimate the ambiguity of each training instance as per Eq. (6).

Results In Table 3, we report summary statistics on the accuracy, reliability, and ambiguity of predictions at a population level and an individual level (see Table 2). These results characterize a single noise draw that is unknown to practitioners. We include results for alternate noise draws to show that these trends generalize (see Appendix C).

Metric	Definition	Description
TrueError(f)	$\frac{1}{n}\sum_{i\in[n]} e^{\text{true}}(f(\boldsymbol{x}_i), y_i)$	Error rate of f on true training labels
ΔError(f)	$\frac{1}{n}\sum_{i\in[n]} e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i) - e^{\text{true}}(f(\boldsymbol{x}_i), y_i)$	Difference in true error between a model trained on true labels and and a model trained on noisy labels. Note: $\Delta\text{Error}(f) \approx 0$ for Hedge
Ambiguity(f)	$\underset{i\in[n]}{\text{Median}}(\hat{\mu}(\boldsymbol{x}_i))$	Median estimate ambiguity across all instances subject to label noise
Regret(f)	$\frac{1}{n}\sum_{i\in[n]} \mathbb{I}\left[e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i) \neq e^{\text{true}}(f(\boldsymbol{x}_i), y_i)\right]$	Average regret over all points. Given any dataset with class-level label noise, we have that $\text{Regret}(f) = \sum_y q_{u\mid y} \cdot \pi_y$
Overreliance(f)	$\frac{1}{n}\sum_{i\in[n]} \mathbb{I}\left[e^{\text{true}}(f(\boldsymbol{x}_i), y_i) = 1, e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i) = 0\right]$	Proportion of predictions from f that are incorrectly perceived as accurate

Table 2: Overview of summary statistics in Table 3. We report these metrics for models that we train from noisy labels using a specific training procedure, model class, noise model, and dataset. We evaluate all models trained on a given dataset and noise model using a fixed noise draw. We assume that the noise model is correctly specified, and that the noise draw is unknown at training time.

On Regretful Predictions Our results in Table 3 highlight several implications of learning from label noise that we describe in Section 2. Our results highlight how we can rely on Prop. 3 to gauge the *expected* number of regretful predictions in practice. In particular, we see that average regret is roughly equal to effective noise rate in Prop. 3. We observe that Prop. 3 only characterizes the expected prevalence of regretful predictions – meaning that it cannot help us tell which predictions incur regret or how regretful predictions may be distributed across examples. In practice, regretful predictions can be affect *any* instance that is subject to label noise. In Table 3, we show results for a class-level noise model where we flip positive instances ($y_i = 1$). Thus, every instance where $\tilde{y} = 0$ would takes part in the lottery of mistakes.

Our results show we may fail to reap benefits from models as a result of such distributional effects – e.g., though overreliance or disparate impact. In Table 3, we highlight these effects by reporting *overreliance* – i.e., the fraction of instances where incorrectly assume that a model assigns a correct prediction. Overreliance is a key measure for decision support: in clinical applications, for example, we wish to expect physicians to rely on predictions that are correct. On the `lungcancer` dataset, under 40% noise, 19.7% of individuals are assigned a regretful prediction by a standard LR model. Among them, 33.0% to 73.1% correspond to mistakes that would lead to overreliance. In the `mortality` dataset, for example, we find that regret is not evenly distributed across subgroups

		$p_{u\|y=1} = 5\%$				$p_{u\|y=1} = 20\%$				$p_{u\|y=1} = 40\%$			
		LR		DNN		LR		DNN		LR		DNN	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	24.0%	23.0%	13.1%	11.9%	26.7%	24.6%	18.3%	26.3%	38.6%	25.1%	28.3%	25.8%
	ΔError	-1.4%	-1.7%	-1.4%	-1.3%	-1.7%	-5.0%	-2.1%	2.3%	10.0%	-8.3%	6.5%	-5.0%
	Ambiguity	6.0%	6.0%	6.0%	6.0%	20.5%	20.5%	20.5%	20.5%	36.0%	36.0%	36.0%	36.0%
	Regret	3.0%	3.0%	3.0%	3.0%	10.1%	10.1%	10.1%	10.1%	19.7%	19.7%	19.7%	19.7%
	Overreliance	0.8%	0.6%	0.8%	0.8%	4.2%	2.6%	4.0%	6.2%	14.8%	5.7%	13.1%	7.3%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	21.9%	21.3%	15.3%	15.2%	24.3%	20.9%	18.5%	16.5%	33.1%	21.2%	29.1%	25.9%
	ΔError	-1.2%	-1.3%	-1.9%	-1.8%	-2.4%	-6.0%	-6.5%	-6.7%	5.5%	-11.8%	2.2%	-11.7%
	Ambiguity	5.5%	5.5%	5.5%	5.5%	18.5%	18.5%	18.5%	18.5%	33.0%	33.0%	33.0%	33.0%
	Regret	2.5%	2.5%	2.5%	2.5%	10.2%	10.2%	10.2%	10.2%	19.8%	19.8%	19.8%	19.8%
	Overreliance	0.7%	0.6%	0.3%	0.4%	3.9%	2.1%	1.8%	1.8%	12.7%	4.0%	11.0%	4.1%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	31.6%	31.2%	30.0%	29.5%	32.5%	31.3%	31.4%	30.2%	39.3%	31.6%	43.2%	29.6%
	ΔError	-0.5%	-0.7%	-1.1%	-0.7%	-0.1%	-3.0%	-0.3%	-3.3%	9.0%	-6.7%	13.6%	-5.4%
	Ambiguity	5.5%	5.5%	5.5%	5.5%	18.5%	18.5%	18.5%	18.5%	31.5%	31.5%	31.5%	31.5%
	Regret	2.5%	2.5%	2.5%	2.5%	10.0%	10.0%	10.0%	10.0%	19.7%	19.7%	19.7%	19.7%
	Overreliance	1.0%	0.9%	0.7%	0.9%	4.9%	3.5%	4.8%	3.3%	14.4%	6.5%	16.7%	7.2%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	20.1%	20.1%	17.6%	18.0%	21.2%	19.7%	19.2%	18.1%	30.6%	19.9%	27.1%	18.7%
	ΔError	-1.3%	-1.5%	-1.4%	-1.3%	-3.9%	-6.2%	-4.1%	-5.9%	3.0%	-10.9%	0.1%	-10.6%
	Ambiguity	5.0%	5.0%	5.0%	5.0%	18.0%	18.0%	18.0%	18.0%	31.5%	31.5%	31.5%	31.5%
	Regret	2.2%	2.2%	2.2%	2.2%	9.8%	9.8%	9.8%	9.8%	19.5%	19.5%	19.5%	19.5%
	Overreliance	0.5%	0.4%	0.4%	0.5%	2.9%	1.8%	2.9%	1.9%	11.2%	4.3%	9.8%	4.4%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	33.7%	33.5%	28.7%	29.3%	35.4%	33.5%	32.0%	35.4%	42.7%	34.1%	41.2%	42.1%
	ΔError	-0.2%	-0.5%	0.5%	-0.0%	1.5%	-2.4%	3.2%	1.5%	12.4%	-4.5%	14.9%	14.4%
	Ambiguity	6.0%	6.0%	6.0%	6.0%	20.5%	20.5%	20.5%	20.5%	36.5%	36.5%	36.5%	36.5%
	Regret	2.6%	2.6%	2.6%	2.6%	10.0%	10.0%	10.0%	10.0%	19.6%	19.6%	19.6%	19.6%
	Overreliance	1.2%	1.1%	1.6%	1.3%	5.8%	3.8%	6.6%	5.7%	16.0%	7.6%	17.3%	17.0%

Table 3: Accuracy, reliability, and ambiguity of models across model classes, training procedures, and noise regimes. We show results when learning from a noisy dataset where under a class-level noise model where we flip 5%, 20% and 40% of instances (e.g., diagnostic error). We include results for other draws of the noise in Appendix C.

defined by `age` or `sex`. Specifically, we find that 40% label noise leads to twice the regret in older patients than younger patients, despite noise rates being uniform across both subgroups (Fig. 5 in Appendix C). We find similar effects across datasets, model classes, and noise regimes. Overall, these results underscore the need to measure the effects of label noise empirically – especially in tasks where we care about how a model performs over subclasses and subpopulations.

On Learning by Hedging Our results highlight can learn models that are robust to noise at a population level but but that assign mistakes by lottery. As shown in Table 3, we observe that $\Delta\text{Error} \approx 0$ and $\text{Regret}(f) > 0$ across experimental conditions. In general, we find that Hedge can moderate the impact of label noise at a population level – leading to lower values of $\Delta\text{Error} \approx 0$. On the `mortality` dataset, for example, Hedge reduces the error rate by almost 10% compared to Ignore for a DNN model under 40% label noise. As shown, these issues do not resolve regret.

On Promoting Safety by Anticipating Mistakes The only way to flag regretful predictions is by obtaining clean labels, which is often impossible or infeasible in practice. Our method in Algorithm 1 flags these points using the noise model and noisy dataset, without clean labels. We train plausible models on plausible versions of the clean dataset to flag "mistakes". Our results show this reliably detects regretful instances. As seen in Table 3, median ambiguity correlates with regret across datasets and noise rates. This holds for multiple label noise draws (see Appendix C). In practice, our approach supports tasks like selective classification or active learning. For example, in clinical predictions, we can abstain from uncertain predictions using ambiguity estimates. We use a confidence threshold rule $\mathbb{I}\left[\text{conf}(\boldsymbol{x}_i) \leq \tau\right]$ where $\text{conf}(\boldsymbol{x}_i)$ is the confidence score (either $1 - \mu(\boldsymbol{x}_i)$ or $\hat{p}(\tilde{y}_i \mid \boldsymbol{x}_i)$). Fig. 3 shows that abstaining on 20% of the dataset (keeping 80% coverage) reduces regret by 5% and risk by 6%. By contrast, the standard approach requires abstaining from all predictions to achieve comparable regret.

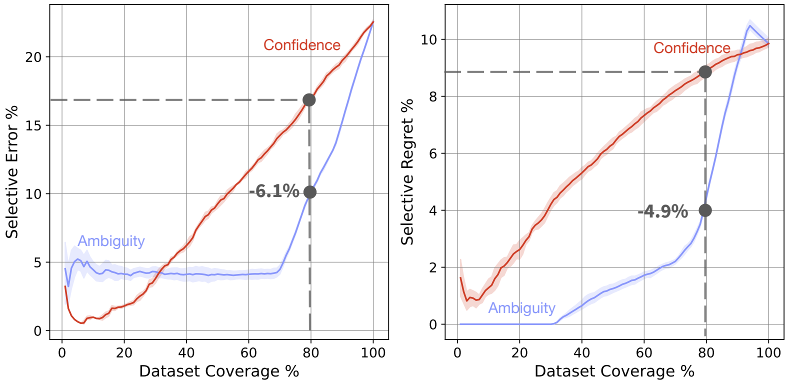


Figure 3: Risk-coverage curves for an LR model on the `shock_mimic` dataset when we abstain from uncertain predictions using predicted probabilities (red) and ambiguity (blue). We show the selective error (left) and selective regret (right) when we abstain from predictions using a confidence-based threshold rule $\mathbb{I}\left[\text{conf}(\boldsymbol{x}_i) \leq \tau\right]$ as vary $\tau \in (0, 1)$, setting $\text{conf}(\boldsymbol{x}_i) := 1 - \boldsymbol{x}_i$ for our approach, and $\text{conf}(\boldsymbol{x}_i) := \hat{p}(\tilde{y}_i \mid \boldsymbol{x}_i)$ for the standard approach. As shown, due to the ability of Ambiguity to effectively identify uncertain predictions, we can achieve lower error and lower regret by abstaining on fewer instances.

5 Demonstration

We now demonstrate how our approach can benefit data-driven gene-enhancer pair identification. We consider a classification task where our dataset encodes the conditions and outcomes of expensive in-vitro experiments. These datasets exhibit label noise due to hypothesis test outcomes with known Type I and II errors. Our goal is to train a reliable model on these experimental outcomes to predict results for new experiments, enabling the prioritization of experiments. This improves the discovery rate of enhancers, a crucial step in the drug discovery pipeline.

Setup We work with a noisy dataset that summarizes the conditions and outcomes of $n = 9,372$ in-vitro experiments. Here, each experiment is associated with a noisy example $(\boldsymbol{x}_i, \tilde{y}_i)$, where $\boldsymbol{x}_i$ encodes $d = 13$ characteristics of the experimental unit, and $\tilde{y}_i$ represents the outcome of a hypothesis test – i.e., $\tilde{y}_i = 1$ if we reject a null hypothesis. Here, each label is subject to label noise as a result of Type I and Type II error of each hypothesis test: We can consider this scenario as class level noise: Type 1 occurs when $\tilde{y} = 1, y = 0$, and Type 2 when $\tilde{y} = 0, y = 1$. Type 1 error is controlled at 5%, while Type 2 varies by the statistical power of the experiment. Our dataset contains these values for each instance. We use these values to specify the parameters of our noise model. In this case, the resulting noise model exhibits label noise across labels and subgroups.

We split our dataset into a training sample (80%) and a test sample (20%). We use the training sample to train a classifier using ERM, and the test sample to estimate its performance. In this case, we are specifically interested in evaluating the reliability of predictions for “successful" experiments. We identify these cases using test instances where the true experimental outcome was a significant result, and evaluate the performance of our model using test AUPRC and Accuracy. Using this setup we compare two different approaches: (1) a standard approach where we would abstain on uncertain experiments according to $\hat{p}(\tilde{y}_i \mid x_i)$ or (2) our proposed approach where we identify and abstain on ambiguous experiments using Algorithm 1.

Results We report the results in Fig. 4. As shown, we can improve accuracy (+1.4%) and AUPRC (+19.5%) compared to standard confidence-based abstention, with a modest 4% abstention rate (Fig. 4). This demonstrates a real-world scenario where our methods can identify mistakes to improve model performance. Our methods can enhance data-driven discovery by accurately predicting experimental outcomes before they take place, accounting for inherent Type I and Type II error rates. This can help optimize laboratory resource allocation and increase the discovery rate of EG regulatory elements.

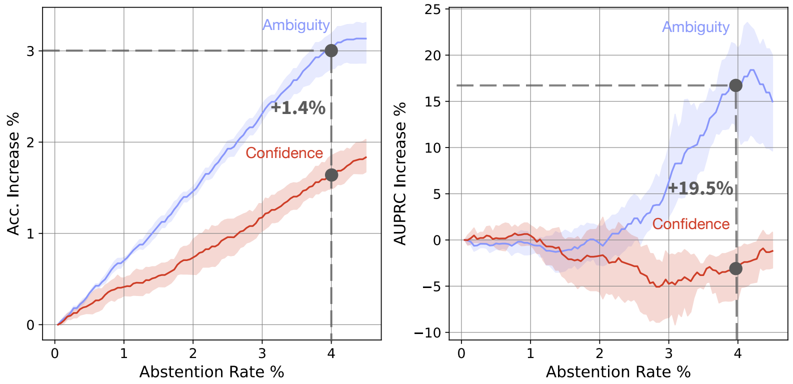


Figure 4: Demonstration of selective classification performance for an LR model on the enhancer dataset. When we abstain from uncertain instances according to ambiguity rates (blue), we can improve both accuracy (+1.4%) and AUPRC (+19.5%) compared to a standard approach (red) by abstaining on only 4% of instances. The noise model here comes from Type 1 and Type 2 errors from statistical hypothesis testing.

6 Concluding Remarks

Learning under label noise presents many hurdles to practitioners. Even if we can learn models that perform well on average, individuals may be subject to a lottery of mistakes – e.g., even with a model boasting 99% accuracy, a small amount of label noise could subject *anyone* to mistakes.

These instance level effects are often overlooked in favor of population level performance. In this work, we studied these limitations through the lens of regret for learning under label noise. Our results highlighted the prevalence of regret in various healthcare decision-support tasks and the inherent limitations of existing label noise learning strategies in mitigating for regret. We then demonstrate an abstention procedure using our proposed measures of ambiguity which can capture instance level uncertainty and lead to safer decisions.

Our work shows that even as regret is inevitable – we can understand and mitigate its effects through uncertainty quantitation. In particular, we can flag regretful predictions by estimating their ambiguity. This analysis can calibrate our reliance on individual predictions – signaling the need to collect more data or avoid prediction altogether – or be used to support formal approaches such as selective classification and active learning. By magnifying the instance level impact of label noise through the lens of regret, we can perform more reliable and safer predictions on individuals in critical tasks.

References

[1] Adebayo, Julius, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429*, 2020.

[2] Bhagwat, Atharva M, Kadija S Ferryman, and Jason B Gibbons. Mitigating algorithmic bias in opioid risk-score modeling to ensure equitable access to pain relief. *Nature medicine*, 29(4):769–770, 2023.

[3] Black, Emily, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 850–863, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533149. URL https://doi.org/10.1145/3531146.3533149.

[4] Brunet, Marc-Etienne, Ashton Anderson, and Richard Zemel. Implications of model indeterminacy for explanations of automated decisions. *Advances in Neural Information Processing Systems*, 35:7810–7823, 2022.

[5] Chiang, Chun-Wei and Ming Yin. You'd better stop! understanding human reliance on machine learning models under covariate shift. In *Proceedings of the 13th ACM Web Science Conference 2021*, pages 120–129, 2021.

[6] Coston, Amanda, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. *CoRR*, abs/2101.00352, 2021. URL https://arxiv.org/abs/2101.00352.

[7] Cover, Thomas M. *Elements of Information Theory*. John Wiley & Sons, 1999.

[8] El-Yaniv, Ran and Yair Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13(2), 2012.

[9] Franc, Vojtech, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023.

[10] Frénay, Benoît and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on neural networks and learning systems*, 25(5):845–869, 2013.

[11] Geifman, Yonatan and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

[12] Gianattasio, Kan Z, Christina Prather, M Maria Glymour, Adam Ciarleglio, and Melinda C Power. Racial disparities and temporal trends in dementia misdiagnosis risk in the united states. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5:891–898, 2019.

[13] Giannakis, Athanasios, Dorottya Móré, Stella Erdmann, Laurent Kintzelé, Ralph Michael Fischer, Monika Nadja Vogel, David Lukas Mangold, Oyunbileg Stackelbergvon , Paul Schnitzler, Stefan Zimmermann, et al. Covid-19 pneumonia and its lookalikes: How radiologists perform in differentiating atypical pneumonias. *European Journal of Radiology*, 144:110002, 2021.

[14] Gschwind, Andreas R, Kristy S Mualim, Alireza Karbalayghareh, Maya U Sheth, Kushal K Dey, Evelyn Jagoda, Ramil N Nurtdinov, Wang Xi, Anthony S Tan, Hank Jones, et al. An encyclopedia of enhancer-gene regulatory interactions in the human genome. *bioRxiv*, 2023.

[15] Hamman, Faisal, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust algorithmic recourse under model multiplicity with probabilistic guarantees. *IEEE Journal on Selected Areas in Information Theory*, 2024.

[16] Hendrickx, Kilian, Lorenzo Perini, Dries PlasVan der , Wannes Meert, and Jesse Davis. Machine learning with a reject option: A survey. *Machine Learning*, pages 1–38, 2024.

[17] Hollenberg, SM. Cardiogenic shock. In *Intensive Care Medicine: Annual Update 2003*, pages 447–458. Springer, 2003.

[18] Hsu, Hsiang and Flavio du Pin Calmon. Rashomon capacity: A metric for predictive multiplicity in probabilistic classification, 2022. URL https://arxiv.org/abs/2206.01295.

[19] Johnson, Alistair EW, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[20] Kanji, Jamil N, Nathan Zelyas, Clayton MacDonald, Kanti Pabbaraju, Muhammad Naeem Khan, Abhaya Prasad, Jia Hu, Mathew Diggle, Byron M Berenger, and Graham Tipples. False negative rate of covid-19 pcr testing: a discordant testing analysis. *Virology journal*, 18:1–6, 2021.

[21] Knaus, William A, Frank E Harrell, Joanne Lynn, Lee Goldman, Russell S Phillips, Alfred F Connors, Neal V Dawson, William J Fulkerson, Robert M Califf, Norman Desbiens, et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203, 1995.

[22] Koh, Pang Wei and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[23] Kulynych, Bogdan, Hsiang Hsu, Carmela Troncoso, and Flavio P Calmon. Arbitrary decisions are a hidden cost of differentially private training. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1609–1623, 2023.

[24] Le Gall, Jean-Roger, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963, 1993.

[25] Lee, John D and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[26] Li, Dana, Lea Marie Pehrson, Lea Tøttrup, Marco Fraccaro, Rasmus Bonnevie, Jakob Thrane, Peter Jagd Sørensen, Alexander Rykkje, Tobias Thostrup Andersen, Henrik Steglich-Arnholm, et al. Inter-and intra-observer agreement when using a diagnostic labeling scheme for annotating findings on chest x-rays—an early step in the development of a deep learning-based decision support system. *Diagnostics*, 12(12): 3112, 2022.

[27] Li, Xuefeng, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. In *International conference on machine learning*, pages 6403–6413. PMLR, 2021.

[28] Liu, Yang and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. *ICML*, 2020.

[29] Lu, Zhuoran and Ming Yin. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. 2021.

[30] Lundberg, Scott M, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[31] Lundberg, Scott M, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[32] Marx, Charles, Flavio P. Calmon, and Berk Ustun. Predictive Multiplicity in Classification, 2019.

[33] Marx, Charles, Youngsuk Park, Hilaf Hasson, Yuyang Wang, Stefano Ermon, and Luke Huan. But are you sure? an uncertainty-aware perspective on explainable ai. In *International Conference on Artificial Intelligence and Statistics*, pages 7375–7391. PMLR, 2023.

[34] Menon, Aditya, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pages 125–134, 2015.

[35] Meyer, Anna P, Aws Albarghouthi, and Loris D'Antoni. The dataset multiplicity problem: How unreliable data impacts predictions. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 193–204, 2023.

[36] Natarajan, Nagarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

[37] NCI, Surveillance Research Program, DCCPS. Surveillance, epidemiology, and end results (seer) program research data (1975-2016), 2019. URL www.seer.cancer.gov.

[38] Oyen, Diane, Michal Kucer, Nicolas Hengartner, and Har Simrat Singh. Robustness to label noise depends on the shape of the noise distribution. *Advances in Neural Information Processing Systems*, 35:35645–35656, 2022.

[39] Patrini, Giorgio, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

[40] Pollard, Tom J, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[41] Ravi, Ramamoorthi and Amitabh Sinha. Hedging uncertainty: Approximation algorithms for stochastic optimization problems. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 101–115. Springer, 2004.

[42] Reed, Scott, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

[43] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[44] Simonyan, Karen. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[45] Song, Hwanjun, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on neural networks and learning systems*, 2022.

[46] Sylolypavan, Aneeta, Derek Sleeman, Honghan Wu, and Malcolm Sim. The impact of inconsistent human annotations on ai driven clinical decision making. *NPJ Digital Medicine*, 6(1):26, 2023.

[47] Ustun, Berk, Lenard A Adler, Cynthia Rudin, Stephen V Faraone, Thomas J Spencer, Patricia Berglund, Michael J Gruber, and Ronald C Kessler. The world health organization adult attention-deficit/hyperactivity disorder self-report screening scale for dsm-5. *Jama psychiatry*, 74(5):520–526, 2017.

[48] Watson-Daniels, Jamelle, David C Parkes, and Berk Ustun. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10306–10314, 2023.

[49] Watson-Daniels, Jamelle, Flavio du Pin Calmon, Alexander D'Amour, Carol Long, David C Parkes, and Berk Ustun. Predictive churn with the set of good models. *arXiv preprint arXiv:2402.07745*, 2024.

[50] Wei, Jiaheng, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*, 2021.

[51] Wei, Jiaheng, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Yang Liu. To smooth or not? when label smoothing meets noisy labels. In *International Conference on Machine Learning*, 2022.

[52] Wei, Yishu, Yu Deng, Cong Sun, Mingquan Lin, Hongmei Jiang, and Yifan Peng. Deep learning with noisy labels in medical prediction problems: a scoping review. *arXiv preprint arXiv:2403.13111*, 2024.

Appendices

A Omitted Proofs

A.1 Results from Section 2

Proof of Prop. 3

Proof. Consider any classification task with label noise. Let $\rho_{X,\tilde{Y}} := \Pr\left(U = 1 \mid X, \tilde{Y}\right)$ denote the noise rate for a point with $(X, \tilde{Y})$ and let $\ell_{01}(f(X), \tilde{Y}) := \mathbb{I}\left[f(X) \neq \tilde{Y}\right]$ denote the zero-one loss.

We first start by showing that using the unbiasedness property of Hedging algorithms such as Natarajan et al. [36], we can achieve zero error in expectation. That is, $\mathbb{E}_{X,Y,U}[e^{\text{pred}}(f(X), \tilde{Y}) - e^{\text{true}}(f(X), Y)] = 0$:

$$
\begin{aligned}
&\mathbb{E}_{X,Y,U}\left[e^{\text{pred}}(f(X), \tilde{Y}) - e^{\text{true}}(f(X), Y)\right] \\
&= \mathbb{E}_{X,Y} E_{U|X,Y}\left[e^{\text{pred}}(f(X), \tilde{Y}) - e^{\text{true}}(f(X), Y)\right] = 0
\end{aligned}
$$

The last line follows from the fact that $\tilde{Y}$ is deterministic from U given Y, and the unbiasedness property: $\mathbb{E}_{U|X,Y}[e^{\text{pred}}(f(X), \tilde{Y})] = e^{\text{true}}(f(X), Y)$

We are now ready to show that despite achieving zero error, we can still incur regret. We begin by expressing the expected regret for any point $(X, \tilde{Y})$ and any noise draw U as:

$$
\begin{aligned}
&\mathbb{E}_{X,\tilde{Y},U}\left[\text{Regret}(X, \tilde{Y}, U)\right] \\
&= \mathbb{E}_{X,\tilde{Y}}\left[(1 - 2q_u) \cdot (e^{\text{pred}}(f(X), \tilde{Y}) + \ell_{01}(f(X), \tilde{Y})) + 2(q_u - 1) \cdot e^{\text{pred}}(f(X), \tilde{Y}) \cdot \ell_{01}(f(X), \tilde{Y}) + q_u\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{X,\tilde{Y},U}\left[\text{Regret}(X, \tilde{Y}, U)\right] &= \mathbb{E}_{X,\tilde{Y},U}\left[\mathbb{I}\left[e^{\text{pred}}(f(X), \tilde{Y}) \neq \mathbb{I}\left[f(X) \neq \tilde{Y}(1-U) + (1-\tilde{Y})U\right]\right]\right] \\
&= \mathbb{E}_{X,\tilde{Y}}\mathbb{E}_{U|X,\tilde{Y}}\left[\mathbb{I}\left[e^{\text{pred}}(f(X), \tilde{Y}) \neq \mathbb{I}\left[f(X) \neq \tilde{Y}(1-U) + (1-\tilde{Y})U\right]\right]\right] \\
&= \mathbb{E}_{X,\tilde{Y}}\mathbb{E}_{U|X,\tilde{Y}}\left[e^{\text{pred}}(f(X), \tilde{Y})(1 - \mathbb{I}\left[f(X) \neq \tilde{Y}(1-U) + (1-\tilde{Y})U\right])\right. \\
&\quad \left. + (1 - e^{\text{pred}}(f(X), \tilde{Y}))\mathbb{I}\left[f(X) \neq \tilde{Y}(1-U) + (1-\tilde{Y})U\right]\right] \\
&= \mathbb{E}_{X,\tilde{Y}}\mathbb{E}_{U|X,\tilde{Y}}\left[e^{\text{pred}}(f(X), \tilde{Y})(1 - \mathbb{I}\left[f(X) \neq \tilde{Y}\right](1-U) - \mathbb{I}\left[f(X) \neq 1 - \tilde{Y}\right]U)\right. \\
&\quad \left. + (1 - e^{\text{pred}}(f(X), \tilde{Y}))(\mathbb{I}\left[f(X) \neq \tilde{Y}\right](1-U) + \mathbb{I}\left[f(X) \neq 1 - \tilde{Y}\right]U)\right]
\end{aligned}
$$

Letting $q_u = \Pr\left(U = 1 \mid X, \tilde{Y}\right)$ and $\ell_{01}(f(X), \tilde{Y}) = \mathbb{I}\left[f(X) \neq \tilde{Y}\right]$, we have:

$$
\begin{aligned}
&= \mathbb{E}_{X,\tilde{Y}}\left[(1 - q_u)(e^{\text{pred}}(f(X), \tilde{Y})(1 - \ell_{01}(f(X), \tilde{Y})) + (1 - e^{\text{pred}}(f(X), \tilde{Y}))\ell_{01}(f(X), \tilde{Y}))\right. \\
&\quad \left. + q_u(e^{\text{pred}}(f(X), \tilde{Y})(1 - \ell_{01}(f(X), 1 - \tilde{Y})) + (1 - e^{\text{pred}}(f(X), \tilde{Y}))\ell_{01}(f(X), 1 - \tilde{Y}))\right] \\
\mathbb{E}_{X,\tilde{Y},U}\left[\text{Regret}(X, \tilde{Y}, U)\right] &= \mathbb{E}_{X,\tilde{Y}}\left[(1 - 2q_u) \cdot (e^{\text{pred}}(f(X), \tilde{Y}) + \ell_{01}(f(X), \tilde{Y}))\right. \\
&\quad \left. + 2(q_u - 1) \cdot e^{\text{pred}}(f(X), \tilde{Y}) \cdot \ell_{01}(f(X), \tilde{Y}) + q_u\right].
\end{aligned}
$$

When there is no label noise, we have that $q_u = 0$ and $e^{\text{pred}}(f(X), \tilde{Y}) = \ell_{01}(f(X), \tilde{Y})$ for all $X, \tilde{Y}$. Because they are binary terms, in this regime, we have:

$$
\mathbb{E}_{X,\tilde{Y},U}\left[\text{Regret}(X, \tilde{Y}, U)\right] = \mathbb{E}_{X,\tilde{Y}}\left[0\right] = 0
$$

When there is label noise, we have that $q_u > 0$ for some $X, \tilde{Y}$. In this regime, we have:

$$\mathbb{E}_{X,\tilde{Y},U}\left[\text{Regret}(X,\tilde{Y},U)\right] = \mathbb{E}_{X,\tilde{Y}}\left[q_u\right] > 0.$$

□

We now introduce Prop. 8 to setup the proof for Prop. 4:

Proposition 8. Minimizing the expected risk under the clean label distribution is equivalent to minimizing a noise-corrected risk under the noisy label distribution

$$\mathbb{E}_{X,Y}\left[\mathbb{I}\left[f(X) \neq Y\right]\right] = \mathbb{E}_{X,\tilde{Y}}\left[(1 - q_u\mathbb{I}\left[f(X) \neq \tilde{Y}\right] + q_u\mathbb{I}\left[f(X) \neq 1 - \tilde{Y}\right]\right] \tag{7}$$

Here:

- $q_u = \frac{(1-\pi_{\tilde{y},\boldsymbol{x}})\cdot p_{u|1-\tilde{y},\boldsymbol{x}}}{p_{u|\tilde{y},\boldsymbol{x}}\cdot(1-\pi_{\tilde{y},\boldsymbol{x}})+(1-p_{u|\tilde{y},\boldsymbol{x}})\cdot\pi_{\tilde{y},\boldsymbol{x}}}$
- $\pi_{\tilde{y},\boldsymbol{x}} = \Pr\left(Y = \tilde{y}|X = \boldsymbol{x}\right)$ is the clean class prior an observed noisy label,
- $p_u = \Pr\left(U = 1 \mid Y = y, X = \boldsymbol{x}\right)$ is the class-level noise probability.

Proof of Prop. 8

Proof. The result is analogous to Lemma 1 in Natarajan et al. [36]. In what follows, we include an additional proof for the sake of completeness.

$$\begin{aligned}
\text{ExpectedRisk}(f) &= \mathbb{E}_{X,Y}\left[\mathbb{I}\left[f(X) \neq Y\right]\right] \\
&= \mathbb{E}_{X,\tilde{Y},U}\left[\mathbb{I}\left[f(X) \neq \tilde{Y}(1-U) + U(1-\tilde{Y})\right]\right] \\
&= \mathbb{E}_{X,\tilde{Y}}\mathbb{E}_{U|X,\tilde{Y}}\left[\mathbb{I}\left[f(X) \neq \tilde{Y}(1-U) + U(1-\tilde{Y})\right]\right] \\
&= \mathbb{E}_{X,\tilde{Y}}\mathbb{E}_{U|X,\tilde{Y}}\left[\mathbb{I}\left[f(X) \neq \tilde{Y}\right](1-U) + \mathbb{I}\left[f(X) \neq 1-\tilde{Y}\right]U\right] \\
&= \mathbb{E}_{X,\tilde{Y}}\left[\mathbb{E}_{U|X,\tilde{Y}}[\mathbb{I}\left[f(X) \neq \tilde{Y}\right](1-U)] + \mathbb{E}_{U|X,\tilde{Y}}[\mathbb{I}\left[f(X) \neq 1-\tilde{Y}\right]U]\right] \\
&= \mathbb{E}_{X,\tilde{Y}}\left[\Pr\left(U = 0|\tilde{Y},X\right)\mathbb{I}\left[f(X) \neq \tilde{Y}\right] + \Pr\left(U = 1|\tilde{Y},X\right)\mathbb{I}\left[f(X) \neq 1-\tilde{Y}\right]\right] \\
&= \mathbb{E}_{X,\tilde{Y}}\left[\Pr\left(Y = \tilde{Y}|\tilde{Y},X\right)\mathbb{I}\left[f(X) \neq \tilde{Y}\right] + \Pr\left(Y \neq \tilde{Y}|\tilde{Y},X\right)\mathbb{I}\left[f(X) \neq 1-\tilde{Y}\right]\right] \\
&= \mathbb{E}_{X,\tilde{Y}}\left[(1-q_u\mathbb{I}\left[f(X) \neq \tilde{Y}\right] + q_u\mathbb{I}\left[f(X) \neq 1-\tilde{Y}\right]\right]
\end{aligned}$$

We write q_u in terms of the clean class priors and class-level noise probabilities using Bayes theorem. □

Proof of Prop. 4

Proof. We define u^{mle} as a noise draw $\boldsymbol{u}$ such that using u^{mle} to minimize the Expected Risk implicitly coincides with the true minimizer of the Expected Risk (defined in Prop. 8). That is:

$$\begin{aligned}
&\operatorname*{argmin}_{f\in\mathcal{F}}\mathbb{E}_{X,\tilde{Y}}\left[\mathbb{I}\left[f(X) \neq \tilde{Y}(1-\boldsymbol{u}) + \boldsymbol{u}(1-\tilde{Y})\right]\right] \\
&= \operatorname*{argmin}_{f\in\mathcal{F}}\mathbb{E}_{X,\tilde{Y}}\left[(1-q_u)\mathbb{I}\left[f(X) \neq \tilde{Y}\right] + q_u\mathbb{I}\left[f(X) = \tilde{Y}\right]\right]
\end{aligned}$$

We can express the LHS as:

$$f' \in \operatorname*{argmin}_{f\in\mathcal{F}}\mathbb{E}_{X,\tilde{Y}}\left[\mathbb{I}\left[f(X) \neq \tilde{Y}(1-\boldsymbol{u}) + \boldsymbol{u}(1-\tilde{Y})\right]\right] \tag{8}$$

$$= \operatorname*{argmin}_{f\in\mathcal{F}}\mathbb{E}_{X,\tilde{Y}}\left[(1-\boldsymbol{u})\mathbb{I}\left[f(X) \neq \tilde{Y}\right] + \boldsymbol{u}\mathbb{I}\left[f(X) = \tilde{Y}\right]\right] \tag{9}$$

We can denote the minimizer of the RHS:

$$\hat{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \mathbb{E}_{X,\tilde{Y}} \left[(1-q_u)\mathbb{I}\left[f(X) \neq \tilde{Y}\right] + q_u \mathbb{I}\left[f(X) = \tilde{Y}\right] \right] \tag{10}$$

Observe that:

$$\begin{aligned} q_{u|y,\boldsymbol{x}} < 0.5 &\implies \hat{f}(X) = \tilde{Y} \\ q_{u|y,\boldsymbol{x}} > 0.5 &\implies \hat{f}(X) = 1 - Y \end{aligned}$$

Thus, we have that $\boldsymbol{u} := \mathbb{I}\left[q_u > 0.5\right] \implies \hat{f} = f'$, as desired. Further, we can show that this u^{mle} is likely never u^{true}:

$$\lim_{n\to\infty} \Pr\left(\boldsymbol{u}^{\text{mle}} = \boldsymbol{u}^{\text{true}}\right) = \lim_{n\to\infty} \prod_{i=1}^{n} \Pr\left(u_i^{\text{mle}} = u_i^{\text{true}}\right) = 0 \tag{11}$$

□

A.2 Other Results

On the Sample Size for Typicality and Selection of ϵ

Proof of Prop. 9. Our goal is to show:

$$\Pr\left(u^{\text{true}} \in \mathcal{U}_\epsilon(\tilde{\boldsymbol{y}})\right) \geq 1 - \delta$$

The uncertainty set $\mathcal{U}_\epsilon(\tilde{\boldsymbol{y}})$ defined on $p_{u|\tilde{y}}$ is a strongly-typical set where the true mean $p_{u|y}$ and the empirical mean is $\hat{p}_u := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[u_i = 1\right]$. Thus,

$$u^{\text{true}} \in \mathcal{U}_\epsilon(\tilde{\boldsymbol{y}}) \Leftrightarrow |\hat{p}_u - p_{u|\tilde{y}}| \leq p_{u|\tilde{y}} \cdot \epsilon \tag{12}$$

We will derive conditions to satisfy the left-hand side of Eq. (12)

Observe that we can write

$$\begin{aligned} |\hat{p}_u - p_{u|\tilde{y}}| &= |(\hat{p}_u - p_u) + (p_u - p_{u|\tilde{y}})| \\ &\leq |\hat{p}_u - p_u| + |p_u - p_{u|\tilde{y}}| && \text{(by the triangle inequality)} \end{aligned}$$

We require $|\hat{p}_u - p_{u|\tilde{y}}| \leq p_{u|\tilde{y}} \cdot \epsilon |\hat{p}_u - p_u|$. Therefore we need $|\hat{p}_u - p_u| + |p_u - p_{u|\tilde{y}}| \leq p_{u|\tilde{y}} \cdot \epsilon$ which implies that $|\hat{p}_u - p_u| \leq p_{u|\tilde{y}} \cdot \epsilon - |p_u - p_{u|\tilde{y}}|$

We can now apply Hoeffding's inequality as u^{true} is a sequence of bounded, independently sampled random variables, let $\alpha = p_{u|\tilde{y}} \cdot \epsilon - |p_u - p_{u|\tilde{y}}|$:

$$\Pr\left(|\hat{p}_u - p_u| \geq \alpha\right) \leq 2 \cdot \exp(-2n\alpha^2)$$

Rearranging, we have that:

$$\Pr\left(u^{\text{true}} \in \mathcal{U}_\epsilon(\tilde{\boldsymbol{y}})\right) = \Pr\left(|\hat{p}_u - p_u| \leq \alpha\right) \geq 1 - 2 \cdot \exp(-2n\alpha^2) = 1 - 2 \cdot \exp(-2n(p_{u|\tilde{y}} \cdot \epsilon - |p_u - p_{u|\tilde{y}}|)^2)$$

We can invert this bound to obtain the following statement: with probability at least $1 - \delta$, $\Pr\left(u^{\text{true}} \in \mathcal{U}_\epsilon(\tilde{\boldsymbol{y}})\right)$ if we the number of samples n obeys:

$$n \geq \frac{-\ln\left(\frac{\delta}{2}\right)}{2(p_{u|\tilde{y}} \cdot \epsilon - |p_u - p_{u|\tilde{y}}|)^2}$$

To conclude the proof, we rearrange for ϵ, that is, given a dataset:

$$\epsilon \geq \frac{1}{p_{u|\tilde{y}}} \left(\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}} + |p_u - p_{u|\tilde{y}}| \right)$$

□

B Supporting Material for Section 3

In this Appendix, we present theoretical results related to our proposed approach.

B.1 On Ambiguity and Regret

To see our intuition, since $u \sim q_{u|\tilde{y}}$, we have $\hat{y} \sim \Pr\left(Y \mid \tilde{Y} = \tilde{y}\right)$, that is $\hat{y}$ represents the Bayes-optimal estimate of the true label given the noisy label.

Thus ambiguity captures the uncertainty in estimating the true label and the model's prediction. Specifically,

$$\begin{aligned}
&\Pr\left(f(\boldsymbol{x}) \neq Y \mid \tilde{Y} = \tilde{y}\right) \\
=&\sum_y \Pr\left(f(\boldsymbol{x}) = y \mid \tilde{Y} = \tilde{y}\right) \cdot \Pr\left(Y \neq y \mid \tilde{Y} = \tilde{y}\right)
\end{aligned}$$

The above is due to the conditional independence between f and $\hat{Y}$. Furthermore,

$$\begin{aligned}
&\sum_y \Pr\left(f(\boldsymbol{x}) = y \mid \tilde{Y} = \tilde{y}\right) \cdot \Pr\left(Y \neq y \mid \tilde{Y} = \tilde{y}\right) \\
=&\sum_y \Pr\left(f(\boldsymbol{x}) = y \mid \tilde{Y} = \tilde{y}\right) \cdot \left(1 - \Pr\left(Y = y \mid \tilde{Y} = \tilde{y}\right)\right) \\
=&1 - \sum_y \Pr\left(f(\boldsymbol{x}) = y \mid \tilde{Y} = \tilde{y}\right) \cdot \Pr\left(Y = y \mid \tilde{Y} = \tilde{y}\right)
\end{aligned}$$

Note that

$$\begin{aligned}
&\sum_y \Pr\left(f(\boldsymbol{x}) = y \mid \tilde{Y} = \tilde{y}\right) \cdot \Pr\left(Y = y \mid \tilde{Y} = \tilde{y}\right) \\
\leq&\frac{1}{2}\left(\sum_y \left(\Pr\left(f(\boldsymbol{x}) = y \mid \tilde{Y} = \tilde{y}\right)\right)^2 + \sum \left(\Pr\left(Y = y \mid \tilde{Y} = \tilde{y}\right)\right)^2\right)
\end{aligned}$$

Here the inequality holds with equality when $\Pr\left(f(\boldsymbol{x}) = y \mid \tilde{Y} = \tilde{y}\right) = \Pr\left(Y = y \mid \tilde{Y} = \tilde{y}\right)$. The term $\sum_y \left(\Pr\left(f(\boldsymbol{x}) = y \mid \tilde{Y} = \tilde{y}\right)\right)^2 + \sum \left(\Pr\left(Y = y \mid \tilde{Y} = \tilde{y}\right)\right)^2$ maximizes when there exists only one y, y' such that $\Pr\left(f(\boldsymbol{x}) = y \mid \tilde{Y} = \tilde{y}\right) = 1, \Pr\left(Y = y' \mid \tilde{Y} = \tilde{y}\right) = 1$ – i.e., the model prediction and the inferred true label have no ambiguity. More generally $\frac{1}{2}\left(\sum_y \left(\Pr\left(f(\boldsymbol{x}) = y \mid \tilde{Y} = \tilde{y}\right)\right)^2 + \sum \left(\Pr\left(Y = y \mid \tilde{Y} = \tilde{y}\right)\right)^2\right)$ is smaller when f and Y carry more ambiguity, achieving minimum when $f(\boldsymbol{x}|\tilde{Y})$ and $Y|\tilde{Y}$ are uniformly distributed.

B.2 On Choosing an Atypicality Parameter

Proposition 9. Given a set of n_p instances $(\boldsymbol{x}, \tilde{y})$ subject to noise rate p_u, we can determine the minimum ϵ to ensure with that any draw of noise falls within our set of plausible draws $\mathcal{F}_\epsilon^{\text{plaus}}$ with high probability. That is, with probability at least $1 - \delta$, $\boldsymbol{u} \in \mathcal{U}_\epsilon(\tilde{\boldsymbol{y}})$ if ϵ obeys:

$$\epsilon \geq \frac{1}{q_{u|\tilde{y}}}\left(\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n_p}} + |p_u - q_{u|\tilde{y}}|\right).$$

Here n_p represents the number of instances under the same noise model. For example, under class level noise, this bound would need to be evaluated separately using the number of instances for each class.

In practice, we can use this bound to set the atypicality parameter ϵ. For example, given a dataset with $n = 10,000$ instances under 20% uniform label noise, for example, a practitioner must set $\epsilon \geq 6\%$ to ensure that the $\boldsymbol{u} \in \mathcal{F}_\epsilon^{\text{plaus}}$ with probability at least 90%.

C Supporting Material for Section 4

C.1 Datasets

lungcancer We used a cohort of 120,641 lung cancer patients diagnosed between 2004-2016 who were monitored in the National Cancer Institute SEER study [37]. The outcome variable is death within five years from any cause, with 16.9% dying within this period. The cohort includes patients across the USA (California, Georgia, Kentucky, New Jersey, and Louisiana), excluding those lost to follow-up. Features include measures of tumor morphology and histology (e.g., size, metastasis, stage, node count and location), as well as clinical interventions at the time of diagnoses (e.g., surgery, chemotherapy, radiology).

shock_eicu & shock_mimic Cardiogenic shock is an acute cardiac condition where the heart fails to sufficiently pump enough blood [17] leading to under-perfusion of vital organs. These datasets are designed to build algorithms to predict cardiogenic shock in ICU patients. Both datasets contain identical features, group attributes, and outcome variables but they capture different patient populations. The shock_eicu dataset includes records from the EICU Collaborative Research Database V2.0 [40], while the shock_mimic dataset includes records from the MIMIC-III database [19]. The target variable is whether a patient with cardiogenic shock will die in the ICU. Features include vital signs and routine lab tests (e.g., systolic BP, heart rate, hemoglobin count) collected within 24 hours before the onset of cardiogenic shock.

saps The Simplified Acute Physiology Score II (SAPS II) score is a risk-score designed to predict the risk of death in ICU patients [24]. The data contains records of 7,797 patients from 137 medical centers in 12 countries. The outcome variable indicates whether a patient dies in the ICU, with 12.8% patient of patients dying. Similar to the other datasets, saps contains features reflecting comorbidities, vital signs, and lab measurements.

support This dataset comprises 9,105 ICU patients from five U.S. medical centers, collected during 1989-1991 and 1992-1994 [21]. Each record pertains to patients across nine disease categories: acute respiratory failure, chronic obstructive pulmonary disease, congestive heart failure, liver disease, coma, colon cancer, lung cancer, multiple organ system failure with malignancy, and multiple organ system failure with sepsis. The aim is to determine the individual-level 2- and 6-month survival rates based on physiological, demographic, and diagnostic data.

C.2 Results for Additional Noise Draws

		5				20				40			
		LR		NN		LR		NN		LR		NN	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	23.3%	23.4%	11.8%	14.0%	27.7%	22.6%	17.9%	16.8%	38.5%	23.2%	35.2%	27.3%
	ΔError	-1.2%	-1.2%	-1.2%	-0.9%	-1.3%	-4.9%	-2.1%	-4.7%	9.4%	-11.3%	12.3%	-12.7%
	Ambiguity	6.0%	6.0%	6.0%	6.0%	20.5%	20.5%	20.5%	20.5%	36.5%	36.5%	36.5%	36.5%
	Regret	2.1%	2.1%	2.1%	2.1%	10.2%	10.2%	10.2%	10.2%	20.1%	20.1%	20.1%	20.1%
	Overreliance	0.5%	0.5%	0.5%	0.6%	4.4%	2.6%	4.1%	2.7%	14.8%	4.4%	16.2%	3.7%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	21.0%	20.2%	15.2%	15.5%	23.7%	20.2%	17.9%	16.2%	33.8%	20.3%	32.6%	25.0%
	ΔError	-1.3%	-1.4%	-1.8%	-1.6%	-2.7%	-6.0%	-6.5%	-5.6%	5.6%	-12.8%	5.5%	-11.2%
	Ambiguity	5.5%	5.5%	5.5%	5.5%	18.5%	18.5%	18.5%	18.5%	33.0%	33.0%	33.0%	33.0%
	Regret	2.3%	2.3%	2.3%	2.3%	9.7%	9.7%	9.7%	9.7%	19.8%	19.8%	19.8%	19.8%
	Overreliance	0.5%	0.4%	0.2%	0.3%	3.5%	1.9%	1.6%	2.1%	12.7%	3.5%	12.6%	4.3%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	31.2%	31.2%	29.7%	29.9%	33.6%	31.0%	31.3%	29.6%	43.0%	31.4%	49.8%	30.3%
	ΔError	-0.6%	-0.7%	-1.1%	-1.1%	0.6%	-3.4%	-0.9%	-4.5%	13.2%	-6.5%	19.8%	-7.9%
	Ambiguity	5.5%	5.5%	5.5%	5.5%	18.5%	18.5%	18.5%	18.5%	31.5%	31.5%	31.5%	31.5%
	Regret	2.4%	2.4%	2.4%	2.4%	9.9%	9.9%	9.9%	9.9%	19.8%	19.8%	19.8%	19.8%
	Overreliance	0.9%	0.8%	0.7%	0.6%	5.3%	3.2%	4.5%	2.7%	16.5%	6.7%	19.8%	6.0%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	19.4%	19.6%	17.4%	17.8%	22.0%	19.8%	19.1%	18.2%	28.2%	19.9%	26.2%	18.6%
	ΔError	-1.3%	-1.4%	-1.4%	-1.3%	-3.1%	-5.6%	-3.9%	-5.5%	1.4%	-11.0%	-0.4%	-11.5%
	Ambiguity	5.0%	5.0%	5.0%	5.0%	18.0%	18.0%	18.0%	18.0%	31.5%	31.5%	31.5%	31.5%
	Regret	2.3%	2.3%	2.3%	2.3%	9.7%	9.7%	9.7%	9.7%	19.8%	19.8%	19.8%	19.8%
	Overreliance	0.5%	0.4%	0.4%	0.5%	3.3%	2.1%	2.9%	2.1%	10.6%	4.4%	9.7%	4.2%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	33.6%	33.6%	28.5%	28.8%	36.4%	33.9%	31.9%	29.9%	43.7%	35.3%	41.6%	38.6%
	ΔError	-0.7%	-0.8%	-0.6%	-0.3%	1.6%	-2.5%	1.8%	0.5%	13.1%	-3.4%	15.3%	7.0%
	Ambiguity	6.0%	6.0%	6.0%	6.0%	21.0%	21.0%	21.0%	21.0%	37.5%	37.5%	37.5%	37.5%
	Regret	2.5%	2.5%	2.5%	2.5%	10.0%	10.0%	10.0%	10.0%	19.9%	19.9%	19.9%	19.9%
	Overreliance	0.9%	0.9%	1.0%	1.1%	5.8%	3.7%	5.9%	5.2%	16.5%	8.3%	17.6%	13.5%

Table 4: Overview of performance and regret for models trained on all datasets, training procedures, and model classes. Noise draw 2.

		5				20				40			
		LR		NN		LR		NN		LR		NN	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	23.8%	22.7%	12.1%	12.6%	27.3%	23.4%	14.3%	21.3%	36.6%	24.5%	27.4%	26.2%
	ΔError	-0.9%	-1.2%	-0.4%	-1.1%	-1.3%	-5.0%	-5.5%	-3.3%	6.9%	-10.9%	6.0%	-4.9%
	Ambiguity	6.0%	6.0%	6.0%	6.0%	20.5%	20.5%	20.5%	20.5%	37.0%	37.0%	37.0%	37.0%
	Regret	2.3%	2.3%	2.3%	2.3%	10.2%	10.2%	10.2%	10.2%	18.9%	18.9%	18.9%	18.9%
	Overreliance	0.7%	0.5%	0.9%	0.6%	4.4%	2.6%	2.3%	3.4%	12.9%	4.0%	12.4%	7.0%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	21.6%	20.8%	16.0%	16.3%	24.2%	21.0%	15.5%	16.5%	32.1%	20.5%	33.5%	26.8%
	ΔError	-1.0%	-1.2%	-1.8%	-1.7%	-2.5%	-5.3%	-7.1%	-6.2%	4.3%	-11.5%	6.0%	-10.2%
	Ambiguity	5.5%	5.5%	5.5%	5.5%	18.5%	18.5%	18.5%	18.5%	33.0%	33.0%	33.0%	33.0%
	Regret	2.4%	2.4%	2.4%	2.4%	9.8%	9.8%	9.8%	9.8%	19.3%	19.3%	19.3%	19.3%
	Overreliance	0.7%	0.6%	0.3%	0.3%	3.6%	2.2%	1.3%	1.8%	11.8%	3.9%	12.7%	4.5%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	31.4%	31.1%	30.1%	30.5%	33.5%	30.9%	31.7%	29.2%	43.3%	31.4%	49.8%	29.4%
	ΔError	-0.5%	-0.7%	-1.2%	-0.8%	0.8%	-3.3%	-0.4%	-4.9%	13.2%	-6.5%	20.0%	-6.0%
	Ambiguity	5.5%	5.5%	5.5%	5.5%	18.5%	18.5%	18.5%	18.5%	31.5%	31.5%	31.5%	31.5%
	Regret	2.6%	2.6%	2.6%	2.6%	10.0%	10.0%	10.0%	10.0%	20.0%	20.0%	20.0%	20.0%
	Overreliance	1.0%	0.9%	0.7%	0.9%	5.4%	3.4%	4.8%	2.5%	16.6%	6.7%	20.0%	7.0%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	19.7%	19.6%	18.0%	17.5%	21.9%	19.9%	19.5%	18.4%	27.0%	20.0%	29.4%	20.0%
	ΔError	-1.4%	-1.5%	-1.5%	-1.4%	-3.4%	-5.9%	-4.4%	-6.0%	0.3%	-11.8%	3.1%	-9.9%
	Ambiguity	5.0%	5.0%	5.0%	5.0%	18.0%	18.0%	18.0%	18.0%	31.5%	31.5%	31.5%	31.5%
	Regret	2.6%	2.6%	2.6%	2.6%	10.1%	10.1%	10.1%	10.1%	20.1%	20.1%	20.1%	20.1%
	Overreliance	0.6%	0.6%	0.6%	0.6%	3.4%	2.1%	2.9%	2.0%	10.2%	4.2%	11.6%	5.1%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	33.7%	33.4%	28.2%	28.0%	36.2%	33.8%	31.4%	34.7%	43.9%	33.9%	39.2%	43.2%
	ΔError	-0.4%	-0.6%	-0.6%	-0.3%	2.0%	-2.2%	2.1%	-1.4%	13.3%	-4.7%	12.9%	13.9%
	Ambiguity	6.0%	6.0%	6.0%	6.0%	20.5%	20.5%	20.5%	20.5%	37.0%	37.0%	37.0%	37.0%
	Regret	2.6%	2.6%	2.6%	2.6%	10.3%	10.3%	10.3%	10.3%	19.6%	19.6%	19.6%	19.6%
	Overreliance	1.1%	1.0%	1.0%	1.1%	6.1%	4.0%	6.2%	4.4%	16.5%	7.5%	16.3%	16.8%

Table 5: Overview of performance and regret for models trained on all datasets, training procedures, and model classes. Noise draw 3.

		5				20				40			
		LR		NN		LR		NN		LR		NN	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	24.0%	23.7%	12.5%	13.0%	27.6%	23.7%	19.2%	24.8%	36.9%	24.8%	27.2%	27.9%
	ΔError	-0.9%	-1.1%	-1.3%	-2.0%	-2.0%	-5.8%	-1.3%	-1.2%	8.3%	-9.8%	3.4%	-1.8%
	Ambiguity	6.0%	6.0%	6.0%	6.0%	20.5%	20.5%	20.5%	20.5%	36.0%	36.0%	36.0%	36.0%
	Regret	2.5%	2.5%	2.5%	2.5%	9.9%	9.9%	9.9%	9.9%	19.8%	19.8%	19.8%	19.8%
	Overreliance	0.8%	0.7%	0.6%	0.3%	4.0%	2.1%	4.3%	4.4%	14.0%	5.0%	11.6%	9.0%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	21.3%	20.8%	14.8%	15.6%	23.8%	20.5%	18.1%	17.7%	36.4%	20.8%	24.2%	30.4%
	ΔError	-1.1%	-1.4%	-1.9%	-1.9%	-2.1%	-5.6%	-5.9%	-5.1%	8.0%	-11.0%	-2.0%	-15.8%
	Ambiguity	5.5%	5.5%	5.5%	5.5%	18.5%	18.5%	18.5%	18.5%	33.0%	33.0%	33.0%	33.0%
	Regret	2.5%	2.5%	2.5%	2.5%	9.7%	9.7%	9.7%	9.7%	19.6%	19.6%	19.6%	19.6%
	Overreliance	0.7%	0.5%	0.3%	0.3%	3.8%	2.0%	1.9%	2.3%	13.8%	4.3%	8.8%	1.9%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	31.5%	31.1%	29.9%	30.1%	33.7%	31.1%	31.6%	30.0%	42.8%	31.4%	43.7%	30.2%
	ΔError	-0.5%	-0.6%	-0.6%	-1.1%	0.6%	-3.1%	-0.6%	-3.8%	12.8%	-5.8%	14.3%	-6.2%
	Ambiguity	5.5%	5.5%	5.5%	5.5%	18.5%	18.5%	18.5%	18.5%	31.5%	31.5%	31.5%	31.5%
	Regret	2.6%	2.6%	2.6%	2.6%	10.0%	10.0%	10.0%	10.0%	20.0%	20.0%	20.0%	20.0%
	Overreliance	1.1%	1.0%	1.0%	0.8%	5.3%	3.5%	4.7%	3.1%	16.4%	7.1%	17.1%	6.9%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	19.7%	19.7%	17.6%	18.0%	21.2%	19.9%	18.2%	18.4%	29.4%	19.8%	25.1%	18.9%
	ΔError	-1.3%	-1.3%	-1.3%	-1.3%	-3.7%	-5.5%	-4.6%	-4.9%	2.0%	-11.1%	-0.9%	-10.0%
	Ambiguity	5.0%	5.0%	5.0%	5.0%	18.0%	18.0%	18.0%	18.0%	31.5%	31.5%	31.5%	31.5%
	Regret	2.3%	2.3%	2.3%	2.3%	9.5%	9.5%	9.5%	9.5%	19.6%	19.6%	19.6%	19.6%
	Overreliance	0.5%	0.5%	0.5%	0.5%	2.9%	2.0%	2.4%	2.3%	10.8%	4.3%	9.4%	4.8%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	33.3%	33.4%	28.6%	27.9%	36.5%	33.5%	32.3%	29.9%	43.2%	33.6%	40.3%	36.5%
	ΔError	-0.4%	-0.7%	0.0%	-0.2%	2.3%	-2.4%	2.2%	-0.0%	12.7%	-5.0%	13.0%	3.9%
	Ambiguity	6.0%	6.0%	6.0%	6.0%	20.5%	20.5%	20.5%	20.5%	36.5%	36.5%	36.5%	36.5%
	Regret	2.6%	2.6%	2.6%	2.6%	9.9%	9.9%	9.9%	9.9%	19.9%	19.9%	19.9%	19.9%
	Overreliance	1.1%	1.0%	1.3%	1.2%	6.1%	3.8%	6.1%	5.0%	16.3%	7.5%	16.4%	11.9%

Table 6: Overview of performance and regret for models trained on all datasets, training procedures, and model classes. Noise draw 4.

		5				20				40			
		LR		NN		LR		NN		LR		NN	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	22.9%	22.6%	13.1%	13.3%	28.1%	23.0%	17.6%	20.2%	38.5%	23.0%	35.5%	26.0%
	ΔError	-1.0%	-1.2%	-1.2%	-1.0%	-1.0%	-5.4%	-3.0%	-1.7%	10.3%	-11.2%	11.3%	-4.0%
	Ambiguity	6.0%	6.0%	6.0%	6.0%	20.0%	20.0%	20.0%	20.0%	36.2%	36.2%	36.2%	36.2%
	Regret	2.7%	2.7%	2.7%	2.7%	10.7%	10.7%	10.7%	10.7%	21.1%	21.1%	21.1%	21.1%
	Overreliance	0.8%	0.8%	0.8%	0.8%	4.8%	2.7%	3.9%	4.5%	15.7%	5.0%	16.2%	8.5%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	21.4%	20.6%	15.5%	15.6%	24.6%	20.8%	17.4%	17.1%	33.2%	21.2%	29.2%	25.7%
	ΔError	-1.0%	-1.1%	-1.6%	-1.7%	-1.7%	-5.4%	-6.4%	-6.8%	5.7%	-11.3%	2.4%	-9.1%
	Ambiguity	5.5%	5.5%	5.5%	5.5%	18.5%	18.5%	18.5%	18.5%	33.0%	33.0%	33.0%	33.0%
	Regret	2.3%	2.3%	2.3%	2.3%	9.8%	9.8%	9.8%	9.8%	19.8%	19.8%	19.8%	19.8%
	Overreliance	0.7%	0.6%	0.4%	0.3%	4.0%	2.2%	1.7%	1.5%	12.7%	4.2%	11.1%	5.3%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	31.7%	31.0%	30.4%	30.0%	35.1%	31.1%	31.4%	30.1%	44.0%	31.3%	38.7%	30.0%
	ΔError	-0.5%	-0.7%	-1.1%	-0.8%	1.6%	-2.9%	-0.7%	-4.9%	14.2%	-5.5%	9.2%	-6.9%
	Ambiguity	5.5%	5.5%	5.5%	5.5%	18.5%	18.5%	18.5%	18.5%	31.5%	31.5%	31.5%	31.5%
	Regret	2.5%	2.5%	2.5%	2.5%	10.2%	10.2%	10.2%	10.2%	20.0%	20.0%	20.0%	20.0%
	Overreliance	1.0%	0.9%	0.7%	0.9%	5.9%	3.6%	4.7%	2.7%	17.1%	7.2%	14.6%	6.6%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	19.7%	19.7%	17.7%	17.6%	22.2%	19.6%	18.6%	18.1%	32.6%	19.6%	25.6%	19.3%
	ΔError	-1.3%	-1.4%	-1.6%	-1.4%	-3.1%	-6.0%	-4.3%	-5.6%	4.9%	-12.2%	-1.0%	-12.4%
	Ambiguity	5.0%	5.0%	5.0%	5.0%	18.0%	18.0%	18.0%	18.0%	31.5%	31.5%	31.5%	31.5%
	Regret	2.4%	2.4%	2.4%	2.4%	10.1%	10.1%	10.1%	10.1%	20.3%	20.3%	20.3%	20.3%
	Overreliance	0.5%	0.5%	0.4%	0.5%	3.5%	2.1%	2.9%	2.2%	12.6%	4.0%	9.7%	3.9%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	33.4%	33.6%	28.5%	28.9%	35.5%	33.7%	31.2%	30.2%	44.5%	34.1%	41.9%	39.6%
	ΔError	-0.4%	-0.5%	-0.2%	-0.1%	1.0%	-2.7%	1.1%	-0.4%	14.5%	-4.8%	14.5%	9.5%
	Ambiguity	6.0%	6.0%	6.0%	6.0%	20.5%	20.5%	20.5%	20.5%	35.5%	35.5%	35.5%	35.5%
	Regret	2.7%	2.7%	2.7%	2.7%	10.0%	10.0%	10.0%	10.0%	20.3%	20.3%	20.3%	20.3%
	Overreliance	1.1%	1.1%	1.2%	1.3%	5.5%	3.6%	5.5%	4.8%	17.4%	7.8%	17.4%	14.9%

Table 7: Overview of performance and regret for models trained on all datasets, training procedures, and model classes. Noise draw 5.

C.3 Additional Experimental Results

We include additional experimental results for the `mortality` dataset using a LR model and class level label noise. These results are aggregated across different initial noise draws, and also show regret and overreliance (fnr) conditioned on class and subgroup identifiers.

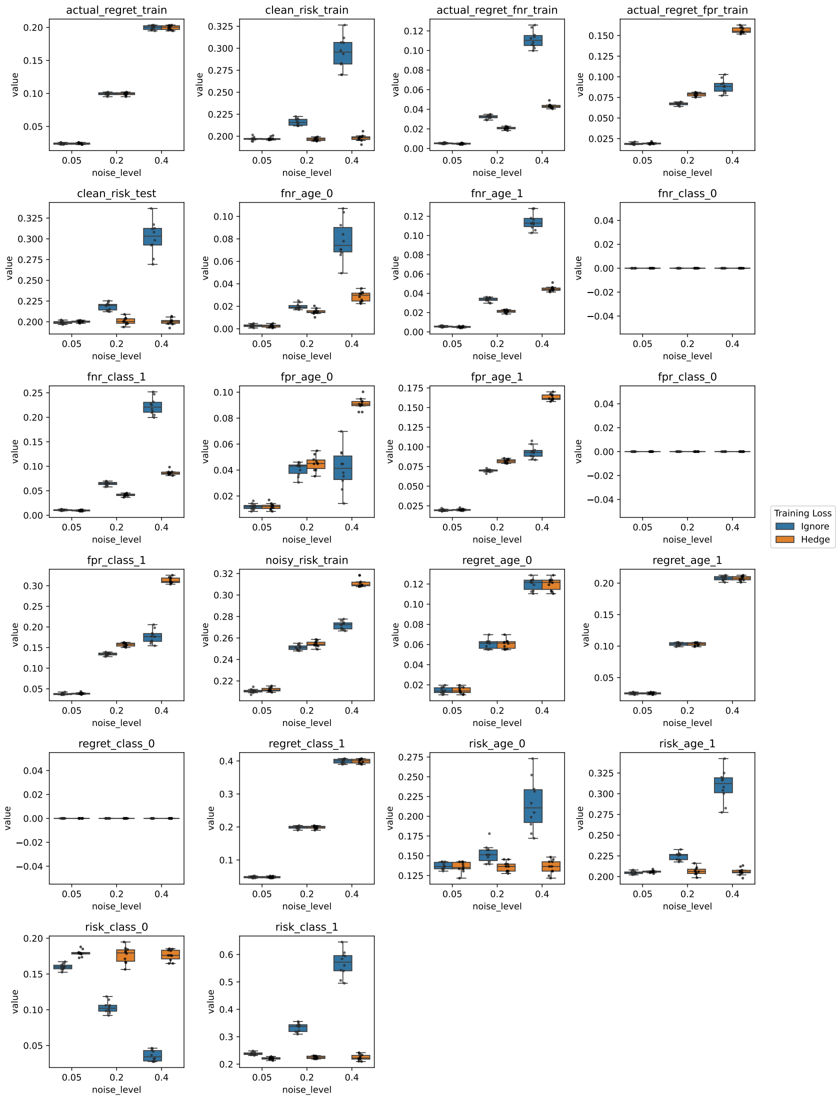


Figure 5: Complete Results for `mortality` Class Level Noise