REGRETFUL DECISIONS UNDER LABEL NOISE

Sujay Nagaraj
University of Toronto

Yang Liu
UC Santa Cruz

Flavio P. Calmon
Harvard SEAS

Berk Ustun
UC San Diego

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 with noisy labels. In this paper, we study the instance-level impact of learning under label noise. We introduce a notion of *regret* for this regime which measures the number of unforeseen mistakes due to noisy labels. We show that standard approaches to learning under label noise can return models that perform well at a population level while subjecting individuals to a *lottery of mistakes*. We present a versatile approach to estimate the likelihood of mistakes at the individual level from a noisy dataset by training models over plausible realizations of datasets without label noise. This is supported by a comprehensive empirical study of label noise in clinical prediction tasks. Our results reveal how failure to anticipate mistakes can compromise model reliability and adoption, and demonstrate how we can address these challenges by anticipating and avoiding 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 assess their risk for an adverse treatment response [3]. In such tasks, we train models with labels that reflect noisy observations of the true outcome we wish to predict. In practice, such noise may arise due to measurement error [e.g., 20, 35], human annotation [26], or inherent ambiguity [35]. In all these cases, label noise can have detrimental effects on model performance [10]. Over the past decade, these issues have led to extensive work on *learning from noisy datasets* [see e.g., 10, 28, 36, 39, 45]. As a result, we have developed foundational results that characterize when label noise can be ignored and algorithms to mitigate its detrimental effects.

By and large, this work has focused on the impact of label noise at the population level. In contrast, studying the effects of label noise at the instance level has received limited attention. This oversight reflects the fact that we cannot provide meaningful guarantees on individual predictions under label noise [28]. In a best-case scenario, where we have perfectly specified distributional assumptions on label noise, *we can learn a model that performs well on average, but we cannot identify where it makes mistakes*; as a result, individuals are subject to a "lottery of mistakes".

These effects undermine the utility of models in major real-world applications, as label noise arises in many settings where models are used to support or automate individual decisions [see, e.g., 51, for a recent meta-review of 72 cases in medicine]. In medical decision support tasks, our inability to identify mistakes can lead to overreliance, where physicians rely on predictions that may be incorrect [6, 25]. In automation tasks, our failure to assess the confidence of predictions can prevent us from reaping broader benefits – e.g., by abstention [9, 16].

In this work, we study how label noise affects individual predictions. Our motivation stems from the fact that, even if we cannot fully resolve the effects of label noise at the instance level, we can mitigate harm while leveraging the benefits of predictive models through uncertainty quantification. Our goal is to uncover these effects and develop methods to address them. Our main contributions are:

1. We introduce a notion of *regret* for learning from noisy datasets, capturing how label uncertainty affects individual predictions. We show that learning under label noise leads to inevitable regret, characterizing key limitations in a wide class of methods for learning from label noise.

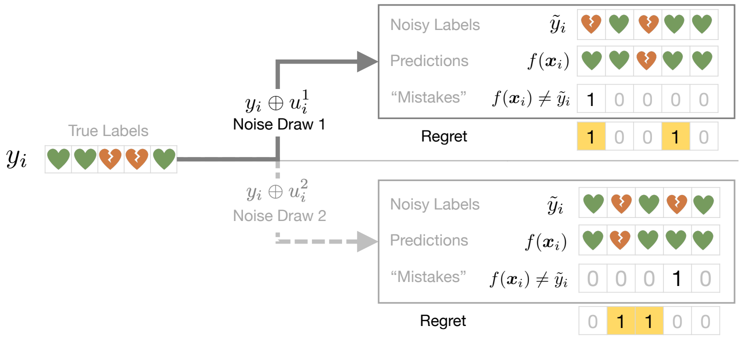

True Labels
y_i
$y_i \oplus u_i^1$
Noise Draw 1
$y_i \oplus u_i^2$
Noise Draw 2
Noisy Labels $\tilde{y}_i$
Predictions $f(\boldsymbol{x}_i)$
"Mistakes" $f(\boldsymbol{x}_i) \neq \tilde{y}_i$
1 0 0 0 0
Regret
1 0 0 1 0
Noisy Labels $\tilde{y}_i$
Predictions $f(\boldsymbol{x}_i)$
"Mistakes" $f(\boldsymbol{x}_i)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

2. We develop a method to flag regretful predictions by training models on plausible realizations of a clean dataset. Our approach can measure the sensitivity of individual predictions under label noise and incorporates common noise assumptions while controlling for plausibility.
3. We conduct a comprehensive empirical study on clinical prediction tasks. Our findings highlight the instance-level impact of label noise, and we demonstrate how our approach can support safer inference 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 by hedging for uncertainty in labels [29, 36, 39]. As we show in Section 2, such approaches are robust to label noise at a population level while subjecting individuals 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 under general distributional assumptions.

We propose to mitigate these issues through a principled approach to uncertainty quantification. Our approach relates to recent work on model multiplicity, which shows how changes in the machine learning pipeline can produce models that assign conflicting predictions [see e.g., 4, 7, 18, 31, 34, 48, 49] and lead to downstream effects on fairness, explanations, and recourse [5, 15, 23, 32]. 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 PRELIMINARIES

We consider a classification task where we wish to learn a model $f : \mathcal{X} \rightarrow \mathcal{Y}$ to predict a label $y \in \mathcal{Y}$ from a feature vector $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$. In a standard regime, we would be given a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where each $(\boldsymbol{x}_i, y_i)$ is drawn from a joint distribution of random variables X and Y. Given the dataset, we would learn a model 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 learn a model from a *noisy dataset* $\tilde{\mathcal{D}} = \{(\boldsymbol{x}_i, \tilde{y}_i)\}_{i=1}^n$, where each *noisy label* $\tilde{y}_i$ represents a potentially corrupted *true label* y_i. In what follows, we refer to this corruption as a *flip* and represent it using a binary variable $u_i := \mathbb{I}[y_i \neq \tilde{y}_i]$. Given the flip u_i, we can express noisy labels in terms of true labels as $\tilde{y}_i := y_i \oplus u_i$ and vice-versa as $y_i := \tilde{y}_i \oplus u_i$. Here, $a \oplus b := a + b - 2ab$ is the XOR operator. Given a noisy dataset, we represent all flips as a vector called 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 $[U_1, \ldots, U_n] \subseteq \{0, 1\}^n$.

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

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

In what follows, we assume that the values $p_{u|y_i,\boldsymbol{x}_i}$ are determined by a general *noise model* that can take on different forms – including uniform, class-level, or instance-level noise (see Table 1). We write p_u instead of $p_{u|y_i,\boldsymbol{x}_i}$ when the noise model is clear from context. We assume that the model is correctly specified and that $p_u < 0.5$ for all points to ensure there are more clean than noisy labels [c.f., 1, 36, 39].

Given a noisy dataset, we denote the noise draw over all instances as the *true draw* $\boldsymbol{u}^{\text{true}} := [u_1^{\text{true}}, \ldots, u_n^{\text{true}}]$. In practice, the true draw $\boldsymbol{u}^{\text{true}}$ is fixed but unknown. From the practitioner's perspective, $\boldsymbol{u}^{\text{true}}$ could be any realization of random variables U. If they knew $\boldsymbol{u}^{\text{true}}$, they could recover the true labels as $y_i = \tilde{y}_i \oplus u_i^{\text{true}}$ and learn without label noise. As this is infeasible, given the noise model and a set of priors, practitioners can estimate the posterior noise model $q_{u|\tilde{y}_i,\boldsymbol{x}_i} := \Pr(U_i = 1 \mid X = \boldsymbol{x}_i, \tilde{Y} = \tilde{y}_i)$ to infer clean labels from observed noisy labels.

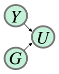
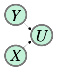
Noise Type	**PGM**	**Noise Model**	**Posterior Model**	**Inference Requirements**	**Sample Use Case**
Uniform		$p_u = \Pr(U = 1)$	$q_u = \Pr(U = 1)$	–	Diagnostic tests with a fixed failure rate (e.g., COVID19 rapid tests [2])
Class-Level		$p_{u\|y} = \Pr(U = 1 \mid Y = y)$	$q_{u\|\tilde{y}} = \Pr\left(U = 1 \mid \tilde{Y} = \tilde{y}\right)$	$\pi_y = \Pr(Y = y)$	Chest X-ray diagnosis where label noise $\tilde{Y}$ changes based on the disease Y [e.g., pneunomia vs COVID 13].
Group-Level		$p_{u\|y,g} = \Pr(U = 1 \mid Y = y, G = g)$	$q_{u\|\tilde{y},g} = \Pr\left(U = 1 \mid \tilde{Y} = \tilde{y}, G = g\right)$	$\pi_{y,g} = \Pr(Y = y \mid G = g)$	Diagnostic tasks where the incidence of label noise changes across subpopulations [e.g., racial bias in diagnosis 12].
Instance-Level		$p_{u\|y,\boldsymbol{x}} = \Pr(U = 1 \mid Y = y, X = \boldsymbol{x})$	$q_{u\|\tilde{y},\boldsymbol{x}} = \Pr\left(U = 1 \mid \tilde{Y} = \tilde{y}, X = \boldsymbol{x}\right)$	$\pi_{y,\boldsymbol{x}} = \Pr(Y = y, X = \boldsymbol{x})$	Data-driven discovery tasks where $\tilde{Y}$ is an experimental outcome confirmed by a hypothesis test with type I/II error [14]

Table 1: Examples of noise models used in the literature. We represent each noise model as a probability distribution with parameters $p_{u|y,\boldsymbol{x}}$ and show the corresponding probabilistic graphical model (PGM). Given a noisy dataset, a noise model, and a prior distribution π_y, we can infer noise draws from a posterior distribution with parameters $q_{u|\tilde{y},\boldsymbol{x}}$.

3 REGRET

Consider a practitioner who learns a model $f : \mathcal{X} \to \mathcal{Y}$ from a noisy dataset. In practice, they may learn a model that performs well on average. However, they will be unable to determine where it makes mistakes. In such tasks, individuals are subject to a *lottery of mistakes*. We characterize this effect in terms of *regret*.

Definition 2. Given a classification task where we learn a model $f : \mathcal{X} \to \mathcal{Y}$ from a noisy dataset, we define the *regret* for an instance $(\boldsymbol{x}_i, \tilde{y}_i)$ 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:

- $e^{\text{true}}(f(\boldsymbol{x}_i), y_i(U_i)) := \mathbb{I}[f(\boldsymbol{x}_i) \neq y_i(U_i)]$ indicates an *actual mistake* with respect to the true label. We write the true label as $y_i(U_i) := \tilde{y}_i \oplus U_i$ to show that it is a random variable.
- $e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i)$ indicates the model has made an *anticipated mistake* – i.e., that it appears to have made a mistake based on what we can tell during training.

In practice, $e^{\text{pred}}(\cdot)$ is determined by how we account for noise, if at all. If we ignore label noise and fit a model via standard ERM on the noisy dataset, then $e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i) := \mathbb{I}[f(\boldsymbol{x}_i) \neq \tilde{y}_i]$. If we fit a model via 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 an unbiased loss defined such that $\mathbb{E}_U[\tilde{\ell}_{01}(\boldsymbol{x}_i, \tilde{y}_i)] = \ell_{01}(f(\boldsymbol{x}_i), y_i)$.

Regret captures the irreducible error we incur due to randomness. In online learning, 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. In this case, randomness undermines our ability to improve the model since we cannot determine which predictions are correct. It also limits our ability to rely on its predictions for decision-making, as individual predictions cannot be assumed to be accurate. As a result, we compromise any downstream applications that depend on the correctness of downstream individual predictions – e.g., model explanations [43, 44] or post-hoc analyses [22, 30]. In Prop. 3, we explore the relationship between these effects and label noise.

Proposition 3. In a classification task where we learn a classifier f from a noisy dataset $\tilde{\mathcal{D}}$:

$$\mathbb{E}_{U|X,\tilde{Y}}\left[\text{Regret}(f(X),\tilde{Y},U)\right] = \Pr(U = 1 \mid \tilde{Y}, X).$$

Prop. 3 provides an opportunity to discuss several implications of learning from label noise. On the one hand, this result implies that regret is *unavoidable* when learning under label noise. In practice, we can only avoid it by “predicting less” (e.g., via selective classification) or by “removing noise” (e.g., via relabeling). On the other hand, the result also implies that we can estimate the *expected* number of regretful predictions in terms of the posterior noise rate. In practice, however, we cannot tell *how* these mistakes are distributed over all instances.

One of the key issues in this regime is that the value of a prediction may be compromised, as each instance where $q_{u|\boldsymbol{x},\tilde{y}} > 0$ is subject to a lottery of mistakes. Consider screening for a rare genetic disease using a diagnostic test. In such cases, we can view the presence of the disease as a clean label y_i and the test's outcome as a noisy label $\tilde{y}_i$. Given a disease that affects 10% of patients and a class-level noise model that flips 10% of positive cases, an average draw of label noise should affect 1% of predictions. In practice, however, these conditions heavily undermine the value of screening because any patient with a negative test may have the disease. We characterize these effects by measuring the proportion of instances in a dataset that are susceptible to regret – i.e., that take part in the lottery of mistakes. Given a noisy dataset $\tilde{\mathcal{D}}$ and a posterior noise model $\Pr\left(U = 1 \mid X, \tilde{Y}\right)$, the number of points susceptible to regret is:

$$\text{Susceptibility}(\tilde{\mathcal{D}}) := \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[\Pr\left(U = 1 \mid X = x_i, \tilde{Y} = \tilde{y}_i\right) > 0\right] \tag{2}$$

On the Regret of Hedging One of the benefits of studying regret in this regime is that we can characterize when learning is feasible at both the population and instance levels. Many algorithms for learning from noisy labels are designed to *hedge* against label noise [41]. Given a noisy dataset and a noise model, hedging minimizes the *expected risk over all possible noise draws*. In some cases, algorithms may implement this strategy explicitly via ERM with a modified loss [see e.g., 33, 36]. In others, algorithms may hedge implicitly – e.g., by assigning sample weights to training instances and setting the weights in a way to minimize expected risk over all possible draws [see e.g., 29, 39, 50].

In the best-case scenario, where we correctly specify the noise model and recover a model that minimizes the average number of mistakes over all noise draws, the resulting model would still incur regret. Formally, we would expect $\mathbb{E}_{U|X,Y}[\Delta\text{Error}(f,\tilde{\mathcal{D}},U)] = 0$ where:

$$\Delta\text{Error}(f,\tilde{\mathcal{D}};U) := \underbrace{\sum_{i=1}^{n} e^{\text{pred}}(f(\boldsymbol{x}_i),\tilde{y}_i)}_{\text{Predicted Training Error}} - \underbrace{\sum_{i=1}^{n} e^{\text{true}}(f(\boldsymbol{x}_i),y_i)}_{\text{True Training Error}} \tag{3}$$

However, the resulting model f would still incur regret $\mathbb{E}_{U|X,\tilde{Y}}\left[\text{Regret}(f,\tilde{\mathcal{D}},U)\right] > 0$. In Prop. 4, we show that the classical hedging algorithm of Natarajan et al. [36] exhibits this behavior.

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 u_i^{mle} corresponds to most likely outcome under the posterior noise model $q_{u|\tilde{y}_i,\boldsymbol{x}_i}$.

Prop. 4 implies that hedging will incur regret unless the noise in the dataset $\boldsymbol{u}^{\text{true}}$ matches the implicit noise draw $\boldsymbol{u}^{\text{mle}}$. In practice, this event is unlikely as $\lim_{n\to\infty}\Pr\left(\boldsymbol{u}^{\text{mle}} = \boldsymbol{u}^{\text{true}}\right) = 0$ (see Appendix A).

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
    $u_i \sim \text{Bernouilli}(q_{u|\tilde{y},\boldsymbol{x}})$ for $i \in [n]$        generate noise draw by posterior inference
    if $[u_1, \ldots, u_n] \in \mathcal{U}_\epsilon$ then        check if draw is plausible using Def. 6
        $\hat{y}_i \leftarrow \tilde{y}_i \oplus u_i$ for $i \in [n]$
        $\hat{\mathcal{D}} \leftarrow \{(\boldsymbol{x}_i, \hat{y}_i)\}_{i=1}^n$        construct plausible clean dataset
        $\hat{f} \leftarrow \text{argmin}_{f \in \mathcal{F}} \hat{R}(f; \hat{\mathcal{D}})$        train plausible model
        $\hat{\mathcal{F}}_\epsilon^{\text{plaus}} \leftarrow \hat{\mathcal{F}}_\epsilon^{\text{plaus}} \cup \{\hat{f}\}$        update plausible models
    end if
until $|\hat{\mathcal{F}}_\epsilon^{\text{plaus}}| = m$
Output  $\hat{\mathcal{F}}_\epsilon^{\text{plaus}}$, sample of $m$ models from the set of plausible models $\mathcal{F}_\epsilon^{\text{plaus}}$
```

4 ANTICIPATING MISTAKES WITH PLAUSIBLE MODELS

Our results in Section 2 show how a model we learn under label noise will inevitably output regretful predictions. In this section, we discuss how to sidestep this limitation while still benefiting from models by estimating the likelihood of assigning a regretful prediction.

4.1 MOTIVATION

Our goal is to evaluate the correctness of individual predictions for models learned from noisy data. In a standard classification task, we apply an algorithm for ERM to a clean dataset, recover the model $\hat{f} \in \text{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \mathbb{I}\left[f(\boldsymbol{x}_i) \neq y_i\right]$, and evaluate the correctness of predictions on the training dataset deterministically using the measure: $\text{Mistake}(\boldsymbol{x}_i, y_i, \hat{f}) = \mathbb{I}\left[\hat{f}(\boldsymbol{x}_i) \neq y_i\right]$. When we repeat this procedure to learn from a noisy dataset $\tilde{D} = \{(\boldsymbol{x}_i, \tilde{y}_i)\}_{i=1}^n$, the corresponding measure is no longer deterministic:

$$\text{Mistake}(\boldsymbol{x}_i, Y_i, \hat{F}) = \mathbb{I}\left[\hat{F}(\boldsymbol{x}_i) \neq Y_i\right]. \tag{4}$$

In this case, the randomness stems from two of the inputs: (1) the true label Y_i, which is a random variable that can only be inferred from the observed noisy label $\tilde{y}_i$ and the noise model; (2) the model $\hat{F} : \mathcal{X} \to \mathcal{Y}$, which is the output of a learning algorithm on the noisy dataset.

Our proposed measure, which we call *ambiguity*, quantifies the expected likelihood of a learning algorithm making a mistake on the training data – i.e., the expected value of (4).

$$\text{Ambiguity}(\boldsymbol{x}_i, \tilde{y}_i) := \mathbb{E}_{Y_i, \hat{F}|\tilde{D}}\left[\text{Mistake}(\boldsymbol{x}_i, Y_i, \hat{F})\right] = \mathbb{E}_{\boldsymbol{u} \sim U|\tilde{D}}\left[\mathbb{I}[\hat{F}(\boldsymbol{x}_i) \neq (\tilde{y}_i \oplus U_i)]\right] \tag{5}$$

This measure uses all the information we have at hand: a noisy dataset and a noise model. We formalize the relationship between ambiguity and regret in Prop. 5.

Proposition 5. Given a classification task, denote the clean label error of a model $\hat{F}(\boldsymbol{x}_i)$ on an instance $\boldsymbol{x}_i$ as e, that is $e := \Pr\left(\hat{F}(\boldsymbol{x}_i) \neq y_i\right)$. When $e < 0.5$, we can claim that a higher label noise rate for instance $\boldsymbol{x}_i$ corresponds to higher $\text{Ambiguity}(\boldsymbol{x}_i, \tilde{y}_i)$.

Since Prop. 3 establishes that regret corresponds to the posterior noise rate, Prop. 5 suggests that ambiguity serves as a viable measure of regret, given its correspondence to the posterior noise rate.

4.2 ESTIMATION

We can construct unbiased estimates of ambiguity using Algorithm 1. Given a noisy dataset and a noise model, this procedure generates plausible realizations of a clean dataset, and then trains a set of plausible models that can be used to estimate ambiguity. In what follows, we describe this procedure in greater detail.

Sampling Plausible Draws Given a noisy dataset $\tilde{\mathcal{D}}$, class-level noise model p_u, and prior distribution $\pi_y := \Pr(Y = y)$, we can sample noise draws from the posterior distribution:

$$q_{u|\tilde{y}} = \frac{(1 - \pi_{\tilde{y}}) \cdot p_{u|1-\tilde{y}}}{p_{u|\tilde{y}} \cdot (1 - \pi_{\tilde{y}}) + (1 - p_{u|\tilde{y}}) \cdot \pi_{\tilde{y}}} \tag{6}$$

This generalizes to different types of noise models (see e.g., Table 1). We can use these samples from the posterior distribution to estimate ambiguity directly. In practice, however, it may lead to biased estimates by returning *atypical draws* – unlikely noise draws under a given noise model (e.g., a noise draw that flips 30% of labels under a uniform noise model with a noise rate of 10%). In settings where we wish to estimate ambiguity using a limited number of draws, an atypical draw can bias our estimates and undermine their utility. Although we could moderate this bias by increasing the number of draws, this would require training a separate model for each draw. We address these issues by sampling from a set of plausible draws.

Definition 6. Given a noise draw $\boldsymbol{u} \in \{0, 1\}^n$, denote its posterior noise rate as $q_{u|\tilde{y}} := \Pr(U = 1 \mid \tilde{Y} = \tilde{y})$ and its empirical estimate as $\hat{q}_{u|\tilde{y}} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[u_i = 1 \mid \tilde{y}_i = y]$. For any $\epsilon \in [0, 1]$, the *set of plausible draws* contains all draws whose empirical noise rate is within ϵ of the posterior rate:

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

The set of plausible draws is a strongly typical set [see 8]. In a classification task where n is large, we can expect most (but not all) draws to concentrate in $\mathcal{U}_\epsilon$ [see Theorem 3.1.2 in 8]. We can limit atypical draws by setting the *atypicality parameter* ϵ, which represents the relative deviation between the true noise rate $q_{u|\tilde{y}}$ and the noise rate of sampled draws. Given a uniform noise model where $q_{u|\tilde{y}} = 0.1$, we would set $\epsilon = 0.2$ to only consider draws that flip between 8% to 12% of instances. Alternatively, we can set ϵ to ensure that $\mathcal{U}_\epsilon(\tilde{\boldsymbol{y}})$ will include a particular noise draw $\boldsymbol{u}_0 \in \mathcal{F}_\epsilon^{\text{plaus}}$ with high probability (see Prop. 9 in **??**). By default, we set $\epsilon = 0.1$ to consider draws within 10% of what we would expect.

Estimating Ambiguity Given a plausible noise draw $\boldsymbol{u}^k \in \mathcal{U}_\epsilon(\tilde{\boldsymbol{y}})$, we construct a *plausible* clean dataset by pairing each $\boldsymbol{x}_i$ with a *plausible* value of true label $\hat{y}_i^k = u^k \oplus \tilde{y}_i$.

Definition 7. The *set of ϵ-plausible models* contains all models trained using ϵ-plausible datasets:

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

We repeat this process m times and use the m plausible models to estimate ambiguity for each point in our noisy dataset as: $\hat{\mu}(\boldsymbol{x}, \tilde{y}) := \frac{1}{m}\sum_{k \in [m]} \mathbb{I}\left[\hat{f}^k(\boldsymbol{x}) \neq \hat{y}^k\right]$.

In practice, we can use ambiguity as a confidence score to operationalize techniques to learn or predict reliably. We propose a few examples and demonstrate how these perform in Section 5:

- *Data Cleaning*: We can use ambiguity to flag regretful instances in a training dataset to drop or relabel. Given the correspondence between regret and noise (Prop. 3), this approach can be used to "de-noise" a dataset to train models that generalize better on clean test data.
- *Selective Prediction*: We can use ambiguity to abstain from potentially regretful predictions at test time via selective prediction [11]. This approach can be used e.g., in clinical decision support, where we only show sufficiently reliable predictions and defer uncertain predictions to a clinician.

Discussion The main limitation of this approach is that we assume access to a correctly specified noise model. We view this assumption as a practical limitation and discuss potential solutions to validate and mitigate this in Section 7. We also note that the reliability of ambiguity depends on several factors, including assuming that the true noise draw, $\boldsymbol{u}^{\text{true}}$, is typical. In practice, although $\boldsymbol{u}^{\text{true}}$ is unknown, most draws can be shown to be typical – this follows from a standard application of a Chernoff bound [8].

5 EXPERIMENTS

In this section, we present an empirical study on clinical prediction tasks. Our goals are to document the effects of label noise on individual-level predictions and to demonstrate the validity and utility of our ambiguity measure. Supporting material and code can be found in Appendix B and GitHub.

Setup We work with 5 classification datasets from clinical applications where models support individual medical decisions (see Table 3). We treat the labels in each dataset as true labels. We create noisy datasets by corrupting the labels using a noise draw sampled according to three class-level noise models with noise rates $[5\%, 20\%, 40\%]$ where label noise only affects positive instances ($y_i = 1$). We split each dataset into a training sample (80%), which we use to train a logistic regression model (LR) and a neural network (DNN) using noisy labels, and a test sample (20%), which we use to measure out-of-sample performance using true labels. We train these models using the following methods:

1. Ignore, where we ignore label noise and fit a model to predict noisy training labels; and
2. Hedge where we hedge against label noise using the method of Natarajan et al. [36].

Our setup yields 12 models for each dataset (3 noise regimes × 2 model classes × 2 training procedures). For each model, we estimate the ambiguity for each point in the training dataset using Section 4 to sample $m = 100$ plausible models with atypicality $\epsilon = 10\%$. We assume that the noise model is correctly specified and the noise draw is unknown at training time.

Results We characterize the accuracy and reliability of predictions from each model using the measures in Table 2. We report our results for LR models in Table 3 and defer results for DNN to Appendix B for clarity. We first describe how regret manifests in real-world prediction tasks. We then use the ambiguity scores on individual instances to guide *data cleaning* and *selective prediction* for safer inference. In what follows, we discuss all results.

Metric	**Definition**	**Description**
TrueError$(f, \tilde{\mathcal{D}})$	$\frac{1}{n}\sum_{i\in[n]} e^{\text{true}}(f(\boldsymbol{x}_i), y_i)$	Error rate of f on the *clean* training labels.
AnticipatedError$(f, \tilde{\mathcal{D}})$	$\frac{1}{n}\sum_{i\in[n]} e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i) - e^{\text{true}}(f(\boldsymbol{x}_i), y_i)$	Error rate of f on the *noisy* labels.
Susceptibility$(\tilde{\mathcal{D}})$	$\frac{1}{n}\sum_{i\in[n]} \mathbb{I}\left[\text{Pr}\left(U = 1 \mid X = x_i, \tilde{Y} = \tilde{y}_i\right) > 0\right]$	Proportion of instances in $\tilde{\mathcal{D}}$ subject to regret.
Regret$(f, \tilde{\mathcal{D}})$	$\frac{1}{n}\sum_{i\in[n]} \mathbb{I}\left[e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i) \neq e^{\text{true}}(f(\boldsymbol{x}_i), y_i)\right]$	Mean regret across all instances in $\tilde{\mathcal{D}}$. We expect Regret$(f, \tilde{\mathcal{D}}) \approx \sum_y q_{u\mid y} \cdot \pi_y$ under class-level label noise.
Overreliance$(f, \tilde{\mathcal{D}})$	$\frac{1}{n}\sum_{i\in[n]} \mathbb{I}\left[e^{\text{true}}(f(\boldsymbol{x}_i), y_i) = 1, e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i) = 0\right]$	Proportion of predictions in $\tilde{\mathcal{D}}$ 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.

On Label Noise and Regret Our results in Table 3 highlight several implications of learning from label noise. We find that our analytical results in Prop. 3 hold empirically: that the average number of regretful individual predictions corresponds to the effective noise rate in each dataset. Though regret is proportional to the noise rate, how it is distributed across instances is uncertain – it can affect *any* instance subject to label noise. As this is a class-level model where only positive instances ($y = 1$) experience label noise, all negative ($\tilde{y} = 0$) are subject to a lottery of mistakes because $\text{Pr}\left(\tilde{y} = 1 \mid y = 0\right) = 0$. We use susceptibility (2) to measure the proportion of points vulnerable to this lottery. In Table 3, even when label noise is as low as 5%, the proportion of susceptible points is $> 50\%$ across all five datasets. This suggests that, even when label noise rates are low, a disproportionately large number of instances are susceptible to regret. Most practitioners assume that their training data is clean and ignore label noise, but most real-world datasets are not perfectly labeled – this inevitably leads to regretful predictions on individuals.

We observe similar effects across all datasets, model classes, and noise regimes, underscoring the need to quantify the effect of label noise on individual-level predictions – especially in tasks where models make important decisions.

		5		20		40	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	24.4%	23.5%	27.1%	24.6%	41.0%	24.3%
	Anticipated Error	25.7%	25.2%	28.3%	29.4%	28.2%	33.5%
	Regret	3.0%	3.0%	10.1%	10.1%	19.7%	19.7%
	Overreliance	1.1%	0.9%	6.3%	3.8%	22.6%	7.9%
	Susceptibility	52.6%	52.6%	59.7%	59.7%	69.3%	69.3%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	20.8%	20.2%	25.0%	20.3%	34.9%	20.1%
	Anticipated Error	22.1%	21.7%	26.8%	26.4%	27.4%	32.5%
	Regret	2.5%	2.5%	10.2%	10.2%	19.8%	19.8%
	Overreliance	0.8%	0.6%	5.8%	2.8%	18.8%	5.5%
	Susceptibility	52.5%	52.5%	60.2%	60.2%	69.8%	69.8%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	31.7%	30.8%	33.7%	30.8%	43.0%	31.1%
	Anticipated Error	32.2%	31.5%	32.7%	33.6%	30.0%	36.5%
	Regret	2.5%	2.5%	10.0%	10.0%	19.7%	19.7%
	Overreliance	1.5%	1.3%	8.1%	5.4%	23.4%	11.3%
	Susceptibility	52.7%	52.7%	60.2%	60.2%	69.9%	69.9%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	19.5%	19.0%	23.2%	19.1%	33.2%	19.4%
	Anticipated Error	20.7%	20.4%	25.7%	25.0%	27.7%	30.9%
	Regret	2.2%	2.2%	9.8%	9.8%	19.5%	19.5%
	Overreliance	0.6%	0.5%	4.9%	2.6%	17.3%	5.8%
	Susceptibility	52.2%	52.2%	59.8%	59.8%	69.5%	69.5%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	33.1%	33.7%	36.7%	33.7%	44.2%	33.9%
	Anticipated Error	33.4%	34.1%	34.1%	36.0%	29.7%	38.6%
	Regret	2.6%	2.6%	10.0%	10.0%	19.6%	19.6%
	Overreliance	1.8%	1.7%	9.6%	6.0%	24.3%	12.1%
	Susceptibility	52.6%	52.6%	60.0%	60.0%	69.6%	69.6%

Table 3: Accuracy and reliability of predictions for LR models trained on noisy datasets where we flip 5%, 20% and 40% of positive instances. We defer results for DNN models to Appendix B for clarity.

On the Consequences of Regret To demonstrate how regretful predictions can negatively impact individuals, we consider a particular flavor of regret – *overreliance* – the fraction of instances where a practitioner would incorrectly assume that a model assigned a correct prediction (i.e., $e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i) = 0$ and $e^{\text{true}}(f(\boldsymbol{x}_i), y_i) = 1$). From Table 3, we consider overreliance on the `lungcancer` dataset, under 40% noise. We observe up to 23.4% of individuals are assigned this type of prediction – in a model that aims to predict cancer, these represent patients who have cancer but are mistakenly classified as cancer-free based on the prediction of a seemingly accurate model. By analyzing how regretful predictions are distributed, we can quantify the impact of label noise, identify high-risk cases, and adjust our reliance on model predictions – ensuring that practitioners do not blindly trust or explain away incorrect outputs in important model decisions.

On Learning by Hedging Our results highlight that we can learn models robust to noise at a population level yet assign mistakes by lottery. As shown in Table 3, we observe that Hedge can moderate the impact of label noise at a population level by reducing the true (clean label) error compared to Ignore. Even with more noise robustness, regret is unchanged and remains high across all experimental conditions. On the `mortality` dataset, for example, Hedge reduces the error rate by over 13% compared to Ignore for a LR model under 40% label noise. However, regret is unchanged and continues to affect 19.5% of instances. It is interesting to note that Hedge *is* able to moderate the effects of overreliance, as these predictions are a subset of clean label errors. This suggests that Hedge is only able to moderate the flavor of regret but not the regret rate itself; Hedge only *redistributes* overreliant predictions to cases where $e^{\text{pred}}(f(\boldsymbol{x}_i), \tilde{y}_i) = 1$ and $e^{\text{true}}(f(\boldsymbol{x}_i), y_i) = 0$ – i.e., where a practitioner may ignore seemingly incorrect predictions and fail to reap the benefits of a correct prediction.

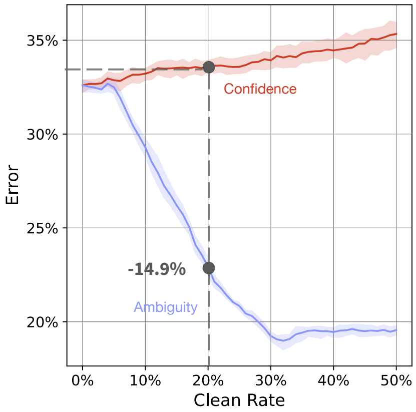


Figure 2: Clean test error for a LR model on the `shock_mimic` dataset under 40% class-level label noise when dropping instances from the training dataset using a confidence-based threshold rule. We plot the clean test error vs percent of data cleaned guided by either predicted probabilities $\text{conf}(\boldsymbol{x}_i) := \hat{p}(\tilde{y}_i \mid \boldsymbol{x}_i)$ or ambiguity $\text{conf}(\boldsymbol{x}_i) := 1 - \hat{\mu}(\boldsymbol{x}_i, \tilde{y}_i)$.

On Safety via Data Cleaning Our proposed approach in Section 4 can *clean* noisy datasets by using estimates of ambiguity to drop regretful instances from a training dataset – denoised datasets can then be used to train models that generalize better on clean test data. In Fig. 2, we demonstrate the effectiveness of this approach on the `shock_mimic` dataset. Here, we drop instances in the training dataset using a confidence-based threshold rule of the form $\mathbb{I}\left[\text{conf}(\boldsymbol{x}_i) \leq \tau\right]$ where $\text{conf}(\boldsymbol{x}_i)$ is a confidence score for a given instance in $\tilde{\mathcal{D}}$ and τ is a threshold set to control the number of instances we wish to drop (i.e., set τ to the k^{th}-percentile of $\text{conf}(\cdot)$ to drop $k\%$ of instances). We consider confidence scores from training data computed using Algorithm 1 based on ambiguity (i.e., $\text{conf}(\boldsymbol{x}_i) = 1 - \hat{\mu}(\boldsymbol{x}_i, \tilde{y}_i)$) or mean predicted probabilities (i.e., $\text{conf}(\boldsymbol{x}_i) = \hat{p}(y_i \mid \boldsymbol{x}_i)$) across plausible models. We show how removing instances with high ambi-

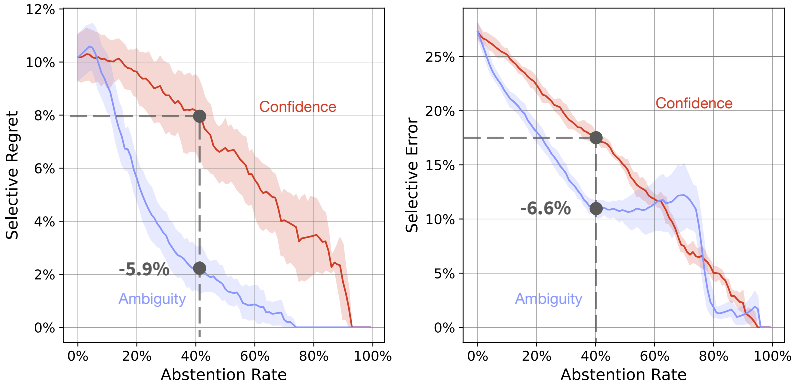


Figure 3: Selective classification frontiers for a LR model on the `shock_mimic` dataset under 20% class-level noise when abstaining from uncertain predictions at test time using a confidence-based threshold rule. We plot the selective regret (left) and selective error (right) as we vary the percent of abstained predictions for confidence measures based on predicted probabilities $\text{conf}(\boldsymbol{x}_i) := \hat{p}(\tilde{y}_i \mid \boldsymbol{x}_i)$ and ambiguity $\text{conf}(\boldsymbol{x}_i) := 1 - \hat{\mu}(\boldsymbol{x}_i)$.

guity from the training dataset prior to training a final model on the cleaned data leads to improved test error on clean labels. In Fig. 2, using ambiguity to drop uncertain instances reduces test error by 14.9% when dropping only 20% of noisy training data compared to the standard approach on `cshock_mimic` under 40% class-level label noise.

On Safety via Selective Prediction Our approach can also reduce regret by abstaining from uncertain predictions at test time. We demonstrate the effectiveness of this approach on the `shock_mimic` dataset in Fig. 3. Here, we use the same confidence-based threshold rule of the form $\mathbb{I}\left[\text{conf}(\boldsymbol{x}_i) \leq \tau\right]$ where $\text{conf}(\boldsymbol{x}_i)$ is a confidence score and τ is a threshold value. We consider confidence scores based on standard predicted probabilities and ambiguity, where ambiguity can be measured using cheaply acquired test instances (e.g., noisy test data). We show how the selective test error on clean labels and selective regret change as we vary the confidence threshold value $\tau \in (0, 1)$. In Fig. 3, using ambiguity reduces selective regret by -5.9% and selective error by -6.6% by abstaining on 40% of test instances compared to the standard approach on `cshock_mimic` under 20% label noise.

6 DEMONSTRATION

We demonstrate how our approach can support a modern scientific discovery task in biotechnology. In such tasks, we would usually run an in vitro experiment to discover instances with a desired property [e.g., identifying new antibiotics 46]. Given a dataset of previous successful and unsuccessful experiments, we can train a model to predict future experimental outcomes based on experimental characteristics without having to acquire resources to actually run the experiment – accelerating discovery by prioritizing experiments that are likely to succeed.

Setup We work with the `enhancer` dataset from Gschwind et al. [14] to predict the outcome of experiments to discover *enhancers* – i.e., segments of DNA that regulate gene expression. The dataset contains $n = 992$ noisy instances $(\boldsymbol{x}_i, \tilde{y}_i)$, where each feature vector $\boldsymbol{x}_i$ encodes $d = 13$ characteristics of a previously performed experiment (e.g., cell type, gene location, etc.). The noisy label takes $\tilde{y}_i = 1$ if the experiment revealed a statistically significant discovery (i.e., reject null hypothesis: H_0 = “no significant effect”). In this case, we have an instance-level noise model where the noise rates are set by the Type I error for each experiment – which varies based on the p-value for each experiment’s hypothesis test:

$$\Pr\left(\tilde{y}_i = 1 \mid y_i = 0\right) = \Pr\left(\text{False rejecting of null hypothesis } H_0\right) = p\text{-value for experiment } i$$

We split our dataset into a noisy training sample (80%) and a clean test sample (20%). Usingn ERM, we fit a LR model f and estimate performance on the clean test sample. Just as in Section 5, we use a confidence-based threshold rule of the form $\mathbb{I}\left[\text{conf}(\boldsymbol{x}_i) \leq \tau\right]$ where $\text{conf}(\boldsymbol{x}_i)$ is a confidence score and τ is a threshold value. We operationalize our approach without labels (noisy or otherwise) on a test instance $\boldsymbol{x}_i$, by computing the *disagreement* between the prediction of model f and m plausible

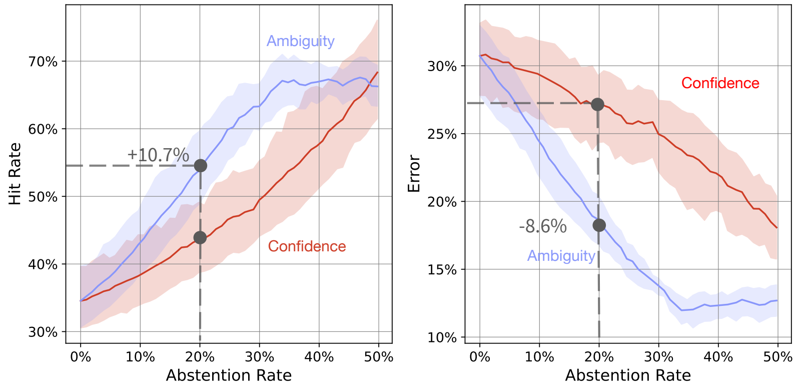


Figure 4: Selective classification frontiers for a LR model on the `enhancer` dataset when abstaining from uncertain predictions at test time using a confidence-based threshold rule. We plot the selective hit rate (left) and selective error (right) as we increase the percent of abstained predictions for confidence measures based on predicted probabilities $\text{conf}(\boldsymbol{x}_i) := \hat{p}(\tilde{y}_i \mid \boldsymbol{x}_i)$ and disagreement $\text{conf}(\boldsymbol{x}_i) := 1 - \text{Disagreement}(\boldsymbol{x}_i)$.

models $\hat{\mathcal{F}}_\epsilon^{\text{plaus}}$ from Algorithm 1:

$$\text{Disagreement}(\boldsymbol{x}_i) := \frac{1}{m} \sum_{k \in [m]} \mathbb{I}\left[\hat{f}^k(\boldsymbol{x}_i) \neq f(\boldsymbol{x}_i)\right] \tag{7}$$

Our goal is to identify which future experiments would succeed or fail given a noisy dataset of prior experimental results using the confidence measure $\text{conf}(\boldsymbol{x}_i) := 1 - \text{Disagreement}(\boldsymbol{x}_i)$.

Results We report our results in Fig. 4 where we show how reliability of predictions for successful experiments. Here, we compare selective prediction using measures: (1) a standard approach where we would abstain on uncertain experiments according to $\hat{p}(\tilde{y}_i \mid \boldsymbol{x}_i)$, or (2) where we abstain according to disagreement rates (7). We evaluate the performance of our model using test *Hit Rate* (i.e., recall – future successful experiments correctly predicted). In addition, we show the test error on clean labels at varying abstention rates.

Our results show that our approach in Section 4 can enhance data-driven discovery by accurately predicting experimental outcomes before taking place. In this case, we can improve the Hit Rate (+10.7%) compared to standard confidence-based abstention, with a modest 20% abstention rate (Fig. 4). Since Hit Rate represents the proportion of future *successful* experiments, we can optimize laboratory resource allocation and increase the discovery rate of enhancers by forgoing 20% of experiments, saving time and resources.

7 CONCLUDING REMARKS

Learning under label noise is a major challenge for practitioners. While models may perform well on average, individuals still face a lottery of mistakes; even models with 99% accuracy can misclassify *anyone* due to label noise. In this work, we studied these limitations through the lens of regret. Our results highlighted the prevalence of regret in various decision-support tasks and the inherent limitations of existing label noise learning strategies in mitigating regret. We introduce a measure of ambiguity to measure instance-level uncertainty in predictions. We then operationalize this measure to reduce regretful decisions and lead to safer predictions via data cleaning and abstention. Our work shows that, even as regret is inevitable, we can use uncertainty quantification to calibrate our reliance on individual predictions – signaling the need for 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 safer and more reliable predictions on individuals in critical tasks.

Limitations Our estimates assume that we have correctly specified our noise model and prior distributions. In practice, we can validate these assumptions by comparing them against distributions that we estimate from the noisy dataset [see e.g., 27, 29, 39]. When working with simple noise models (e.g., uniform or class-level), we may be conservative and assume a higher noise rate. Alternatively, we can hedge against misspecification by setting ϵ to capture a larger set of plausible draws.

ACKNOWLEDGMENTS

This work was supported by funding from the National Science Foundation IIS 2040880, IIS 2313105, IIS-2007951, IIS-2143895, and the NIH Bridge2AI Center Grant U54HG012510. Sujay Nagaraj was supported by a CIHR Vanier Scholarship. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

REFERENCES

[1] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine learning*, 2:343–370, 1988.

[2] Mayara Lisboa Bastos, Gamuchirai Tavaziva, Syed Kunal Abidi, Jonathon R Campbell, Louis-Patrick Haraoui, James C Johnston, Zhiyi Lan, Stephanie Law, Emily MacLean, Anete Trajman, et al. Diagnostic accuracy of serological tests for covid-19: systematic review and meta-analysis. *bmj*, 370, 2020.

[3] Atharva M Bhagwat, 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.

[4] Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 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`.

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

[6] Chun-Wei Chiang 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*, pp. 120–129, 2021.

[7] Amanda Coston, 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`.

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

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

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

[12] Kan Z Gianattasio, 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] Athanasios Giannakis, Dorottya Móré, Stella Erdmann, Laurent Kintzelé, Ralph Michael Fischer, Monika Nadja Vogel, David Lukas Mangold, Oyunbileg von Stackelberg, 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] Andreas R Gschwind, 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] Faisal Hamman, 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] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: A survey. *Machine Learning*, pp. 1–38, 2024.

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

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

[19] Alistair EW Johnson, 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] Jamil N Kanji, 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] William A Knaus, 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] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

[23] Bogdan Kulynych, 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*, pp. 1609–1623, 2023.

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

[26] Dana Li, 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] Xuefeng Li, 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*, pp. 6403–6413. PMLR, 2021.

[28] Yang Liu. Understanding instance-level label noise: Disparate impacts and treatments. In *International Conference on Machine Learning*, pp. 6725–6735. PMLR, 2021.

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

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

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

[32] Charles Marx, 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*, pp. 7375–7391. PMLR, 2023.

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

[34] Anna P Meyer, 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*, pp. 193–204, 2023.

[35] Sujay Nagaraj, Sarah Goodday, Thomas Hartvigsen, Adrien Boch, Kopal Garg, Sindhu Gowda, Luca Foschini, Marzyeh Ghassemi, Stephen Friend, and Anna Goldenberg. Dissecting the heterogeneity of "in the wild" stress from multimodal sensor data. *NPJ Digital Medicine*, 6(1):237, 2023.

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

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

[38] Diane Oyen, 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] Giorgio Patrini, 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*, pp. 1944–1952, 2017.

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

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

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

[45] Hwanjun Song, 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] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

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

[49] Jamelle Watson-Daniels, 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] Jiaheng Wei, 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.

[51] Yishu Wei, 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.

Supplementary Materials

A OMITTED PROOFS

A.1 RESULTS FROM SECTION 3

Proof of Prop. 3. Consider any classification task with label noise. Let $\rho_{X,\tilde{Y}} := \Pr\left(U = 1 \mid X, \tilde{Y}\right)$ denote the posterior 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 start by showing that, by 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:

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

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

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

□

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

$$\mathbb{E}_{X,Y}\left[\mathbb{I}\left[f(X) \neq Y\right]\right] = \mathbb{E}_{X,\tilde{Y}}\left[(1 - q_u \mathbb{I}\left[f(X) \neq \tilde{Y}\right] + q_u \mathbb{I}\left[f(X) \neq 1 - \tilde{Y}\right]\right] \tag{8}$$

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. 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. 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:

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

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

We can denote the minimizer of the RHS:

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_{X,\tilde{Y}}\left[(1-q_u)\mathbb{I}\left[f(X) \neq \tilde{Y}\right] + q_u \mathbb{I}\left[f(X) = \tilde{Y}\right]\right] \tag{11}$$

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}:

$$\lim_{n \to \infty} \Pr\left(\boldsymbol{u}^{\text{mle}} = \boldsymbol{u}^{\text{true}}\right) = \lim_{n \to \infty} \prod_{i=1}^{n} \Pr\left(u_i^{\text{mle}} = u_i^{\text{true}}\right) = 0 \tag{12}$$

□

A.2 RESULTS FROM SECTION 4

The goal of an ambiguity measure is to evaluate the reliability of an algorithm that learns from data at the instance level. In what follows, we express the learning algorithm as a function $\mathcal{A} : \mathcal{D} \to \mathcal{F}$ that takes as input a clean dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and returns as output a model $\hat{f} \in \mathcal{F}$. This analysis will assume that each dataset maps to a unique model.

When we learn from a clean dataset, our algorithm returns the model

$$\hat{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[f(\boldsymbol{x}_i) \neq y_i\right].$$

In this case, a mistake is a measure of the reliability of learning at a given point i:

$$\text{Mistake}(\boldsymbol{x}_i, y_i, \hat{f}) = \mathbb{I}\left[\hat{f}(\boldsymbol{x}_i) \neq y_i\right]. \tag{13}$$

When we learn from a noisy dataset $\tilde{D} = \{(\boldsymbol{x}_i, \tilde{y}_i)\}_{i=1}^n$, we cannot compute this metric because *two* inputs are random.

$$\text{Mistake}(\boldsymbol{x}_i, Y_i, \hat{F}) = \mathbb{I}\left[\hat{F}(\boldsymbol{x}_i) \neq Y_i\right]. \tag{14}$$

Here:

- The true label Y_i is a random variable that can only be inferred based on the noise at point i and the observed noisy label $\tilde{y}_i$.
- The model is also random $\hat{F} : \mathcal{X} \to \mathcal{Y}$ because it reflects the output of a learning algorithm $\hat{F} := \mathcal{A}(\{(\boldsymbol{x}_i, Y_i)_{i=1}^n)$ on dataset with n noisy labels.

Ambiguity is the expected value of the quantity in (14) when we learn a model by empirical risk minimization on a plausible realization of the clean dataset. Our procedure in Algorithm 1 returns an estimate of this expectation using m noise draws, m plausible datasets and m plausible models.

$$\begin{aligned}
\text{Ambiguity}(\boldsymbol{x}_i) &= \mathbb{E}_{Y_i, \hat{F} | \tilde{D}}\left[\text{Mistake}(\boldsymbol{x}_i, Y_i, \hat{F})\right] && (15)\\
&= \mathbb{E}_{Y_i, \hat{F} | \tilde{D}}\left[\mathbb{I}\left[\hat{F}(\boldsymbol{x}_i) \neq Y_i\right]\right] && (16)\\
&= \mathbb{E}_{\boldsymbol{u} \sim U | \tilde{D}}\left[\mathbb{I}\left[\hat{F}(\boldsymbol{x}_i) \neq (\tilde{y}_i \oplus U_i)\right]\right] && (17)\\
&\approx \frac{1}{m} \sum_{k=1}^{m} \mathbb{I}\left[\hat{f}^k(\boldsymbol{x}_i) \neq \hat{y}_i^k\right] && (18)\\
&= \frac{1}{m} \sum_{k=1}^{m} \mathbb{I}\left[\hat{f}^k(\boldsymbol{x}_i) \neq (\tilde{y}_i \oplus u_i^k)\right] && (19)
\end{aligned}$$

Here the final estimate is based on:

- m plausible clean labels: $\hat{y}_i^1, \ldots \hat{y}_i^m$. These are m samples from the distribution of noise at i: $\hat{y}_i^k = \tilde{y}_i^k \oplus u_i^k$ where $u_i^k \sim U_i$.
- m plausible models $\hat{f}^1, \ldots \hat{f}^m$. These are m samples of a random function $\hat{F}^k = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n \mathbb{I}\left[f(\boldsymbol{x}_i) \neq \hat{Y}_i^k\right]$ that we fit applying the learning algorithm $\mathcal{A}$ to a plausible dataset. Here, $\hat{F} = \mathcal{A}(D) = \mathcal{A}((\boldsymbol{x}_i, Y_i)_{i=1}^n)$. Thus, $\hat{F}$ is a function of the plausible labels $\hat{y}_1^k ... \hat{y}_n^k$. Given the noisy labels $\tilde{y}_1, \ldots \tilde{y}_n$, we can write $\hat{y}_1^k \ldots \hat{y}_n^k = (\tilde{y}_1^k \oplus u_1^k) ... (\hat{y}_n^k \oplus u_n^k)$ where $\boldsymbol{u}^k = (u_1^k, \ldots, u_n^k) \sim [U_1, \ldots, U_n]$.

Proof of Prop. 5. Denote the noise rate of $\hat{F}(\boldsymbol{x}_i)$ as e, that is $\Pr\left(\hat{F}(\boldsymbol{x}_i) \neq y_i\right) = e$.

$$\begin{aligned}\mathbb{E}_{Y_i,\hat{F}|\tilde{D}}\left[\mathbb{I}\left[\hat{F}(\boldsymbol{x}_i)\neq\hat{Y}_i\right]\right] &= \mathbb{E}_{Y_i,\hat{F}|\tilde{D}}\left[\mathbb{I}\left[\hat{F}(\boldsymbol{x}_i)\neq\hat{Y}_i \mid \hat{F}(\boldsymbol{x}_i)=y_i\right]\right]\cdot(1-e)\\ &+\mathbb{E}_{Y_i,\hat{F}|\tilde{D}}\left[\mathbb{I}\left[\hat{F}(\boldsymbol{x}_i)\neq\hat{Y}_i \mid \hat{F}(\boldsymbol{x}_i)\neq y_i\right]\right]\cdot e\\ &=\mathbb{E}_{Y_i,\hat{F}|\tilde{D}}\left[\mathbb{I}\left[Y_i\neq y_i\right]\right]\cdot(1-e)+(1-\mathbb{E}_{Y_i,\hat{F}|\tilde{D}}\left[\mathbb{I}\left[\hat{Y}_i\neq y_i\right]\right])\cdot e\\ &=(1-2e)\cdot\mathbb{E}_{Y_i,\hat{F}|\tilde{D}}\left[\mathbb{I}\left[\hat{Y}_i\neq y_i\right]\right]+e\end{aligned}$$

When $e < 0.5$, we can claim that the higher the $\mathbb{E}_{Y_i,\hat{F}|\tilde{D}}\left[\mathbb{I}\left[\hat{Y}_i\neq y_i\right]\right]$, the higher the $\mathbb{E}_{Y_i,\hat{F}|\tilde{D}}\left[\mathbb{I}\left[\hat{F}(\boldsymbol{x}_i)\neq\hat{Y}_i\right]\right]$, the ambiguity measure. If we assume that $\mathbb{E}_{Y_i,\hat{F}|\tilde{D}}\left[\mathbb{I}\left[\hat{Y}_i\neq y_i\right]\right]$ is monotonic in the noise rates in u_i, which is intuitively true, we then establish that the higher the noise, the higher the ambiguity measure.

□

A.3 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%.

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,

$$u^{\text{true}}\in\mathcal{U}_\epsilon(\tilde{\boldsymbol{y}}) \Leftrightarrow |\hat{p}_u-p_{u|\tilde{y}}|\leq p_{u|\tilde{y}}\cdot\epsilon \tag{20}$$

We will derive conditions to satisfy the left-hand side of Eq. (20)

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:

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

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:

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

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

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

$\square$

B SUPPORTING MATERIAL FOR SECTION 5

B.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.

`mortality` 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, `mortality` 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.

B.2 ADDITIONAL RESULTS

		5		20		40	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	23.4%	23.0%	27.7%	22.9%	39.4%	23.5%
	Anticipated Error	24.2%	24.1%	29.1%	27.7%	29.4%	34.3%
	Regret	2.1%	2.1%	10.2%	10.2%	20.1%	20.1%
	Overreliance	0.8%	0.6%	6.2%	3.8%	21.3%	7.1%
	Susceptibility	51.7%	51.7%	59.8%	59.8%	69.7%	69.7%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	20.7%	19.8%	23.6%	20.2%	32.9%	20.1%
	Anticipated Error	22.1%	21.2%	26.3%	26.2%	28.0%	32.5%
	Regret	2.3%	2.3%	9.7%	9.7%	19.8%	19.8%
	Overreliance	0.6%	0.5%	4.7%	2.5%	17.1%	5.5%
	Susceptibility	52.3%	52.3%	59.7%	59.7%	69.8%	69.8%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	31.5%	30.9%	33.4%	31.0%	43.9%	31.2%
	Anticipated Error	32.0%	31.6%	32.9%	34.0%	29.9%	36.8%
	Regret	2.4%	2.4%	9.9%	9.9%	19.8%	19.8%
	Overreliance	1.4%	1.2%	7.8%	5.3%	24.1%	11.2%
	Susceptibility	52.6%	52.6%	60.1%	60.1%	70.0%	70.0%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	19.4%	18.9%	21.9%	19.0%	32.5%	19.5%
	Anticipated Error	20.5%	20.2%	24.7%	24.8%	27.1%	30.7%
	Regret	2.3%	2.3%	9.7%	9.7%	19.8%	19.8%
	Overreliance	0.7%	0.6%	4.6%	2.6%	17.3%	6.2%
	Susceptibility	52.3%	52.3%	59.7%	59.7%	69.8%	69.8%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	33.4%	33.6%	37.5%	33.8%	46.5%	34.1%
	Anticipated Error	34.0%	34.3%	35.1%	36.6%	30.4%	39.5%
	Regret	2.5%	2.5%	10.0%	10.0%	19.9%	19.9%
	Overreliance	1.5%	1.3%	9.6%	5.7%	25.9%	12.0%
	Susceptibility	52.5%	52.5%	60.0%	60.0%	69.9%	69.9%

Table 4: Overview of performance and regret for LR model trained on all datasets and training procedures for noise draw 2.

		5		20		40	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	23.3%	22.7%	27.2%	23.4%	36.4%	24.3%
	Anticipated Error	24.1%	23.9%	28.7%	28.3%	29.6%	35.0%
	Regret	2.3%	2.3%	10.2%	10.2%	18.9%	18.9%
	Overreliance	1.0%	0.8%	6.1%	3.7%	18.3%	6.3%
	Susceptibility	51.9%	51.9%	59.8%	59.8%	68.5%	68.5%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	20.6%	20.0%	24.7%	20.2%	33.5%	20.2%
	Anticipated Error	21.8%	21.3%	26.9%	25.9%	27.7%	32.0%
	Regret	2.4%	2.4%	9.8%	9.8%	19.3%	19.3%
	Overreliance	0.8%	0.7%	5.2%	2.8%	17.3%	5.5%
	Susceptibility	52.4%	52.4%	59.8%	59.8%	69.3%	69.3%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	31.6%	31.0%	33.3%	31.0%	42.1%	31.2%
	Anticipated Error	32.0%	31.6%	32.6%	33.7%	29.9%	36.6%
	Regret	2.6%	2.6%	10.0%	10.0%	20.0%	20.0%
	Overreliance	1.6%	1.4%	8.0%	5.5%	22.9%	11.5%
	Susceptibility	52.8%	52.8%	60.2%	60.2%	70.2%	70.2%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	19.4%	19.1%	23.4%	19.2%	32.3%	19.3%
	Anticipated Error	20.7%	20.6%	25.3%	25.0%	27.0%	30.2%
	Regret	2.6%	2.6%	10.1%	10.1%	20.1%	20.1%
	Overreliance	0.8%	0.7%	5.5%	2.9%	17.4%	6.6%
	Susceptibility	52.6%	52.6%	60.1%	60.1%	70.1%	70.1%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	33.2%	33.8%	37.1%	33.7%	44.8%	34.0%
	Anticipated Error	33.8%	34.6%	34.3%	36.1%	30.2%	39.2%
	Regret	2.6%	2.6%	10.3%	10.3%	19.6%	19.6%
	Overreliance	1.5%	1.3%	9.9%	6.2%	24.6%	11.9%
	Susceptibility	52.6%	52.6%	60.2%	60.2%	69.6%	69.6%

Table 5: Overview of performance and regret for LR model trained on all datasets and training procedures for noise draw 3.

		5		20		40	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	24.8%	23.3%	28.4%	23.8%	38.9%	24.6%
	Anticipated Error	25.8%	24.5%	29.8%	29.6%	29.2%	34.0%
	Regret	2.5%	2.5%	9.9%	9.9%	19.8%	19.8%
	Overreliance	1.1%	0.9%	6.1%	2.9%	20.8%	7.8%
	Susceptibility	52.1%	52.1%	59.5%	59.5%	69.4%	69.4%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	21.1%	20.1%	24.4%	20.0%	34.9%	20.2%
	Anticipated Error	22.4%	21.5%	26.0%	26.0%	27.8%	32.5%
	Regret	2.5%	2.5%	9.7%	9.7%	19.6%	19.6%
	Overreliance	0.8%	0.6%	5.5%	2.5%	18.5%	5.4%
	Susceptibility	52.5%	52.5%	59.7%	59.7%	69.6%	69.6%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	31.5%	31.1%	33.7%	31.3%	41.9%	31.3%
	Anticipated Error	31.9%	31.6%	32.8%	33.8%	29.8%	36.5%
	Regret	2.6%	2.6%	10.0%	10.0%	20.0%	20.0%
	Overreliance	1.6%	1.5%	8.1%	5.6%	22.9%	11.6%
	Susceptibility	52.8%	52.8%	60.2%	60.2%	70.2%	70.2%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	19.3%	19.0%	21.7%	19.1%	33.3%	19.5%
	Anticipated Error	20.5%	20.3%	24.8%	24.6%	27.6%	30.4%
	Regret	2.3%	2.3%	9.5%	9.5%	19.6%	19.6%
	Overreliance	0.7%	0.6%	4.3%	2.6%	17.5%	6.3%
	Susceptibility	52.3%	52.3%	59.5%	59.5%	69.6%	69.6%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	33.4%	33.5%	36.7%	33.5%	45.6%	33.7%
	Anticipated Error	33.6%	34.0%	34.1%	35.9%	29.9%	38.8%
	Regret	2.6%	2.6%	9.9%	9.9%	19.9%	19.9%
	Overreliance	1.8%	1.6%	9.5%	5.9%	25.4%	12.1%
	Susceptibility	52.6%	52.6%	59.9%	59.9%	69.9%	69.9%

Table 6: Overview of performance and regret for LR model trained on all datasets and training procedures for noise draw 4.

		5		20		40	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	23.3%	23.1%	28.8%	23.2%	39.5%	24.1%
	Anticipated Error	24.1%	24.2%	29.6%	28.3%	28.0%	34.3%
	Regret	2.7%	2.7%	10.7%	10.7%	21.1%	21.1%
	Overreliance	1.2%	1.1%	7.1%	3.9%	22.6%	8.3%
	Susceptibility	52.3%	52.3%	60.3%	60.3%	70.7%	70.7%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	20.7%	20.0%	23.8%	20.4%	32.8%	21.2%
	Anticipated Error	21.8%	21.2%	25.9%	26.2%	27.3%	32.6%
	Regret	2.3%	2.3%	9.8%	9.8%	19.8%	19.8%
	Overreliance	0.8%	0.7%	5.2%	2.7%	17.4%	6.3%
	Susceptibility	52.3%	52.3%	59.8%	59.8%	69.8%	69.8%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	31.6%	30.8%	33.6%	30.9%	43.9%	31.2%
	Anticipated Error	32.1%	31.6%	32.8%	33.8%	29.6%	36.1%
	Regret	2.5%	2.5%	10.2%	10.2%	20.0%	20.0%
	Overreliance	1.6%	1.3%	8.2%	5.5%	24.4%	11.7%
	Susceptibility	52.7%	52.7%	60.4%	60.4%	70.2%	70.2%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	19.1%	19.1%	22.1%	19.4%	29.8%	19.6%
	Anticipated Error	20.4%	20.5%	25.0%	25.1%	27.4%	30.4%
	Regret	2.4%	2.4%	10.1%	10.1%	20.3%	20.3%
	Overreliance	0.7%	0.6%	4.8%	2.9%	15.6%	6.8%
	Susceptibility	52.4%	52.4%	60.1%	60.1%	70.3%	70.3%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	33.7%	34.0%	37.2%	34.0%	46.3%	33.7%
	Anticipated Error	33.9%	34.5%	34.8%	36.2%	29.4%	38.1%
	Regret	2.7%	2.7%	10.0%	10.0%	20.3%	20.3%
	Overreliance	1.9%	1.7%	9.5%	6.1%	26.3%	12.9%
	Susceptibility	52.6%	52.6%	60.0%	60.0%	70.3%	70.3%

Table 7: Overview of performance and regret for LR model trained on all datasets and training procedures for noise draw 5.

		5		20		40	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	13.3%	12.8%	18.6%	19.2%	37.5%	26.2%
	Anticipated Error	14.4%	14.0%	20.3%	22.0%	25.1%	26.7%
	Regret	3.0%	3.0%	10.1%	10.1%	19.7%	19.7%
	Overreliance	1.1%	1.0%	5.3%	4.7%	21.4%	13.1%
	Susceptibility	52.6%	52.6%	59.7%	59.7%	69.3%	69.3%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	15.6%	15.9%	18.8%	16.8%	32.7%	22.1%
	Anticipated Error	17.4%	17.5%	23.9%	23.2%	26.6%	25.9%
	Regret	2.5%	2.5%	10.2%	10.2%	19.8%	19.8%
	Overreliance	0.4%	0.5%	3.4%	2.5%	17.7%	10.8%
	Susceptibility	52.5%	52.5%	60.2%	60.2%	69.8%	69.8%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	29.8%	29.7%	31.5%	30.0%	37.7%	29.5%
	Anticipated Error	30.4%	30.4%	31.8%	33.4%	29.7%	36.7%
	Regret	2.5%	2.5%	10.0%	10.0%	19.7%	19.7%
	Overreliance	1.4%	1.3%	7.1%	5.0%	19.8%	9.9%
	Susceptibility	52.7%	52.7%	60.2%	60.2%	69.9%	69.9%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	17.7%	17.9%	19.2%	18.3%	24.0%	18.9%
	Anticipated Error	19.1%	19.4%	23.4%	24.0%	26.2%	29.5%
	Regret	2.2%	2.2%	9.8%	9.8%	19.5%	19.5%
	Overreliance	0.6%	0.5%	3.7%	2.7%	11.7%	6.3%
	Susceptibility	52.2%	52.2%	59.8%	59.8%	69.5%	69.5%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	28.4%	28.6%	31.0%	30.3%	39.4%	35.7%
	Anticipated Error	28.2%	28.2%	28.6%	28.7%	25.2%	27.8%
	Regret	2.6%	2.6%	10.0%	10.0%	19.6%	19.6%
	Overreliance	2.0%	2.1%	8.7%	8.1%	22.6%	19.1%
	Susceptibility	52.6%	52.6%	60.0%	60.0%	69.6%	69.6%

Table 8: Overview of performance and regret for DNN model trained on all datasets and training procedures for noise draw 1.

		5		20		40	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	13.8%	13.0%	17.1%	16.0%	38.6%	24.2%
	Anticipated Error	14.5%	14.1%	20.3%	21.1%	25.1%	26.5%
	Regret	2.1%	2.1%	10.2%	10.2%	20.1%	20.1%
	Overreliance	0.8%	0.5%	4.4%	3.2%	22.5%	12.1%
	Susceptibility	51.7%	51.7%	59.8%	59.8%	69.7%	69.7%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	16.3%	16.4%	16.9%	17.3%	33.5%	22.1%
	Anticipated Error	18.1%	18.0%	23.2%	23.4%	27.3%	35.7%
	Regret	2.3%	2.3%	9.7%	9.7%	19.8%	19.8%
	Overreliance	0.3%	0.4%	2.2%	2.3%	17.9%	4.8%
	Susceptibility	52.3%	52.3%	59.7%	59.7%	69.8%	69.8%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	30.1%	29.6%	30.8%	29.4%	38.1%	29.7%
	Anticipated Error	30.8%	30.6%	31.9%	33.9%	30.1%	36.7%
	Regret	2.4%	2.4%	9.9%	9.9%	19.8%	19.8%
	Overreliance	1.2%	1.1%	6.5%	4.0%	19.9%	10.2%
	Susceptibility	52.6%	52.6%	60.1%	60.1%	70.0%	70.0%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	17.9%	17.7%	18.7%	18.6%	26.7%	19.1%
	Anticipated Error	19.2%	19.0%	23.2%	23.8%	26.7%	29.3%
	Regret	2.3%	2.3%	9.7%	9.7%	19.8%	19.8%
	Overreliance	0.6%	0.6%	3.4%	3.0%	13.5%	6.8%
	Susceptibility	52.3%	52.3%	59.7%	59.7%	69.8%	69.8%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	28.8%	29.7%	32.7%	31.6%	40.5%	34.1%
	Anticipated Error	29.1%	29.9%	29.6%	30.2%	25.9%	28.4%
	Regret	2.5%	2.5%	10.0%	10.0%	19.9%	19.9%
	Overreliance	1.6%	1.6%	9.3%	8.1%	23.3%	17.9%
	Susceptibility	52.5%	52.5%	60.0%	60.0%	69.9%	69.9%

Table 9: Overview of performance and regret for DNN model trained on all datasets and training procedures for noise draw 2

		5		20		40	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	13.4%	13.2%	16.7%	16.7%	31.6%	25.3%
	Anticipated Error	14.3%	13.9%	19.2%	21.6%	23.4%	32.1%
	Regret	2.3%	2.3%	10.2%	10.2%	18.9%	18.9%
	Overreliance	0.8%	0.9%	4.7%	3.3%	17.7%	8.9%
	Susceptibility	51.9%	51.9%	59.8%	59.8%	68.5%	68.5%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	15.7%	16.5%	16.1%	16.1%	26.2%	23.3%
	Anticipated Error	17.5%	18.1%	22.2%	22.2%	26.6%	35.9%
	Regret	2.4%	2.4%	9.8%	9.8%	19.3%	19.3%
	Overreliance	0.3%	0.5%	2.3%	2.4%	12.9%	5.2%
	Susceptibility	52.4%	52.4%	59.8%	59.8%	69.3%	69.3%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	29.8%	29.8%	31.7%	29.4%	47.9%	29.6%
	Anticipated Error	30.3%	30.3%	31.6%	32.5%	29.7%	37.8%
	Regret	2.6%	2.6%	10.0%	10.0%	20.0%	20.0%
	Overreliance	1.5%	1.4%	7.4%	5.1%	27.1%	9.4%
	Susceptibility	52.8%	52.8%	60.2%	60.2%	70.2%	70.2%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	18.4%	17.7%	20.3%	18.3%	26.2%	19.3%
	Anticipated Error	19.8%	19.2%	23.9%	24.3%	25.8%	28.4%
	Regret	2.6%	2.6%	10.1%	10.1%	20.1%	20.1%
	Overreliance	0.7%	0.6%	4.3%	2.8%	13.8%	7.7%
	Susceptibility	52.6%	52.6%	60.1%	60.1%	70.1%	70.1%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	28.8%	28.3%	32.1%	30.1%	41.4%	36.5%
	Anticipated Error	29.0%	28.9%	29.2%	30.5%	26.1%	27.7%
	Regret	2.6%	2.6%	10.3%	10.3%	19.6%	19.6%
	Overreliance	1.6%	1.4%	9.3%	7.0%	23.6%	19.7%
	Susceptibility	52.6%	52.6%	60.2%	60.2%	69.6%	69.6%

Table 10: Overview of performance and regret for DNN model trained on all datasets and training procedures for noise draw 3

		5		20		40	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	14.5%	13.8%	20.5%	20.0%	27.1%	24.4%
	Anticipated Error	15.5%	14.8%	21.7%	20.2%	22.0%	30.6%
	Regret	2.5%	2.5%	9.9%	9.9%	19.8%	19.8%
	Overreliance	0.9%	0.9%	5.5%	6.1%	15.9%	9.8%
	Susceptibility	52.1%	52.1%	59.5%	59.5%	69.4%	69.4%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	16.0%	16.3%	17.9%	18.3%	24.9%	21.0%
	Anticipated Error	17.9%	18.3%	23.2%	23.6%	26.7%	36.3%
	Regret	2.5%	2.5%	9.7%	9.7%	19.6%	19.6%
	Overreliance	0.3%	0.3%	2.9%	2.9%	12.1%	3.4%
	Susceptibility	52.5%	52.5%	59.7%	59.7%	69.6%	69.6%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	29.6%	29.4%	31.6%	29.5%	40.8%	30.1%
	Anticipated Error	30.2%	30.2%	31.9%	33.2%	29.7%	35.8%
	Regret	2.6%	2.6%	10.0%	10.0%	20.0%	20.0%
	Overreliance	1.4%	1.3%	7.1%	4.7%	22.1%	11.1%
	Susceptibility	52.8%	52.8%	60.2%	60.2%	70.2%	70.2%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	18.3%	17.9%	19.1%	17.9%	23.1%	18.5%
	Anticipated Error	19.4%	19.2%	23.1%	23.2%	26.0%	30.4%
	Regret	2.3%	2.3%	9.5%	9.5%	19.6%	19.6%
	Overreliance	0.7%	0.6%	3.6%	2.8%	11.3%	5.5%
	Susceptibility	52.3%	52.3%	59.5%	59.5%	69.6%	69.6%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	29.1%	28.6%	31.5%	30.2%	40.7%	32.4%
	Anticipated Error	29.3%	29.0%	28.6%	30.4%	25.6%	29.3%
	Regret	2.6%	2.6%	9.9%	9.9%	19.9%	19.9%
	Overreliance	1.8%	1.6%	9.0%	6.9%	23.5%	16.3%
	Susceptibility	52.6%	52.6%	59.9%	59.9%	69.9%	69.9%

Table 11: Overview of performance and regret for DNN model trained on all datasets and training procedures for noise draw 4

		5		20		40	
Dataset	Metrics	Ignore	Hedge	Ignore	Hedge	Ignore	Hedge
`shock_eicu` $n = 3,456$ $d = 104$ Pollard et al. [40]	True Error	12.8%	12.7%	18.7%	18.5%	36.4%	22.1%
	Anticipated Error	13.7%	13.9%	19.6%	20.5%	24.4%	31.8%
	Regret	2.7%	2.7%	10.7%	10.7%	21.1%	21.1%
	Overreliance	1.0%	0.9%	6.1%	5.5%	21.9%	8.3%
	Susceptibility	52.3%	52.3%	60.3%	60.3%	70.7%	70.7%
`shock_mimic` $n = 15,254$ $d = 104$ Johnson et al. [19]	True Error	16.5%	16.3%	18.3%	16.5%	29.8%	21.4%
	Anticipated Error	18.3%	17.7%	23.8%	23.6%	26.9%	27.3%
	Regret	2.3%	2.3%	9.8%	9.8%	19.8%	19.8%
	Overreliance	0.3%	0.5%	2.8%	1.8%	15.5%	9.5%
	Susceptibility	52.3%	52.3%	59.8%	59.8%	69.8%	69.8%
`lungcancer` $n = 62,916$ $d = 40$ NCI [37]	True Error	30.2%	29.6%	31.6%	29.7%	38.6%	29.7%
	Anticipated Error	30.7%	30.5%	31.7%	33.4%	29.4%	37.4%
	Regret	2.5%	2.5%	10.2%	10.2%	20.0%	20.0%
	Overreliance	1.4%	1.2%	7.4%	4.9%	20.6%	9.8%
	Susceptibility	52.7%	52.7%	60.4%	60.4%	70.2%	70.2%
`mortality` $n = 20,334$ $d = 84$ Le Gall et al. [24]	True Error	18.1%	17.8%	19.5%	18.3%	27.3%	18.8%
	Anticipated Error	19.5%	19.3%	23.5%	24.1%	26.0%	29.4%
	Regret	2.4%	2.4%	10.1%	10.1%	20.3%	20.3%
	Overreliance	0.6%	0.5%	4.0%	2.8%	14.6%	6.9%
	Susceptibility	52.4%	52.4%	60.1%	60.1%	70.3%	70.3%
`support` $n = 9,696$ $d = 114$ Knaus et al. [21]	True Error	28.4%	28.6%	32.4%	30.0%	41.7%	35.2%
	Anticipated Error	28.5%	28.8%	29.3%	29.9%	25.9%	33.6%
	Regret	2.7%	2.7%	10.0%	10.0%	20.3%	20.3%
	Overreliance	1.8%	1.8%	9.3%	7.2%	24.4%	16.5%
	Susceptibility	52.6%	52.6%	60.0%	60.0%	70.3%	70.3%

Table 12: Overview of performance and regret for DNN model trained on all datasets and training procedures for noise draw 5