Predictive Multiplicity in Probabilistic Classification

Jamelle Watson-Daniels,[1] **David C. Parkes,**[1] **Berk Ustun**[2]

[1]Harvard University, [2]UC San Diego
jwatsondaniels@g.harvard.edu, parkes@eecs.harvard.edu, berk@ucsd.edu

Abstract

There may exist multiple models that perform almost equally well for any given prediction task. We examine how predictions change across these competing models. In particular, we study *predictive multiplicity* – in probabilistic classification. We formally define measures for our setting and develop optimization-based methods to compute these measures for convex empirical risk minimization problems. We apply our methodology to gain insight into why predictive multiplicity arises. We demonstrate the incidence and prevalence of predictive multiplicity in real-world risk assessment tasks. Our results emphasize the need to report multiplicity more widely.

1 Introduction

Probabilistic classification is often incorporated into real-world risk assessment tasks to inform decisions. For instance, probabilistic classifiers that predict consumer default risk are used by lenders to underwrite loans (Bekhet and Eletter 2014; Attigeri, Pai, and Pai 2017). Similarly in clinical applications, physicians make treatment decisions using models that predict whether a person suffers from a serious illness (Than et al. 2014; Khand et al. 2017; Chen et al. 2021). And in criminal justice, judges often make parole and sentencing decisions guided by models that predict the probability that a person will fail to appear in court (Austin, Ocker, and Bhati 2010; Latessa et al. 2010; Christin, Rosenblat, and Boyd 2015; Zeng, Ustun, and Rudin 2017).

The standard approach to selecting a probabilistic classifier involves optimizing a loss function via empirical risk minimization. But for a given prediction task, there may exist multiple models that perform almost equally well, which is referred to in machine learning as model *multiplicity* (Breiman 2001). These near-optimal, *competing models*, have similar performance but characteristic differences - e.g. their interpretability (Semenova, Rudin, and Parr 2019), explainability (Fisher, Rudin, and Dominici 2019; Dong and Rudin 2020), counterfactual invariance (D'Amour et al. 2020), or fairness (Coston, Rambachan, and Chouldechova 2021; Black and Fredrikson 2021; Ali, Lahoti, and Gummadi 2021). These differences can drastically change how we develop, choose, and use models (Black, Raghavan, and Barocas 2022).

We investigate how predictions change across competing models by studying *predictive multiplicity*: the prevalence of conflicting predictions over competing models (Marx, Calmon, and Ustun 2019). To understand our motivation, consider the significance of competing models assigning vastly different predictions in practice. In mortality prediction, a conflicting risk prediction would alter treatment decisions and health outcomes (Moreno et al. 2005). In drug discovery, a conflicting risk prediction could switch the compounds chosen for confirmatory experiments (Stokes et al. 2020). By measuring and reporting the prevalence of conflicts, we can improve how we choose and use machine learning models. If end-users know that an individual risk estimate conflicts over the set of competing models, they could abstain from prediction (Black, Leino, and Fredrikson 2022; Hamid et al. 2017) or defer a decision to a human expert (Mozannar and Sontag 2020; Kompa, Snoek, and Beam 2021a). If model developers know that many risk estimates conflict when compared across competing models, they might reconsider deployment and dedicate time to contend with multiplicity. These implications underline the importance of measuring and reporting predictive multiplicity more widely.

Our main contributions are:

1. We formally introduce measures of predictive multiplicity in our setting. The Viable Prediction Range examines how multiplicity affects predictions. Ambiguity and discrepancy reflect the proportion of individuals assigned conflicting risk estimates by competing models.

2. We develop optimization-based methods to compute our measures of predictive multiplicity for convex empirical risk minimization problems. This includes employing mixed-integer non-linear programming and outer-approximation algorithms. Whereas previous work defines competing models over a single performance metric, our methods enable developers to generalize for additional near-optimal metrics.

3. We offer insights into why predictive multiplicity arises via systematic experiments on synthetic data. We find that predictive multiplicity is more prevalent for examples that are both outliers and close to the discriminant boundary, for datasets that are less separable, and for minority groups when a dataset has a majority-minority structure.

4. We present an empirical study of predictive multiplicity on seven real-world risk assessment tasks. We show

that probabilistic classification tasks can in fact admit competing models that assign substantially different risk estimates. Our results also demonstrate how multiplicity can disproportionately impact marginalized individuals.

5. We provide a software implementation of our toolkit.

Related Work. Our work is positioned alongside research on *model multiplicity*. This effect has been referenced in the statistics literature. For example, Chatfield (1995) calls for performing a sensitivity analysis over competing models, while Breiman (2001) cites multiplicity as a reason to avoid explaining a single model to draw conclusions about the broader data-generating process. Recent advances in computation make multiplicity analysis possible, leading to a stream of research on how competing models differ (Fisher, Rudin, and Dominici 2019; Dong and Rudin 2020; Semenova, Rudin, and Parr 2019; D'Amour et al. 2020; Veitch et al. 2021; Pawelczyk, Broelemann, and Kasneci 2020; Coston, Rambachan, and Chouldechova 2021; Black and Fredrikson 2021; Ali, Lahoti, and Gummadi 2021)

Our work is distinctly focused on how multiplicity affects prediction. Our approach builds on Marx, Calmon, and Ustun (2019), who study this effect in classification tasks with yes-or-no predictions. As shown in Figure 1, their measures and methods do not extend to our setting. Measuring multiplicity in probabilistic classification is complicated by the need to clarify the meaning of "conflicting". In effect, what constitutes a conflicting risk prediction can change across applications (e.g., predictions that vary by 5% or 30%). Likewise, what constitutes a "competing" model can change across applications. The present work addresses both of these problems by introducing methods that allow users to specify what is "competing" (near-optimal metric) and what is "conflicting" (deviation threshold). Also, previous work has yet to examine why predictive multiplicity arises, which we contribute to.

One way we compute predictive multiplicity is by constructing a range of individual risk predictions as a way to quantify pointwise uncertainty resulting from an underspecified empirical risk minimization problem. This relates to methods for evaluating predictive uncertainty such as conformal prediction (Shafer and Vovk 2008; Romano et al. 2020) as well as Bayesian approaches (see e.g., Dusenberry et al. 2020; Lum, Dunson, and Johndrow 2021). However, conformal prediction focuses on uncertainty that arises due to non-conformity between historical data and new data, which is orthogonal to our goal. We focus on a non-Bayesian approach, recognizing that non-Bayesian methods are very typical in applied machine learning. Our goals relate also to a line of work that aims to quantify and communicate uncertainty in machine learning (Hofman, Goldstein, and Hullman 2020; Kale, Kay, and Hullman 2020; McGrath et al. 2020; Soyer and Hogarth 2012; Kompa, Snoek, and Beam 2021b) and calibrate trust among stakeholders (Joslyn and LeClerc 2013). Other complementary work seeks interventions to resolve multiplicity (Ali, Lahoti, and Gummadi 2021) or ensembling (Black, Leino, and Fredrikson 2022).

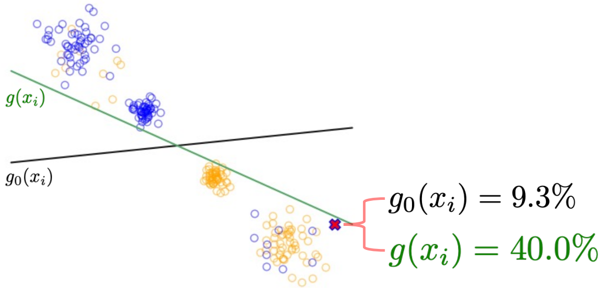


Figure 1: Classification models that make the same yes-or-no predictions can still assign conflicting risk predictions. Here, we show a 2D classification task with $n^+ = 200$ positive examples (blue) and $n^- = 200$ negative examples (orange). We plot the decision boundary of a baseline model g_0 (black; log-loss/AUC/calibration = 0.41/0.88/17%) and a competing model that performs almost equally well $g(\boldsymbol{x}_i)$ (green; log-loss/AUC/calibration = 0.42/0.89/16%). As shown, both classifiers make the same yes-or-no predictions, but assign conflicting risk estimates to individual examples – e.g., example $\boldsymbol{x}_i$ is assigned a risk estimate of $g_0(\boldsymbol{x}_i) = 9.3\%$ by the baseline model but $g(\boldsymbol{x}_i) = 40.0\%$ by the competing model.

2 Framework

We consider a probabilistic classification task with a dataset of n examples $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Each example consists of a feature vector $\boldsymbol{x}_i = [1, x_{i1}, \ldots, x_{id}] \in \mathcal{X} \subseteq \mathbb{R}^{d+1}$ and a label $y_i \in \mathcal{Y} = \{-1, +1\}$, where $y_i = +1$ is an event of interest (e.g., default on a loan). With the dataset, we train a probabilistic classifier $g : \mathcal{X} \to [0, 1]$ – i.e., a model that assigns a risk estimate to example $\boldsymbol{x}_i$ as: $g(\boldsymbol{x}_i) := \Pr(y_i = +1|\boldsymbol{x}_i)$. We refer to this model as the *baseline model*, g_0, because it is the optimal solution to an empirical risk minimization (ERM) problem of the form:

$$\min_{g \in \mathcal{H}} L(g; \mathcal{D}), \tag{1}$$

where $\mathcal{H}$ is a family of probabilistic classifiers, and $L(\ \cdot\ ; \mathcal{D})$ is a loss function evaluated on the dataset $\mathcal{D}$. In what follows, we write $L(g)$ instead of $L(g; \mathcal{D})$ for conciseness. We evaluate the performance of a model in terms of $L(g)$, as well as the following metrics:

1. *Risk Calibration*: A risk-calibrated model assigns risk predictions that match observed frequencies (Naeini, Cooper, and Hauskrecht 2015). We measure risk calibration in terms of *expected calibration error*:

$$\text{ECE}(g) = \sum_{b=1}^{B} \frac{n_b}{n} |\hat{p}_b(g) - \bar{p}_b|. \tag{2}$$

Here: I_b is the index set of n_b examples in bin $b \in [B]$; and $\hat{p}_b(g) := \frac{1}{n_b} \sum_{i \in I_b} g(\boldsymbol{x}_i)$ and $\bar{p}_b = \frac{1}{n_b} \sum_{i \in I_b} \mathbb{1}[y_i = +1]$ are the mean predicted risk and mean observed risk of examples in bin $b \in [B]$, respectively.

2. *Rank Accuracy*: A rank-accurate model outputs risk predictions that can be used to correctly order examples in terms of true risk. We assess rank accuracy using the *area under the ROC curve*:

$$\mathrm{AUC}(g) = \frac{1}{n^+n^-} \sum_{\substack{i:y_i=+1 \\ k:y_k=-1}} \mathbb{1}[g(\boldsymbol{x}_i) > g(\boldsymbol{x}_k)], \qquad (3)$$

where $n^+ = |\{i : y_i = +1\}|$ and $n^- = |\{i : y_i = -1\}|$.

In what follows, we let $M(g; \mathcal{D}) \in \mathbb{R}_+$ denote the performance of $g \in \mathcal{H}$ over a dataset $\mathcal{D}$ in regards to *performance metric* $M(g)$, where the convention is that lower values of $M(g)$ are better; when working with AUC, we measure the *AUC error*: $M(g) = 1 - \mathrm{AUC}(g)$.

2.1 Competing Models

Competing models are classifiers with near-optimal performance compared to the baseline model. A *competing model* is any model $g \in \mathcal{H}$ whose performance is within ϵ of the baseline model g_0.

Definition 1 (ϵ-Level Set) *Given a baseline model g_0, metric M, and error tolerance $\epsilon > 0$, the* set of competing models *(ϵ-level set) is the set:*

$$\mathcal{H}_\epsilon(g_0) := \{g \in \mathcal{H} : M(g) \leq M(g_0) + \epsilon\}.$$

Our methods consider multiplicity over a range of ϵ values. In practice, a suitable choice of ϵ should reflect the epistemic uncertainty in the performance of the baseline model. For instance, one could employ bootstrap re-sampling to measure the model uncertainty due to sample variation or consider worst-case uncertainty through generalization bounds.

2.2 Measuring Viable Risk Predictions

To examine how multiplicity affects predictions, we define a range of viable risk estimates that can be assigned by competing models.

Definition 2 (Viable Prediction Range) *The* viable prediction range *is the smallest and largest risk estimate assigned to example i over competing models in the ϵ-level set:*

$$V_\epsilon(\boldsymbol{x}_i) := [\min_{g \in \mathcal{H}_\epsilon(g_0)} g(\boldsymbol{x}_i), \max_{g \in \mathcal{H}_\epsilon(g_0)} g(\boldsymbol{x}_i)]. \qquad (4)$$

For a prediction task, computing the viable prediction ranges over a sample illuminates the extent to which competing models assign different risk estimates to individuals. Although we express the prediction range over an ϵ-level set using $[\cdot, \cdot]$ interval notation, not all predictions between the min and the max may be attainable by a competing model.

2.3 Measuring Predictive Multiplicity

We say that a risk estimate is *conflicting* if it differs from the baseline risk estimate by at least some deviation threshold, $\delta \in (0, 1)$. The appropriate value of δ will depend on the application; i.e. a conflicting risk prediction in a clinical decision support task may differ from that which constitutes a conflicting risk prediction in recidivism prediction. The following examples illustrate the importance of reporting predictive multiplicity over a range of δ values.

Recidivism Prediction: Consider predicting an individual's risk of failing to appear in court using past arrest data (Lum, Dunson, and Johndrow 2021). Suppose there are four risk categories partitioned as follows– low: 0-10%, medium-low: 10-20%, medium-high: 20-30%, high: 30-100%. In this example, a deviation threshold $\delta = 10\%$ is informative because it would flag a change in risk that is large enough for any individual to go from "low" risk to "high" risk.

Medical Risk Prediction: Consider the task of predicting stroke risk for patients with atrial fibrillation (see e.g., the $CHADS_2$ risk score at MDCalc.com). The individual risk estimates can be used to inform blood thinner prescription decisions. One recommended usage suggests the following partitioning– 0% - 0.3%: do not prescribe blood thinner, 0.3-2.8%: maybe prescribe blood thinner, 2.9%+: prescribe blood thinner. If we study predictive multiplicity for this model, a value such as $\delta = 1\%$ is informative because a risk estimate shift by 1% could change the decision to prescribe a blood thinner for many individuals.

With a better understanding of the deviation threshold, we now define measures of predictive multiplicity. Ambiguity and discrepancy reflect the proportion of examples in a sample S assigned conflicting risk estimates by competing models. These definitions follow Marx, Calmon, and Ustun (2019), who give analogous definitions for the problem of multiplicity with binary predictions (see Figure 1 for an illustration of the difference between this problem and the multiplicity of risk estimates).

Definition 3 (Ambiguity) *The* (ϵ, δ)-ambiguity *of a probabilistic classification task over a sample S is the proportion of examples in S whose baseline risk estimate changes by at least δ over the ϵ-level set:*

$$A_{\delta,\epsilon}(g_0; S) := \frac{1}{|S|} \sum_{i \in S} \mathbb{1}[\max_{g \in \mathcal{H}_\epsilon(g_0)} |g(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)| \geq \delta].$$

Relative to the baseline model, ambiguity makes a statement about the proportion of individuals whose risk estimate is uncertain by at least δ. High ambiguity means more uncertainty in risk predictions. Users may also consult the viable prediction range to guide decisions using the baseline model.

Definition 4 (Discrepancy) *The (ϵ, δ)-discrepancy of a probabilistic classification task over a sample S is the maximum proportion of examples in S whose risk estimates could change by at least δ by switching the baseline model with a competing model in the ϵ-level set:*

$$D_{\delta,\epsilon}(g_0; S) := \max_{g \in \mathcal{H}_\epsilon(g_0)} \frac{1}{|S|} \sum_{i \in S} \mathbb{1}[|g(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)| \geq \delta].$$

Relative to the baseline model, discrepancy reflects the maximum the number of conflicting risk estimates as a result of replacing baseline model with a competing model in the ϵ-level set.

Ambiguity and discrepancy differ in the stance they take in regard to the worst case. Discrepancy measures the worst-case number of predictions that will change by switching the baseline model with a competing model. In contrast, ambiguity focuses on the worst case for prediction variation over the set of competing models. If we were to abstain from prediction on points that are assigned a conflicting prediction by a competing model (using e.g., selective classification methods Black, Leino, and Fredrikson 2022), then ambiguity would reflect the abstention rate.

Computing Ambiguity with Viable Prediction Ranges. As shown in Figure 2, we can use the viable prediction ranges of all points in a sample to compute ambiguity. Given the viable prediction range for each example, we can calculate the maximum difference between the baseline risk and that assigned by competing models. We can then compute ambiguity by measuring the proportion of examples where this difference exceeds the deviation threshold.

3 Methodology

In this section, we detail the procedure for computing measures of predictive multiplicity. This methodology can be applied to any convex loss function $L(\cdot)$, and together with a training problem that employs a convex regularization term. We illustrate the methodology on the classification task described in §2 by training a probabilistic classifier via logistic regression, with $g(\boldsymbol{x}_i) = \frac{1}{1+\exp(-\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)}$, where $\boldsymbol{w} = [w_0, w_1, \ldots, w_d]^\top \in \mathbb{R}^{d+1}$ is a coefficient vector. We train this baseline model by solving Eq. (1) to minimize normalized *logistic loss*: $L(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-\langle \boldsymbol{w}, y_i \boldsymbol{x}_i \rangle))$.

3.1 Measuring Ambiguity

We first present a method for computing ambiguity for different choices of ϵ and δ. The method also gives a conservative approximation of the viable prediction range for each example. We construct a pool of *candidate models* that assign a specific risk estimate to each example. From these models, we select those with performance within ϵ of the baseline model as the set of competing models.

Definition 5 (Candidate Model) *Given a baseline model g_0, a finite set of user-specified threshold probabilities $P \subseteq [0,1]$, then for each $p \in P$ a* candidate model *for example $\boldsymbol{x}_i$ is an optimal solution to the following constrained ERM:*

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^{d+1}} \quad & L(\boldsymbol{w}) \\ \text{s.t.} \quad & g(\boldsymbol{x}_i) \leq p, \quad \textit{if } p < g_0(\boldsymbol{x}_i) \\ & g(\boldsymbol{x}_i) \geq p. \quad \textit{if } p > g_0(\boldsymbol{x}_i) \end{aligned} \tag{5}$$

For each threshold probability $p \in P$, we train a candidate model g such that the probability assigned to the example is constrained to the threshold p. In this way, by training for each example and threshold probability $p \in P$, we obtain the set of candidate models $\mathcal{G} := \{g : i \in S, p \in P\}$. We choose to solve the instances in order of increasing values of threshold probability p, which allows us to warm-start the optimization using previous solutions.

Given the set of candidate models, we define a *candidate ϵ-level set* as

$$\tilde{\mathcal{H}}_\epsilon(g_0) = \{g \in \mathcal{G} : M(g_0) \leq M(g_0) + \epsilon\}. \tag{6}$$

We can use the candidate ϵ-level set to compute measures of predictive multiplicity.

This method is exact for ambiguity defined in terms of near-optimal loss when the grid of threshold probabilities $P \subseteq [0,1]$ aligns with $g_0(x_i) \pm \delta$ (i.e., is selected as appropriate to the baseline prediction for an example and the value of δ). For other metrics, such as AUC, this approach to compute ambiguity gives a conservative estimate (i.e., lower bound)—the training of a candidate model does not directly optimize for AUC, but we can retain only those candidate models that are competitive for the appropriate ϵ-level set definition. Since $\tilde{\mathcal{H}}_\epsilon(g_0) \subseteq \mathcal{H}_\epsilon(g_0)$, the candidate-model approach also provides a conservative estimate of the viable prediction range (Eq. (4)) for an example.

3.2 Measuring Discrepancy

Discrepancy is the maximum proportion of examples assigned conflicting risk estimates by a single competing model, $g \in \mathcal{H}_\epsilon(g_0)$. Recall that a conflicting risk estimate differs from the baseline risk estimate $g_0(\boldsymbol{x}_i)$ by at least some deviation threshold, $\delta > 0$. Therefore, measuring discrepancy with respect to a baseline model corresponds to solving the following maximization problem:

$$\max_{g \in \mathcal{H}_\epsilon(g_0)} \quad \sum_{i \in S} \mathbb{1}[|g(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)| \geq \delta]. \tag{7}$$

Given a sample S, the baseline loss L_0, error tolerance ϵ, and deviation threshold δ, we can formulate Eq. (7) as a mixed-integer non-linear program (MINLP):

$$\begin{aligned} \max_{\boldsymbol{w} \in \mathbb{R}^{d+1}} \quad & \sum_{i \in S} d_i \\ \text{s.t.} \quad & L(\boldsymbol{w}) \leq L_0 + \epsilon && && \text{(8a)} \\ & d_i = v_{i,\delta} + z_{i,\delta} && \forall i \in S && \text{(8b)} \\ M_{z,i}(1 - z_{i,\delta}) & \geq \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - U_{i,\delta} && \forall i \in S && \text{(8c)} \\ M_{v,i}(1 - v_{i,\delta}) & \geq -\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + B_{i,\delta} && \forall i \in S && \text{(8d)} \\ & d_i, z_{i,\delta}, v_{i,\delta} \in \{0,1\} && \forall i \in S \end{aligned}$$

The MINLP in (8) fits the parameters of a linear classifier that maximizes discrepancy . Here, the objective maximizes number of examples assigned a conflicting risk estimate using the indicator variables $d_i := \mathbb{1}[|g(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)| \geq \delta]$. Each d_i is set to $z_{i,\delta} := \mathbb{1}[g(\boldsymbol{x}_i) \leq (g_0(\boldsymbol{x}_i) - \delta)]$ (or $v_{i,\delta} := \mathbb{1}[g(\boldsymbol{x}_i) \geq (g_0(\boldsymbol{x}_i) + \delta)]$) when the model assigns a risk estimate to example i that exceeds δ on the low-side (or high-side) of the baseline risk estimate, respectively. We ensure the indicator behavior of $z_{i,\delta}$ and $v_{i,\delta}$ through the "Big-M" constraints (8d) and (8c), which flag deviations in score space. The Big-M parameters can be set as $M_{z,i} := -U_{i,\delta} + \max_{\boldsymbol{w}} \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle$ and $M_{v,i} := B_{i,\delta} - \min_{\boldsymbol{w}} \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle$, where $U_{i,\delta} := \text{logit}(g_0(\boldsymbol{x}_i) - \delta)$, and $B_{i,\delta} := \text{logit}(g_0(\boldsymbol{x}_i) + \delta)$. When the values of $U_{i,\delta}$ and $B_{i,\delta}$ lie outside of the $[0,1]$ domain of the logit, we can drop the relevant indicator variable from the formulation. We provide additional details on our MINLP formulation in Appendix A.

Outer-Approximation Algorithm. The challenge in solving (8) is that constraint (8a) is non-linear. We construct a linear approximation of the loss (see e.g., Franc and Sonnenburg 2008; Joachims, Finley, and Yu 2009) using an iterative,

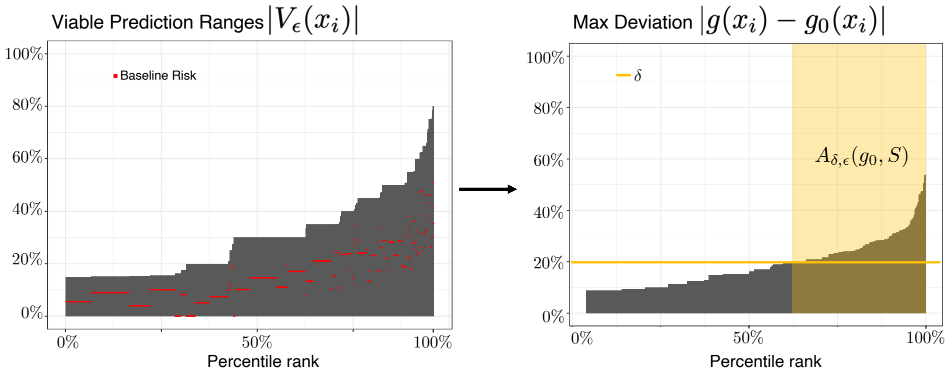


Figure 2: How viable prediction can be used to compute ambiguity. Left, we plot the magnitude of the viable prediction range, $V_\epsilon(\boldsymbol{x}_i)$ as well as the baseline risk estimate. From the viable prediction range, we can take the max difference from the baseline (right). Given a deviation threshold, δ, the examples assigned conflicting risk estimates can be identified to compute ambiguity, $A_{\delta,\epsilon}(g_0; S)$, highlighted in yellow.

outer-approximation method (see e.g., Ustun and Rudin 2017; Bertsimas et al. 2016; Bertsimas and King 2017) to solve. The algorithm recovers a globally optimal solution to the MINLP in (8), and can be implemented using a mixed-integer programming solver with callback functions (see e.g., Ustun and Rudin 2017; Bertsimas et al. 2016; Bertsimas and King 2017). The procedure builds a branch-and-bound tree to discover integer-feasible solutions that obey all constraints other than (8a). For each feasible solution identified, the procedure computes its loss to determine if it is feasible with respect to constraint (8a). If feasible, the procedure retains the solution. Otherwise, it updates the loss function approximation by adding a new linear constraint.

This method is exact for computing discrepancy in terms of near-optimal loss. For other metrics, we can again treat the intermediate solutions to the outer-approximation algorithm as candidate models and use these candidates to recover a lower bound similar to the method used in § 3.1.

4 Numerical Experiments

In this section, we present experiments on synthetic and real-world data. Our goals are to: (1) reveal dataset characteristics that impact predictive multiplicity; and (2) determine the extent to which real risk assessment tasks exhibit predictive multiplicity in practice.

4.1 Synthetic Datasets

Linear Separability. To demonstrate how separability informs predictive multiplicity, we compute ambiguity while varying the degree of separability and show results in Figure 3 column **(A)**. We set $\delta = 20\%$ and $\epsilon = 5\%$ and control separability by increasing the variance of the data from $\sigma = 4$ (top) to $\sigma = 10$ (bottom). A clear trend is that ambiguity increases as the data becomes less separable from 1% to 21%. Notice, also that the ambiguous examples tend to be those near the discriminant boundary and outliers.

Outliers and Margin Distance. We examine how predictive multiplicity relates to outlier distance from the discriminant boundary. We position outliers near and far from the discriminant boundary and compute ambiguity. As shown in Figure 3 column **(B)**, a clear trend is that examples that are outliers but far from the discriminant boundary (high margin) are less susceptible to predictive multiplicity.

Majority-Minority Structure. We consider the effect of systematically varying the majority-minority structure of data. For this, we generate a majority class that has a different statistical pattern of features than a minority class. Given the two groups, the model is faced with a tradeoff between correctly predicting one group or the other. In Figure 3 column **(C)**, we vary the ratio in a majority-minority structure revealing that the minority group is more prone to predictive multiplicity. The ambiguity of the minority group at 10:1 is substantially larger than for the majority group. This shows the importance of evaluating multiplicity across subgroups.

4.2 Real-World Datasets

In this section, we evaluate predictive multiplicity in risk prediction tasks from medicine, lending, and criminal justice.[1] Altogether, we consider seven datasets that exhibit variations in sample size, number of features, and class imbalance (see Table 1 in the Appendix). For each dataset, we compute viable prediction ranges, ambiguity and discrepancy using the methods outlined in §3. When training candidate models, we adopt a grid of target predictions: $P = \{0.01, 0.1, 0.2, \ldots, 0.9, 0.99\}$. We compute discrepancy by solving the MINLP Eq. (7) with CPLEX v20.1 (Diamond and Boyd 2016) on a single CPU with 16GB RAM. Our results are shown in Figure 4, and additional results are in Appendix E.

Viable Prediction Ranges. Our results show that competing models can assign risk estimates that vary substantially.

[1] This is not an endorsement of current usage of risk assessment tools in criminal justice. The use of prediction software raises serious concerns in this domain. And we do not condone building models on arrest data to inform or justify increased policing.

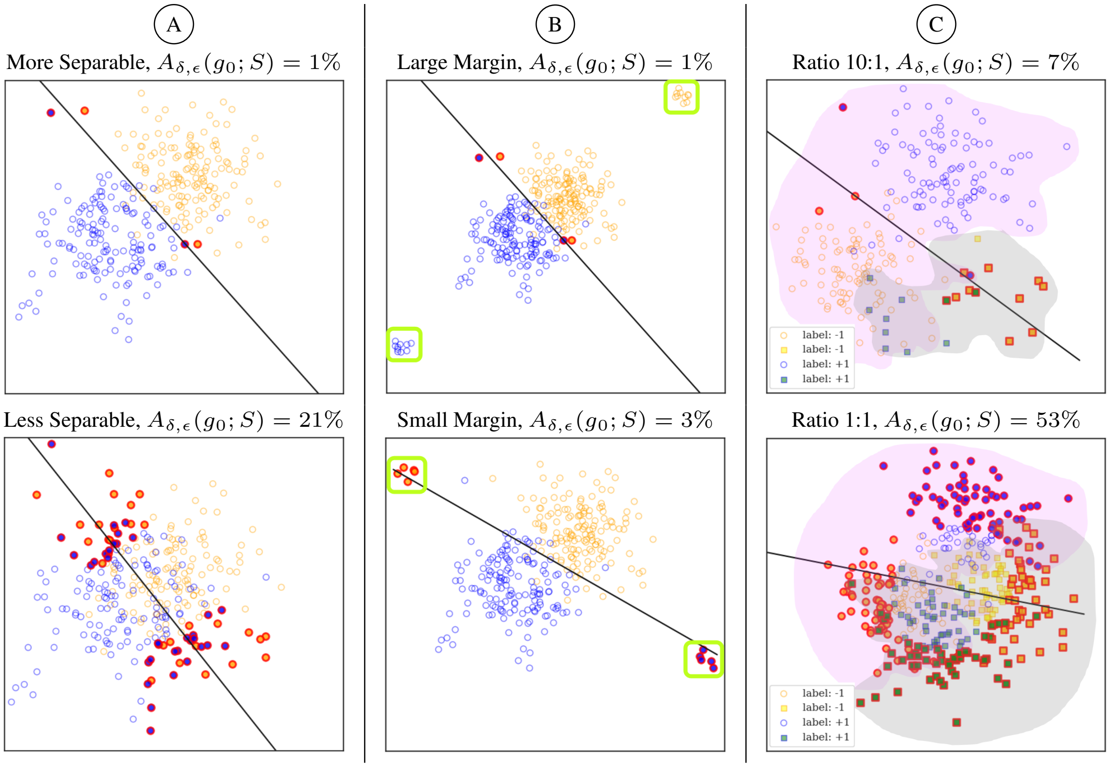


Figure 3: Experiments on synthetic data. In **(A)**, we vary the degree of **separability** and find that ambiguity increases as separability decreases. In **(B)**, we position **outliers** near and away from the discriminant boundary finding that outliers closer to the boundary are more prone to ambiguity. In **(C)**, we vary the ratio in a **majority-minority** structured dataset: magenta shading-majority group (circles), grey shading- minority group (squares) revealing that the minority group is more prone to ambiguity. In the figures, $Y = +1$ examples are blue, $Y = -1$ examples are orange, and ambiguous examples are highlighted red and we set $\delta = 20\%$ and $\epsilon = 5\%$.

Viable prediction ranges are plotted in columns **(A)** and **(B)** of Figure 4, and we see non-zero viable prediction ranges for all examples across all datasets. The viable ranges for `apnea` and `mammo` appear much larger compared to `compas_arrest`. In terms of near-optimal loss, `apnea` has the most variation, while `mammo` has the most variation in terms of AUC. This points to the value in varying near-optimal metric.

Ambiguity and Discrepancy. Ambiguity and discrepancy are shown in columns **(C)** and **(D)** of Figure 4, respectively. For $\epsilon = 1\%$ and $\delta = 20\%$, we see ambiguity values at 35.3% (`mammo`), 95.8% (`apnea`), and 51.4% (`compas_arrest`). This means that 35.3% of breast cancer risk estimates vary by at least 20% over near-optimal models. We see discrepancy values at 3.6% (`mammo`), 1.2% (`apnea`), and 5.4% (`compas_arrest`) for $\epsilon = 1\%$ and $\delta = 20\%$. `compas_arrest` is the worst in terms of discrepancy, while `apnea` has the most severe ambiguity. Thus, ambiguity and discrepancy are not always coupled.

On the Choice of Performance Metric. In settings where we would want a model that performs well in terms of AUC, we should measure predictive multiplicity over a set of competing models with near-optimal AUC. In practice, it is often convenient to measure predictive multiplicity over a set of competing models that attain near-optimal loss (since the loss can be encoded into an optimization problem). This is a problem because small variations in loss can lead to large variations in AUC – thus models with near-optimal loss may not match models with near-optimal AUC.

Our results show that measures of predictive multiplicity vary considerably based on the performance metric used to define the set of competing models. In particular, we find that measures like discrepancy and ambiguity will vary when measured over competing models that attain near-optimal loss, AUC, or ECE. For example, if we want to estimate the prevalence of samples whose predictions can change by over $\delta = 20\%$ on the `mammo` dataset, we find that ambiguity $= 35\%$ for competing models with loss within 1% of the baseline loss, but ambiguity $= 45\%$ over models with AUC within 0.5% of the baseline AUC. These differences highlight the need for approaches that measure predictive multiplicity in the terms of performance metric that we use to evaluate the model (e.g., AUC or ECE).

On Samples Prone to Ambiguity. Our results reveal a relationship between ambiguity and individual *uniqueness* (number of duplicates), *class* imbalance, and *baseline risk estimate*. For uniqueness, we find that across datasets, less than 10% of examples with more than 20 duplicates are am-

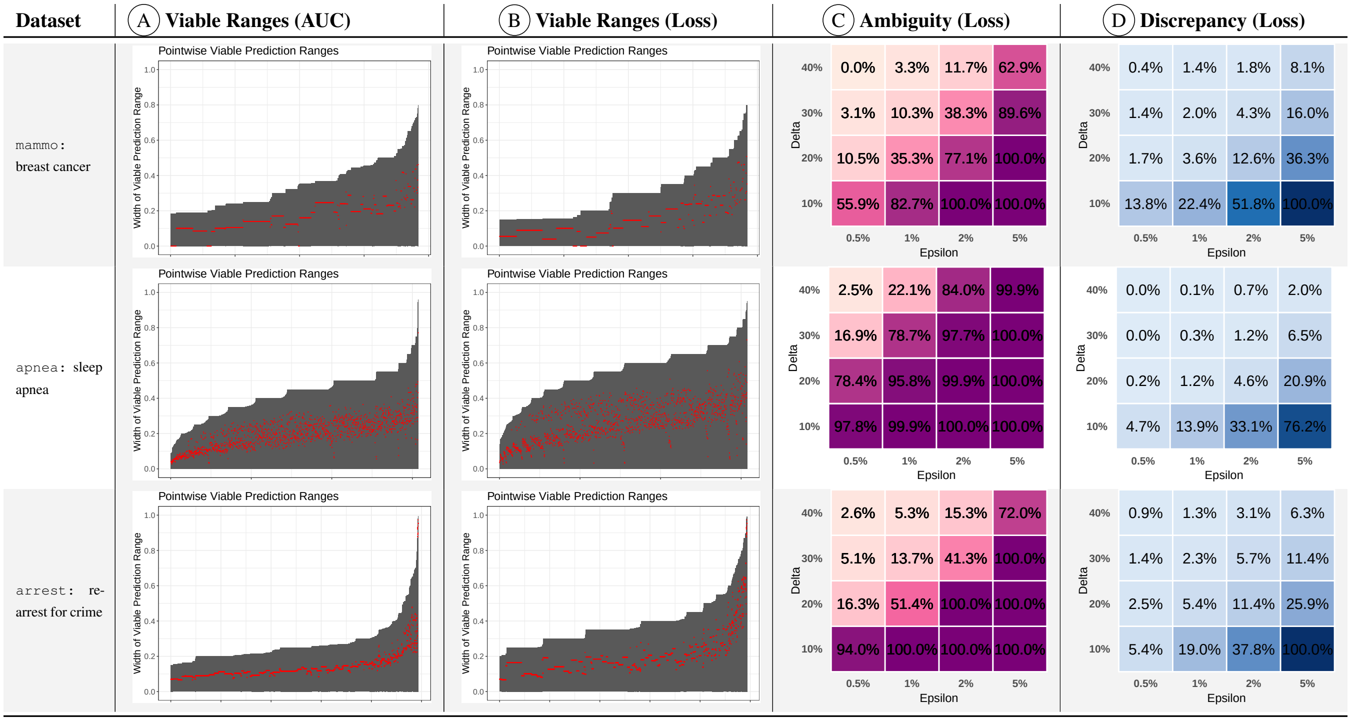


Figure 4: Predictive multiplicity in probabilistic classification on mammo, apnea and arrest. We show the distribution of viable prediction ranges ($|V_\epsilon(\boldsymbol{x}_i)|$ on the y-axis, the x-axis percentile rank and relative baseline estimates in red) for competing models with near-optimal training AUC **(A)** and training loss **(B)**. See illustration in Figure 2. We also show ambiguity **(C)** and discrepancy **(D)** for competing models with respect to training loss. We include similar results for the four other datasets in Appendix E.

biguous. That unique examples are more prone to ambiguity is related to our findings on outliers (see §4.1).

In terms of class imbalance, we find datasets with class imbalance skewed negative (adult, bank) often exhibit multiplicity on positive examples. In comparison, datasets that are roughly balanced by class (e.g., mammo, compas_arrest) have the same level of ambiguity for each class. This can be interpreted in light of the majority-minority effect from §4.1.

In terms of the baseline risk estimate, we see high ambiguity for examples with baseline risk near 50% on all datasets. For instance, all examples with baseline risk between 45% and 55% are ambiguous for the mammo dataset ($\epsilon = 0.5\%$ AUC, $\delta = 20\%$). There is no reason to believe that high ambiguity is less problematic for these samples. Rather, the importance of ambiguity will depend on the risk thresholds that drive decisions in a particular domain.

On the Disparate Impact of Multiplicity. Given the implications of risk prediction tasks, our results demonstrate how multiplicity can disproportionately impact individuals from historically marginalized subpopulations. For example, our results for predicting the risk of rearrest show that individuals who are ethnically Hispanic are disproportionately affected by predictive multiplicity: ambiguity is 39% for African Americans and 49% for Caucasians, compared to 98% for Hispanics ($\epsilon = 1\%$ and $\delta = 20\%$). Hence, reporting predictive multiplicity for subgroups can reveal important ethical considerations when testing risk assessment models deployed throughout society.

5 Concluding Remarks

In our focus on predictive multiplicity, we developed methods to evaluate the effect of slightly perturbing optimal model performance, revealing that similar models do not always assign similar predictions. We studied how competing models can assign conflicting predictions in probabilistic classification tasks. The proposed optimization-based methods compute our simple measures reliably. Compared to previous work, our methods allow for flexibility in choosing near-optimal metric and deviation threshold. Using synthetic data, we also present the first study providing insight into the kinds of data characteristics that give rise to predictive multiplicity.

As for future work, more research is needed to examine predictive multiplicity for other loss functions (our methods immediately generalize to convex loss functions) and to extend to other model classes. Also, it will be important to study how to effectively communicate these effects to practitioners and decision makers. Given predictive multiplicity metrics, practitioners can make better decisions in model selection while end-users can adjust their reliance on individual risk predictions. Concisely, analyzing predictive multiplicity promotes accountability and transparency in machine learning.

References

Ali, J.; Lahoti, P.; and Gummadi, K. P. 2021. *Accounting for Model Uncertainty in Algorithmic Discrimination*, volume 1. Association for Computing Machinery. ISBN 9781450384735.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias — ProPublica.

Attigeri, G. V.; Pai, M. M.; and Pai, R. M. 2017. Credit risk assessment using machine learning algorithms. *Advanced Science Letters*, 23(4): 3649–3653.

Austin, J.; Ocker, R.; and Bhati, A. 2010. Kentucky Pretrial Risk Assessment Instrument Validation. *The JFA Institute*, 5.

Bekhet, H. A.; and Eletter, S. F. K. 2014. Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. *Review of Development Finance*, 4(1): 20–28.

Bertsimas, D.; and King, A. 2017. Logistic regression: From art to science. *Statistical Science*, 367–384.

Bertsimas, D.; King, A.; Mazumder, R.; et al. 2016. Best subset selection via a modern optimization lens. *Annals of statistics*, 44(2): 813–852.

Black, E.; and Fredrikson, M. 2021. Leave-One-out Unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 285–295. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Black, E.; Leino, K.; and Fredrikson, M. 2022. Selective Ensembles for Consistent Predictions. In *International Conference on Learning Representations*.

Black, E.; Raghavan, M.; and Barocas, S. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 850–863. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Breiman, L. 2001. Statistical modeling: The two cultures. *Statistical Science*, 16(3): 199–215.

Chatfield, C. 1995. Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3): 419.

Chen, I. Y.; Joshi, S.; Ghassemi, M.; and Ranganath, R. 2021. Probabilistic machine learning for healthcare. *Annual Review of Biomedical Data Science*, 4: 393–415.

Christin, A.; Rosenblat, A.; and Boyd, D. 2015. Courts and predictive algorithms. *Data & civil rights: A new era of policing and justice*, 13.

Coston, A.; Rambachan, A.; and Chouldechova, A. 2021. Characterizing Fairness Over the Set of Good Models Under Selective Labels. *CoRR*, abs/2101.00352.

D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M. D.; Hormozdiari, F.; Houlsby, N.; Hou, S.; Jerfel, G.; Karthikesalingam, A.; Lucic, M.; Ma, Y.; McLean, C.; Mincu, D.; Mitani, A.; Montanari, A.; Nado, Z.; Natarajan, V.; Nielson, C.; Osborne, T. F.; Raman, R.; Ramasamy, K.; Sayres, R.; Schrouff, J.; Seneviratne, M.; Sequeira, S.; Suresh, H.; Veitch, V.; Vladymyrov, M.; Wang, X.; Webster, K.; Yadlowsky, S.; Yun, T.; Zhai, X.; and Sculley, D. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv*.

Diamond, S.; and Boyd, S. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17: 1–5.

Dong, J.; and Rudin, C. 2020. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12): 810–824.

Dusenberry, M. W.; Tran, D.; Choi, E.; Kemp, J.; Nixon, J.; Jerfel, G.; Heller, K.; and Dai, A. M. 2020. Analyzing the role of model uncertainty for electronic health records. *ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning*, 204–213.

Elter, M.; Schulz-Wendtland, R.; and Wittenberg, T. 2007. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Medical Physics*, 34(11): 4164–4172.

Fisher, A.; Rudin, C.; and Dominici, F. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(Vi).

Franc, V.; and Sonnenburg, S. 2008. Optimized cutting plane algorithm for support vector machines. In *Proceedings of the 25th International Conference on Machine Learning*, 320–327. ACM.

Hamid, K.; Asif, A.; Abbasi, W.; Sabih, D.; et al. 2017. Machine learning with abstention for automated liver disease diagnosis. In *2017 International Conference on Frontiers of Information Technology (FIT)*, 356–361. IEEE.

Hofman, J. M.; Goldstein, D. G.; and Hullman, J. 2020. How Visualizing Inferential Uncertainty Can Mislead Readers about Treatment Effects in Scientific Results. *Conference on Human Factors in Computing Systems - Proceedings*.

Joachims, T.; Finley, T.; and Yu, C.-N. J. 2009. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1): 27–59.

Joslyn, S.; and LeClerc, J. 2013. Decisions With Uncertainty: The Glass Half Full. *Current Directions in Psychological Science*, 22(4): 308–315.

Kale, A.; Kay, M.; and Hullman, J. 2020. Visual Reasoning Strategies for Effect Size Judgments and Decisions. *IEEE Transactions on Visualization and Computer Graphics*, 1–1.

Khand, A.; Frost, F.; Grainger, R.; Fisher, M.; Chew, P.; Mullen, L.; Patel, B.; Obeidat, M.; Albouaini, K.; Dodd, J.; Goldstein, S. A.; Newby, L. K.; Cyr, D. D.; Neely, M.; Lüscher, T. F.; Brown, E. B.; White, H. D.; Ohman, E. M.; Roe, M. T.; Hamm, C. W.; Six, A. J.; Backus, B. E.; and Kelder, J. C. 2017. Heart Score Value. *Netherlands Heart Journal*, 10(6): 1–10.

Kohavi, R. 1996. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In *KDD*, 202–207.

Kompa, B.; Snoek, J.; and Beam, A. L. 2021a. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1): 1–6.

Kompa, B.; Snoek, J.; and Beam, A. L. 2021b. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4(1).

Latessa, E. J.; Lemke, R.; Makarios, M.; Smith, P.; and Lowenkamp, C. T. 2010. The creation and validation of the ohio risk assessment system (ORAS). *Federal Probation*, 74(1): 16–22.

Lum, K.; Dunson, D. B.; and Johndrow, J. 2021. Closer than they appear: A Bayesian perspective on individual-level heterogeneity in risk assessment. *ArXiv*.

Mangasarian, O. L.; Street, W. N.; and Wolberg, W. H. 1995. Breast Cancer Diagnosis and Prognosis Via Linear Programming. *Operations Research*, 43(4): 570–577.

Marx, C.; Calmon, F. P.; and Ustun, B. 2019. Predictive multiplicity in classification.

McGrath, S.; Mehta, P.; Zytek, A.; Lage, I.; and Lakkaraju, H. 2020. When Does Uncertainty Matter?: Understanding the Impact of Predictive Uncertainty in ML Assisted Decision Making. *CoRR*, abs/2011.06167.

Moreno, R. P.; Metnitz, P. G.; Almeida, E.; Jordan, B.; Bauer, P.; Campos, R. A.; Iapichino, G.; Edbrooke, D.; Capuzzo, M.; and Le Gall, J. R. 2005. SAPS 3 - From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Medicine*, 31(10): 1345–1355.

Moro, S.; Cortez, P.; and Rita, P. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62: 22–31.

Mozannar, H.; and Sontag, D. 2020. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, 7076–7087. PMLR.

Naeini, M. P.; Cooper, G. F.; and Hauskrecht, M. 2015. Binary classifier calibration using a Bayesian non-parametric approach. *SIAM International Conference on Data Mining 2015, SDM 2015*, 208–216.

Pawelczyk, M.; Broelemann, K.; and Kasneci, G. 2020. On counterfactual explanations under predictive multiplicity. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence, UAI 2020*, 124: 839–848.

Romano, Y.; Barber, R. F.; Sabatti, C.; and Candès, E. 2020. With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 1–14.

Salvatier, J.; Wiecki, T. V.; and Fonnesbeck, C. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2016(4): 1–20.

Semenova, L.; Rudin, C.; and Parr, R. 2019. A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *ArXiv*, 1–64.

Shafer, G.; and Vovk, V. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9: 371–421.

Soyer, E.; and Hogarth, R. M. 2012. The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting*, 28(3): 695–711.

Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; et al. 2020. A deep learning approach to antibiotic discovery. *Cell*, 180(4): 688–702.

Than, M.; Flaws, D.; Sanders, S.; Doust, J.; Glasziou, P.; Kline, J.; Aldous, S.; Troughton, R.; Reid, C.; Parsonage, W. A.; Frampton, C.; Greenslade, J. H.; Deely, J. M.; Hess, E.; Sadiq, A. B.; Singleton, R.; Shopland, R.; Vercoe, L.; Woolhouse-Williams, M.; Ardagh, M.; Bossuyt, P.; Bannister, L.; and Cullen, L. 2014. Development and validation of the emergency department assessment of chest pain score and 2h accelerated diagnostic protocol. *EMA - Emergency Medicine Australasia*, 26(1): 34–44.

Ustun, B.; and Rudin, C. 2017. Optimized Risk Scores. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

Ustun, B.; Westover, M. B.; Rudin, C.; and Bianchi, M. T. 2016. Clinical prediction models for sleep apnea: The importance of medical history over symptoms. *Journal of Clinical Sleep Medicine*, 12(2): 161–168.

Veitch, V.; D'Amour, A.; Yadlowsky, S.; and Eisenstein, J. 2021. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*.

Yeh, I. C.; and Lien, C. h. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2 PART 1): 2473–2480.

Zeng, J.; Ustun, B.; and Rudin, C. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3): 689–722.

A MIP formulation for discrepancy ERM

To train a competing model that optimizes discrepancy, we solve a maximization problem of the form:

$$\max_{g \in \mathcal{H}_\epsilon(g_0)} \quad \sum_{i=1}^{n} d_i \tag{9}$$

Here, $d_i = \mathbb{1}[|g(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)| \leq \delta]$ can also be rewritten in terms of score $d_i = \mathbb{1}[s_w(\boldsymbol{x}_i) \geq \text{logit}(\delta + g_0(\boldsymbol{x}_i))] + \mathbb{1}[s_w(\boldsymbol{x}_i) \leq \text{logit}(g_0(\boldsymbol{x}_i) - \delta)]$. We recover the solution to (9) by solving the following integer program:

$$\begin{array}{lrll}
\max\limits_{\boldsymbol{w} \in \mathbb{R}^{d+1}} & \sum\limits_{i=0}^{n} d_i & & \\
\text{s.t.} & L(\boldsymbol{w}) \leq L(\boldsymbol{w}_0) + \epsilon & & \text{(10a)} \\
& d_i = v_{i,\delta} + z_{i,\delta} & i = 1, ..., n & \text{(10b)} \\
& M_{z,i}(1 - z_{i,\delta}) \geq (s_w(\boldsymbol{x}_i) - U_{i,\delta}) & i = 1, ..., n & \text{(10c)} \\
& M_{v,i}(1 - v_{i,\delta}) \geq -(s_w(\boldsymbol{x}_i) - B_{i,\delta}) & i = 1, ..., n & \text{(10d)} \\
& s_w(\boldsymbol{x}_i) = \sum\limits_{j=0}^{d} w_j x_{i,j} & i = 1, ..., n & \text{(10e)} \\
& d_i \in \{0, 1\} & i = 1, ..., n & \text{(10f)} \\
& z_{i,\delta} \in \{0, 1\} & i = 1, ..., n & \text{(10g)} \\
& v_{i,\delta} \in \{0, 1\} & i = 1, ..., n & \text{(10h)} \\
& w_j \in \mathbb{R} & j = 0, ..., d & \text{(10i)}
\end{array}$$

Here:

- $L(\boldsymbol{w}_0) := \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-\langle \boldsymbol{w}_0, y_i \boldsymbol{x}_i \rangle))$ is the log-loss of the baseline classifier on the training data
- $\epsilon \geq 0$ is the loss tolerance (i.e., the maximum additional loss of any competing classifier)
- $U_{i,\delta}$ is a parameter that we set as $U_{i,\delta} := \text{logit}(g_0(\boldsymbol{x}_i) - \delta)$
- $B_{i,\delta}$ is a parameter that we set as $B_{i,\delta} := \text{logit}(g_0(\boldsymbol{x}_i) + \delta)$
- $M_{z,i}$ is a Big-M parameter that we set as $M_{z,i} = -U_{i,\delta} + \max_{\boldsymbol{w}} \sum_{j=0}^{d} w_j x_{ij}$
- $M_{v,i}$ is a Big-M parameter that we set as $M_{v,i} = B_{i,\delta} - \min_{\boldsymbol{w}} \sum_{j=0}^{d} w_j x_{ij}$
- W^{max} and W^{min} are user-defined coefficient bounds

Big-M Derivations Recall that by definition,

$$g(\boldsymbol{x}_i) := \Pr(y_i = +1 | \boldsymbol{x}_i) = \frac{1}{1 + \exp(-\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)} \tag{11}$$

Therefore, $s_w(\boldsymbol{x}) = \text{logit}(g(\boldsymbol{x}_i))$. Our goal is to write the objective, $|g(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)| \geq \delta$, in terms of score, $s_w(\boldsymbol{x}_i)$.

$$\begin{aligned}
d_i &= \mathbb{1}[|g(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)| \geq \delta], \\
&= \mathbb{1}[g(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i) \geq \delta] + \mathbb{1}[g_0(\boldsymbol{x}_i) - g(\boldsymbol{x}_i) \geq \delta] \\
&= \mathbb{1}[g(\boldsymbol{x}_i) \geq \delta + g_0(\boldsymbol{x}_i)] + \mathbb{1}[-g(\boldsymbol{x}_i) \geq \delta - g_0(\boldsymbol{x}_i)] \\
&= \mathbb{1}[g(\boldsymbol{x}_i) \geq g_0(\boldsymbol{x}_i) + \delta] + \mathbb{1}[g(\boldsymbol{x}_i) \leq g_0(\boldsymbol{x}_i) - \delta]
\end{aligned}$$

Now we transform into score space

$$\begin{aligned}
&= \mathbb{1}[\text{logit}(g(\boldsymbol{x}_i)) \geq \text{logit}(g_0(\boldsymbol{x}_i) + \delta)] + \mathbb{1}[\text{logit}(g(\boldsymbol{x}_i)) \leq \text{logit}(g_0(\boldsymbol{x}_i) - \delta)] \\
&= \mathbb{1}[s_w(\boldsymbol{x}_i) \geq \text{logit}(g_0(\boldsymbol{x}_i) + \delta)] + \mathbb{1}[s_w(\boldsymbol{x}_i) \leq \text{logit}(g_0(\boldsymbol{x}_i) - \delta)]
\end{aligned}$$

Let $U_{i,\delta} = \text{logit}(g_0(\boldsymbol{x}_i) - \delta)$ and $B_{i,\delta} = \text{logit}(g_0(\boldsymbol{x}_i) + \delta)$.

$$\begin{aligned}
&= \mathbb{1}[s_w(\boldsymbol{x}_i) \geq B_{i,\delta}] + \mathbb{1}[s_w(\boldsymbol{x}_i) \leq U_{i,\delta}] \\
&= v_{i,\delta} + z_{i,\delta}
\end{aligned}$$

To ensure that $z_{i,\delta} = 1$ whenever $\mathbb{1}[s_w(\boldsymbol{x}_i) \leq U_{i,\delta}] = 1$, and $z_{i,\delta} = 0$ whenever $\mathbb{1}[s_w(\boldsymbol{x}_i) \leq U_{i,\delta}] = 0$, we add the following Big-M constraint:

$$M_{z,i}(1 - z_{i,\delta}) \geq s_w(\boldsymbol{x}_i) - U_{i,\delta}$$

Here we can set the Big-M parameter as:

$$\begin{aligned} M_{z,i} &= \max_{\boldsymbol{w}}(s_w(\boldsymbol{x}_i) - U_{i,\delta}), \\ &= -U_{i,\delta} + \max_{\boldsymbol{w}} s_w(\boldsymbol{x}_i), \\ &= -U_{i,\delta} + \max_{\boldsymbol{w}}\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle, \\ &= -U_{i,\delta} + \max_{\boldsymbol{w}} \sum_{j=0}^{d} w_j x_{ij} \\ &= -U_{i,\delta} + W^{\max} \sum_{j=0}^{d} x_{ij} \end{aligned}$$

Next, to ensure that $v_{i,\delta} = 1$ whenever $\mathbb{1}[s_w(\boldsymbol{x}_i) \geq B_{i,\delta}] = 1$, and that $v_{i,\delta} = 0$ whenever $\mathbb{1}[s_w(\boldsymbol{x}_i) \geq B_{i,\delta}] = 0$, we add the following Big-M constraint:

$$M_{v,i}(1 - v_{i,\delta}) \geq -(s_w(\boldsymbol{x}_i) - B_{i,\delta})$$

Here, we can set the Big-M parameter as:

$$\begin{aligned} M_{v,i} &= \max_{\boldsymbol{w}}(B_{i,\delta} - s_w(\boldsymbol{x}_i)), \\ &= B_{i,\delta} + \max_{\boldsymbol{w}} -s_w(\boldsymbol{x}_i), \\ &= B_{i,\delta} - \min_{\boldsymbol{w}} s_w(\boldsymbol{x}_i), \\ &= B_{i,\delta} - \min_{\boldsymbol{w}}\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle, \\ &= B_{i,\delta} - \min_{\boldsymbol{w}} \sum_{j=0}^{d} w_j x_{ij}, \\ &= B_{i,\delta} - W^{\min} \sum_{j=0}^{d} x_{ij} \end{aligned}$$

When performing experiments using CPLEX software, we set the MIP gap $= 0.0$ and the time limit to 600 seconds.

B Outer Approximation Algorithm

Loss Callback Formulation We let $L_\epsilon^{\max} := L^0 + \epsilon$. This allows us to write the loss constraint $L(\boldsymbol{w}) \leq L^0 + \epsilon$ as follows.

$$L(\boldsymbol{w}) \leq L_\epsilon^{\max} \tag{12}$$

$$L(\boldsymbol{w}) - L_\epsilon^{\max} \leq 0 \tag{13}$$

$$c(\boldsymbol{w}) \leq 0 \tag{14}$$

We will present an algorithm where we approximate $c(\cdot)$ by a linear approximation at a fixed point $\boldsymbol{w}^k \in \mathbb{R}^d$. The linear approximation has the form:

$$\hat{c}^k(\boldsymbol{w}) := c(\boldsymbol{w}^k) + \nabla L(\boldsymbol{w}^k)(\boldsymbol{w} - \boldsymbol{w}^k) \tag{15}$$

$$= c(\boldsymbol{w}^k) + \sum_{j=1}^{d} \nabla L(w_j^k)(w_j - w_j^k) \tag{16}$$

Recall that $L(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-\langle \boldsymbol{w}^k, y_i \boldsymbol{x}_i\rangle))$. The derivative evaluated at $\boldsymbol{w}^k$ is therefore,

$$\nabla_j L(w_j^k) = \frac{1}{n}\sum_{i=1}^{n} \nabla_j \log(1 + \exp(-\langle \boldsymbol{w}^k, y_i \boldsymbol{x}_i\rangle)) \tag{17}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left[\frac{1}{1 + \exp(-\langle \boldsymbol{w}^k, y_i \boldsymbol{x}_i\rangle)} * \exp(-\langle \boldsymbol{w}^k, y_i \boldsymbol{x}_i\rangle) * -y_i \boldsymbol{x}_i\right] \tag{18}$$

To perform the outer approximation, we add the following loss cut if $L(\boldsymbol{w}^k) - L_\epsilon^{\text{max}} > 0$

$$0 \geq L(\boldsymbol{w}^k) - L_\epsilon^{\text{max}} + \sum_{j=1}^{d} \nabla L(w_j^k)(w_j - w_j^k) \tag{19}$$

$$0 \geq L(\boldsymbol{w}^k) - L_\epsilon^{\text{max}} + \sum_{j=1}^{d} \nabla L(w_j^k) * w_j - \sum_{j=1}^{d} \nabla L(w_j^k) * w_j^k \tag{20}$$

$$-\sum_{j=1}^{d} \nabla L(w_j^k) * w_j \geq L(\boldsymbol{w}^k) - L_\epsilon^{\text{max}} - \sum_{j=1}^{d} \nabla L(w_j^k) * w_j^k \tag{21}$$

C Bayesian Comparison

To illustrate briefly how our approach compares with Bayesian methods, we compute the maximum loss in the 90% credible region, and set the multiplicity loss tolerance ϵ to that value. For a single example selected from the `breastcancer` dataset, we plot the viable prediction range and the posterior predictive for models within the loss tolerance. We see in Figure 5 that the viable prediction range is significantly wider than the Bayesian 90% credible region, highlighting a difference between the two frameworks.

For this demonstration, we consider a logistic regression model, along with normal priors for parameters; we adopt a Gaussian prior on parameters $\boldsymbol{w}_j$, and set the mean to 0. We assume weak information regarding the true values of the parameters by adopting a large variance, $\boldsymbol{w}_j \sim N(0, 10000)$ for each parameter. The likelihood is

$$\mathcal{L}(\boldsymbol{w}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i)^{y_i}(1 - p(\boldsymbol{x}_i))^{(1-y_i)}.$$

We use Metropolis-Hastings to sample from the posterior using the MAP as a starting point and with $50,000$ samples and two chains. After sampling, a trace object is returned that contains samples from the posterior distribution. We use the PyMC3 Python package (Salvatier, Wiecki, and Fonnesbeck 2016).

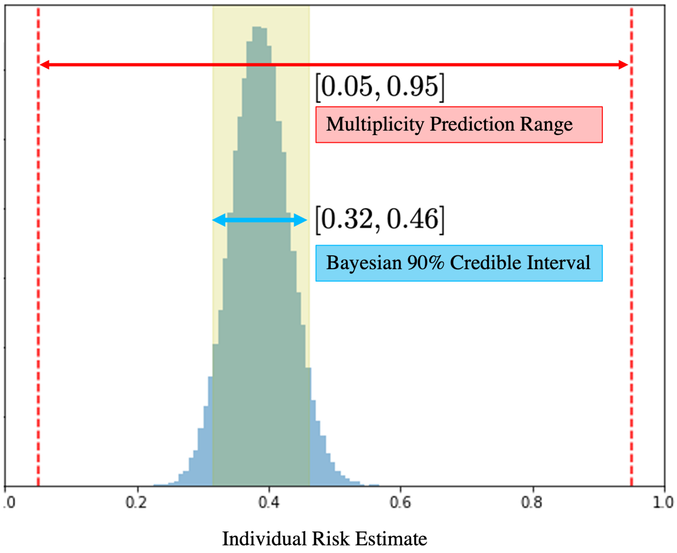


Figure 5: A comparison of a Bayesian 90% credible interval and the viable prediction range for a single example selected from a dataset. We perform this study on a small `breastcancer` dataset (Mangasarian, Street, and Wolberg 1995) for predicting whether patient breastcancer biopsy is malignant. Here, we compute the maximum loss in the 90% credible region, and set the multiplicity loss tolerance ϵ to that value. The viable prediction range is substantially wider than the Bayesian 90% credible region, illustrating the difference between the two frameworks.

D Datasets

Name	Outcome Variable	n	d	Class Imbalance	Train Loss	Train AUC	Train ECE
mammo (Elter, Schulz-Wendtland, and Wittenberg 2007)	mammogram shows breast cancer	961	12	0.86	0.471	85%	2.4%
credit (Yeh and Lien 2009)	customer default on loan	30,000	23	3.50	0.453	74%	1.6%
bank (Moro, Cortez, and Rita 2014)	person opens bank account after marketing call	41,188	57	0.12	0.268	82%	0.9%
adult (Kohavi 1996)	person in 1994 US census earns over $50,000	32,561	36	0.31	0.332	90%	0.8%
compas_arrest (Angwin et al. 2016)	rearrest for any crime	5,380	18	0.84	0.612	72%	1.1%
compas_violent (Angwin et al. 2016)	rearrest for violent crime	8,768	18	0.13	0.332	67%	0.3%
apnea (Ustun et al. 2016)	patient diagnosed with obstructive sleep apnea	1,537	36	0.70	0.565	76%	3.3%

Table 1: Publicly available datasets used to train risk assessment models in §4.2. For each dataset, we report n, d, the class imbalance ratio, $|n^+|/|n^-|$, and the performance metrics of the baseline model on training data. We work with sub-sampled versions of credit, bank and adult by randomly sampling $n = 5000$ points from each dataset.

D.1 Synthetic Datasets

To study the causes of predictive multiplicity, we generate small, synthetic datasets to conduct the studies shown in Figure 3. For this, we generate isotropic Gaussian blobs for clustering, using sklearn functions. For the separability study, we generate a dataset with $N = 200$ randomly generated samples with varying standard deviation. For the outliers study, we generate a dataset with $N = 320$ randomly generated samples with two clusters of outliers at varied positions. For the majority-minority structure study, we generate a dataset with variation in N_{majority} and N_{minority}, with a complete population of size $N = 150$ for each group, and this subsampled to lead to group imbalance.

E Additional Results

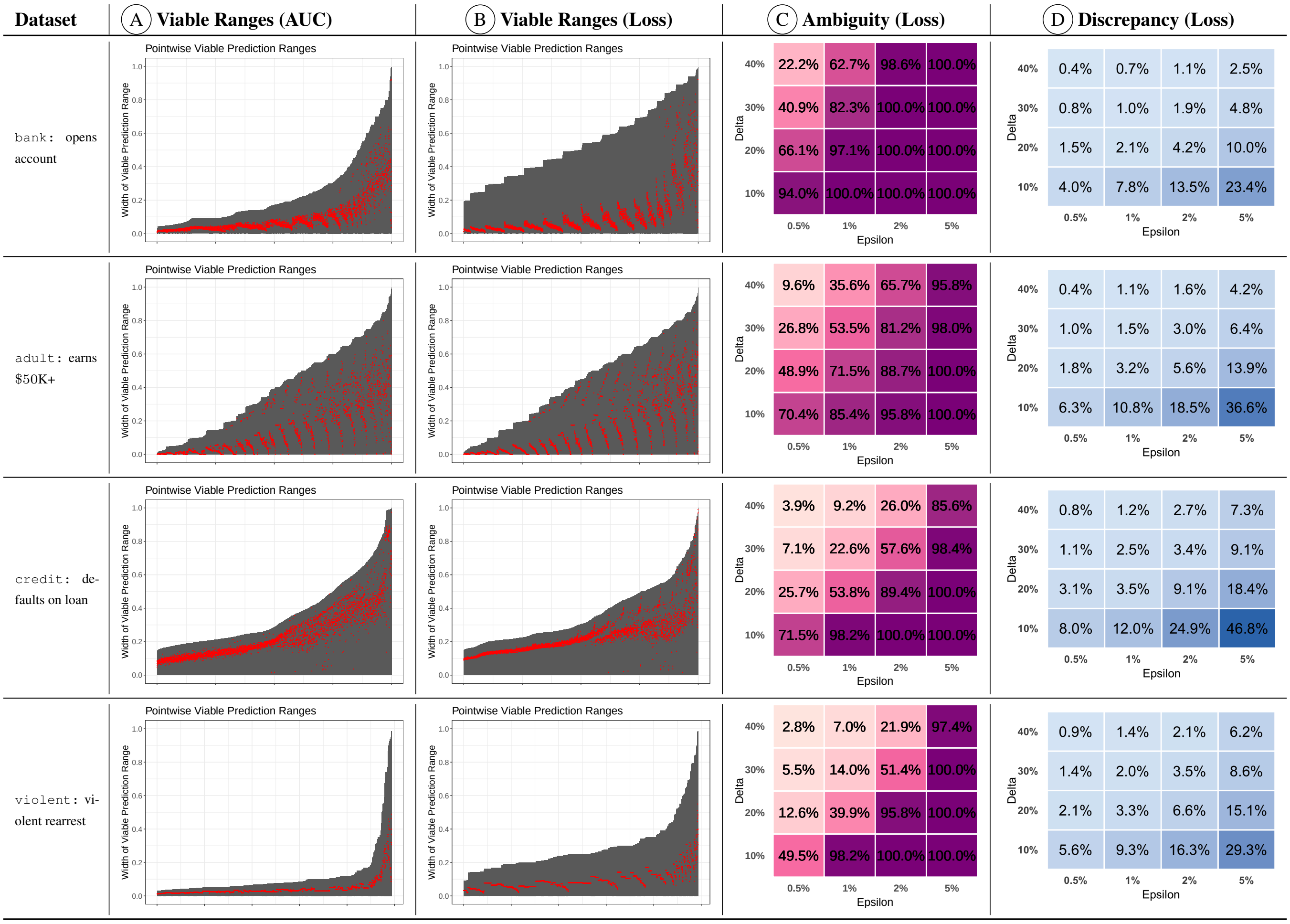
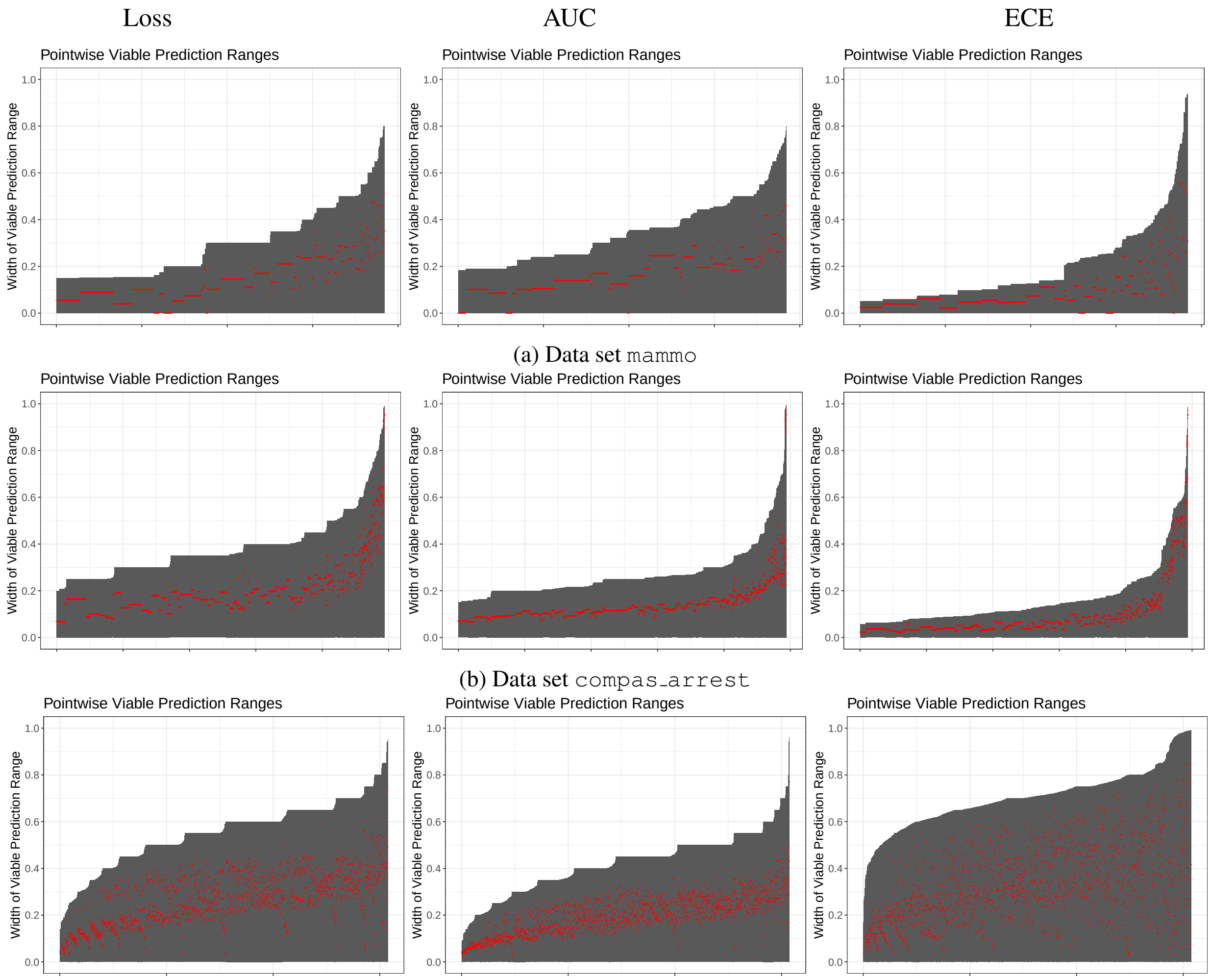


Figure 6: Predictive multiplicity in probabilistic classification on the additional datasets: `bank`, `adult`, `credit` and `violent`. We show the distribution of viable prediction ranges ($|V_\epsilon(\boldsymbol{x}_i)|$ on the y-axis, the x-axis percentile rank and relative baseline estimates in red) for competing models with near-optimal training AUC **(A)** and training loss **(B)**. See illustration in Figure 2. We also show ambiguity **(C)** and discrepancy **(D)** for competing models with respect to training loss.

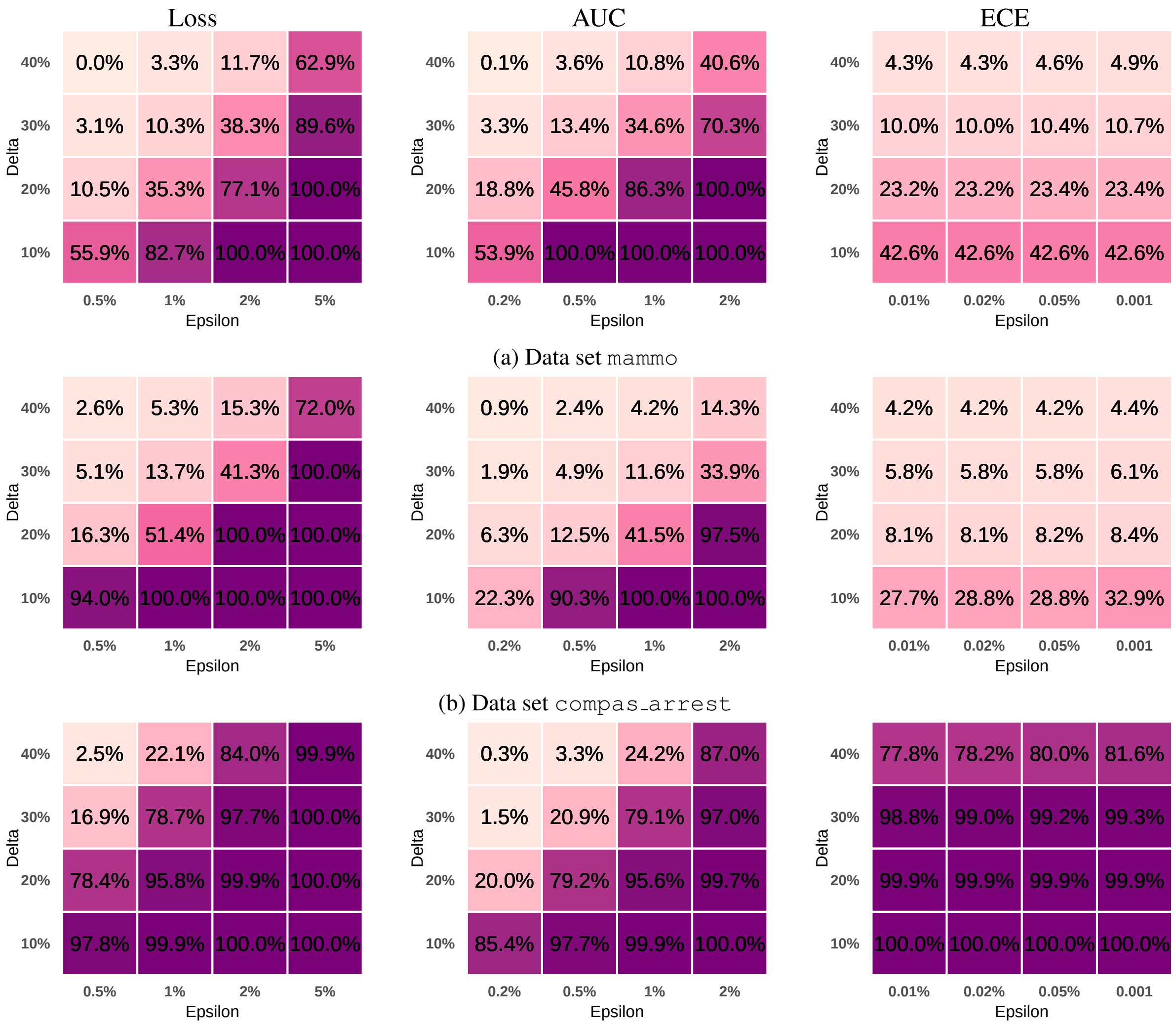


(a) Data set `mammo`

(b) Data set `compas_arrest`

(c) Data set `apnea`

Figure 7: Additional viable prediction ranges ($|V_\epsilon(\boldsymbol{x}_i)|$ for each near-optimal metric. Results shown for the datasets `mammo`, `compas_arrest`, `apnea`, and for ϵ-level sets defined on loss (1%), AUC (0.5%), and ECE (0.02%).



(a) Data set `mammo`

(b) Data set `compas_arrest`

(c) Data set `apnea`

Figure 8: Additional ambiguity heatmaps for each near-optimal metric. Results shown for the datasets `mammo`, `compas_arrest`, `apnea`, and for ϵ-level sets defined on loss (1%), AUC (0.5%), and ECE (0.02%).