Reliable Decisions with Threshold Calibration

Anonymous Author(s)
Affiliation
Address
email

Abstract

Decision makers rely on probabilistic forecasts to predict the loss of different decision rules before deployment. When the forecasted probabilities match the true frequencies, predicted losses will be accurate. Although perfect forecasts are typically impossible, probabilities can be calibrated to match the true frequencies on average. However, we find that this *average* notion of calibration, which is typically used in practice, does not necessarily guarantee accurate decision loss prediction. Specifically in the regression setting, the loss of threshold decisions, which are decisions based on whether the forecasted outcome falls above or below a cutoff, might not be predicted accurately. We propose a stronger notion of calibration called threshold calibration, which is exactly the condition required to ensure that decision loss is predicted accurately for threshold decisions. We provide an efficient algorithm which takes an uncalibrated forecaster as input and provably outputs a threshold-calibrated forecaster. Our procedure allows downstream decision makers to confidently estimate the loss of any threshold decision under any threshold loss function. Empirically, threshold calibration improves decision loss prediction without compromising on the quality of the decisions in two real-world settings: hospital scheduling decisions and resource allocation decisions.

1 Introduction

Decision makers need to understand the consequences of their decisions prior to making them. When decisions are based on predictions from a machine learning model, the decision loss – the loss incurred under a decision rule based on the predictions – summarizes the consequences of these decisions. As an example, suppose a machine learning practitioner develops a model to predict patient length-of-stay in the hospital [17, 3]. A hospital decides whether they have capacity to admit new patients based on the model's predictions of current patients' length-of-stay (e.g. for each current patient who is predicted to have a length-of-stay that is less than k days, the hospital schedules a new patient). Incorrect decisions due to the model's predictions cause the hospital to accrue costs from under-utilizing resources or overbooking procedures. The decision loss is an aggregation of the costs incurred from incorrect decisions. To determine whether a decision rule is safe to use, the hospital would like to have an accurate estimate of the decision loss under different choices of k and different costs associated with errors. This type of decision-making scenario occurs in many high-stakes settings such as designing interventions for adverse weather events [32, 9] and resource allocation decisions using economic estimates [16, 31].

Probabilistic predictions (probabilistic forecasts) can be used to estimate decision loss prior to deployment. In this work, we consider the regression setup, where a forecast is represented by a cumulative distribution function over the possible outcomes. If the forecasted probabilities of incorrect decisions match the true frequencies of these events, then the average decision loss can be accurately predicted from the forecasts. However, forecasted probabilities of incorrect decisions do not typically match the true ones, yielding inaccurate decision loss predictions. We refer to the

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

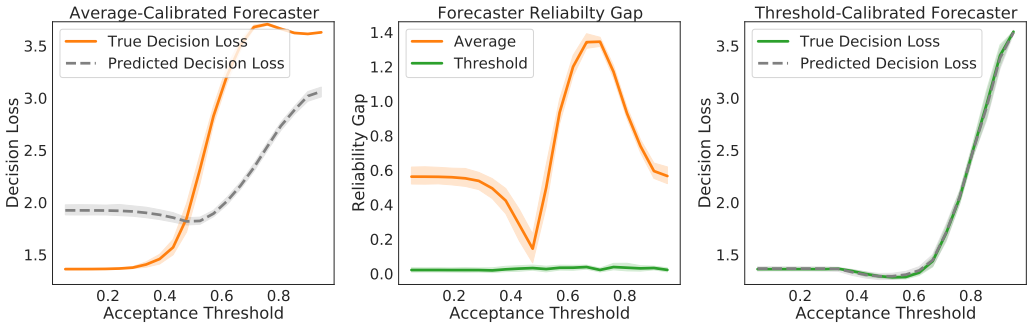

Figure 1: We evaluate average-calibrated and threshold-calibrated patient length-of-stay forecasters across a range of threshold decision rules. **Left**: The average-calibrated forecaster underestimates the true decision loss for some decision rules and overestimates it on others, resulting in a nonzero reliability gap. **Middle**: The reliability gap is minimized under the threshold-calibrated forecaster but not under the average-calibrated forecaster. **Right**: The threshold-calibrated forecaster accurately predicts the true decision loss across a range of decision rules.

absolute difference between the average loss predicted by forecaster and the true average loss as the **reliability gap**.

Many previous works in calibration and uncertainty quantification are motivated by the assumption that calibrated uncertainty estimates will yield safer or more reliable downstream decisions [30, 2, 21, 23, 24]. However, we find that the standard notion of calibration, average calibration, does not guarantee zero reliability gap for even a simple class of decision rules: threshold decision rules (left, Figure 1). In a threshold decision, a decision maker takes one of two possible actions depending on whether an outcome falls above or below a cutoff (e.g. the hospital schedules a new patient if a current patient's length-of-stay is less than 3 days, otherwise the hospital does not schedule a new patient). Stronger calibration properties, such as distribution calibration [30], are theoretically guaranteed to yield zero reliability gap but are difficult to achieve in practice. In particular, flexible distribution families can better approximate the true distribution than simple ones and yield lower decision loss, but applying distribution calibration to such forecasters can increase the decision loss and enlarge the reliability gap compared to the uncalibrated forecaster. Thus, existing calibration definitions are either insufficient or impractical for minimizing the reliability gap under threshold decision rules.

To address these shortcomings, we propose a new notion of calibration called **threshold calibration**. Threshold calibration strikes a balance between average and distribution calibration; it is exactly the condition required to guarantee zero reliability gap under threshold decisions and is practical to enforce (Figure 1, middle, right). First, we establish that threshold calibration is the necessary and sufficient condition to guarantee zero reliability gap for any threshold decision under any threshold loss. Second, we design an *efficient* algorithm that takes an uncalibrated forecaster as input and provably outputs a threshold-calibrated forecaster. Third, we show that empirically, threshold calibration is a *practical* solution; in two real-world settings and a suite of benchmark regression tasks, we find that threshold calibration minimizes the reliability gap across decision makers with different threshold loss functions while achieving similar or improved decision loss compared to the baselines.

2 Preliminaries

2.1 Notation and Forecasting Setup

We consider the regression setup with a feature space $\mathcal{X}$ and a label space $\mathcal{Y} \subset \mathbb{R}$. The input is a random variable $X \in \mathcal{X}$ and the label is a random variable $Y \in \mathcal{Y}$. We use capital letters to denote random variables X, Y and lower case letters to denote their values x, y.

Let $\mathcal{F}(\mathcal{Y})$ be the space of cumulative distribution functions (CDFs) over $\mathcal{Y}$. A forecaster $h : \mathcal{X} \to \mathcal{F}(\mathcal{Y})$ is a function that maps an input from the feature space to a CDF on $\mathcal{Y}$. In other words, given a fixed input $x \in \mathcal{X}$, the forecaster outputs the predicted CDF $h[x] \in \mathcal{F}(\mathcal{Y})$. Ideally, the forecaster aims to predict the CDF of Y given X.

To further clarify the notation, for a fixed input-label pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $h[x]$ is a CDF over the predicted label values and $h[x](y) \in [0, 1]$ is the value of the CDF $h[x]$ at the point y. We note that $h[X]$ is a random variable that takes values in $\mathcal{F}(\mathcal{Y})$ and $h[X](Y)$ is a random variable that takes values in $[0, 1]$.

Let $h^*[x]$ be the true conditional CDF of Y given x. We use $\sim$ to denote the distribution of a random variable. We have that $Y \sim h^*[x]$. We introduce a new random variable $\tilde{Y}$ to represent a label distributed according to the $h[x]$, the forecasted conditional distribution, so $\tilde{Y} \sim h[x]$.

2.2 Decision-Making

Let $\mathcal{A}$ be a countable action space. A decision rule $\delta : \mathcal{X} \to \mathcal{A}$ is any map from an input x (e.g. a current patient's attributes) to an action a (e.g. admit a new patient). We assume that a decision maker has a loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, describing the loss incurred when choosing an action a on an input-label pair (x, y). Because the labels y are unobserved, the decision maker often wants to minimize their expected loss assuming that the labels are distributed according to the forecasted distribution. As a result, they use the Bayes decision rule with respect to h.

Definition 1 (Bayes Decision Rule)**.** *Given a space of decision rules Δ, the Bayes decision rule δ^* with respect to the forecaster h is the decision rule in Δ that minimizes the expected loss under the forecasted distribution*

$$\delta_h^* = \arg\inf_{\delta \in \Delta} \mathbb{E}_X \mathbb{E}_{\tilde{Y} \sim h[x]}[\ell(X, \tilde{Y}, \delta(X))]]$$

2.3 Threshold Decisions

We focus on the setting where the decision maker aims to minimize a threshold loss function. The action space $\mathcal{A}$ consists of two actions so $\mathcal{A} = \{0, 1\}$. A threshold loss function ℓ is defined as follows

$$\ell(x, y, a) = \sum_{i \in \{0,1\}} c_{1,i} \mathbb{I}(y \leq y_0, a = i) + \sum_{i \in \{0,1\}} c_{0,i} \mathbb{I}(y > y_0, a = i),$$

where $c_{i,j} \in \mathbb{R}$. The $c_{i,j}$'s denote *decision costs*, costs associated with different outcomes, and y_0 is a *decision threshold*. Let $\mathcal{L}$ be the space of threshold loss functions, which are all losses of this form with any $c_{i,j} \in \mathbb{R}$ and $y_0 \in \mathbb{R}$.

Given a threshold loss function ℓ, the decision maker can use the Bayes decision rule δ_h^* in Definition 1 to select which action to take. We show that the resulting decision rules always take the form of

$$\delta_h^*(x) = \mathbb{I}(h[x](y_0) \geq \alpha) \text{ or } \delta_h^*(x) = \mathbb{I}(h[x](y_0) \leq \alpha)$$

for some parameters $\alpha \in [0, 1]$ and $y_0 \in \mathcal{Y}$ that depends on the loss function (proved in Appendix B). We call such decision rules **threshold decision rules** because intuitively, they choose the action based on whether $h[x](y_0)$ is greater (or less) than a threshold α. We denote the space of such decision rules as Δ_h. Since the decision maker's loss function is a threshold loss function, the decision maker can restrict the space of decision rules they consider to threshold decision rules on the forecasted CDFs.

3 Reliable Decision-Making with Threshold Calibration

3.1 Problem Setup

Forecasts are often produced by one group, such as machine learning practitioners or scientists, and consumed by another, such as policy makers or private agents [15]. Motivated by this paradigm, we model these two entities separately:

1. A forecaster h takes inputs $x \in \mathcal{X}$ and produces CDFs $h[x]$ over the possible outcomes in $\mathcal{Y}$. The provider of h does not know the specific downstream tasks for which h is used.
2. A decision maker with a dataset of unlabeled inputs $\mathcal{D} = \{x_i\}_{i=1}^n$, binary action space $\mathcal{A} = \{0, 1\}$, and a threshold loss function $\ell \in \mathcal{L}$ of interest. The decision maker must take an action $a_i \in \mathcal{A}$ for each unlabeled input x_i. The decision maker uses the forecaster h to select $\{a_i\}_{i=1}^n$ because (1) the decision maker does not have enough labeled data to build their own model locally or (2) building the model requires a domain expert.

Multiple decision makers may rely on the same forecaster but have different loss functions. Further, a decision maker's loss function can change if their decision costs or decision threshold change. In this setting, we identify the conditions on h that the provider can enforce to ensure reliable decision-making under threshold decisions.

3.2 Reliability Gap

Decision makers often need to accurately estimate the average decision loss incurred under a decision rule prior to deployment. To quantify the accuracy of these decision loss predictions, we define the reliability gap.

Definition 2 (Reliability Gap)**.** *Given a forecaster h, we define the the reliability gap $\gamma(\delta, \ell)$ of a particular decision rule δ under a loss function ℓ as*

$$\gamma(\delta, \ell) = |\mathbb{E}_X \mathbb{E}_{\tilde{Y} \sim h[x]}[\ell(X, \tilde{Y}, \delta(X))] - \mathbb{E}_X \mathbb{E}_{Y \sim h^*[x]}[\ell(X, Y, \delta(X))]|.$$

The first term in the equation is the average decision loss predicted by the forecaster. Under the forecasted distribution, the labels $\tilde{Y}$ are distributed according to $h[X]$. As a result, the first term does not depend on the true labels and can be computed by the decision maker using the unlabeled data prior to deployment. The second term is the true average decision loss. Under the true conditional distribution, the labels Y are distributed according to $h^*[X]$. So, the second term can be thought of as the loss that is incurred at test-time. One caveat is that the reliability gap quantifies the reliability of *average* decision loss prediction and obtaining zero reliability gap does not imply any instance-based guarantees for individual decisions.

When the forecaster perfectly matches the true distribution (i.e. $h = h^*$), we have $\gamma(\delta, \ell) = 0$ for any decision rule δ and any loss function ℓ. However, in practice, we cannot assume that the forecaster predicts the true distribution. In addition, we would like the forecaster to be applicable for different downstream decision makers. As a result, we study the necessary and sufficient conditions on the forecaster that guarantee zero reliability gap for any threshold decision on the forecasted CDFs under any threshold loss function.

3.3 Threshold Calibration

We define the property of threshold calibration and show that it is necessary and sufficient to ensure zero reliability gap under any threshold decisions on the forecasted CDFs and any threshold loss function. The lemma and theorem in this section are proven in Appendix B.

We define the property of threshold calibration below.

Definition 3 (Threshold Calibration)**.** *A forecaster h satisfies threshold calibration if*

$$\Pr[h[x](Y) \leq c \mid h[x](y_0) \leq \alpha] = c \quad \forall x \in \mathcal{X}, \forall y_0 \in \mathcal{Y}, \alpha \in [0,1], \forall c \in [0,1]. \tag{1}$$

In other words, when we condition on predicted CDFs which have α-quantile equal to y_0, we require that the true label y should be below the c-th quantile of the predicted CDF exactly c percent of the time. We require this to hold for any choice of $x \in \mathcal{X}, y_0 \in \mathcal{Y}, \alpha \in [0,1]$, and $c \in [0,1]$. We make the following observation about conditioning on the complementary predicted CDFs.

Lemma 1. *Given a forecaster h that satisfies Definition 3, then we have that $\forall x \in \mathcal{X}, \forall y_0 \in \mathcal{Y}, \alpha \in [0,1], \forall c \in [0,1], \Pr[h[x](Y) \leq c \mid h[x](y_0) > \alpha] = c$.*

Using Lemma 1 and Definition 3, we define the threshold calibration error (TCE) to measure deviation from threshold calibration at a threshold $y_0 \in \mathcal{Y}$ and quantile $\alpha \in [0,1]$.

Definition 4 (Threshold Calibration Error)**.**

$$\begin{aligned} \textit{TCE}(h, y_0, \alpha) = &\int_0^1 |\Pr[h[x](Y) \leq c \mid h[x](y_0) \leq \alpha] - c| \quad dc \\ &+ \int_0^1 |\Pr[h[x](Y) \leq c \mid h[x](y_0) > \alpha] - c| \quad dc. \end{aligned}$$

Threshold calibration is a specific type of group calibration [27], where calibration across the collection of groups $\mathcal{G} = \{(X, Y) \in \mathcal{X} \times \mathcal{Y} \mid h[x](y_0) \leq \alpha\}_{y_0 \in \mathcal{Y}, \alpha \in [0,1]}$ is desired. Since threshold calibration requires achieving calibration on intersecting groups, it is also related to the notion of multicalibration [18]. In Section 4, we give an efficient algorithm for achieving threshold calibration that is inspired by previous work on multicalibration.

Threshold calibration is a desirable property due to its connection to achieving zero reliability gap.

Theorem 1. *Let $\mathcal{L}$ be the space of threshold loss functions. Given a forecaster h, let Δ_h be the space of threshold decision rules on the forecasted CDFs of h. A forecaster h satisfies threshold calibration if and only if $\gamma(\delta, \ell) = 0 \quad \forall \delta \in \Delta_h, \forall \ell \in \mathcal{L}$.*

In other words, under a threshold-calibrated forecaster, we achieve zero reliability gap under any threshold decision on the forecasted CDFs and any threshold loss function.

3.4 Comparison to Existing Calibration Definitions

We compare threshold calibration to other methods for calibrating probabilistic forecasts. Average calibration is the standard definition of calibration for regression [22, 13].

Definition 5 (Average Calibration)**.** *A forecaster h satisfies average calibration if*

$$\Pr[h[x](Y) \leq c] = c \quad \forall x \in \mathcal{X}, \forall c \in [0, 1].$$

In other words, a forecaster is average-calibrated if the true label Y is below the c-th quantile of the forecasted CDF $h[x]$ exactly c percent of the times.

In contrast, distribution calibration is a much stronger definition of calibration [30]. Intuitively, distribution calibration requires a forecaster to be calibrated for every distribution in the forecaster's model family.

Definition 6 (Distribution Calibration)**.** *A forecaster h satisfies distribution calibration if*

$$\Pr[h[x](Y) \leq c \mid h[x] = g] = c \quad \forall x \in \mathcal{X}, \forall g \in \mathcal{F}(\mathcal{Y}),$$

where $\mathcal{F}$ is space of CDFs corresponding to the forecaster's model family.

We outline the relationship between average, threshold, and distribution calibration in the following proposition.

Proposition 1. *If a forecaster satisfies distribution calibration, then it satisfies threshold calibration. If a forecaster satisfies threshold calibration, then it satisfies average calibration.*

We note that the converses of the statements in Proposition 1 are not necessarily true. A threshold-calibrated forecaster does not necessarily satisfy distribution calibration. An average-calibrated forecaster does not necessarily satisfy threshold calibration or distribution calibration. This implies that an average-calibrated forecaster does not satisfy the necessary condition of Theorem 1, meaning that the reliability gap under threshold decisions may not be zero. So, decision makers who rely on a forecaster that only satisfies average calibration (but not threshold calibration) are not guaranteed to accurately estimate their decision loss under threshold decisions.

From Proposition 1, we have that a distribution-calibrated forecaster satisfies the necessary condition of Theorem 1. However, distribution calibration can be challenging to achieve in practice because the same CDF is rarely predicted more than one time on the training samples, making it difficult to guarantee calibration without compromising the *sharpness* of the forecasts. Sharpness corresponds to the width of the prediction intervals generated from the forecasts, and sharp forecasts yield short prediction intervals. Although distribution calibration is theoretically guaranteed to yield zero reliability gap, we observe that achieving distribution calibration is challenging when the model family is complex (Section 5).

Finally, we emphasize the threshold calibration is exactly the condition needed to guarantee the reliability gap is zero in Theorem 1.

4 Achieving Threshold Calibration

We design a recalibration algorithm that takes an uncalibrated forecaster as input and provably outputs a threshold-calibrated forecaster. Our algorithm is an iterative procedure that terminates when the

maximum TCE is less than a user specified threshold ϵ. Our key result is that the algorithm must terminate after $O(1/\epsilon^2)$ iterations.

Pseudo-code for the algorithm is shown in Algorithm 1. Intuitively, at each iteration of the algorithm, we find the y_0^t and α^t where the TCE in Definition 4 is maximized. This partitions the input $\mathcal{X}$ into two parts: those where $h[x](y_0^t) \leq \alpha^t$ and those where $h[x](y_0^t) > \alpha^t$. For each partition, we use a standard recalibration algorithm (Isotonic regression [22]) to achieve average calibration. Intuitively, after the recalibration step, the forecaster should satisfy average calibration for each partition, and hence the TCE in Definition 4 must be (close to) 0 for y_0^t and α^t. We repeat this procedure until the TCE is less than ϵ for every possible y_0 and α.

Algorithm 1: Threshold Recalibration

Input: Forecaster $h : \mathcal{X} \to \mathcal{F}(\mathcal{Y})$, maximum error $\epsilon > 0$
Output: A threshold-calibrated forecaster
Set $h^0 \leftarrow h$
for $t = 1, 2, \cdots$ *until maximum threshold calibration error* $\sup_{y_0,\alpha}$ *TCE*$(h^{t-1}, y_0, \alpha) \leq \epsilon$ **do**
  Find the y_0 and α that maximize threshold calibration error.
  $y_0^t, \alpha^t \leftarrow \arg\sup_{(y_0,\alpha)\in\mathcal{Y}\times[0,1]} \text{TCE}(h^{t-1}, y_0, \alpha)$
  Partition input features $\mathcal{X}$ into $\mathcal{X}_0 \leftarrow \{x \in \mathcal{X} \mid h^{t-1}[x][y_0^t] \leq \alpha^t\}$ and $\mathcal{X}_1 = \mathcal{X} \setminus \mathcal{X}_0$.
  Use Isotonic regression to learn recalibration maps $\phi_0^t, \phi_1^t : \mathcal{F}(\mathcal{Y}) \to \mathcal{F}(\mathcal{Y})$ on $\mathcal{X}_0$ and $\mathcal{X}_1$ respectively.
  Apply the recalibration map to obtain new prediction functions.
  $h^t[x] \leftarrow \begin{cases} \phi_0^t(h^{t-1}[x]) & \text{if } x \in \mathcal{X}_0 \\ \phi_1^t(h^{t-1}[x]) & \text{otherwise} \end{cases}$
end
return h^T where T is the final iteration count.

In practice, the TCE and the recalibration maps are estimated from samples. We describe the finite sample version of the algorithm in Appendix A.

The following theorem shows that our iterative threshold calibration procedure converges in a small number of iterations. The intuition of the proof is that after each iteration, the L2 distance between the prediction functions h and the true CDF h^* must decrease by at least ϵ^2. Therefore, the algorithm must terminate before the L2 distance decreases below 0 (which is impossible). A full proof is provided in Appendix B.

Theorem 2. *Algorithm 1 converges after at most $O(1/\epsilon^2)$ iterations and outputs a forecaster with threshold calibration error at most ϵ.*

5 Experiments

In the following experiments, we demonstrate that threshold calibration can minimize the reliability gap (1) across a range of decision costs, (2) across a range of decision thresholds, and (3) in simple and complex model families. Across all datasets and forecaster model families that we consider, we find that threshold calibration outperforms the baselines in reducing the size of the reliability gap while attaining similar or improved decision loss compared to the baselines.

5.1 Datasets

We consider datasets that relate to real-world decision-making tasks and standard benchmarks. In the main paper, we show results on the UCI Protein and the MIMIC-III datasets. All remaining results can be found in Appendix A.

MIMIC-III. Patient length-of-stay predictions are used for hospital scheduling and resource management [17]. We consider a patient length-of-stay forecaster trained on patient admission laboratory values from the MIMIC-III dataset [19]. In our decision task, the hospital decides to schedule a new patient for an elective procedure if a current patient is predicted to have a short length of stay.

Demographic and Health Survey (DHS). Local wealth measurements are used to inform resource allocation decisions. We use the DHS data from Sheehan et al. [29] to predict asset wealth from satellite images as done in Yeh et al. [31] and Sheehan et al. [29]. Our experimental setup is motivated by the decision task defined in Yeh et al. [31], where aid is allocated to regions where the predicted asset wealth falls below a particular threshold.

UCI Regression Datasets. We evaluate on a suite of UCI regression datasets (Naval, Protein, Energy, Crime) [12]. They are common benchmarks in the uncertainty quantification literature [30, 2, 8, 22].

5.2 Experimental Setup and Baselines

Experimental Setup. We consider a forecaster that outputs Gaussian distributions and a forecaster that outputs Gaussian-Laplace mixture distributions. We use a train/validation/test split. The uncalibrated forecaster is a neural network trained on the training set with the validation set used for early stopping. For large datasets (Protein, Energy, Naval, MIMIC-III), the recalibration transform is trained on the validation set. For small datasets (Crime, DHS), the recalibration transform is trained on the training and validation set. On the test set, we evaluate our method and the baselines on the decision-making metrics (Section 5.3). Calibration metrics are also measured and results are provided in Appendix A.

Baselines. We compare the uncalibrated forecaster to the forecaster after enforcing average, threshold, or distribution calibration through a posthoc recalibration procedure. Methods for achieving these properties are described in Appendix A.

5.3 Decision-Making Metrics

We simulate decision makers enumerated $i = 1, 2 \ldots M$ who use a probabilistic forecaster h for their threshold decision tasks. We assume that there is no cost associated with true positives or true negatives, and the total cost of a false positive plus a false negative is equal to 10 for all decision makers. As a result, decision maker i's task is determined by a decision threshold y_0^i and decision cost ratio c_i. Each decision maker has a loss function $\ell^i(x, y, a) = 10c_i\mathbb{I}(a = 1, y \geq y_0^i) + 10(1 - c_i)\mathbb{I}(a = 0, y < y_0^i)$ and a decision rule $\delta^i_{h,\alpha}(x) = \mathbf{1}(h[x](y_0^i) \geq \alpha)$. We consider decision makers with $(y_0^i, c_i) \in \mathcal{Y}_0 \times \mathcal{C}$ where $\mathcal{Y}_0$ and $\mathcal{C}$ each consist of 50 uniformly spaced points that span the label space and [0.05, 0.95], respectively.

Let $\mathcal{C}$ be the set of possible choices for α. For each decision maker i, we compute the decision loss (the loss incurred by the Bayes decision rule $\delta^i_{h,\alpha_*}(X)$) and the reliability gap (averaged over the possible threshold decision rules).

$$\text{Decision Loss} = \mathbb{E}_X \mathbb{E}_{Y \sim h^*[x]}[\ell^i(X, Y, \delta^i_{h,\alpha^*}(X))] \quad \text{Reliability Gap} = \frac{1}{|\mathcal{C}|} \sum_{\alpha \in \mathcal{C}} |\gamma(\delta^i_{h,\alpha}, \ell^i)|.$$

Aggregate statistics can be obtained by averaging over all k decision makers, all decision makers who share the same threshold y_0, or all decision makers who share the same cost ratio c.

5.4 Results

Using the MIMIC-III and UCI Protein datasets, we study the effect of recalibration on the reliability gap and the decision loss achieved by decision makers with different decision thresholds and cost ratios. Furthermore, we examine the effect of recalibration on forecasters that output CDFs from simple (Gaussian) and complex (Gaussian-Laplace mixture) model families.

Threshold Calibration Minimizes Reliability Gap Across Decision Thresholds. We evaluate the effect of recalibrating the Gaussian forecaster on decision makers with different decision thresholds. On both datasets, threshold calibration yields the largest decrease in the reliability gap (left plots, Figure 2). Distribution calibration also decreases the reliability gap across decision thresholds, relative to the uncalibrated Gaussian forecaster. Average calibration does not consistently reduce the reliability gap; on the UCI Protein dataset, the reliability gap of the average-calibrated forecaster enjoys a slight decrease at some decision thresholds but is increased at others, relative to the uncalibrated Gaussian forecaster (middle left, Figure 2). Lastly, these calibration methods achieve similar decision loss to the uncalibrated forecaster (right plots, Figure 2). These trends are consistent

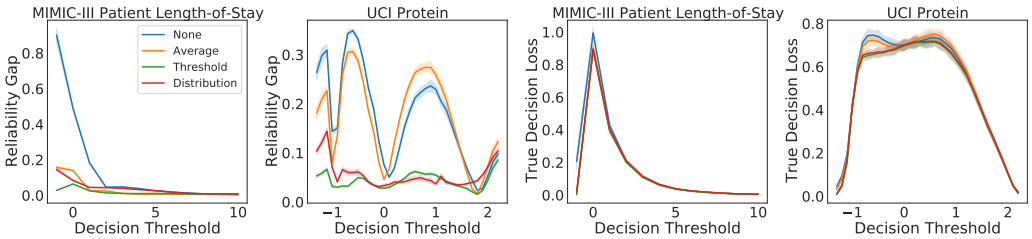

Figure 2: Under the Gaussian forecaster and across different decision thresholds, threshold calibration reduces the reliability gap on both datasets while average calibration does not reduce the reliability gap on the Protein dataset (**Left, Middle Left**), and all calibration methods yield improved or comparable decision loss compared to the uncalibrated forecaster (**Middle Right, Right**). Error bars represent 95% confidence intervals and are generated over 6 random trials.

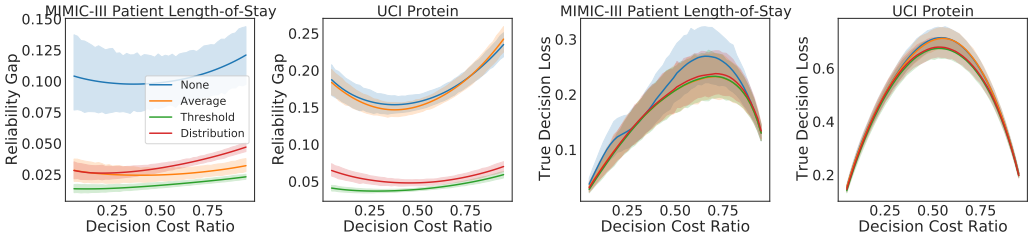

Figure 3: Under the Gaussian forecaster and across different decision cost ratios, threshold calibration reduces the reliability gap on both datasets while average calibration does not reduce the reliability gap on the Protein dataset (**Left, Middle Left**), and all calibration methods yield improved or comparable decision loss compared to the uncalibrated forecaster (**Middle Right, Right**). Error bars represent 95% confidence intervals and are generated over 6 random trials.

with the results obtained on the other datasets under the Gaussian forecaster. Threshold calibration outperforms baselines across different decision thresholds under the Gaussian-Laplace forecaster, as well (Appendix A).

Threshold Calibration Minimizes Reliability Gap Across Decision Cost Ratios. Across decision makers with different cost ratios, distribution and threshold calibration reduce the reliability gap relative to the uncalibrated forecaster, with threshold calibration yielding the largest decreases in the reliability gap (left plots, Figure 3). Meanwhile, average calibration does not consistently reduce the reliability gap; on the UCI Protein dataset, it achieves similar reliability gap to the uncalibrated forecaster (middle left, Figure 3). As before, these calibration methods achieve similar decision loss to the uncalibrated forecaster (right plots, Figure 3). These trends are consistent with results obtained on the other datasets under the Gaussian forecaster. Threshold calibration outperforms baselines across different decision cost ratios under the Gaussian-Laplace forecaster, as well (Figure 4).

Distribution Calibration Degrades Performance under Complex Model Families. Forecasters that can output CDFs from more flexible model families (e.g. Gaussian-Laplace mixture distributions) may be able to better capture the true conditional distribution of Y given x compared to Gaussian forecasters. As a result, we examine the effect of the recalibration procedures when the uncalibrated forecasts follow a more flexible distribution. The uncalibrated Gaussian-Laplace forecaster (Figure 4) yields a smaller reliability gap and smaller decision loss compared to the uncalibrated Gaussian forecaster (Figure 3). Applying threshold calibration to the Gaussian-Laplace forecaster further reduces the reliability gap. However, under the Gaussian-Laplace forecaster, distribution calibration enlarges the size of the reliability gap and increases the decision loss. Although distribution calibration is theoretically guaranteed to minimize the reliability gap, it is challenging to achieve in finite samples without compromising the *sharpness* of the forecasts (in our case, decision loss). So, we find that decision loss and reliability gap increase. We hypothesize that the recalibration dataset may not contain many instances that yield similar distribution parameters, so the recalibration transform does

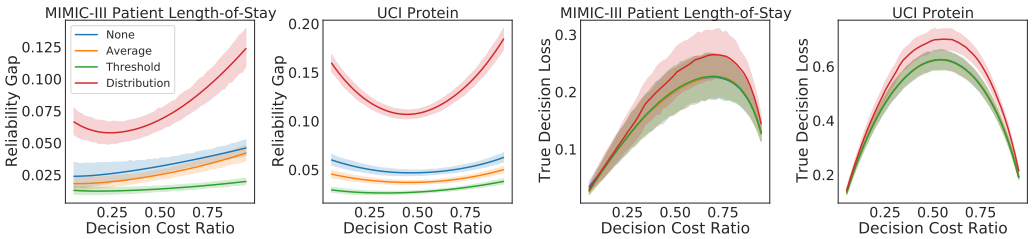

Figure 4: We consider the effect of recalibrating the Gaussian-Laplace forecaster under a range of decision cost ratios. Threshold calibration reduces the reliability gap while distribution calibration can enlarge the reliability gap (**Left, Middle Left**). Average and threshold calibration achieve comparable or lower decision loss as the baseline forecaster, while distribution calibration increases the decision loss. Error bars represent 95% confidence intervals and are generated over 6 random trials.

not generalize well to unseen data. We also observe these trends the UCI Crime, UCI Energy, and DHS datasets.

6 Related Work

Forecasting and Decision Making. The connection between forecasts and decision making was first studied in economics [1, 28]. The development of Bayesian decision analysis connected topics of forecasts and decision-based loss functions [11, 4]. Decision-making under uncertainty with probabilistic forecasts was then studied in econometrics [7].

Calibration. Calibration definitions have been studied in the statistics literature [5, 10, 6, 25]. For the regression setting, methods for ensuring that machine learning models satisfy average calibration have been studied in [22, 8]. In addition, methods for enforcing achieving stronger calibration notions have also been introduced such as distribution calibration [30] and individual calibration [33].

Multicalibration. Our definition of threshold calibration is most related to the line of work on multicalibration [18, 20]. Given a large collection $\mathcal{G}$ of potentially intersecting groups of the data, a predictor is multicalibrated on $\mathcal{G}$ if it is simultaneously calibrated on every sufficiently large group in $\mathcal{G}$ [18]. Previous works give methods for achieving mean and moment multicalibration for predictor functions. Our iterative procedure for achieving threshold calibration is inspired by methods for achieving multicalibration.

7 Limitations and Societal Impact

Our work demonstrates that certain types of calibration enable decision makers to estimate decision loss before deployment, which should not be confused with enabling decision makers to make optimal decisions. Also, our work assumes that predictions of Y do not affect the true label Y. However, when predictions are used to make decisions, they can often influence the outcome they aim to predict [26]. Our work does not account for these performative effects, so the decision loss may not be accurately estimated in these settings. Future work could focus on developing calibration procedures that enable forecasters to be robust to such distribution shifts. In addition, we specifically focus on binary-action threshold decisions. Future work may generalize our results to the setting where decision makers have loss functions involving multiple thresholds and multiple actions.

8 Conclusion

We show that a threshold-calibrated forecaster theoretically guarantees accurate decision loss estimation under threshold decision losses and threshold decision rules. We provide an iterative procedure for achieving threshold calibration and show that in practice it minimizes the reliability gap relative to baselines without compromising the forecaster's decision loss. Such estimates permit decision makers to reason about the consequences of their decisions prior to deployment.

References

[1] H. theil. economic forecasts and policy. assisted by j.s. cramer, h. moerman, a. russchen. contributions to economic analysis, nr xv. amsterdam, north-holland publishing company, 1958, xxxi p. 562 p., fl. 50—. *Bulletin de l'Institut de recherches économiques et sociales*, 25(2): 169–169, 1959. doi: 10.1017/S1373971900078951.

[2] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14927–14937. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/aab085461de182608ee9f607f3f7d18f-Paper.pdf.

[3] S. Barnes, Eric Hamrock, Matthew F. Toerper, S. Siddiqui, and S. Levin. Real-time prediction of inpatient length of stay for discharge prioritization. *Journal of the American Medical Informatics Association : JAMIA*, 23 e1:e2–e10, 2016.

[4] James O. Berger and James O. Berger. *Statistical decision theory and Bayesian analysis*. Springer-Verlag, New York, 1985. ISBN 0387960988 9780387960982 3540960988 9783540960980. URL http://www.amazon.com/Statistical-Decision-Bayesian-Analysis-Statistics/dp/0387960988/ref=sr_1_11?ie=UTF8&qid=1403880466&sr=8-11&keywords=Bayesian+statistics.

[5] GLENN W. BRIER. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

[6] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, USA, 2006. ISBN 0521841089.

[7] Gary Chamberlain. Econometrics and decision theory. *Journal of Econometrics*, 95:255–283, 2000.

[8] Peng Cui, Wenbo Hu, and Jun Zhu. Calibrated reliable regression using maximum mean discrepancy. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17164–17175. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/c74c4bf0dad9cbae3d80faa054b7d8ca-Paper.pdf.

[9] Murray Dale, Jon Wicks, Ken Mylne, Florian Pappenberger, Stefan Laeger, and Steve Taylor. Probabilistic flood forecasting and decision-making: An innovative risk-based approach. *Natural Hazards*, 70, 11 2014. doi: 10.1007/s11069-012-0483-z.

[10] A. Dawid. Present position and potential developments: some personal views. 1984.

[11] Morris H. DeGroot. *Optimal statistical decisions*. McGraw-Hill, New York, NY [u.a], 1970. ISBN 0070162425. URL http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+021834997&sourceid=fbw_bibsonomy.

[12] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[13] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, 69(2):243–268, 2007. URL https://EconPapers.repec.org/RePEc:bla:jorssb:v:69:y:2007:i:2:p:243-268.

[14] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[15] Clive W.J. Granger and Mark J. Machina. Forecasting and decision theory. In G. Elliott, C. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1 of *Handbook of Economic Forecasting*, chapter 2, pages 81–98. Elsevier, 2006. URL https://ideas.repec.org/h/eee/ecofch/1-02.html.

[16] Margaret Grosh, Carlo del Ninno, Emil Tesliuc, and Azedine Ouerghi. *For Protection and Promotion: The Design and Implementation of Effective Safety Nets*. The World Bank, 2008. URL https://EconPapers.repec.org/RePEc:wbk:wbpubs:6582.

[17] H. Harutyunyan, Hrant Khachatrian, David C. Kale, and A. Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6, 2019.

[18] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/hebert-johnson18a.html.

[19] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, 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*, 3:160035, 2016.

[20] Christopher Jung, Changhwa Lee, Mallesh M. Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation, 2020.

[21] Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18237–18248. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d3d9446802a44259755d38e6d163e820-Paper.pdf.

[22] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2801–2809. PMLR, 2018. URL http://proceedings.mlr.press/v80/kuleshov18a.html.

[23] Meelis Kull, Telmo de Menezes e Silva Filho, and Peter A. Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 623–631. PMLR, 2017. URL http://proceedings.mlr.press/v54/kull17a.html.

[24] Ali Malik, Volodymyr Kuleshov, Jiaming Song, Danny Nemer, Harlan Seymour, and Stefano Ermon. Calibrated model-based deep reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4314–4323. PMLR, 2019. URL http://proceedings.mlr.press/v97/malik19a.html.

[25] Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595 – 600, 1973. doi: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/apme/12/4/1520-0450_1973_012_0595_anvpot_2_0_co_2.xml.

[26] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/perdomo20a.html.

[27] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf.

[28] Robert W. Rudd. Theil, henri, applied economic forecasting, chicago, rand mcnally . . . company, 1966, xxv + 474 pp. ($14.00). *American Journal of Agricultural Economics*, 49(1_Part_I):241–243, 1967. doi: https://doi.org/10.2307/1237096. URL https://onlinelibrary.wiley.com/doi/abs/10.2307/1237096.

[29] Evan Sheehan, Chenlin Meng, Matthew Tan, Burak Uzkent, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2698–2706, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330784. URL https://doi.org/10.1145/3292500.3330784.

[30] Hao Song, Tom Diethe, Meelis Kull, and Peter A. Flach. Distribution calibration for regression. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5897–5906. PMLR, 2019. URL http://proceedings.mlr.press/v97/song19a.html.

[31] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications*, 11(1), 5 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-16185-w. URL https://www.nature.com/articles/s41467-020-16185-w.

[32] Weiran Yuchi, Jiayun Yao, Kathleen E. McLean, Roland Stull, Radenko Pavlovic, Didier Davignon, Michael D. Moran, and Sarah B. Henderson. Blending forest fire smoke forecasts with observed data can improve their utility for public health applications. *Atmospheric Environment*, 145:308–317, 2016. ISSN 1352-2310. doi: https://doi.org/10.1016/j.atmosenv.2016.09.049. URL https://www.sciencedirect.com/science/article/pii/S1352231016307592.

[33] Shengjia Zhao, Tengyu Ma, and Stefano Ermon. Individual calibration with randomized forecasting. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11387–11397. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/zhao20e.html.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...

(a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

(b) Did you describe the limitations of your work? [Yes] See Section 7.

(c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 7.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes]

(b) Did you include complete proofs of all theoretical results? [Yes] See Appendix B.

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix A

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 5 and Appendix A

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix A.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] See Appendix A.

(b) Did you mention the license of the assets? [Yes] See Appendix A.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Appendix A.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We did not directly obtain data from anyone but we used publicly available datasets.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The creators of the MIMIC-III dataset, which we use, deanonymized the data to remove any personally identifiable information. To the best of our knowledge, there is no other potential source of personally identifiable or offensive content in the data we use.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not do human subject research.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not do human subject research.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not do human subject research.