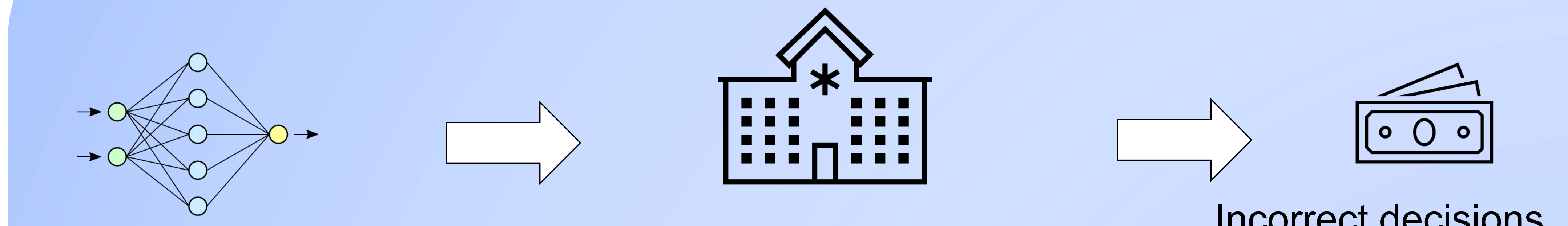
Reliable Decisions with Threshold Calibration

Roshni Sahoo[1], Shengjia Zhao[1], Alyssa Chen[2], Stefano Ermon[1]

[1]rsahoo, sjzhao, ermon @ cs.stanford.edu, [2]alyssa.chen@utsouthwestern.edu

Example: Hospital Scheduling Decisions



Forecaster predicts patient length-of-stay in the hospital.

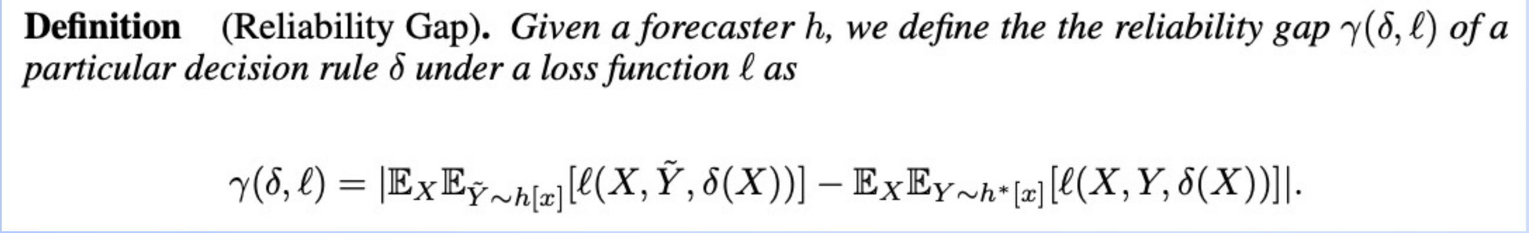
Hospital decides whether they have capacity to admit new patients based on threshold decision on model's predictions of current patients' length of stay.

Incorrect decisions from based on the model's predictions cause the hospitals to accrue costs.

Question: What notion of calibration is necessary and sufficient to guarantee that a forecaster (ML model) enables decision makers to predict their decision loss prior to deployment under threshold decision rules?

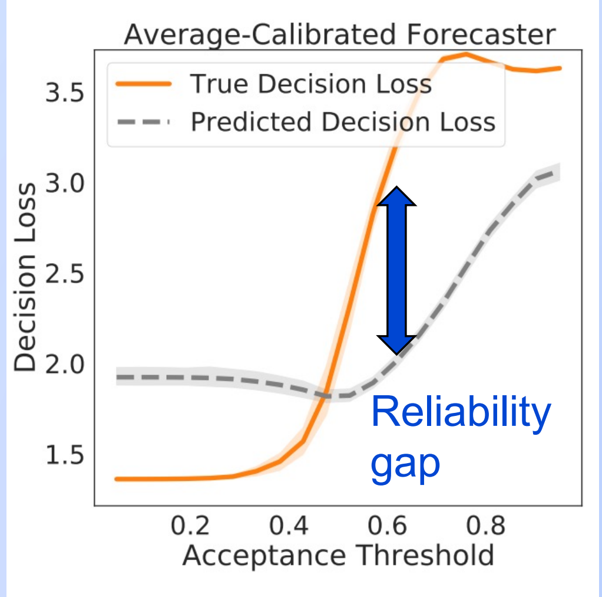
Reliability Gap

We define the reliability gap to be the absolute difference between the predicted decision loss and the true decision loss.

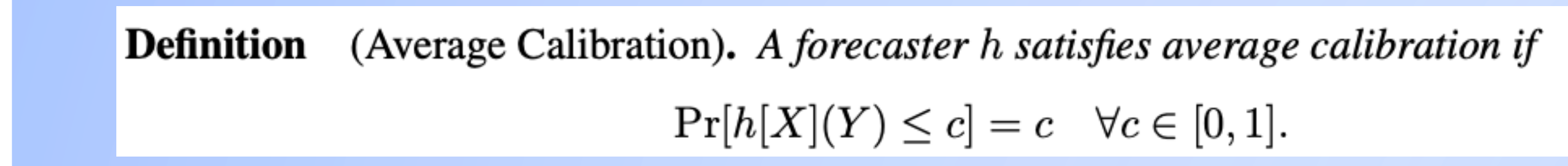


Definition (Reliability Gap). *Given a forecaster h, we define the the reliability gap $\gamma(\delta, \ell)$ of a particular decision rule δ under a loss function ℓ as*

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

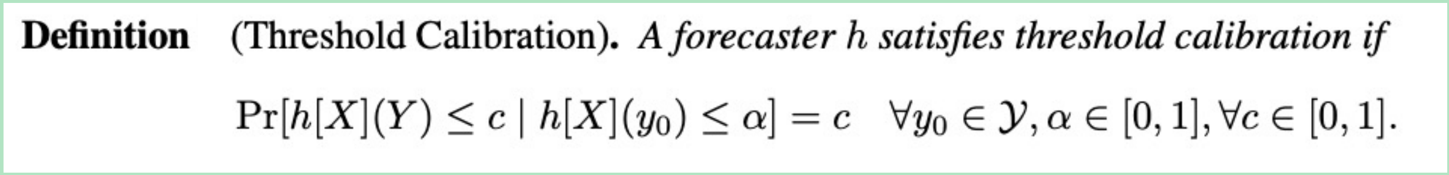


Calibration Definitions



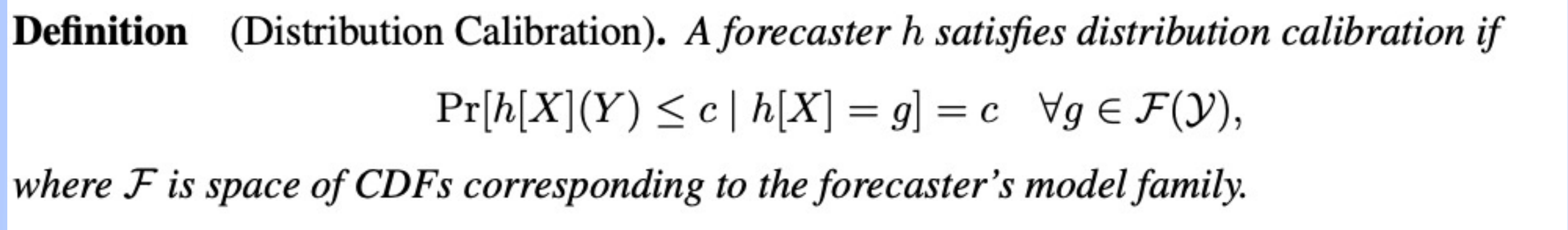
Definition (Average Calibration). *A forecaster h satisfies average calibration if*

$$\Pr[h[X](Y) \le c] = c \quad \forall c \in [0,1].$$



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

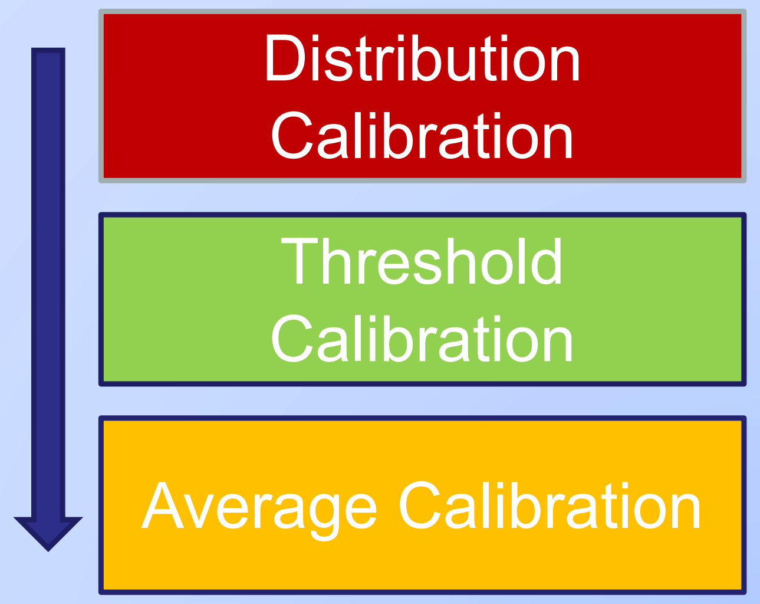
$$\Pr[h[X](Y) \le c \mid h[X](y_0) \le \alpha] = c \quad \forall y_0 \in \mathcal{Y}, \alpha \in [0,1], \forall c \in [0,1].$$



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

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

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



Distribution Calibration

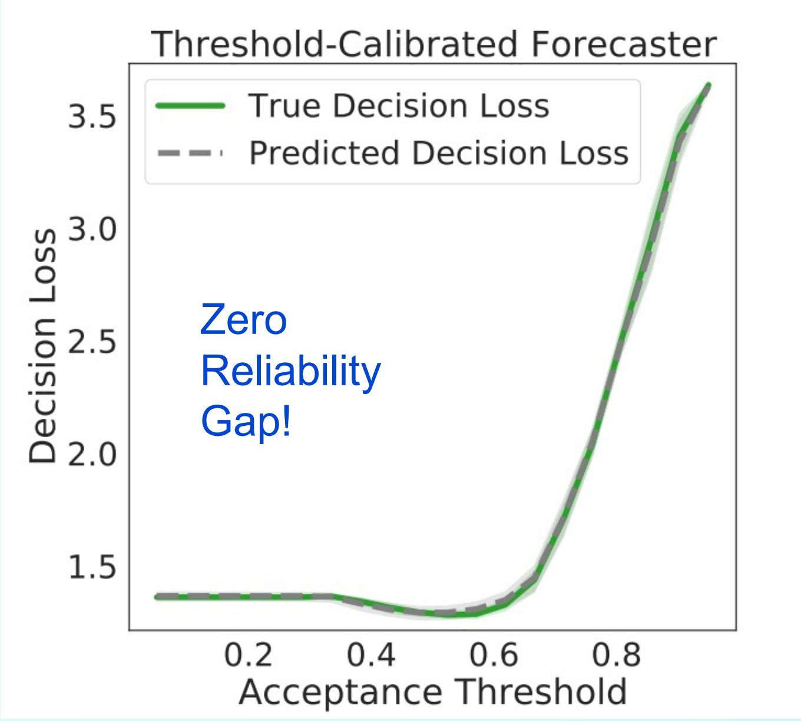
Threshold Calibration

Average Calibration

Threshold Calibration: Necessary and Sufficient Condition for Zero Reliability Gap.

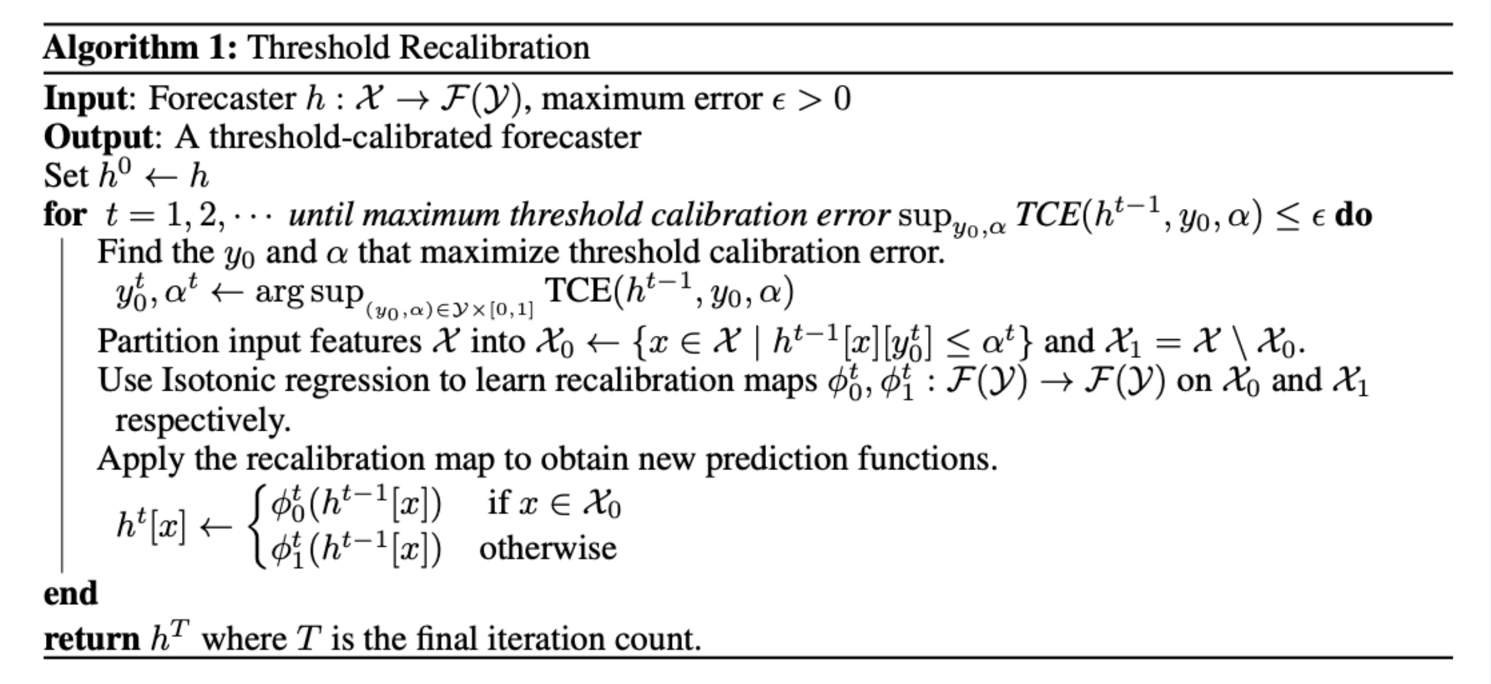
Under a threshold-calibrated forecaster, we achieve zero reliability gap under any threshold decision on the forecasted CDFs and any loss function.

Existing calibration definitions include average and distribution calibration. Threshold calibration addresses shortcomings of average calibration (insufficient for minimizing reliability gap) and distribution calibration (difficult to enforce in practice).



Achieving Threshold Calibration

We provide a recalibration algorithm that takes an uncalibrated forecaster as input and provably outputs a threshold calibrated forecaster.

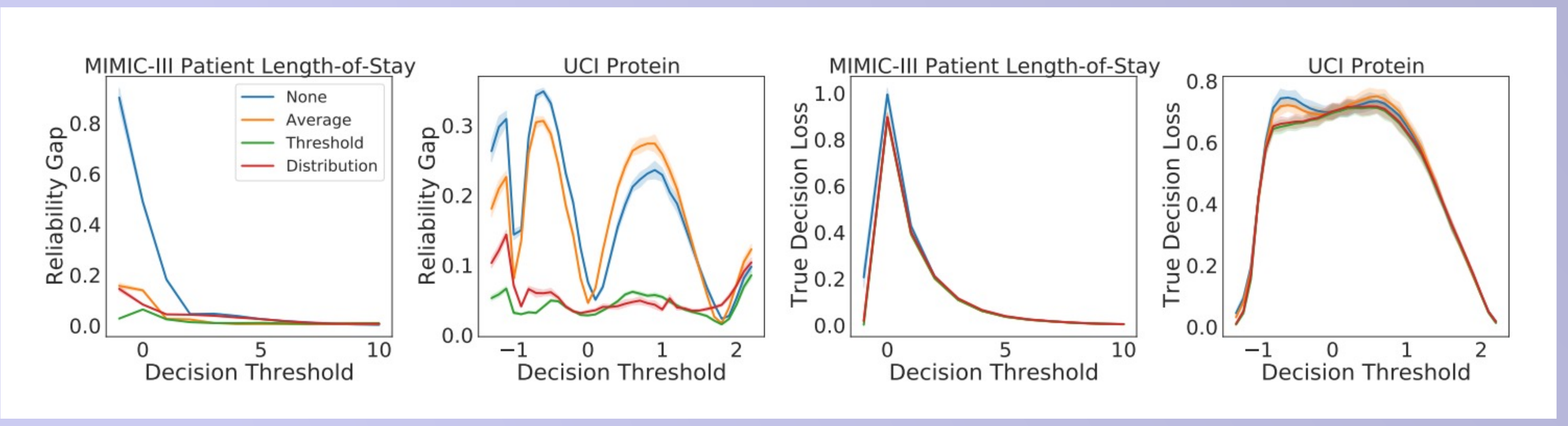


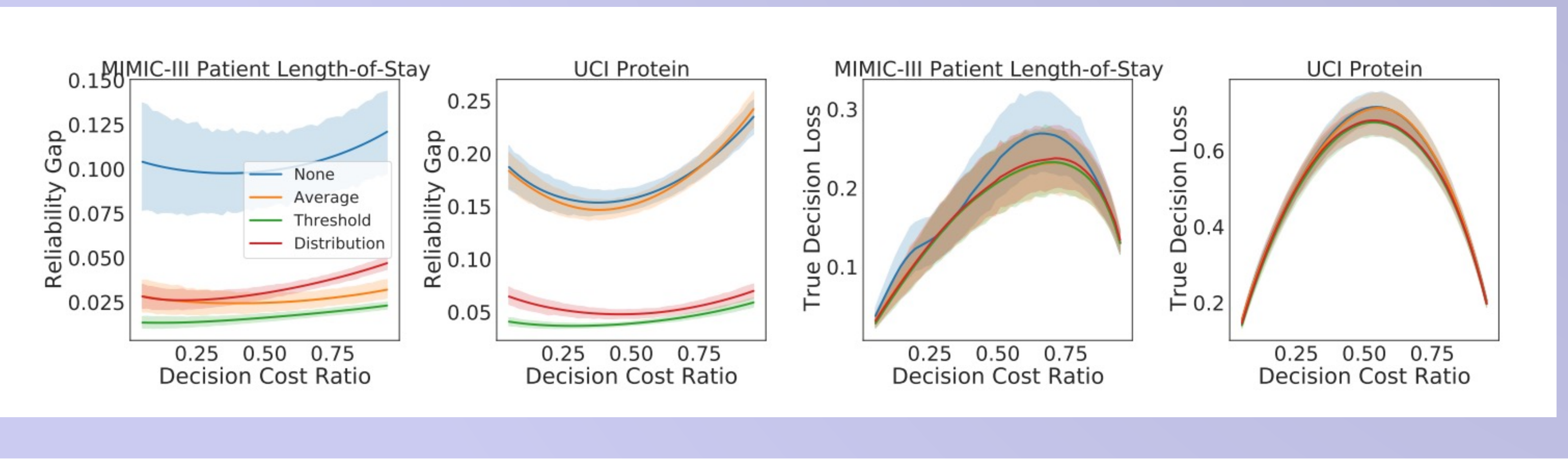
```
Algorithm 1: Threshold Recalibration
Input: Forecaster $h : \mathcal{X} \to \mathcal{F}(\mathcal{Y})$, maximum error $\epsilon > 0$
Output: A threshold-calibrated forecaster
Set $h^0 \leftarrow h$
for $t = 1, 2, \cdots$ until maximum threshold calibration error $\sup_{y_0,\alpha} TCE(h^{t-1}, y_0, \alpha) \le \epsilon$ do
    Find the $y_0$ and $\alpha$ that maximize threshold calibration error.
    $y_0^t, \alpha^t \leftarrow \arg\sup_{(y_0,\alpha)\in\mathcal{Y}\times[0,1]} \mathrm{TCE}(h^{t-1}, y_0, \alpha)$
    Partition input features $\mathcal{X}$ into $\mathcal{X}_0 \leftarrow \{x \in \mathcal{X} \mid h^{t-1}[x][y_0^t] \le \alpha^t\}$ 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.
```

Procedure converges after at most $O\left(\frac{1}{\varepsilon^2}\right)$ iterations and outputs a forecaster with threshold calibration error at most ε.

Inspired by previous work on multicalibration.

Experiments: Threshold calibration outperforms other calibration methods in minimizing the reliability gap across different threshold loss functions without compromising on the accuracy of the decisions.


MIMIC-III Patient Length-of-Stay
None
Average
Threshold
Distribution
Reliability Gap
Decision Threshold
UCI Protein
Reliability Gap
Decision Threshold
MIMIC-III Patient Length-of-Stay
True Decision Loss
Decision Threshold
UCI Protein
True Decision Loss
Decision Threshold


MIMIC-III Patient Length-of-Stay
Reliability Gap
None
Average
Threshold
Distribution
Decision Cost Ratio
UCI Protein
Reliability Gap
Decision Cost Ratio
MIMIC-III Patient Length-of-Stay
True Decision Loss
Decision Cost Ratio
UCI Protein
True Decision Loss
Decision Cost Ratio