Updating Clinical Risk Stratification Models Using Rank-Based Compatibility: Approaches for Evaluating and Optimizing Clinician-Model Team Performance

Erkin Ötleş EOTLES@UMICH.EDU
Medical Scientist Training Program
University of Michigan
Ann Arbor, MI, USA

Brian T. Denton BTDENTON@UMICH.EDU
Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI, USA

Jenna Wiens WIENSJ@UMICH.EDU
Division of Computer Science and Engineering
University of Michigan
Ann Arbor, MI, USA

Abstract

As data shift or new data become available, updating clinical machine learning models may be necessary to maintain or improve performance over time. However, updating a model can introduce compatibility issues when the behavior of the updated model does not align with user expectations, resulting in poor user-model team performance. Existing compatibility measures depend on model decision thresholds, limiting their applicability in settings where models are used to generate rankings based on estimated risk. To address this limitation, we propose a novel rank-based compatibility measure, $\mathcal{C}^{\mathrm{R}}$, and a new loss function that aims to optimize discriminative performance while encouraging good compatibility. Applied to a case study in mortality risk stratification leveraging data from MIMIC, our approach yields more compatible models while maintaining discriminative performance compared to existing model selection techniques, with an increase in $\mathcal{C}^{\mathrm{R}}$ of 0.019 (95% confidence interval: 0.005, 0.035). This work provides new tools to analyze and update risk stratification models used in clinical care.

1. Introduction

As machine learning (ML) models become increasingly integrated into clinical workflows, understanding the impact of model updates on these workflows and users is crucial. Models may be retrained and updated as new data become available to maintain or improve performance over time (Finlayson et al., 2021; Jenkins et al., 2021; Davis et al., 2022). For example, Memorial Sloan Kettering Cancer Center's prostate cancer outcome prediction models are updated annually (Vickers et al., 2017). While primarily intended to improve model performance, model updating can also affect users' expectations, *i.e.*, how users believe a model will perform given specific examples or patients. When models behave in

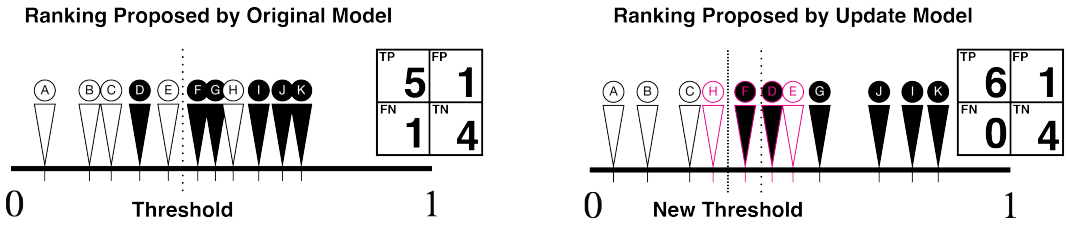

Figure 1: Backwards trust compatibility ($\mathcal{C}^{\text{BT}}$, Bansal et al. (2019b)) vs. Rank-based compatibility ($\mathcal{C}^{\text{R}}$, proposed). A model is used to stratify patients at risk for an outcome (black) from those that are not (white). Both the original and updated models have decision thresholds independently set to maximize accuracy on the validation set (shown). On this set, the original model has an accuracy of $\frac{9}{11}$ and an AUROC of $\frac{26}{30}$. The updated model switches the order of patients highlighted in magenta, resulting in higher accuracy $\frac{10}{11}$ and AUROC $\frac{28}{30}$. Out of the 9 patients correctly labeled by the original model the updated model labeled 8 correctly, this fraction, $\frac{8}{9}$, is $\mathcal{C}^{\text{BT}}$. This measure depends on the model decision thresholds. Our compatibility measure, $\mathcal{C}^{\text{R}}$, evaluates the ordering of patient-pairs. Of the 26 patient-pairs correctly ordered by the original model, the updated model correctly ordered 25 (makes an error on patient-pair E-F), yielding a $\mathcal{C}^{\text{R}}$ of $\frac{25}{26}$.

unexpected ways (*e.g.*, make mistakes in situations where they were previously accurate), user-model team performance can suffer (Bansal et al., 2019b; Guo and Yang, 2020). Thus, selecting updated models based solely on discriminative performance may be insufficient. Model developers may need to consider the potential disruption to existing workflows and alignment with user expectations in addition to discriminative performance (Bansal et al., 2019b; Zahedi and Kambhampati, 2021). This creates a need for practical tools to estimate how updated models might influence user expectations without directly querying users (Bansal et al., 2021). Fundamentally, we would like a way to answer this question: *to what extent does an updated model retain the correct behavior of an original model?*

To this end, *compatibility measures* assess how much an updated model may disrupt a user's mental model compared to the original model and an evaluation dataset. While researchers have proposed compatibility measures for supervised classification, like the backwards trust compatibility measure, these existing measures depend on a decision threshold (Bansal et al., 2019b). However, selecting a single fixed threshold may not be appropriate in many settings. In the context of patient risk stratification tools, decision thresholds can depend on system constraints or user preferences (Wynants et al., 2019; Gorski et al., 2017). Similar to how the receiver operating characteristic curve evaluates discriminative performance across all decision thresholds, there is a need for compatibility measures that are independent of a threshold.

Given this gap, we propose a novel rank-based compatibility measure that estimates the probability that an updated model will correctly rank a pair of discordantly labeled patients (a *patient-pair*), given that the original model was correct. This new measure offers

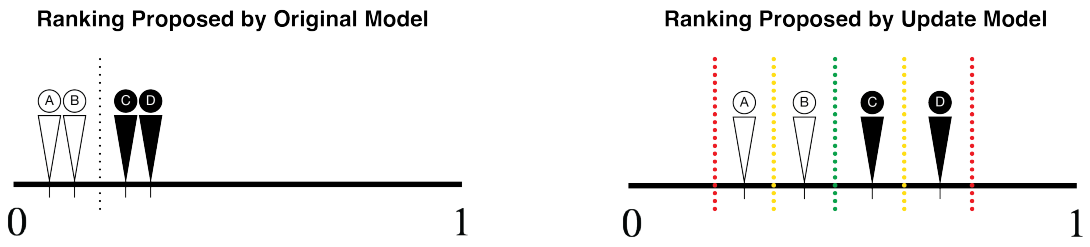

Figure 2: $\mathcal{C}^{\text{BT}}$ is sensitive to the choice of both model decision thresholds. Both models have perfect rank-based discrimination (*i.e.*, AUROC = 1). Depending on the updated decision threshold, $\mathcal{C}^{\text{BT}}$ may be $\frac{1}{2}$ (red), $\frac{3}{4}$ (yellow), or 1 (green). Regardless of the model decision threshold, $\mathcal{C}^{\text{R}}$ is 1 for this example.

a broader evaluation framework for model updates used in risk stratification and ranking, and applies in settings where model outputs are used for clinical resource allocation decisions. By considering the concordance between model rankings, we can proactively detect potentially harmful updates and avoid negative impacts on user-model team performance. **Figure 1** provides an overview of our proposed approach, illustrating its relationship to existing performance and compatibility measures. **Figure 2** illustrates the limitations of backwards trust compatibility compared to rank-based compatibility. In this work, we also demonstrate how our new measure relates to model discriminative performance and develop a loss function that can be used to directly optimize for compatibility.

Generalizable Insights about Machine Learning in the Context of Healthcare

Healthcare has witnessed an explosion of ML models in recent years, and it is a domain in which the task of ranking patients based on risk arises frequently. At the same time, models must be updated to retain clinical utility. For example, the Epic sepsis model, a patient deterioration model used by tens of thousands of clinicians in the United States, was recently updated in light of reports of poor performance (Gerhart and Thayer, 2021; Wong et al., 2021; Ross, 2022). We focus on a similar case study in which we stratify patients according to their risk of in-hospital mortality. While it may seem that discriminative performance must suffer to maintain compatibility, we show that developers can generate compatible updated models without negatively affecting discriminative performance by using our proposed loss function during training. Compared to updating approaches that ignore compatibility, or use existing compatibility measure, this work facilitates model updates that are more consistent with clinicians' expectations and thus may be more readily accepted and adopted in practice.

Our main contributions are as follows:

- To the best of our knowledge, we introduce the first rank-based compatibility measure based on the concordance of risk estimate pairs.

- We characterize the extent to which the new compatibility measure may vary over potential model updates.
- We introduce a loss function that incorporates ranking *incompatibility loss*, which can be used to train model updates with improved rank-based compatibility characteristics.
- Using MIMIC-III, we present empirical results that demonstrate how the proposed loss function leads to improved rank-based compatibility without a significant decrease in AUROC compared to standard model selection approaches.

2. Problem Setup & Background

In the context of learning risk stratification models, a patient i is represented by the tuple $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the feature vector and $y_i \in \{0, 1\}$ represents the binary label (*e.g.*, outcome). Risk stratification model, $f(\cdot)$, outputs risk estimates, $\hat{p}_i \in [0, 1]$ that estimate $\Pr(y_i = 1|\mathbf{x}_i)$. These risk estimates can be converted to predicted labels, $\hat{y}_i = \mathbb{1}(\hat{p}_i > \tau)$, where τ is some decision threshold.

We seek to assess the impact on user expectations when updating an original model, $f^o(\cdot)$, to an updated model, $f^u(\cdot)$. Note that the original and updated models are specific instantiations of the risk stratification models introduced above. They produce risk estimates denoted as $\hat{p}_i^o$ and $\hat{p}_i^u$, respectively. We refer to the combination of an original and updated model as a *model-pair*. Decision thresholds for the original and the updated models are τ^o and τ^u, respectively.

The original and candidate updated risk stratification models are evaluated on a held-out set of patients, denoted as I. This set can be partitioned into two mutually exclusive subsets based on patient labels: 0-labeled patients, I^0, and 1-labeled patients, I^1. The size of these subsets of patients are denoted as n^0 and n^1, respectively, and their sum, n, is the cardinality of I. We formalize the notion of a *patient-pair*, a pair of patients i and j that do not share the same label (*i.e.*, $i \in I^0$ and $j \in I^1$). The total number of patient-pairs, m, is the product $n^0 n^1$. We denote the number of patient-pairs correctly ranked by the original and updated models as m^{o+} and m^{u+}, respectively. Both m^{o+} and m^{u+} are integers taking on values between 0 and m inclusively. Given an original model, we aim to select an updated model that achieves good discriminative performance and compatibility.

2.1. Discriminative Performance

Discriminative performance measures a model's ability to separate patients with different labels (Harrell Jr et al., 1996). The area under the receiver operating characteristic curve (AUROC) is widely used to evaluate the discriminative performance of risk stratification models since it evaluates performance across all decision thresholds τ. The AUROC corresponds to the probability of correctly ranking two patients with differing labels based on the risk estimates produced by the model. It may be estimated by counting the number of patient-pairs ranked correctly by a model, m^{o+}, and then normalizing by the total number of patient-pairs m (Hanley and Mcneil, 1982):

$$\text{AUROC}(f^o) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbb{1}(\hat{p}_i^o < \hat{p}_j^o)}{m} = \frac{m^{o+}}{m} \tag{1}$$

The AUROC ranges between 0 and 1; a value of 0.5 corresponds to an ordering that is no better than random. The AUROC is the binary case of the concordance index (c-index), and both are related to the Wilcoxon-Mann-Whitney U statistic (Harrell, 1982; Kendall, 1938; Harrell Jr et al., 1996).

2.2. Backwards Trust Compatibility

Currently, *backwards trust compatibility* ($\mathcal{C}^{\text{BT}}$) is the primary compatibility measure described in the literature (Bansal et al., 2019b,a). $\mathcal{C}^{\text{BT}}$ measures the agreement between the true label and the predicted labels produced by the original and updated models by counting the number of patients both labeled correctly and normalizing by the number of patients the original model labeled correctly:

$$\mathcal{C}^{\text{BT}}(f^o, f^u) = \frac{\sum_{i \in I} \mathbb{1}(y_i = \hat{y}_i^o) \cdot \mathbb{1}(y_i = \hat{y}_i^u)}{\sum_{i \in I} \mathbb{1}(y_i = \hat{y}_i^o)} \tag{2}$$

$\mathcal{C}^{\text{BT}}$ depends on an evaluation set of patients, I, and values range between 0 and 1. $\mathcal{C}^{\text{BT}} = 0$ when the updated model mislabels all the patients labeled correctly by the original model, and $\mathcal{C}^{\text{BT}} = 1$ when the updated model correctly labels all the patients the original model got correct. $\mathcal{C}^{\text{BT}}$ is not symmetric, as $\mathcal{C}^{\text{BT}}(f^o, f^u)$ does not necessarily equal $\mathcal{C}^{\text{BT}}(f^u, f^o)$. $\mathcal{C}^{\text{BT}}$ is expected to decrease in settings with dataset shifts as the feature-label relationships captured by the model-pairs differ (Srivastava et al., 2020).

In the context of patient risk stratification models, calculating $\mathcal{C}^{\text{BT}}$ requires first thresholding risk scores to produce binary predictions. However, many settings in healthcare do not use a decision threshold (Wynants et al., 2019). For example, patients in the emergency department may be stratified by continuous risk estimates, and surgeons may use different risk thresholds to recommend surgery. In use cases where there are multiple thresholds, $\mathcal{C}^{\text{BT}}$ may be computed multiple times; however, this is problematic for several reasons. First, the evaluation grows proportionally with the number of thresholds being considered. Second, there is limited utility in doing so for cases with a class imbalance (see **Appendix Section C.6**) Third, $\mathcal{C}^{\text{BT}}$ is sensitive to the selection of thresholds, and poorly chosen thresholds could lead to a model with good discrimination being evaluated poorly, as shown in **Figure 2**. These suggest a need for a compatibility measure that applies directly to continuous risk estimates without thresholding.

3. Methods

We present our proposed rank-based compatibility measure, $\mathcal{C}^{\text{R}}$, which measures compatibility independent of a decision threshold by examining the ranking concordance of patient-pairs. While related to the AUROC, we hypothesize that optimizing discriminative model

Table 1: Relationship between original and updated model AUROC, proportion of patient-pairs and count variables.

	Original Model Ranks Correctly	**Original Model Ranks Incorrectly**	
Updated Model Ranks Correctly	$\phi^{++} = \frac{m^{++}}{m}$	$\phi^{-+} = \frac{m^{-+}}{m}$	$\text{AUROC}(f^u) = \frac{m^{u+}}{m}$
Updated Model Ranks Incorrectly	$\phi^{+-} = \frac{m^{+-}}{m}$	$\phi^{--} = \frac{m^{--}}{m}$	$1 - \text{AUROC}(f^u)$
	$\text{AUROC}(f^o) = \frac{m^{o+}}{m}$	$1 - \text{AUROC}(f^o)$	

performance by minimizing binary cross-entropy loss when training models may not necessarily lead to high $\mathcal{C}^{\text{R}}$. Thus, we propose a new loss function based on a differentiable approximation of $\mathcal{C}^{\text{R}}$ that can be used when training updated models.

3.1. Rank-Based Compatibility

The rank-based compatibility, presented in **Equation 3**, compares the ranking produced by the updated model against the ranking produced by the original model.

$$\mathcal{C}^{\text{R}}(f^o, f^u) := \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbb{1}(\hat{p}_i^o < \hat{p}_j^o) \cdot \mathbb{1}(\hat{p}_i^u < \hat{p}_j^u)}{\sum_{i \in I^0} \sum_{j \in I^1} \mathbb{1}(\hat{p}_i^o < \hat{p}_j^o)} \tag{3}$$

Given a set of evaluation patients, I, $\mathcal{C}^{\text{R}}$ corresponds to the number of patient-pairs that both models rank correctly normalized by the number of patient-pairs that the original model ranked correctly. In contrast with $\mathcal{C}^{\text{BT}}$, which operates by counting patients, $\mathcal{C}^{\text{R}}$ operates on patient-pairs produced by the mutually disjoint subsets I^0 and I^1. $\mathcal{C}^{\text{R}}$ measures the concordance of ranking patient-pairs and ranges from 0 to 1. In contrast, $\mathcal{C}^{\text{BT}}$ measures concordance with respect to binary patient predictions.

Although this work focuses on risk stratification models that operate over patients with binary outcomes, $\mathcal{C}^{\text{R}}$ is not limited to this setting; we present a general form of $\mathcal{C}^{\text{R}}$ in **Appendix Equation 6**.

Relationship to AUROC. Both $\mathcal{C}^{\text{R}}$ and AUROC involve counting correct patient-pair rankings. We introduce several ancillary rank-based compatibility variables to clarify how $\mathcal{C}^{\text{R}}$ relates to AUROC. Four proportion of patient-pairs (POP) variables measure how two models rank (correctly vs. incorrectly) patient-pairs.

The POP variables, ϕ^{ab}, follow a convention where a represents how the original model ranks patient-pairs correctly ($+$) vs. incorrectly ($-$), and b represents the same information for the updated model. For example, the POP variable for patient-pairs *correctly* ordered by both models is denoted by ϕ^{++}, and the proportion of patient-pairs *incorrectly* ordered by both models is ϕ^{--}. The four POP variables sum to 1. From the POP variables, one can

calculate the AUROC of each model (*e.g.*, $\text{AUROC}(f^o) = \phi^{++} + \phi^{+-}$). Each POP variable is proportional to a patient-pair count variable: m^{++}, m^{+-}, m^{-+}, and m^{--}, which follow the same $\cdot^{ab}$ notation. The relationships among the POP variables, the count variables, and discriminative performances can be expressed in a tabular manner, as depicted in **Table 1**. From these relationships, $\mathcal{C}^{\text{R}} = \frac{m^{++}}{m^{o+}} = \frac{\phi^{++}}{\text{AUROC}(f^o)}$.

Rank-Based Compatibility Lower Bound. Given $\text{AUROC}(f^o)$ and $\text{AUROC}(f^u)$, we can bound all POP variables (see Appendix Section B.2). Here, we assume that $0.5 < \text{AUROC}(f^o) \leq \text{AUROC}(f^u) \leq 1$, yielding the following lower bound for the rank-based compatibility:

$$\frac{\text{AUROC}(f^o)+\text{AUROC}(f^u)-1}{\text{AUROC}(f^o)} \leq \mathcal{C}^{\text{R}}(f^o, f^u)$$

This bound can be used to contextualize the $\mathcal{C}^{\text{R}}$ of an update, as the range of $\mathcal{C}^{\text{R}}$ changes depending on the model AUROCs being considered. The lower bound of $\mathcal{C}^{\text{R}}$ increases with respect to the AUROC of the updated model (shown graphically in **Appendix Section B.3**). We note that the upper bound is always 1 for the model updating region we are interested in.

3.2. Optimizing for Rank-Based Compatibility

While standard model training and selection procedures that typically focus on discriminative performance will result in a larger lower bound for $\mathcal{C}^{\text{R}}$, one may choose to optimize directly for $\mathcal{C}^{\text{R}}$. However, as defined, $\mathcal{C}^{\text{R}}$ is non-differentiable due to the *ranking indicator function*, $\mathbb{1}(\hat{p}_i < \hat{p}_j)$. To facilitate the use of rank-based incompatibility loss in gradient-based optimization, we introduce a differentiable approximation of rank-based compatibility:

$$\widetilde{\mathcal{C}^{\text{R}}}(f^o, f^u) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \sigma(\hat{p}^o_j - \hat{p}^o_i) \cdot \sigma(\hat{p}^u_j - \hat{p}^u_i)}{\sum_{i \in I^0} \sum_{j \in I^1} \sigma(\hat{p}^o_j - \hat{p}^o_i)}$$

This approximation replaces the ranking indicator function used to evaluate patient pairs with a *ranking sigmoid function*:

$$\sigma(\hat{d}_{ji}) = \frac{1}{1 + \exp(-s \cdot \hat{d}_{ji})}$$

Where $\hat{d}_{ji}$ is the difference in risk estimates produced for a patient pair (*i.e.*, $\hat{d}_{ji} = \hat{p}_j - \hat{p}_i$ and ranges between -1 and 1). A correct ranking corresponds to $\hat{d}_{ji} > 0$ and an incorrect ranking corresponds to $\hat{d}_{ji} < 0$. The sigmoid function maps this to a value between 0 and 1, closer to the behavior of the ranking indicator function (Han and Moraga, 1995). A hyperparameter, s, controls the spread of this mapping. Note that using a sigmoid to overcome discontinuity in the loss function is similar to work introduced to optimize for the AUROC directly (Yan et al., 2003).

Risk stratification models are often trained by minimizing the binary cross-entropy loss $\mathcal{L}^{BCE}$. This attempts to optimize the discriminative performance of the model by reducing

the probability estimates for 0-labeled patients and increasing them for 1-labeled patients, and indirectly optimizes the correct ranking of patient-pairs, the AUROC (Cortes and Mohri, 2003). However, $\mathcal{L}^{BCE}$ only examines the relationship between a patient's label and the risk estimates produced by a model. To incorporate rank-based compatibility, we augment model update training to incentivize rank-based compatibility, using a weighted combination of binary cross-entropy and $\widetilde{\mathcal{L}^R} = 1 - \widetilde{\mathcal{C}^{\mathrm{R}}}$:

$$\alpha\mathcal{L}^{BCE} + (1-\alpha)\widetilde{\mathcal{L}^R} \text{ where } \alpha \in [0,1] \tag{4}$$

Hyperparameter α controls the trade-off between discriminative performance and compatibility. During training, the predictions produced by the original model are incorporated into the loss function.

4. Experiments & Results

We focus on understanding and engineering model updates in terms of $\mathcal{C}^{\mathrm{R}}$ using a real-world benchmark dataset. While $\mathcal{C}^{\mathrm{R}}$ could be used as a validation metric when selecting among candidate models during an update procedure, we hypothesize that by including $\widetilde{\mathcal{C}^{\mathrm{R}}}$ in the loss function, we can achieve better compatibility without paying a penalty in terms of AUROC. To test this hypothesis, we generated and analyzed model updates on the MIMIC-III mortality prediction dataset.

Questions. Our experiments seek to answer two related questions:

1. *What is the empirical distribution of $\mathcal{C}^{\mathrm{R}}$ achieved using standard model updates when using real data?* (**Section 4.2**, **Figure 4**)

2. *Compared to standard model update generation and selection approaches, can we use the rank-based incompatibility loss, $\widetilde{\mathcal{L}^R}$, to generate updates with better $\mathcal{C}^{\mathrm{R}}$? Can this be accomplished without a loss of AUROC?* (**Section 4.3**, **Figures 5**, **6**, **7**, **13**, and **14**)

4.1. Data & Model Updating Setup

Dataset & Task. We use Bansal et al. (2019b)'s work as foundation for our experimental setup. Their experimental work analyzing $\mathcal{C}^{\mathrm{BT}}$ in the setting of updating an in-hospital mortality prediction model served as a template for our main analyses. In order to maintain consistency and enable comparisons between $\mathcal{C}^{\mathrm{BT}}$ and $\mathcal{C}^{\mathrm{R}}$ we modeled our predictive task, dataset splits, and model architectures considered off of their initial experiments.

Like Bansal et al. (2019b), we employed the MIMIC-III dataset (Johnson et al., 2016), with the goal of predicting in-hospital mortality based on the first 48 hours of a patient's ICU stay, with the population and task defined by Harutyunyan et al. (2019). The data were transformed using FIDDLE (Tang et al., 2020). For details regarding the data inclusion and transformation, please see the procedures detailed by Tang et al. (2020). Since our goal wasn't to learn the best possible mortality prediction tool, but to investigate the applicability of $\mathcal{C}^{\mathrm{R}}$, we reduced the number of features from $350,832$ to $35,000$, for computational efficiency. This was done by random sampling.

We randomly split the MIMIC-III data into three disjoint datasets. Two of these datasets were allocated for model development and validation. The third dataset was reserved for held-out evaluation. 8,577 patients in the MIMIC-III dataset meet the in-hospital mortality inclusion criteria defined by Harutyunyan et al. (2019). The datasets were split similarly to Bansal et al. (2019b), with 1,000 allocated to the original model dataset, 5,000 were assigned to the updated model dataset, and 2,577 held-out for the evaluation dataset. The two model datasets were used to develop and validate the original and updated models. The model datasets were each split equally (50/50%) into development and validation datasets. The dataset partitions and their sizes are depicted in **Figure 3**.

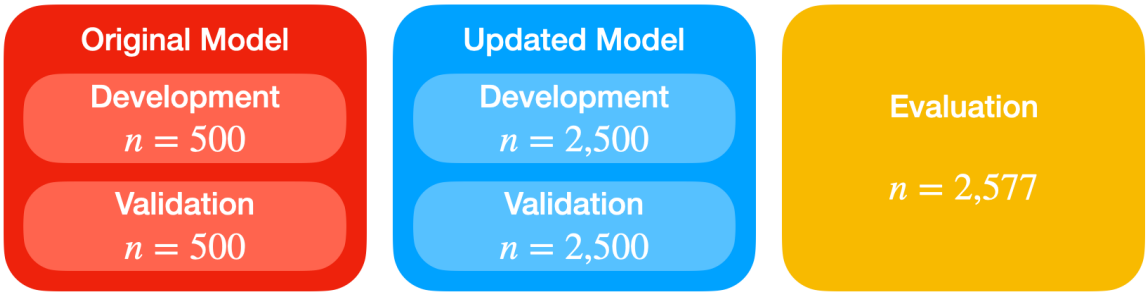

Figure 3: The MIMIC-III mortality data was partitioned into three datasets. Two of these datasets were allocated for model development and validation, and one was held-out for evaluation. Model-pairs were evaluated on the evaluation dataset.

Original model training & selection. Original models were trained using regularized logistic regression. L2 regularization strength $\{0.1, 0.01, 0.001\}$ was selected to maximize validation AUROC.

Updated model training & selection. Two different types of updated models were created to assess standard updating approaches against our proposed loss function. Standard updates, "BCE models", were trained to minimize $\mathcal{L}^{BCE}$ subject to regularization. The same regularization weights used for the original models were available for the updated models.

Using the same original model and data, we generated additional updated models, "RBC models" based on a loss function that incorporates $\mathcal{L}^{R}$ (**Equation 4**), sweeping α in the set $\{0, 0.1, 0.2, ..., 0.9, 1\}$.

Updated models from the "BCE" and "RBC models" were selected based on maximizing the following validation function:

$$\beta \, \text{AUROC}(f^u) + (1 - \beta) \, \mathcal{C}^{\text{R}}(f^o, f^u) \text{ where } \beta \in [0, 1] \tag{5}$$

Evaluation. The selected updated models were evaluated in terms of $\mathcal{C}^{\text{R}}$ and AUROC on the held-out evaluation dataset. The process of splitting the data, training model-pairs, and evaluation was replicated 40 times.

4.2. Rank-Based Compatibility Distribution

We first investigate: *What is the empirical distribution of $\mathcal{C}^{\mathrm{R}}$ achieved using standard model updates (i.e., minimizing the binary cross-entropy loss) when using real data?* Using the experimental setup described above, we created 150 standard updated models for each original model, minimizing $\mathcal{L}^{BCE}$. To introduce variation, these 150 candidate "BCE models" were created by combining dataset resampling, shuffling, and regularization weights. The updated model development dataset was either resampled with replacement (45 of the times) or shuffled (5 of the times, which yields difference in models due to our use of stochastic gradient descent) and then models were trained using binary cross-entropy loss with one of three L2 regularization weights ($(45+5)\cdot 3 = 150$).

We calculated the AUROC of the original model and the resultant $\mathcal{C}^{\mathrm{R}}$ and AUROC across the candidate update models (**Figure 4**). Across the 150 "BCE models," empirical 95% confidence intervals were calculated for AUROC(f^u), and violin plots were generated for $\mathcal{C}^{\mathrm{R}}$.

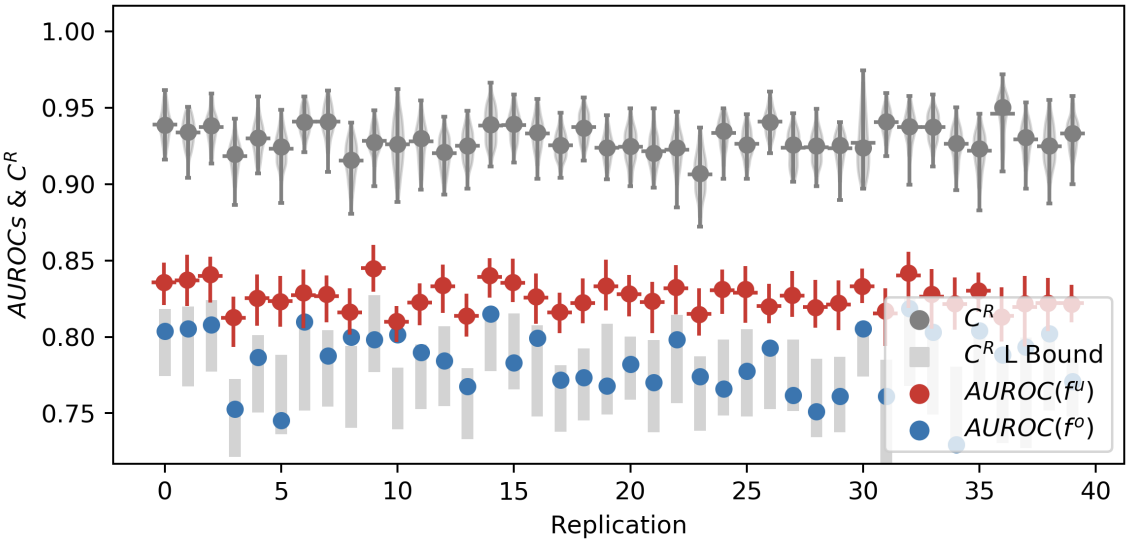

Figure 4: $\mathcal{C}^{\mathrm{R}}$ Distribution For Model Updates on the MIMIC-III Mortality Task. An original model was selected for each replication and 150 "BCE models" were generated as candidate updates. We plot the AUROC of the original model (blue dots) and updated "BCE models" (red, 95% confidence intervals). We also show the expected lower bounds for $\mathcal{C}^{\mathrm{R}}$ (light gray). Finally, the "BCE models" $\mathcal{C}^{\mathrm{R}}$s distribution are plotted as violin plots (gray).

The observed $\mathcal{C}^{\mathrm{R}}$ values for the set of candidate updates vary across a portion of the feasible range (between the lower[1] and upper bounds). We note that the observed distributions of $\mathcal{C}^{\mathrm{R}}$ shifts in relation to the AUROCs of the models. To control for this shift, we

1. Note, that the lower bound is presented as a range. This is because each candidate update model has a separate lower bound depending on its AUROC.

can examine the POP variable ϕ^{++}. We see that the distribution of ϕ^{++} for this experiment tends to a central value; this is shown and discussed in **Appendix Section C.4**. These results show the behavior of $\mathcal{C}^{\mathrm{R}}$ for one data-generating process, where we see some variation in $\mathcal{C}^{\mathrm{R}}$ values that provide limited options for model developers to select among. Additionally, we see that larger $\mathcal{C}^{\mathrm{R}}$ values are possible but not observed through standard update generation procedures (this is the space above the observed $\mathcal{C}^{\mathrm{R}}$ violin plots in **Figure 4**). These findings are underscored in an analytical sketch discussed in **Appendix Section B.5**. *All together, these results mean that model developers may be constrained if they wish to develop updated models that optimize for $\mathcal{C}^{\mathrm{R}}$ using standard update generation procedures.*

4.3. Weighted Loss vs. Standard Updated Model Selection

We now investigate our second question: *Compared to optimizing for $\mathcal{L}^{BCE}$ alone, does incorporating the rank-based incompatibility loss, $\mathcal{L}^{R}$, generate updates with better $\mathcal{C}^{\mathrm{R}}$?*

For each replication, we generated 150 "BCE models" using the generation procedure described above. For each value of $\alpha \in \{0, 0.1, 0.2, ..., 0.9, 1\}$, we also generated 3 "RBC models". This was done by sweeping the regularization strengths used above. Aside from the objective function used during training (and early stopping), other aspects of model training and selection were held constant across approaches. To give the baseline the best chance, we resampled and shuffled the training data while training the BCE models to more fully explore the space of potential updates (resulting in 150 updates instead of 3). The best "BCE" and "RBC models" from these model sets were selected based on validation performance using **Equation 5**. We compare the resulting "BCE" and "RBC models" by calculating the difference in rank-based compatibility, $\Delta\,\mathcal{C}^{\mathrm{R}}$, and difference in AUROC, Δ AUROC (an example of this calculation can be found in **Appendix Section C.2**). We repeated this process 40 times, for every value of α and every value of $\beta \in \{0, 0.1, ..., 0.9, 1\}$ and compared the mean differences in both $\mathcal{C}^{\mathrm{R}}$ and AUROC.

Results are displayed in **Figure 5**. There is a trade-off between AUROC and $\mathcal{C}^{\mathrm{R}}$. For many α-β combinations, there is a significant gain in $\mathcal{C}^{\mathrm{R}}$ (blue) at the cost of lower AUROC (red) when using the proposed objective function during optimization. However, we note many cases in which there is a gain in compatibility without paying a penalty in terms of AUROC. For example, when $\alpha = 0.5$ and $\beta = 0.5$, we achieve a significant gain in compatibility of $\Delta\,\mathcal{C}^{\mathrm{R}} = 0.019$ (95% confidence interval: 0.005, 0.035) with an $\Delta AUROC = -0.009$ (−0.030, 0.011).[2] By incorporating $\mathcal{L}^{R}$ during training, it is possible to achieve improved compatibility without compromising discriminative performance. Out of the 121 α-β combinations, 57 demonstrate statistically significant improvements in $\mathcal{C}^{\mathrm{R}}$ while maintaining AUROC; see **Appendix Section C.5** for further discussion.

Examining results across replications for an $\alpha = 0.6$ while we vary β **Figure 6**, we see that across selection options, the "RBC model" generally provides a better $\mathcal{C}^{\mathrm{R}}$ (statistically significant for $\beta \leq 0.6$) without a significant decrease in AUROC (*i.e.*, $\Delta AUROC$ is at or close to zero). In **Figure 7**, we set $\beta = 0.6$ during the selection process for both the "RBC models" and the "BCE models", and sweep α during training "BCE models". Again,

2. The "RBC models" had the following performance: $\mathcal{C}^{\mathrm{R}} = 0.966$ (0.948, 0.979) AUROC = 0.828 (0.804, 0.855) vs. "BCE models" with $\mathcal{C}^{\mathrm{R}} = 0.947$ (0.932, 0.963)

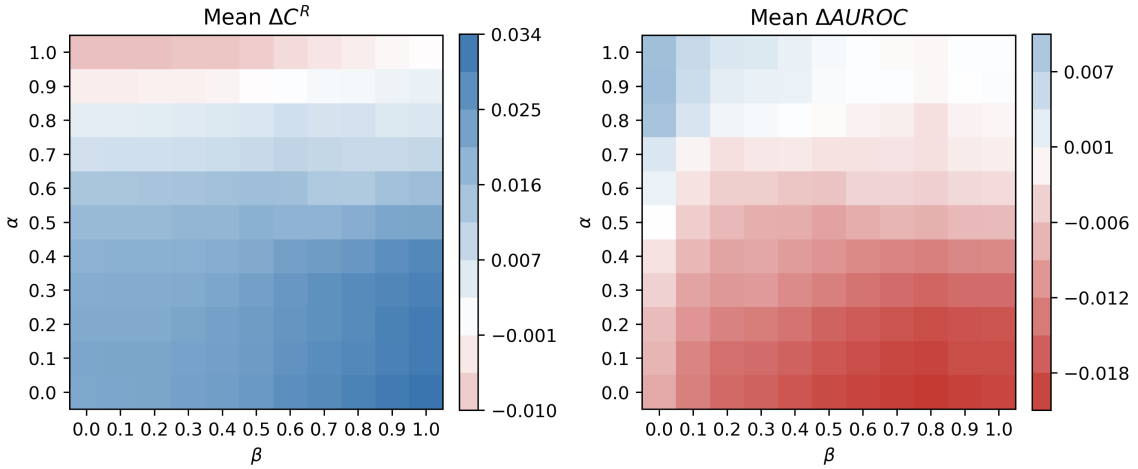

Figure 5: Performance Differences Between "RBC Models" and "BCE Models" With Variation of α and β. Mean value of $\Delta\,\mathcal{C}^{\mathrm{R}}$ on the left and mean value of Δ AUROC on the right, blue shows improvement of "RBC Models" over "BCE models" and red shows degradation. For a large majority of α-β pairs, there is an improvement in mean $\mathcal{C}^{\mathrm{R}}$. For a smaller majority, there is a degradation in mean AUROC. This suggests that there is a trade-off between AUROC and $\mathcal{C}^{\mathrm{R}}$, with improved $\mathcal{C}^{\mathrm{R}}$ coming at the cost of AUROC. Although this trade-off exists, we note that the degradations in AUROC are often not statistically significant, while the improvements in $\mathcal{C}^{\mathrm{R}}$ are. This is shown and discussed in **Appendix Section C.5**.

we observe that for specific α values (e.g., $\alpha = 0.3 - 0.6$), we can significantly improve compatibility without penalizing AUROC performance.

These empirical results suggest that by incorporating rank-based compatibility into the objective function during training, we can generate model updates with larger $\mathcal{C}^{\mathrm{R}}$ *values than obtained through standard update generation procedures (i.e., minimizing for* $\mathcal{L}^{BCE}$ *alone).* Moreover, while there is often a trade-off between $\mathcal{C}^{\mathrm{R}}$ and AUROC, achieving gains in $\mathcal{C}^{\mathrm{R}}$ while maintaining AUROC(f^u) is possible.

5. Discussion & Conclusion

When selecting among potential updated clinical risk stratification models, it may be important to consider compatibility with existing models already in use. In this study, we propose the first rank-based compatibility measure, $\mathcal{C}^{\mathrm{R}}$, which measures the concordance in ranking between two models. We illustrate the connection between $\mathcal{C}^{\mathrm{R}}$ and discriminative model performance. This relationship suggests that increased rank-based compatibility accompanies improved discriminative performance, as the lower bound of rank-based compatibility increases as each model's discriminative performance increases. Despite this relationship, we show empirically that it is improbable to observe very high levels of rank-based compatibility through standard updated model development, which tends to focus on optimizing

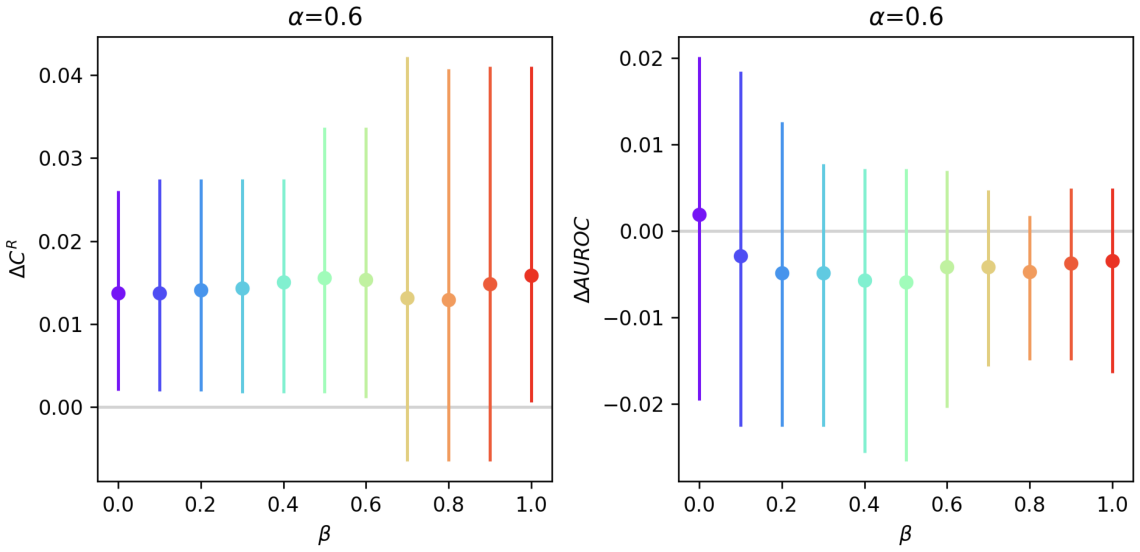

Figure 6: Performance Difference for $\alpha = 0.6$ Sweeping Across β. Comparing $\Delta\,\mathcal{C}^{\mathrm{R}}$ and Δ AUROC for various β values. We see that for all β values, there is no statistically significant degradation in AUROC while for β values less than 0.7 we see improvement in $\mathcal{C}^{\mathrm{R}}$. This suggests that "RBC models" yield a benefit over "BCE models" in this regime.

discriminative performance. These findings motivate methods that enable developers to build models with good discriminative performance and rank-based compatibility. As such, we introduce a new differentiable rank-based incompatibility loss function that can be used when training updated models to further optimize for rank-based compatibility.

We used the MIMIC-III dataset to compare our proposed approach to generating model updates to a standard approach that optimizes for binary cross-entropy alone. Our results highlight standard updated model development's limitations in identifying model updates with very high compatibility. Using our proposed approach, we identify models with equivalent discriminative performance yet significantly better compatibility. However, if rank-based compatibility is greatly emphasized over discriminative performance, then improvements may come at a cost.

The rank-based compatibility measure serves a different role than the original backwards trust compatibility measure proposed by Bansal et al. (2019b). Depending on the use case, one may choose one over the other. Use cases that strongly depend on decision thresholds, such as sending a notification when a patient risk estimate exceeds a specific threshold, may correspond to clinician mental models best represented by $\mathcal{C}^{\mathrm{BT}}$. In settings where the decision may depend on the state of the system, such as hospital admission decisions, which are impacted by the number of patients in the emergency department (Gorski et al., 2017), the $\mathcal{C}^{\mathrm{R}}$ may better represent clinician mental models because it is not tied to a fixed threshold. Additionally, the complexity of this evaluation grows proportionally with

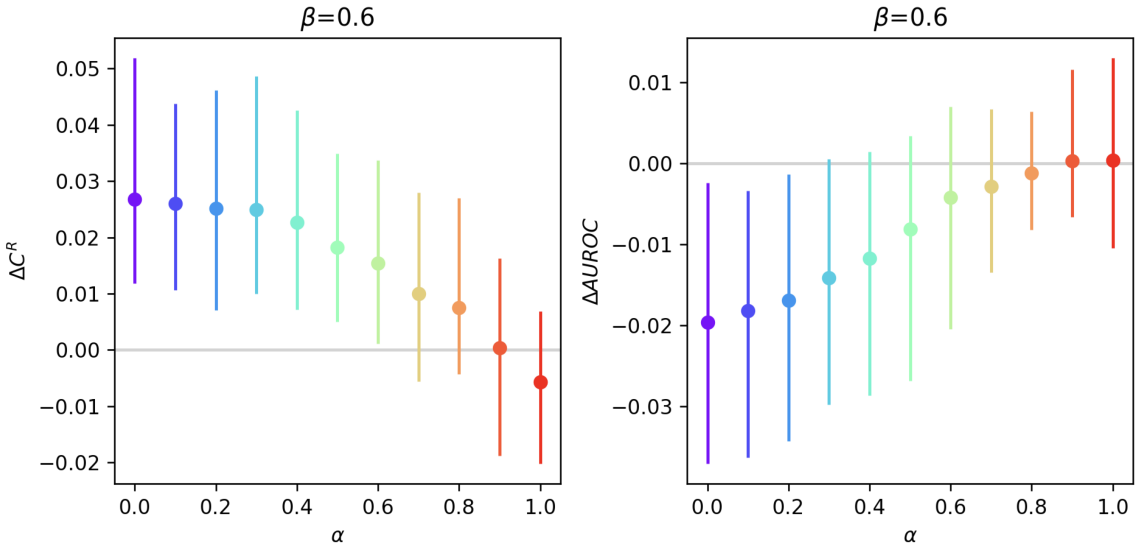

Figure 7: Performance Difference for $\beta = 0.6$ Sweeping Across α. Comparing $\Delta\,\mathcal{C}^{\mathrm{R}}$ and $\Delta\,\mathrm{AUROC}$ for various α values. In this case see a more limited benefit of the "RBC models" over "BCE models", with $\alpha \in [0.3, 0.6]$ showing significant benefit in $\mathcal{C}^{\mathrm{R}}$ and no significant degradation in AUROC.

the number of models and thresholds being considered. Thus, if there are many potential thresholds, it may be more effective to use $\mathcal{C}^{\mathrm{R}}$ directly.

Although we know the absolute scale of rank-based compatibility with 0 denoting "no compatibility" and 1 denoting "perfect compatibility", we do not have a sense of what the numbers in between mean and how they compare across model updates. Ideally, we would like to have a sense of what is an excellent rank-based compatibility value, like we do with the AUROC measure (*e.g.*, $\mathrm{AUROC}(f) > 0.85$). This will likely come with further study of models being updated across different tasks. One advantage $\mathcal{C}^{\mathrm{R}}$ does present is that its improvements can be directly compared against improvements in AUROC by examining the POP variables.

While we discussed the different use cases for $\mathcal{C}^{\mathrm{R}}$ vs. $\mathcal{C}^{\mathrm{BT}}$, we did not explore users' preferences. Although there may be update tasks for which the $\mathcal{C}^{\mathrm{R}}$ measure is better suited, we have not yet characterized the relationship between rank-based compatibility and user mental models. For example, a sepsis detection system that flags patients as at risk (Henry et al., 2022) or sends an alert notification (Wong et al., 2021) may be a good candidate for the $\mathcal{C}^{\mathrm{BT}}$ compatibility measure. Users in these cases would expect consistent correct classification of patients when the underlying model is updated. If users interact with the model to help risk stratify their patients, then the $\mathcal{C}^{\mathrm{R}}$ measure may be a better choice. Tools used for cardiovascular event risk stratification (Lip et al., 2010) and in-hospital deterioration risk stratification (Epic Systems Corporation, 2020; Kamran et al., 2022) may be more effectively updated using $\mathcal{C}^{\mathrm{R}}$.

Like backwards trust compatibility, rank-based compatibility captures the "global" user perspective. Modifying rank-based compatibility to focus on individual user perspectives may lead to better compatibility and parity with user expectations (Martinez et al., 2020, 2021). We have focused our study of rank-based compatibility exclusively on when the updated model continues the correct behaviors established by the original model. Previous user studies have shown that user mental models are influenced by the error behavior of classification models (Bansal et al., 2019a). This may hold for risk stratification models, motivating the study of incorrect ranking in conjunction with rank-based compatibility. We believe there is much work to do with this measure in terms of human user studies.

Finally, the primary analysis we present is based on the experimental setup developed by Bansal et al. (2019b). Although this was intentionally done to enable the comparison of the $\mathcal{C}^{\text{BT}}$ and $\mathcal{C}^{\text{R}}$ it is not an exhaustive evaluation. Notably, future work may benefit from the exploration of different tasks, datasets, and model architectures. Some tasks like survival analysis (Ötleș et al., 2022) may be able to use the general form of $\mathcal{C}^{\text{R}}$, **Equation 6**. Different model architectures may need adaption of the joint optimization of performance and compatibility proposed by this work. Additionally, there are real world complexities that are unaccounted for in this analysis, such as outcome censoring due to clinician interventions based on model predictions (Adam et al., 2020) and the impact of deployment infrastructure changing as models are updated (Ötleș et al., 2021).

These limitations notwithstanding, the new rank-based compatibility measure and incompatibility loss present a novel way to think about model maintenance and updating models, beyond simply optimizing for AUROC. Furthermore, optimizing the rank concordance between the output of two models, rather than thresholded predictions, may be more robust to calibration shifts, a commonly observed phenomenon in healthcare (Hickey et al., 2013; Davis et al., 2017; Minne et al., 2012). We expect this new measure applies in evaluating healthcare risk stratification models. However, there are likely settings in domains beyond healthcare that would similarly benefit from such rank-based measures. Overall, this work enables the evaluation and development of model updates that have the potential to lead to better clinician-model team performance.

Acknowledgments

EÖ was supported by NIH grant T32GM007863 and JW was supported by the Alfred P. Sloan Foundation. The authors would like to thank the anonymous reviewers and editors of the 2023 Machine Learning for Healthcare Conference for their thoughtful feedback.

References

George Alexandru Adam, Chun-Hao Kingsley Chang, Benjamin Haibe-Kains, and Anna Goldenberg. Hidden risks of machine learning applied to healthcare: Unintended feedback loops between models and future data causing model degradation. In *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pages 710–731. PMLR, 07–08 Aug 2020.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance.

In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11, 2019a.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:2429–2437, 2019b. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33012429.

Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11405–11414, 2021. ISBN 2374-3468.

Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, 16, 2003.

Sharon E Davis, Thomas A Lasko, Guanhua Chen, Edward D Siew, and Michael E Matheny. Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association*, 24(6):1052–1061, 2017. ISSN 1067-5027. doi: 10.1093/jamia/ocx030. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6080675.

Sharon E Davis, Colin G Walsh, and Michael E Matheny. Open questions and research gaps for monitoring and updating ai-enabled tools in clinical settings. *Frontiers in Digital Health*, 4:958284, 2022. ISSN 2673-253X.

Epic Systems Corporation. Artificial intelligence triggers fast, life-saving care for covid-19 patients. *News From Epic*, 2020 (06/28/2021), 2020. URL https://www.epic.com/epic/post/artificial-intelligence-epic-triggers-fast-lifesaving-care-covid-19-patients.

Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021. ISSN 0028-4793. doi: 10.1056/nejmc2104626.

Jackie Gerhart and Johnston Thayer. For clinicians, by clinicians: Our take on predictive models. *Cool Things*, 2023(06/28/2021), 2021. URL https://www.epic.com/epic/post/for-clinicians-by-clinicians-our-take-on-predictive-models.

Jillian K. Gorski, Robert J. Batt, Erkin Ötleș, Manish N. Shah, Azita G. Hamedani, and Brian W. Patterson. The impact of emergency department census on the decision to admit. *Academic Emergency Medicine*, 24(1):13–21, 2017. ISSN 1069-6563. doi: 10.1111/acem.13103. URL https://onlinelibrary.wiley.com/doi/pdf/10.1111/acem.13103.

Yaohui Guo and X. Jessie Yang. Modeling and predicting trust dynamics in human–robot teaming: A bayesian inference approach. *International Journal of Social Robotics*, 2020. ISSN 1875-4791. doi: 10.1007/s12369-020-00703-3.

Jun Han and Claudio Moraga. *The influence of the sigmoid function parameters on the speed of backpropagation learning*, pages 195–201. Springer Berlin Heidelberg, 1995. ISBN 0302-9743. doi: 10.1007/3-540-59497-3_175.

J A Hanley and B J Mcneil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982. ISSN 0033-8419. doi: 10.1148/radiology.143.1.7063747.

Frank E. Harrell. Evaluating the yield of medical tests. *JAMA: The Journal of the American Medical Association*, 247(18):2543, 1982. ISSN 0098-7484. doi: 10.1001/jama.1982.03320430047030.

Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996. ISSN 0277-6715.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1), 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9. URL https://doi.org/10.1038/s41597-019-0103-9.

Katharine E. Henry, Roy Adams, Cassandra Parent, Hossein Soleimani, Anirudh Sridharan, Lauren Johnson, David N. Hager, Sara E. Cosgrove, Andrew Markowski, Eili Y. Klein, Edward S. Chen, Mustapha O. Saheed, Maureen Henley, Sheila Miranda, Katrina Houston, Robert C. Linton, Anushree R. Ahluwalia, Albert W. Wu, and Suchi Saria. Factors driving provider adoption of the trews machine learning-based early warning system and its effects on sepsis treatment timing. *Nature Medicine*, 28(7):1447–1454, 2022. ISSN 1078-8956. doi: 10.1038/s41591-022-01895-z.

G. L. Hickey, S. W. Grant, G. J. Murphy, M. Bhabra, D. Pagano, K. Mcallister, I. Buchan, and B. Bridgewater. Dynamic trends in cardiac surgery: why the logistic euroscore is no longer suitable for contemporary cardiac surgery and implications for future risk models. *European Journal of Cardio-Thoracic Surgery*, 43(6):1146–1152, 2013. ISSN 1010-7940. doi: 10.1093/ejcts/ezs584. URL http://europepmc.org/articles/pmc3655624?pdf=render.

David A. Jenkins, Glen P. Martin, Matthew Sperrin, Richard D. Riley, Thomas P. A. Debray, Gary S. Collins, and Niels Peek. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagnostic and Prognostic Research*, 5(1), 2021. ISSN 2397-7523. doi: 10.1186/s41512-020-00090-3.

Alistair E.W. 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(1):160035, 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL http://europepmc.org/articles/pmc4878278?pdf=render.

Fahad Kamran, Shengpu Tang, Erkin Ötleș, Dustin S Mcevoy, Sameh N Saleh, Jen Gong, Benjamin Y Li, Sayon Dutta, Xinran Liu, Richard J Medford, Thomas S Valley, Lauren R West, Karandeep Singh, Seth Blumberg, John P Donnelly, Erica S Shenoy, John Z Ayanian, Brahmajee K Nallamothu, Michael W Sjoding, and Jenna Wiens. Early identification of patients admitted to hospital for covid-19 at risk of clinical deterioration: model development and multisite external validation study. *BMJ*, page e068576, 2022. ISSN 1756-1833. doi: 10.1136/bmj-2021-068576.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81.

Gregory YH Lip, Robby Nieuwlaat, Ron Pisters, Deirdre A Lane, and Harry JGM Crijns. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*, 137(2):263–272, 2010. ISSN 0012-3692.

Jonathan Martinez, Kobi Gal, Ece Kamar, and Levi H. S. Lelis. Personalization in human-ai teams: Improving the compatibility-accuracy tradeoff. *arXiv pre-print server*, 2020.

Jonathan Martinez, Kobi Gal, Ece Kamar, and Levi HS Lelis. Improving the performance-compatibility tradeoff with personalized objective functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5967–5974, 2021. ISBN 2374-3468.

Lilian Minne, Saeid Eslami, Nicolette De Keizer, Evert De Jonge, Sophia E. De Rooij, and Ameen Abu-Hanna. Effect of changes over time in the performance of a customized saps-ii model on the quality of care assessment. *Intensive Care Medicine*, 38(1):40–46, 2012. ISSN 0342-4642. doi: 10.1007/s00134-011-2390-2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3233667.

Erkin Ötleș, Jeeheh Oh, Benjamin Li, Michelle Bochinski, Hyeon Joo, Justin Ortwine, Erica Shenoy, Laraine Washer, Vincent B. Young, Krishna Rao, and Jenna Wiens. Mind the performance gap: Examining dataset shift during prospective validation. *Proceedings of Machine Learning Research*, 2021.

Erkin Ötleș, Jon Seymour, Haozhu Wang, and Brian T Denton. Dynamic prediction of work status for workers with occupational injuries: assessing the value of longitudinal observations. *Journal of the American Medical Informatics Association*, 2022. ISSN 1067-5027. doi: 10.1093/jamia/ocac130.

Casey Ross. Epic overhauls popular sepsis algorithm criticized for faulty alarms. *Stat*, 2022.

Megha Srivastava, Besmira Nushi, Ece Kamar, Shital Shah, and Eric Horvitz. An empirical analysis of backward compatibility in machine learning systems. *KDD*, 2020.

Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 27(12):1921–1934, 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa139.

Andrew J. Vickers, Mathew Kent, and Peter T. Scardino. Implementation of dynamically updated prediction models at the point of care at a major cancer center: Making nomograms more like netflix. *Urology*, 102:1–3, 2017. ISSN 0090-4295. doi: 10.1016/j.urology.2016.10.049. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5376358.

Andrew Wong, Erkin Ötleș, John P. Donnelly, Andrew Krumm, Jeffrey Mccullough, Olivia Detroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 2021. ISSN 2168-6106. doi: 10.1001/jamainternmed.2021.2626.

Laure Wynants, Maarten Van Smeden, David J. Mclernon, Dirk Timmerman, Ewout W. Steyerberg, and Ben Van Calster. Three myths about risk thresholds for prediction models. *BMC Medicine*, 17(1), 2019. ISSN 1741-7015. doi: 10.1186/s12916-019-1425-3.

Lian Yan, Robert H Dodier, Michael Mozer, and Richard H Wolniewicz. Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 848–855, 2003.

Zahra Zahedi and Subbarao Kambhampati. Human-ai symbiosis: A survey of current approaches. *arXiv pre-print server*, 2021.

Appendix A. Background

A.1. Decision Threshold Dependence of $\mathcal{C}^{\mathrm{BT}}$

Like accuracy, $\mathcal{C}^{\mathrm{BT}}$ is highly dependent on model thresholds. We illustrate this in **Figure 2**, which shows the dependence of $\mathcal{C}^{\mathrm{BT}}$ on the updated model decision threshold (τ^u). Depending on the choice of τ^u $\mathcal{C}^{\mathrm{BT}}(f^o, f^u)$ may be $\frac{1}{2}$, $\frac{3}{4}$, or 1. Because every patient-pair is correctly ordered by both models in this example, the $\mathcal{C}^{\mathrm{R}}$ equals 1. If τ^o had been set to a much larger (or smaller) value, then it would have been possible for $\mathcal{C}^{\mathrm{BT}}$ values of 0 to occur for this example. Ultimately, poorly chosen decision thresholds or models miscalibrated with one another may demonstrate poor $\mathcal{C}^{\mathrm{BT}}$ even if both the original and updated models have good discrimination and concordance in their correct patient-pair rankings ($\mathcal{C}^{\mathrm{R}}$).

Appendix B. Methods

B.1. General Form Rank-Based Compatibility

Equation 3 defines rank-based compatibility for risk stratification models operating over binary labels. Rank-based compatibility is not limited to use only in situations where the outcomes are binary. The core concept can be applied to any set of patient labels that can be ordered (*e.g.*, integer or real values). We now present a general form rank-based compatibility equation that can be employed in these situations.

$$\mathcal{C}^{\mathrm{R}}(f^o, f^u) := \frac{\sum_{i\in I}\sum_{j\in I} \mathbb{1}(\hat{p}_i^o < \hat{p}_j^o)\cdot \mathbb{1}(\hat{p}_i^u < \hat{p}_j^u)\cdot \mathbb{1}(y_i < y_j)}{\sum_{i\in I}\sum_{j\in I} \mathbb{1}(\hat{p}_i^o < \hat{p}_j^o)\cdot \mathbb{1}(y_i < y_j)} \tag{6}$$

This equation has several minor changes from **Equation 3**. Differences in the summation indices enable the equation to evaluate every patient-pair and an additional term ($\mathbb{1}(\hat{p}_i^o < \hat{p}_j^o)$) in the numerator and denominator checks if this patient-pair is ordered correctly by the label.

B.2. POP Variable Bounds

Given the assumption that the updated model is at least as good as the original model and that both models are better than random (*i.e.*, $0.5 \leq AUROC(f^o) \leq AUROC(f^u)$) and the relationships established in **Table 1**, several constraints exist on the POP variables. These are:

$$\begin{aligned} \mathrm{AUROC}(f^o) + \mathrm{AUROC}(f^u) - 1 \leq \phi^{++} &\leq \mathrm{AUROC}(f^o) \\ 0 \leq \phi^{+-} &\leq 1 - \mathrm{AUROC}(f^u) \\ 0 \leq \phi^{-+} &\leq 1 - \mathrm{AUROC}(f^o) \\ 0 \leq \phi^{--} &\leq 1 - \mathrm{AUROC}(f^u) \end{aligned}$$

In this study, we focus only on the POP variable that represents both models ranking patient-pairs correctly, ϕ^{++}, as it is the only one used directly in $\mathcal{C}^{\mathrm{R}}$. So we briefly discuss how we derive its bounds. The minimum value ϕ^{++} can take is the smallest proportion of correctly ordered patient-pairs by both models. Since the AUROCs of both models must

be at least 0.5, the smallest this proportion is when there is minimal overlap in the set of correctly ordered patient-pairs for each model. This is the sum of the two AUROCs subtracted by 1. The maximal value for ϕ^{++} is determined by the smaller of the two model's AUROC which is AUROC(f^o).

B.3. Lower-bound of $\mathcal{C}^{\mathrm{R}}$

We produce a plot for the lower bound of the rank-based compatibility measure (**Figure 8**). For the regime of model updating that we are interested in $0.5 < \text{AUROC}(f^o) \leq \text{AUROC}(f^u) \leq 1$), only the lower bound of $\mathcal{C}^{\mathrm{R}}$ varies, increasing as the discriminative performance of the two models grows.

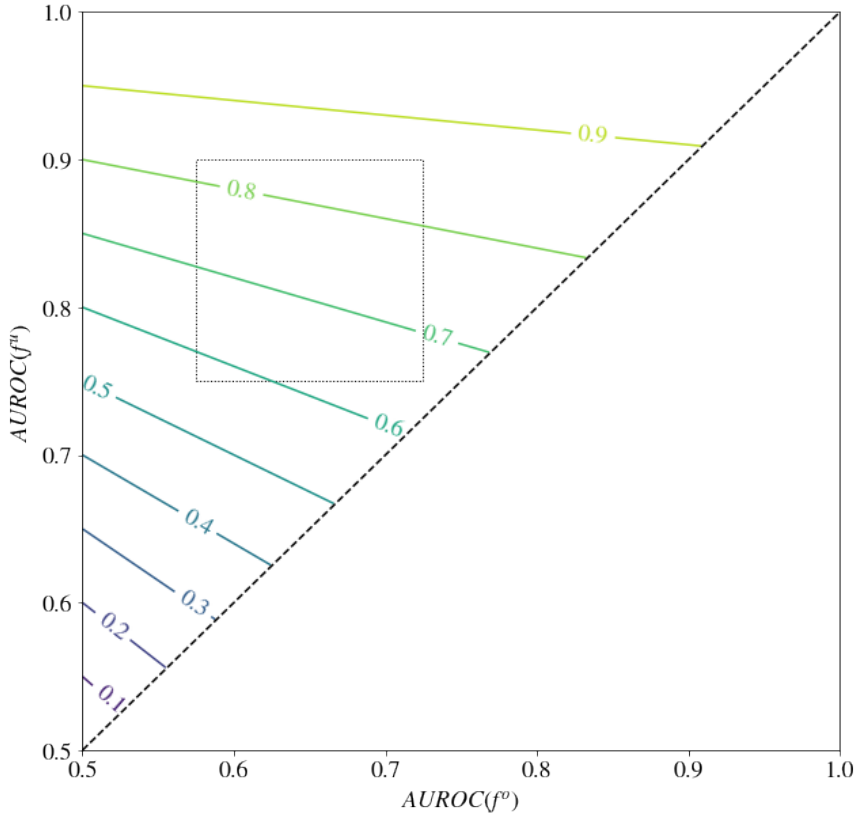

Figure 8: Lower bound of $\mathcal{C}^{\mathrm{R}}$ with respect to the AUROC of the original and updated models. The lower bound increases as both models' performance increases. The boxed region with dotted lines demarcates a typical discriminative performance region. In this region we would expect to observe $\mathcal{C}^{\mathrm{R}}$s no smaller than 0.5.

B.4. Central Tendency of $\mathcal{C}^{\mathrm{R}}$

Though the bounds suggest that higher compatibility is partially correlated with higher discriminative performance, we expect that in practice updated models $\mathcal{C}^{\mathrm{R}}$ values will have

a central tendency. We present a brief analytical sketch to explore this behavior. Note, we do not seek to create a distribution for the $\mathcal{C}^{\mathrm{R}}$ generally; instead, we seek to build intuition for how $\mathcal{C}^{\mathrm{R}}$ may vary with both models' AUROC. This analytical approach is based on a combinatorial argument. We analyze the number of ways a given $\mathcal{C}^{\mathrm{R}}$ can occur given AUROCs for the original and updated models. This analysis is based on how each model ranks each patient-pair. A patient-pair's ranking for a given model is whether that model correctly ranks (*e.g.*, $\hat{p}_i < \hat{p}_j$ for the updated model) or incorrectly ranks that patient pair.

We can use the ranking of all patient-pairs to represent the behavior of original and updated models. All patient-pairs are distributed between two sets: correctly and incorrectly ranked. Suppose we constrain the distribution of patient-pairs between these two sets to align with the discriminative performance of the model being represented. In that case, we can then get a sense of the number of patient-pairs that both models rank correctly. This number is m^{++} and can be directly used to calculate the $\mathcal{C}^{\mathrm{R}}$ as per **Equation 3**. As mentioned in **Appendix Section B.3**, m^{++} may range between $m^{o+} + m^{u+} - m$ and m^{o+}, corresponding to the bounds $\mathcal{C}^{\mathrm{R}}$ introduced in **Equation 3.1**. Assuming models do not have any restrictions on how patient-pairs may be ranked, we count the number of ways that each value of $m^{++} = k$ can be achieved given that each model meets a specific AUROC. We refer to this count as ν, where $\nu = |\{m^{++} = k | m^{o+}, m^{u+}\}|$.

ν is the numerator of the hypergeometric distribution with parameters related to the number of patient-pairs correctly ranked by the original and updated models. The number of patient-pairs that both models ranked correctly, $m^{++} = k$, is defined in relation to the number of total patient-pairs, m, the number of patient-pairs we are interested in selecting, m^{o+}, and the number of selections, m^{u+}. The number of combinations that produce a given $m^{++} = k$ is as follows:

$$\nu = \binom{m^{o+}}{k}\binom{m - m^{o+}}{m^{u+} - k}$$

The location and shape of this function provide us with a sense of the behavior conditional on the two model's AUROC. In **Appendix Section B.5** we plot this function and show where we would expect its maxima to occur. From this analysis, we expect $\mathcal{C}^{\mathrm{R}}$ to be centered around the AUROC of the updated model and should have a strong central tendency behavior.

While we do not believe this specific center to hold for all data generating processes and model updating procedures, we hypothesize that the central tendency of $\mathcal{C}^{\mathrm{R}}$ does. In **Section 4.2**, we investigate the central tendency of $\mathcal{C}^{\mathrm{R}}$ for original and updated models trained using real data. The above analysis is still illuminating as it provides a way to estimate the relative number of combinations between different rank-based compatibility levels. There are many more ways for an updated model to achieve moderate rank-based compatibility (near the value of the AUROC of the updated model) than a very high level of compatibility (*e.g.*, above 0.95). This suggests that achieving high rank-based compatibility may only be possible with directed search efforts.

B.5. Maxima of Central Tendency of $\mathcal{C}^{\mathrm{R}}$

The location of the maxima and shape of this function provides us with a sense of the behavior of $\mathcal{C}^{\mathrm{R}}$ conditional on maintaining a fixed level of discrimination. We would expect

this function's maxima to coincide with the mode of the corresponding hypergeometric distribution. For large values of m^{o+}, m^{u+}, and m we expect the mode of the hypergeometric distribution to be approximately equal to its mean. **Equation B.4** has its maxima at $m^{++} = k^*$, where k^* is the value that provides the largest number of combinations.[3]This is:

$$k^* = \left\lfloor \frac{(m^{o+}+1)(m^{u+}+1)}{m+2} \right\rfloor$$
$$\approx \frac{m^{o+}m^{u+}}{m} \text{ for large } m^{o+}, m^{u+}, \text{ and } m.$$

We can then plot **Equation B.4** to investigate the behavior of $\mathcal{C}^{\mathrm{R}}$ given AUROC(f^o) and AUROC(f^u). **Figure 9** shows the number of combinations for each $\mathcal{C}^{\mathrm{R}}$ value given original-updated model pairs. Each model pair had the same original model performance (AUROC(f^o) $= 0.65$), and the updated performance ranged between (AUROC(f^u) $\in [0.65, 0.95]$). Examination of these curves reveals several findings. First, the k^* for each model pair aligns with the AUROC of the updated model. Second, these curves exhibit a strong central tendency as the number of combinations decreases exponentially (note the logarithmic vertical axis) as $m^{++} = k$ diverges from k^*.

3. This maxima is expressed in terms of m^{++}, which can be converted to be in terms of $\mathcal{C}^{\mathrm{R}}$ by dividing by m^{o+}. This maxima occurs at $\frac{m^{o+}}{m^{o+}}\frac{m^{u+}}{m} = \frac{m^{u+}}{m} = \text{AUROC}(f^u)$.

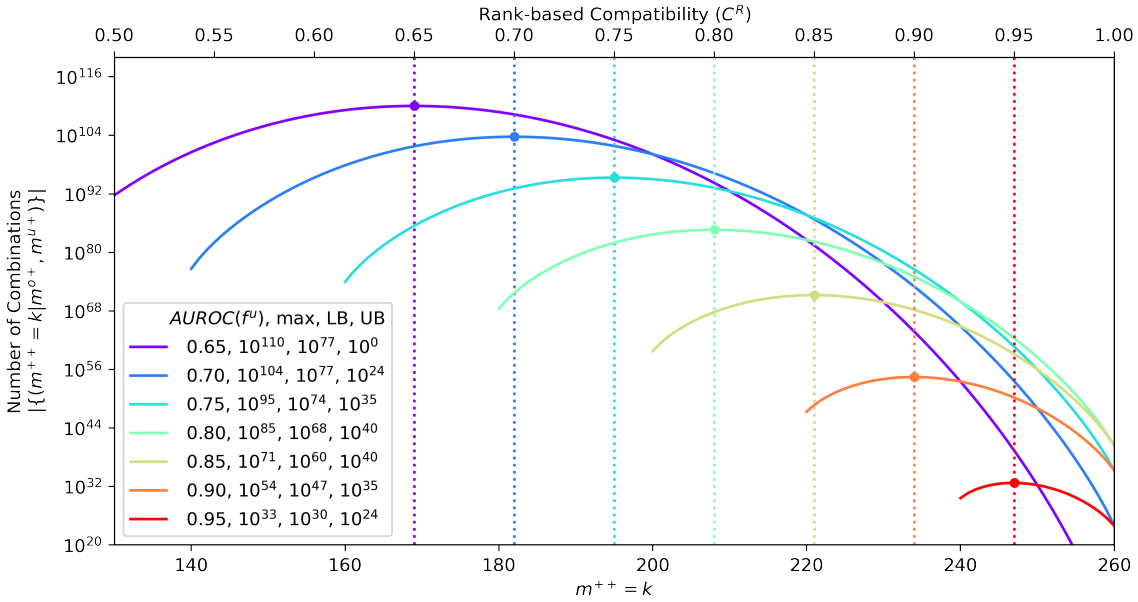

Figure 9: Number of combinations that yield a given $\mathcal{C}^{\mathrm{R}}$ value for an original-updated model pair. All model pairs have $\mathrm{AUROC}(f^o) = 0.65$ and $\mathrm{AUROC} \in [0.65, 0.95]$ ($m = 400$). The updated model's AUROC is plotted as a vertical dotted line. The $m^{++} = k^*$ value that achieves the largest number of combinations is plotted as a dot on the curves. This point aligns with $\mathrm{AUROC}(f^o)$. These curves exhibit a strong central tendency as the number of combinations decreases increasingly (note the logarithmic vertical axis) as $m^{++} = k$ diverges from the k^*.

Appendix C. Experiments & Results

C.1. Computing Environment

This analysis was conducted using Python on a server running Ubuntu 16.04.07 with 112 x86 CPU cores and 503GB of RAM. Experiments and analyses were run using Python version 3.7.4.

C.2. Example Replication Results

In **Figure 10** we show the $\mathcal{C}^{\mathrm{R}}$ and AUROC values calculated on the held-out evaluation data for all of the engineered models and a subset of the selection models. This subset represents the selection models along the pareto frontier of the trade-off between $\mathcal{C}^{\mathrm{R}}$ and AUROC (calculated using the updated model validation data). We also depict how $\Delta\,\mathcal{C}^{\mathrm{R}}$ and $\Delta\,\mathrm{AUROC}$ would be calculated between the engineered model where $\alpha = 0.6$ and the selected candidate update with the best AUROC.

For this example, we note that the circled engineered model induces a positive $\Delta\,\mathcal{C}^{\mathrm{R}}$, which denotes an increase in $\mathcal{C}^{\mathrm{R}}$, and a negative $\Delta\,\mathrm{AUROC}$, which represents a reduction in AUROC. Although the $\Delta\,\mathrm{AUROC}$ is negative, this does not mean that this updated

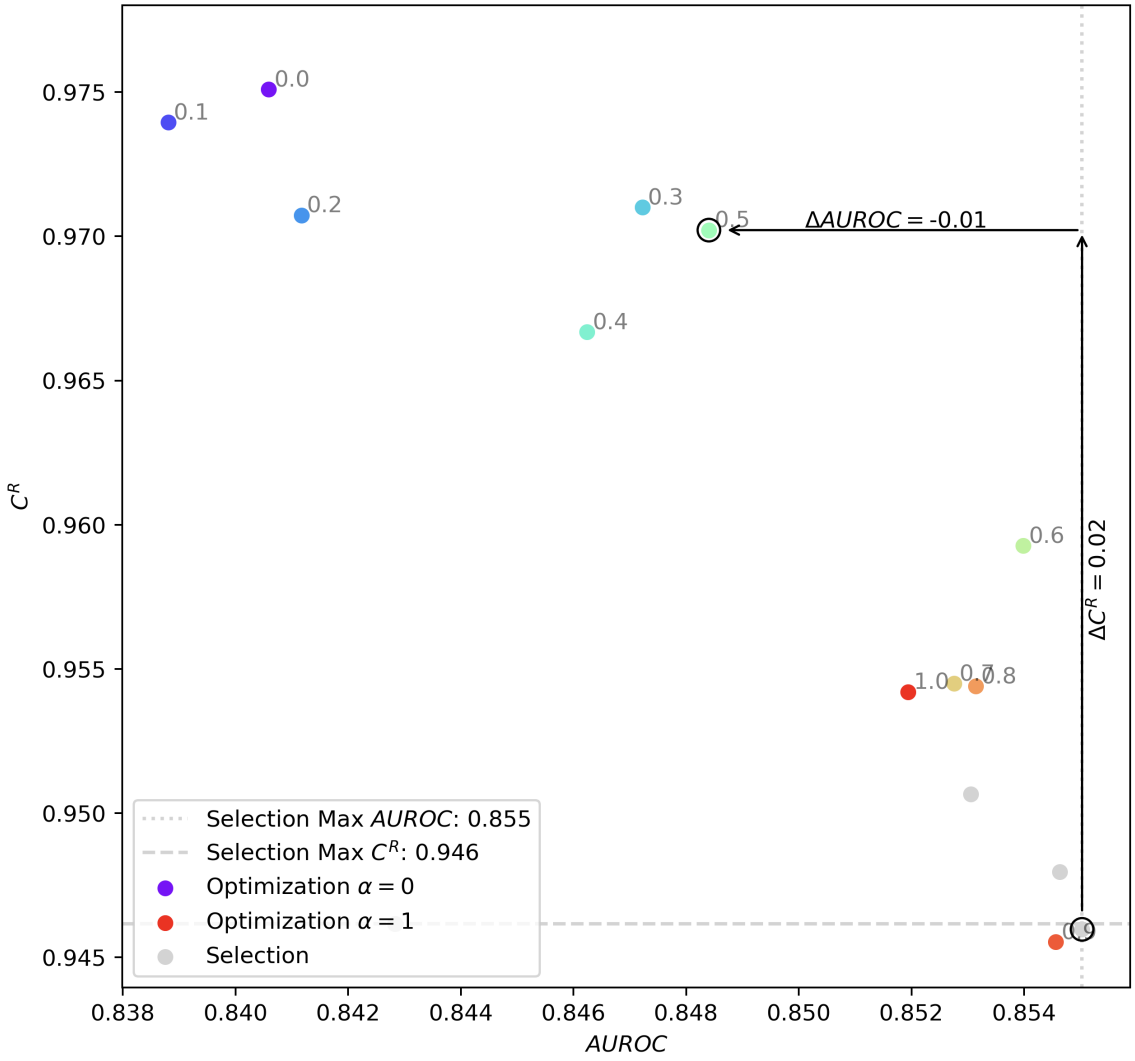

Figure 10: Example of Engineered Model vs. Selection Model Results. The AUROC and $\mathcal{C}^{\mathrm{R}}$ calculated on held-out evaluation dataset are reported for the engineered models and a subset of the selection models. In this example, we note that the circled engineered model ($\alpha = 0.5$) induces a positive $\Delta\,\mathcal{C}^{\mathrm{R}}$, which denotes an increase in $\mathcal{C}^{\mathrm{R}}$, and a negative Δ AUROC which indicates a reduction in AUROC.

model performs worse than the original model, which has an AUROC = 0.805. Instead, the engineered update (AUROC = 0.848) does not perform as well as the best-performing selection model (AUROC = 0.855).

C.3. Improvement in $\mathcal{C}^{\mathrm{R}}$ Compared with Distribution From Standard Model Updating

We show that "RBC Models" can produce $\mathcal{C}^{\mathrm{R}}$ values greater than what is observed through standard model updating procedures with little cost in terms of AUROC. See **Figure 11**.

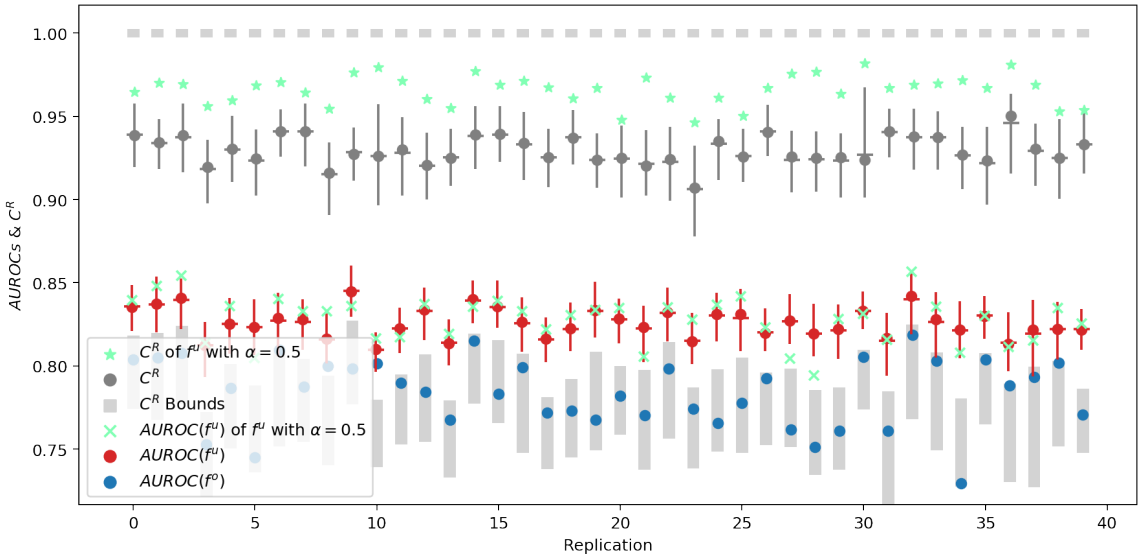

Figure 11: Improvement in $\mathcal{C}^{\mathrm{R}}$ of "RBC Models" at $\alpha = 0.5$. We note how for nearly all of the replications the "RBC Model" produces $\mathcal{C}^{\mathrm{R}}$ values exceeding those produced by the "BCE Models". The AUROC values are relatively in line with one another.

C.4. ϕ^{++} Central Tendency

As mentioned above, the distributions of $\mathcal{C}^{\mathrm{R}}$ shown in **Figure 9** shift in relation to the AUROCs of the models. To control for this shift we examined the POP variable ϕ^{++}. We did this by calculating the ϕ^{++} for each updated model. We then created a histogram for all updated models (histogram bin size=0.01). This procedure was repeated for all 40 replications. We then averaged the bin counts over all the replications. These results are plotted in **Figure 12**

From this plot, we see that each replication has a strongly peaked histogram and that the mean distribution of ϕ^{++} has a robust central tendency.

C.5. Improvement of $\Delta\,\mathcal{C}^{\mathrm{R}}$ and Non-Degradation of Δ AUROC

The graphs presented in **Section 4.3** show the mean $\Delta\,\mathcal{C}^{\mathrm{R}}$ and Δ AUROC. However, we examined the 95% confidence intervals to assess statistically significant differences. We determined if there was a statistically significant improvement in $\Delta\,\mathcal{C}^{\mathrm{R}}$ (*i.e.*, the confidence interval does not include 0) and if there was not a statistically significant degradation in

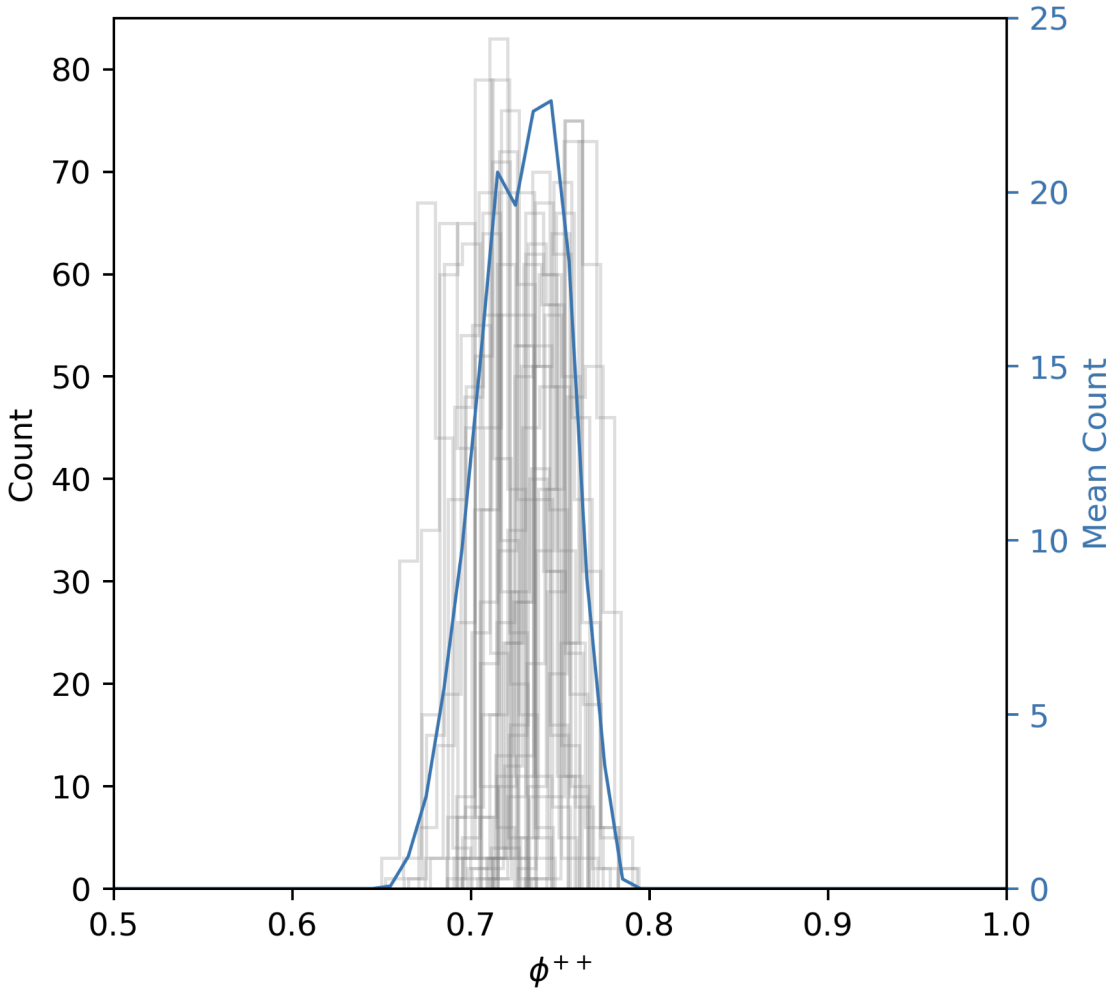

Figure 12: Central Tendency of ϕ^{++}. ϕ^{++} for each replication plotted in gray, bin size=0.01. Note a small amount of uniform jitter was added during plotting. The mean of these histograms across all replications is plotted in blue.

Δ AUROC (*i.e.*, the confidence interval does include 0). The α-β combinations that met these criteria are in blue in **Figure 13**. We note that 57 out of the 121 α-β combinations show an improvement with the inclusion of the $\mathcal{L}^R$ loss.

In order to characterize the α-β combinations where we see this improvement, we plot the critical confidence interval values (the lower bound of $\Delta\,\mathcal{C}^{\mathrm{R}}$ and the upper bound of Δ AUROC) along with their product in **Figure 14**.

From **Figure 14**, we can see there are several "regions" of α-β combinations. When α is high (*i.e.*, $\alpha \geq 0.7$), then we observe that we may not have a statistically significant improvement in Δ AUROC (the red region on top of the left panel). This makes sense. As α increases, we de-emphasize the importance of $\mathcal{C}^{\mathrm{R}}$, and thus, there should be little difference

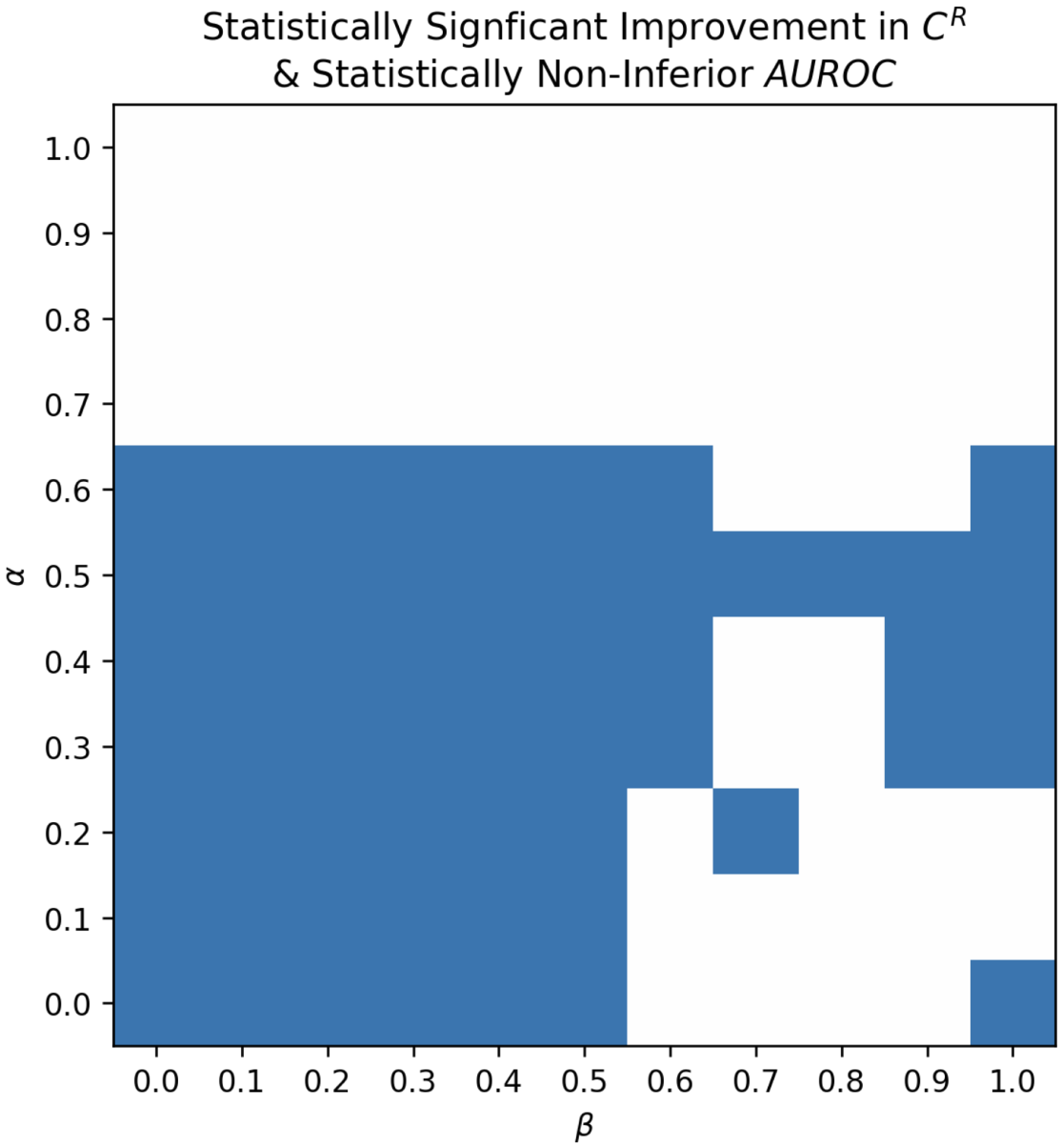

Figure 13: α-β Combinations Showing Improvement. All α-β combinations with a statistically significant improvement in $\mathcal{C}^{\mathrm{R}}$ without a statistically significant degradation in AUROC are depicted in blue.

(in terms of $\mathcal{C}^{\mathrm{R}}$) between the "RBC" and "BCE models". We note that when α is low, and β is high (*i.e.*, $\alpha \leq 0.4$ and $\beta \geq 0.6$) that we may have a statistically significant degradation in Δ AUROC. Again, this makes sense, as this α-β combination represents training "RBC models" to focus on $\mathcal{C}^{\mathrm{R}}$ but then attempting to select models based on AUROC. This

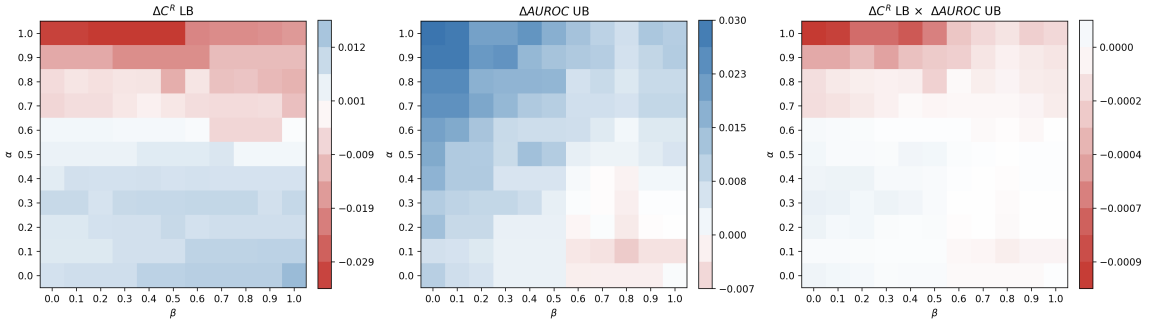

Figure 14: Details of α-β Combinations Showing Improvement. In the left panel, we show the 95% confidence interval lower bound for $\Delta\,\mathcal{C}^{\mathrm{R}}$. The blue areas represent α-β combinations that yield a statistically significant improvement (*i.e.*, $\Delta\,\mathcal{C}^{\mathrm{R}} > 0$). In the middle panel, we show the 95% confidence interval upper bound for Δ AUROC. The blue areas represent combinations without statistically significant degradation (*i.e.*, Δ AUROC ≥ 0). In the right panel, we show the product of the two previous panels to show how we arrived at the above results.

training-selection discrepancy would disadvantage the "RBC models" in terms of AUROC. Thus, when we overlay these areas of interest, we see that we generally tend to observe statistically significant improvements in $\Delta\,\mathcal{C}^{\mathrm{R}}$ that come without an AUROC cost in the region of low α and low β (*i.e.*, $\alpha \leq 0.6$ and $\beta \leq 0.5$). Notably, this region aligns with model developers seeking to emphasize compatibility as a part of their updated model development process.

C.6. $\mathcal{C}^{\mathrm{BT}}$ Across Thresholds

As discussed in **Section 2.2**, $\mathcal{C}^{\mathrm{BT}}$ depends on setting a decision threshold for each model in the model-pair. To help contextualize how various thresholds impact $\mathcal{C}^{\mathrm{BT}}$ we conduct an additional analysis of the main experiment discussed in **Section 4.3**. In this experiment, we sweep the thresholds for the original model, τ^o, and the updated model, τ^u, and find the maximum achievable $\mathcal{C}^{\mathrm{BT}}$ over all of the "BCE models". This analysis can be used to observe $\mathcal{C}^{\mathrm{BT}}$ across multiple thresholds, as per Wynants et al. (2019).

For each replication, we swept both τ^o and τ^u and selected the updated "BCE model" that maximized the validation $\mathcal{C}^{\mathrm{BT}}$. We then computed the $\mathcal{C}^{\mathrm{BT}}$ on the held-out evaluation dataset for the selected updated "BCE model" given the two threshold values and the original model. In **Figure 15**, we show the accuracy of each model, and in **Figure 16**, we show the mean evaluation $\mathcal{C}^{\mathrm{BT}}$ for each τ^o-τ^u pair.

We note that many of τ^o-τ^u pairs corresponding to model-pairs with good accuracy (*i.e.*, $\tau^o, \tau^u \geq 0.1$) yield good maximum achievable $\mathcal{C}^{\mathrm{BT}}$ values ($\mathcal{C}^{\mathrm{BT}}(f^o, f^u) > 0.9$). An area of interest is where τ^o is very low (between 0 and 0.1) and where τ^o is low, and τ^u is very low ($0.1 \leq \tau^o \leq 0.2$ and $0 \leq \tau^u \leq 0.1$). We show fine-grained results for this area in **Figure 17**. In this detailed view, we see that poor $\mathcal{C}^{\mathrm{BT}}$ values ($\mathcal{C}^{\mathrm{BT}}(f^o, f^u) < 0.5$) are achieved under two conditions. The first is when $\tau^o = 0$ and $\tau^u \geq 0.35$. In this case, the

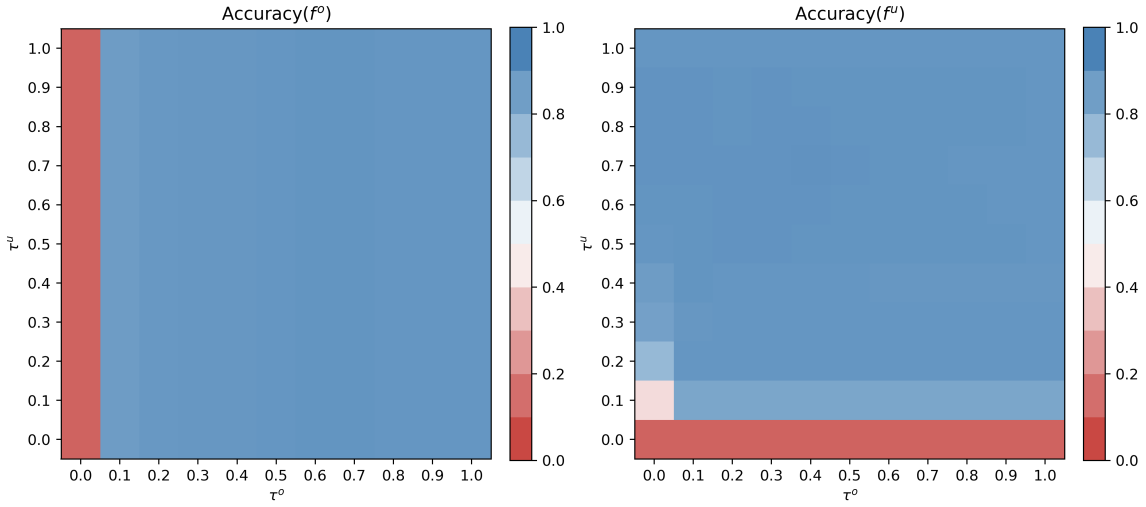

Figure 15: Held-out Evaluation Accuracy for τ^o-τ^u Pairs. Original models display good accuracy ($Accuracy(f^o) > 0.7$) when $\tau^o > 0.1$. Updated models display good accuracy when $\tau^u > 0.1$.

$\mathcal{C}^{\text{BT}}$ value decreases as τ^u increases. The second is when $\tau^o \geq 0$ and $\tau^u \leq 0.01$. In this case, lower τ^u values correspond with lower $\mathcal{C}^{\text{BT}}$ values.

These extreme case threshold values cause the models to tip their classification balances against one another, *e.g.*, the original model and decision threshold may label everyone as 0, and the updated model and threshold label everyone as 1. We note that the regions of large variation in $\mathcal{C}^{\text{BT}}$ only occurs in areas where one of the models has bad accuracy and are an empirical observation of the illustrative example depicted in **Appendix Section A.1**. Additionally, the values in these figures cannot be directly compared to the $\mathcal{C}^{\text{R}}$. However, they underscore the dependence of $\mathcal{C}^{\text{BT}}$ on decision thresholds.

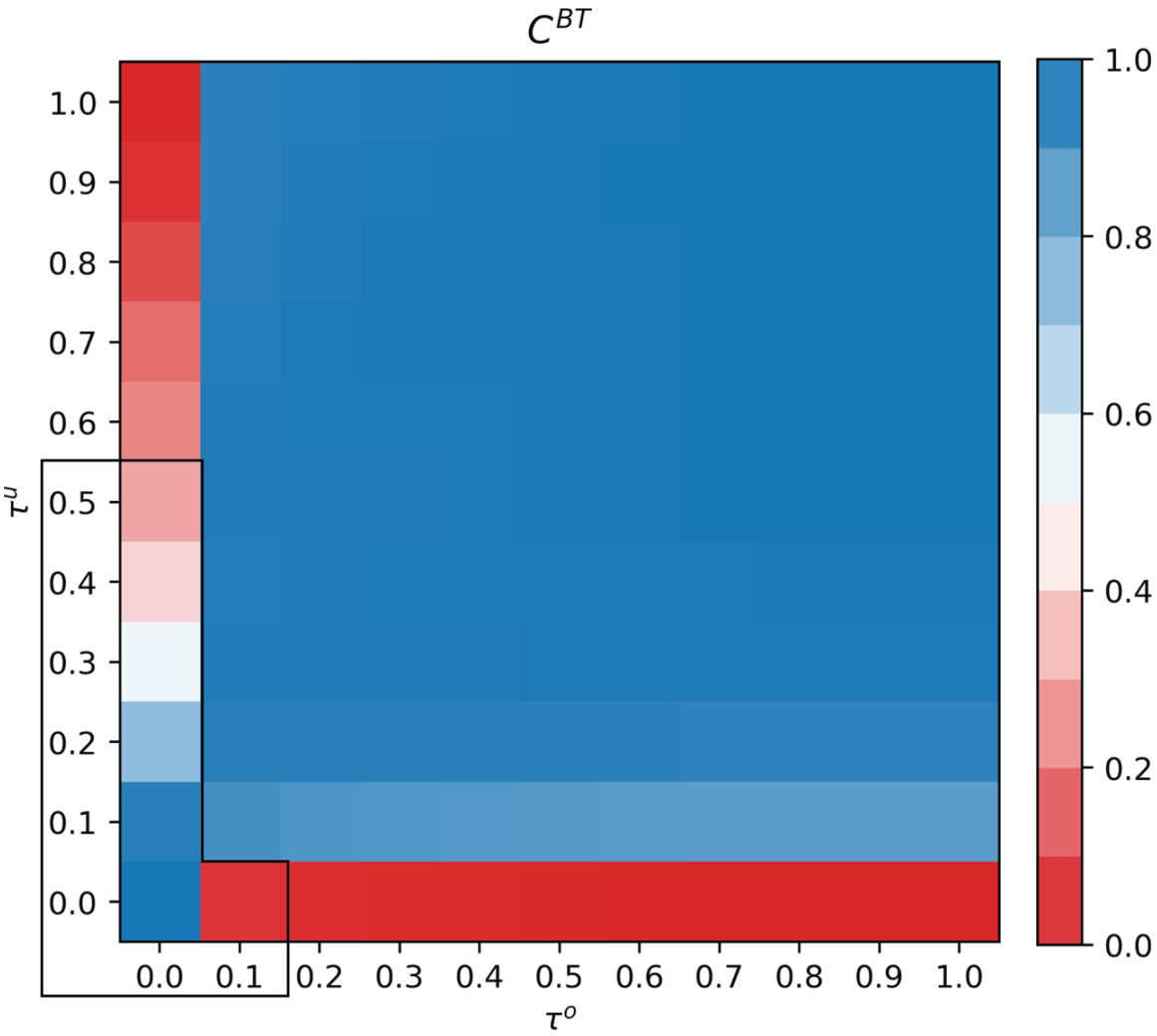

Figure 16: Mean Maximum Achievable Held-out Evaluation $\mathcal{C}^{\mathrm{BT}}$ for τ^o-τ^u Pairs. The majority of model-pairs yield good maximum achievable $\mathcal{C}^{\mathrm{BT}}$ values. The boxed area denotes an area of poor performance, depicted in detail in **Figure 17**.

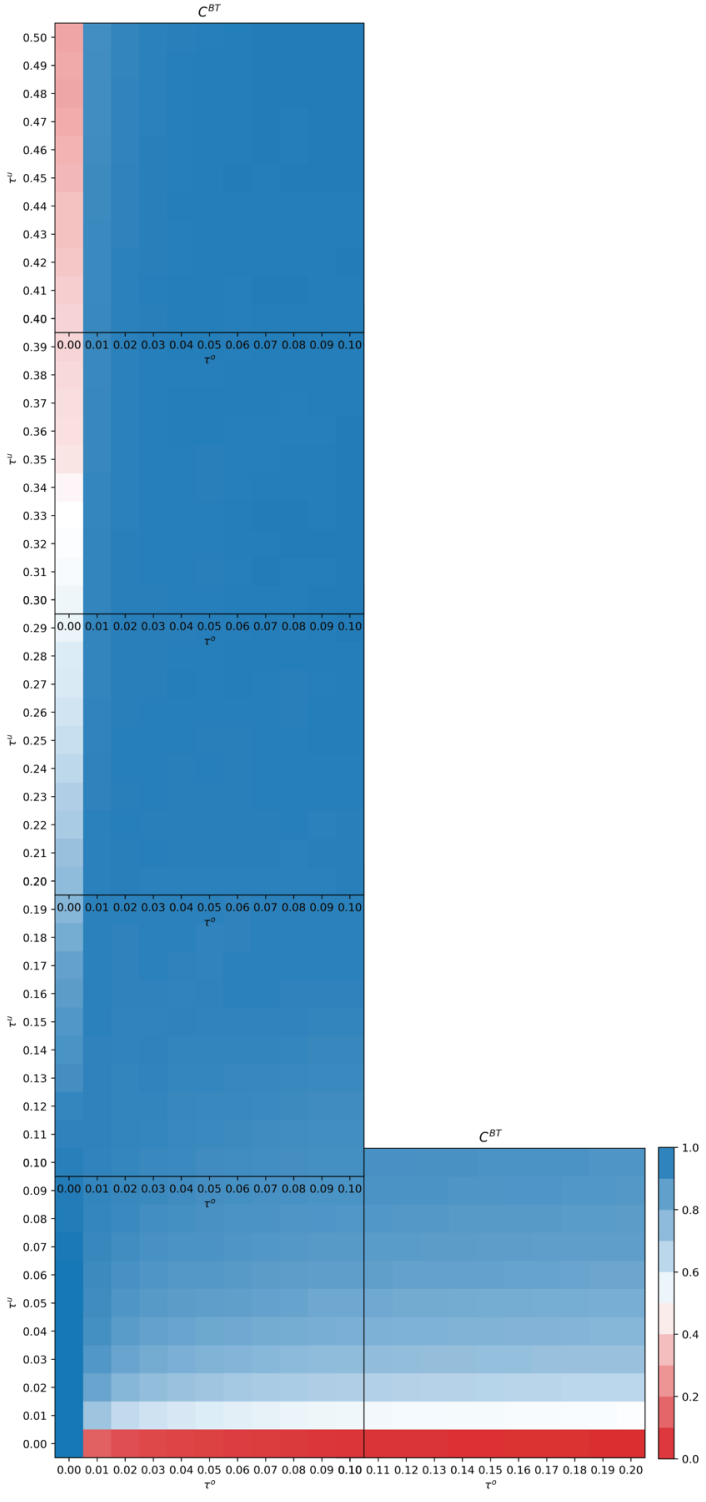

Figure 17: Detailed View of Mean Maximum Achievable Held-out Evaluation $\mathcal{C}^{\mathrm{BT}}$ for τ^o-τ^u Model-Pairs.