LEVERAGING SHARED FEATURE REPRESENTATION IN CROSS-DOMAIN ALIGNMENT OF DECISION THRESHOLDS FOR ELECTRONIC HEALTH RECORDS DATA

Elena Gal, Anshul Thakur, & Soheila Molaei
Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK
{elena.gal,anshul.thakur,soheila.molaei}@eng.ox.ac.uk

Andrew A.S.Soltan
Oxford University Hospitals NHS Foundation Trust, UK
Department of Oncology, University of Oxford, UK
Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK
andrew.soltan@oncology.ox.ac.uk

David A. Clifton
Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK
Oxford-Suzhou Institute of Advanced Research (OSCAR), Suzhou, China
david.clifton@eng.ox.ac.uk

ABSTRACT

The real-world deployment of clinical machine learning models requires adaptability to distributional shifts caused by variations in the patient population and data acquisition mechanisms. However, distributional shifts are known to significantly affect the raw probability scores output by deep learning models and thus compromise performance on clinically important metrics when a threshold must be chosen to generate the final output. We propose a generative learning-based method for threshold setting that utilises unlabelled samples from the target distribution to learn shared feature representation, reduce the distance between domains and improve cross-domain alignment of output probabilities. We demonstrate that the proposed method improves the alignment of decision thresholds for several clinical tasks on real-world electronic health record (EHR) data and derive theoretical bounds on calibration error. Our approach doesn't require ground-truth labels for target data, facilitating its use in EHR-based applications.

1 INTRODUCTION

Machine Learning (ML) models are increasingly used in healthcare, in particular for the analysis of Electronic Health Care Records (EHRs). The availability of large volumes of EHR data offers promising potential for applying ML as a tool for clinical decision support. The incorporation of such models into clinical workflows holds the promise of making a major positive impact by reducing clinical mistakes and alleviating workforce shortages.

However, robustness under domain shift remains a fundamental challenge for successful integration of such models into clinical practice. Domain shift occurs when the "source" data samples used for training and the "target" data samples follow different distributions. In healthcare applications this can be caused by changes in patient population, disease management protocols and medical technology over time, as well as the requirement for the ML models to be portable between different hospital sites with the associated changes in disease prevalence and treatment patterns, such as e.g. procedures, tests and medication ordering habits and site policies (Nestor et al., 2019; Zhang et al., 2022; Kompa et al., 2021). Despite the practical importance of ML models for EHR data, behavior of such models under distributional shift is relatively understudied compared to e.g. imaging data (Avati et al., 2021). In this work we aim to partially address this gap by studying the problem of

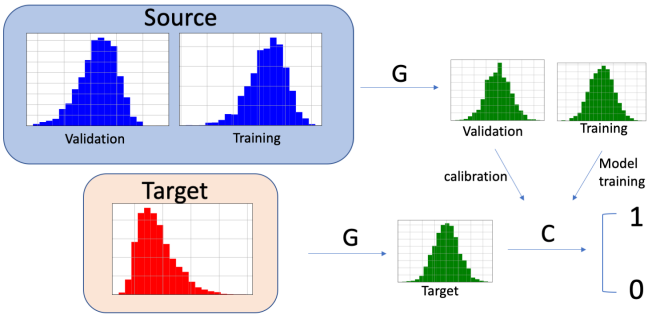

Figure 1: Transformation G reduces the distance between the distributions of source and target samples. The transformed validation subset can now be used for training the classifier C, calibrating the output probabilities and setting thresholds on the target domain.

decision threshold setting and calibration for these models and proposing a practical approach based on generative learning to improve performance in the presence of domain shift.

Although AUC ROC is considered a “gold standard” evaluation criterion and is often used exclusively for evaluating and ranking performance of clinical machine learning models, deployment of models, such as neural networks, in practice requires setting of an *operating point* or a *threshold*. The optimal threshold is decided by taking into account the requirements of the specific problem, and can be dictated by class imbalance, required sensitivity level and feasibility and cost considerations. For example, Youden Index threshold (Youden, 1950) that optimizes the balance between sensitivity and specificity while giving same weight to both metrics is often used in the literature.

When source and target data are drawn from the same distribution, the threshold can be found using labelled validation subset of the source data, however, in the presence of distributional shift, optimization using the subset of the source data will not provide optimal operating point for the target. Using labelled data samples from the target domain is the straightforward alternative, but obtaining such samples in clinical practice is time-consuming, costly and associated with ethical and compliance issues such as the need for informed consent and privacy. As a consequence in some cases such labelled samples might not be available at all, or not be available in the quantity sufficient to represent a true distribution of the target data.

In many situations where collection of labelled data samples is unfeasible, *unlabelled* samples might be available. The intuition behind the use of unlabelled samples to improve performance in the presence of domain shift is that they contain information that can be used to compare the source and target distribution, and hence such information can be used to "adjust" the classification. To implement this idea in practice, we propose to use a generative modeling based approach to learn a transformation of source and target data features that reduces the distance between distributions. The transformed validation subset is used to set thresholds for the target domain (Figure1).

The situation is similar for model *calibration*. The model is said to be calibrated if the predicted probabilities coincide with empirical frequency of the labels. Model calibration was empirically shown to be affected by distributional shift (Ovadia et al., 2019). Our analysis in Section 4 using the theoretical framework introduced in (Ben-David et al., 2006; 2010) shows that the target calibration error of the model is bounded by the distance between source and target domains. This provides a formal theoretical foundation to our approach of reducing this distance by learning a shared representation of source and target data features for better calibration and threshold setting on the target domain.

We conduct experiments on two large real-world Electronic Health Records (EHR) datasets: The eICU Collaborative Research Database (Pollard et al., 2019) containing records from ICU stays in 208 US hospitals and the dataset containing emergency admission records from several UK hospitals during the COVID-19 pandemic that was used in the development of the CURIAL model (Soltan et al., 2022) - a clinical model for early identification of COVID-19 cases among patients presenting to hospitals emergency departments. Our experimental results show strong improvement in performance when the operating point is chosen using the transformed validation data.

2 PREVIOUS WORK

The work of Ben-David et al. (2006; 2010) introduces the framework of empirical estimation of distance ("divergence") between distributions using finite samples. In the present work we use this notion to show that estimated calibration error (ECE) (Naeini et al., 2015) is bounded by the distance between domains, providing theoretical foundations for our approach.

We utilize adversarial domain adaptation approach proposed in (Ganin et al., 2016) for finding distance decreasing transformation of source and target data in practice. This approach is based on the generative adversarial network (GAN) approach of Goodfellow et al. (2014). It uses a discriminator network to measure the divergence between feature representations of source and target distributions. While adversarial domain adaptation has been previously leveraged for machine learning applied to EHR data to ameliorate the drop in performance due to distributional shift as measured by ROC AUC (Purushotham et al., 2016), its effect on the calibration of the probabilities output by the model and on decision threshold setting was not previously studied. Our work is the first to investigate the use of these methods for threshold setting in the presence of distributional shift.

The challenge of threshold setting is of substantial practical significance, however the approaches in the literature are often dataset specific and the research on general approaches is scarce (Hernández-Orallo et al., 2012; Zou et al., 2016; Johnson & Khoshgoftaar, 2021), especially for distributional shift. Of note is the recent work of Roschewitz et al. (2023) that proposes an unsupervised prediction alignment (UPA) method for threshold setting in medical imaging data under acquisition domain shift. UPA applies 'histogram matching' to the models probability outputs on source and target domains. The main limitation of this method is that it requires similar prevalence of positive and negative cases across domains. We apply UPA in the EHR setting and compare it to our approach for threshold setting in our experimental evaluation in Figures 3 and 6.

While a variety of calibration methods exist in the literature, including Bayesian approaches (Kingma et al., 2015; Blundell et al., 2015; Gal & Ghahramani, 2016; Kendall & Gal, 2017) and bootstraping and ensembling methods (Osband et al., 2016; Lakshminarayanan et al., 2017) the *post-hoc* calibration methods involving re-calibration of probabilities on a held-out validation set are more prevalent in healthcare models, due to transparency and simplicity of implementation. Several previous works proposed to combine adversarial domain adaptation with temperature scaling calibration method of Guo et al. (2017) to improve performance in the presence of covariate shift (Wang et al., 2020; Park et al., 2020; Pampari & Ermon, 2020). In the binary classification setting that we consider in this paper the histogram calibration method of Zadrozny & Elkan (2001) is often superior to temperature scaling (which in this setting is equivalent to the Platt method (Platt et al., 1999) without the bias term). Our work is the first to evaluate the performance of this method when combined with adversarial domain adaptation. Moreover the above works do not provide theoretical analysis or bounds, whereas the explicit behavior of the histogram calibration method with respect to the source and target datasets allows us to provide theoretical analysis and guarantees in Section 4.

3 PRELIMINARIES

Formal Setup Let X denote the input space. We will consider the task of binary classification, i.e. label space $Y = \{0, 1\}$. Denote by $\mathcal{D}, \mathcal{D}'$ the distribution from which we sample the *source* and the *target* datasets respectively and let $f_{\mathcal{D}}, f_{\mathcal{D}'} : X \to Y$ be the corresponding labelling functions. We are provided with a labelled sample $\mathcal{S}$ drawn i.i.d according to $\mathcal{D}$ that is used for training of a probabilistic classifier C. We want to analyse threshold optimisation and calibration of C for a sample $\mathcal{T}$ drawn according to $\mathcal{D}'$.

For an input $\mathbf{x}$ the classification procedure has the two following stages:

- C produces classification probability $p(\mathbf{x})$ as an output of its penultimate layer.
- The binary classification outcome $y \in \{0, 1\}$ is computed according to the rule $p(\mathbf{x}) \leq, > A$, where $A \in [0, 1]$ is the chosen threshold.

Calibration error The classifier C is *well-calibrated* when the prevalence of positively labelled instances among the datapoints $\mathbf{x}$ in the test set with predicted probabilities $p(\mathbf{x})$ is (approximately)

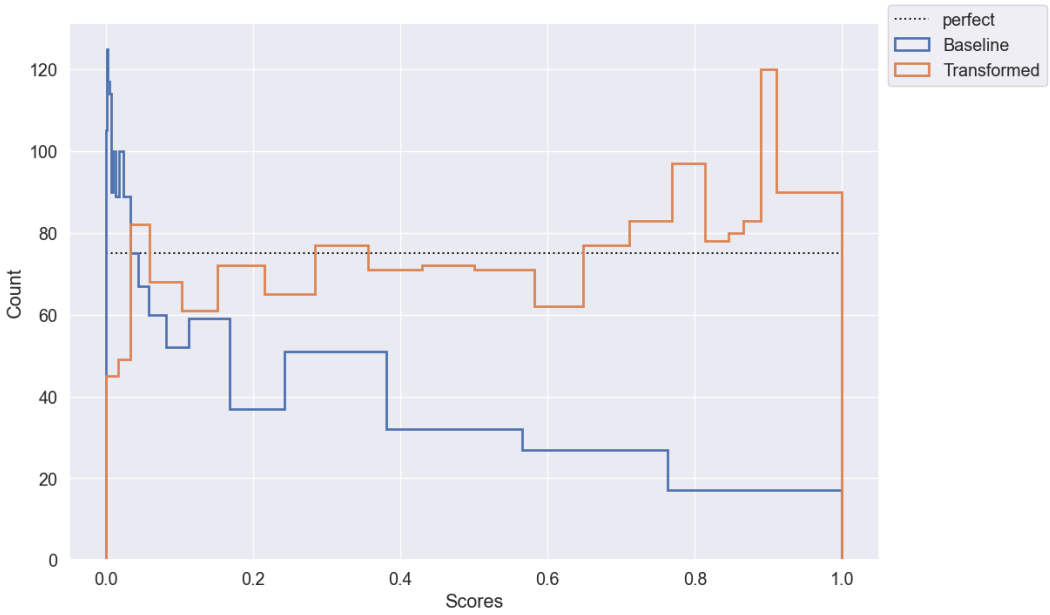

Figure 2: Distribution of predicted probabilities on the target domain after calibration on original and transformed validation sets. The *width* of the bins was determined on the respectively original and transformed validation set so that points were distributed uniformly across bins. The classifier calibrated on the original validation dataset places an excessive amount of target samples into the bins with lowest probabilities, while after calibration on transformed dataset the distribution of probabilities remains close to uniform distribution (indicated by dotted line). In this example the probabilities are output by a classifier trained on the eICU Mortality task and deployed on a new hospital site.

equal to $p(\mathbf{x})$, i.e.

$$P(y = 1|p(\mathbf{x}) = p) \approx p$$

To quantify calibration the following notion of binary expected calibration error (Binary-ECE)(Naeini et al., 2015; Roelofs et al., 2022) is commonly used in the literature:

Definition 3.1. Estimated calibration error is an average gap across all bins in a reliability diagram relative to an ideal reliability diagram weighted by the size of each bin:

$$\text{ECE} = \sum_{i=1}^{M} \frac{|B_i|}{N} |\overline{y}(B_i) - \overline{s}(B_i)| \tag{1}$$

where $\overline{y}(B_i)$ is the proportion of positives in the bin B_i and $\overline{s}(B_i)$ is the average predicted probability, N, M are the total numbers of instances and bins respectively.

The *post-hoc* calibration methods adjust the probabilities of a *trained classifier* using a labelled validation subset of the training set $\mathcal{S}$. One of the simplest methods is *histogram binning* proposed in Zadrozny & Elkan (2001). It is often used in practical applications due to its relative simplicity and competitive performance (Gupta & Ramdas, 2021). In this method the validation dataset is used to divide the interval $[0, 1]$ into a number of bins $\{B_i\}_{i=0}^{M}$. Each point in the training dataset is assigned to a bin according to its predicted probability $p(\mathbf{x})$. The sizes of the bins are chosen so that they contain approximately equal number of data points. Computing actual number of positives in each bin allows to learn calibration coefficients for the classifier and minimize the ECE (Naeini et al., 2015).

As in the case of threshold optimisation, classifiers calibrated in such a way can be expected to stay well-calibrated on the target $\mathcal{T}$ if both the validation and the target are sampled from the same distribution, but not otherwise. In particular, in the presence of distributional shift between the source

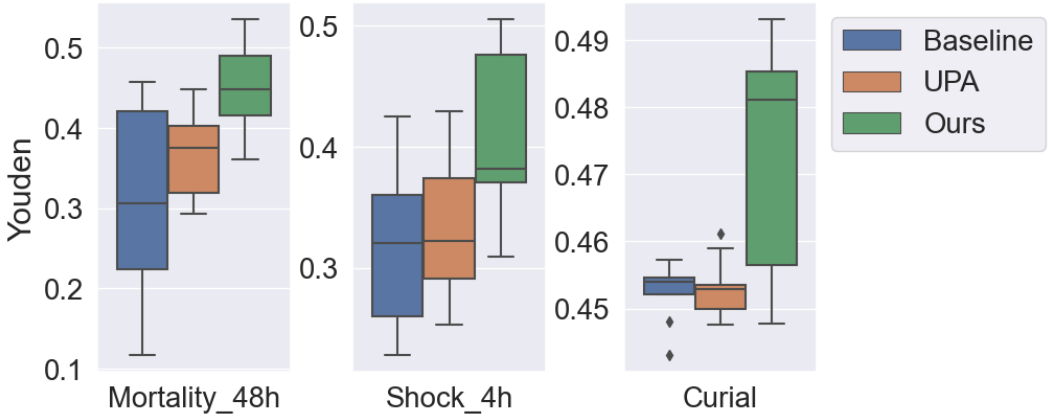

Figure 3: We calibrate threshold to maximise Youden index on the validation dataset with ("Ours") and without("Baseline") transforming the validation dataset, as well as using UPA approach from (Roschewitz et al., 2023) ("UPA"). The plot shows resulting Youden index distributions across several *target* hospital sites on the eICU datasets for Mortality within 48 hours and Shock within 4 hours prediction tasks and on the CURIAL dataset. Each box shows the quartiles summarizing the results across all target sites and the error bars indicate the minimal and maximal values across the experiments. We observe that calibration on transformed validation dataset significantly outperforms the other approaches on all datasets.

and the target domain we can no longer expect that the bins found using the validation subset of the source domain will contain similar numbers of points of the target domain, since the distribution of the probabilities predicted by the classifier can change due to covariate shift. The actual number of positives in each bin can change as well due to label shift. See Figure 2 for an example of distribution of predicted probabilities on the target domain for the case of classifier trained to predict adverse events from electronic health records of hospital patients that is deployed on a previously unseen hospital site.

We analyse how the distance between source and target distributions affects ECE in the following section.

4 Theoretical Analysis

4.1 Notions of Distance Between Domains

The key notion we will use in our analysis of threshold optimisation and calibration in the presence of distributional shift is the measure of the difference between $\mathcal{D}$ and $\mathcal{D}'$. The concept of *$\mathcal{H}$-divergence* based on the L_1 distance between the distributions, introduced in (Kifer et al., 2004; Ben-David et al., 2010) is commonly used in the literature, since it can be estimated from finite samples and allows for focusing on a certain relevant class of *hypothesis* functions $\mathcal{H} := \{h : X \to Y\}$.

Definition 4.1. $\mathcal{H}$-divergence between $\mathcal{D}, \mathcal{D}'$ for the class of hypothesis functions $\mathcal{H}$ is given by the following formula:

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') := 2 \sup_{h \in \mathcal{H}} \left| \Pr_{\mathcal{D}} I(h) - \Pr_{\mathcal{D}'} I(h) \right|$$

where $I(h)$ is the set for which $h \in \mathcal{H}$ is the characteristic function, i.e. $x \in I(h) \iff h(x) = 1$.

Ben-David et al. (2010) show that $\mathcal{H}$-divergence can be approximated by the *empirical $\mathcal{H}$-divergence* between (series) of samples $\mathcal{S}$ and $\mathcal{T}$ from $\mathcal{D}, \mathcal{D}'$:

$$\hat{d}_{\mathcal{H}}(\mathcal{S}, \mathcal{T}) := 2 \sup_{h \in \mathcal{H}} \left| \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} h(\mathbf{x}) - \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} h(\mathbf{x}) \right|$$

The following observation based on Ben-David et al. (2010) allows to compute *empirical $\mathcal{H}$-divergence* using machine learning methods in practice:

Remark 4.2. For a symmetric hypothesis class $\mathcal{H}$, i.e. one for which $h \in \mathcal{H} \implies 1 - h \in \mathcal{H}$ the bound on $\mathcal{H}$-divergence can be estimated from the accuracy of the classifier solving the binary classification problem of distinguishing source and target instances. The divergence which is close to 0 corresponds to the accuracy close to 0.5 and the divergence close to 2 corresponds to the accuracy close to 1, as intuitively expected.

In the following section we will show that smaller $\hat{d}_{\mathbf{H}}$ leads to smaller change in estimated calibration error when passing from $\mathcal{S}$ to $\mathcal{T}$, i.e. for the close enough samples $\mathcal{T}$ and $\mathcal{S}$ the calibration error does not increase significantly on the target.

4.2 Bounds on Estimated Calibration Error with Distributional Shift

Consider a classifier C (and a corresponding hypothesis h_C) with given estimated calibration error (ECE, Definition 1) on a sample $\mathcal{S}$. We would like to analyze ECE on the sample from the target domain $\mathcal{T}$ in terms of empirical $\mathcal{H}$-divergence between the samples $\mathcal{S}$ and $\mathcal{T}$. To compute ECE on $\mathcal{T}$ we need to divide it into the same number of equally sized bins as the source. The factors affecting ECE are the resulting number of datapoints in each bin and the actual proportion of positives in each bin.

The following result provides an upper bound for absolute difference in the amount of data points placed in the bins predicted **by any classifier** C in terms of divergence between $\mathcal{S}, \mathcal{T}$.

Theorem 1. *For any classifier C the absolute difference in the proportion of points contained in bins with same predicted probability range on $\mathcal{S}$ and $\mathcal{T}$ is bounded by the empirical $\mathcal{H}$-divergence between samples $\mathcal{S}$ and $\mathcal{T}$ for big enough hypothesis class $\mathcal{H}$.*

The *actual proportion of positives* in bins with corresponding probability ranges on $\mathcal{S}$ and $\mathcal{T}$ is affected not just by the distance between the domains but also by the difference between labelling functions $f_{\mathcal{D}}, f_{\mathcal{D}'} : X \to Y$. To quantify this difference via the functions in the hyphothesis class $\mathcal{H}$ Ben-David et al. (2010) propose the notion of the *ideal joint hyphothesis*.

Definition 4.3. The *risk* of a hyphothesis h for $\mathcal{D}$ is

$$\epsilon_{\mathcal{D}}(h, f_{\mathcal{D}}) = E_{\mathbf{x}\sim\mathcal{D}} \left| h(x) - f(x) \right|$$

Definition 4.4. The empirical *risk* of a hypothesis h on a sample $\mathcal{S}$ is

$$\hat{\epsilon}_{\mathcal{S}}(h, f_{\mathcal{D}}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \left| h(x) - f(x) \right|$$

Definition 4.5. The ideal joint hypothesis is the hypothesis which minimizes the combined error

$$h^* = \arg\min_{h \in \mathcal{H}} \epsilon_{\mathcal{S}}(h, f_{\mathcal{D}}) + \epsilon_{\mathcal{T}}(h, f_{\mathcal{D}'})$$

Denote the error of the ideal joint hypothesis h^* by λ. For any pair of samples from $\mathcal{D}, \mathcal{D}'$ we can also consider empirical error $\hat{\lambda}$.

Then we have the following

Theorem 2. *Given a classifier C, consider bins B_i and B_i' containing the points with predicted probabilities in the range $[p_1, p_2]$ of $\mathcal{S}$ and $\mathcal{T}$ respectively. The absolute difference between proportion of positively labelled instances in B_i and B_i' is bounded as follows*

$$\sum_{\mathbf{x} \in \mathcal{S}} \left| \frac{1}{|B_i|} f_{\mathcal{D}}(\mathbf{x}) - \frac{1}{|B_i|} f_{\mathcal{D}'}(\mathbf{x}) \right| \leq \hat{d}_{\mathcal{H}} + \hat{\lambda}$$

where $\hat{d}_{\mathcal{H}}$ is the empirical $\mathcal{H}$-divergence between samples $\mathcal{S}$ and $\mathcal{T}$ and $\hat{\lambda}$ is the empirical combined error.

The intuition behind this result is that for "similar" source and target domains, the labelling for any two probability bins chosen with **arbitrary** classifier is related to similarity of labelling for the domains overall.

For proofs see Appendix.

5 Proposed Method

The theoretical analysis of the previous section suggests the following method for calibration and threshold setting: given samples $\mathcal{S}, \mathcal{T}$, find a feature transformation $G : X \to Z$ that reduces the empirical divergence $\hat{d}_{\mathcal{H}}$ between $G(\mathcal{S}), G(\mathcal{T})$, and use the transformed samples $G(\mathcal{S}), G(\mathcal{T})$ for training and calibrating the classifier. If the empirical divergence between $G(\mathcal{S}), G(\mathcal{T})$ is small the classifier calibrated on a subset of $G(\mathcal{S})$ will stay well-calibrated on $G(\mathcal{T})$ (Figure 1).

In practice the labelled source sample $\mathcal{S}$ and the target sample $\mathcal{T}$ from the distributions $\mathcal{D}, \mathcal{D}'$ on the input space X are given and empirical $\mathcal{H}$-divergence between them can be estimated using Remark 4.2. We need to find a transformation G that makes the two distributions similar while preserving the information needed to approximate the pullbacks of labelling functions $f_{\mathcal{D}}, f_{\mathcal{D}'}$ with respect to G. More concretely, we want to find $G : X \to Z$ and a hypothesis h on Z, such that *both* the empirical divergence $\hat{d}_{\mathcal{H}}$ between $G(\mathcal{S}), G(\mathcal{T})$ and the empirical *hypothesis risk* are small. Formally we have

Definition 5.1. Let $G : X \to Z$. We can define the distribution $\mathcal{D}_G$ and the labelling function $\tilde{f}$ by

$$\Pr_{\mathcal{D}_G}[B] = \Pr_{D}\left[G^{-1}(B)\right]$$
$$\tilde{f} = \mathbb{E}_{\mathcal{D}}\left[f(\mathbf{x}) | G(\mathbf{x}) = \mathbf{z}\right]$$

Definition 5.2. Let $G : X \to Z$. The hypothesis risk for the transformed source distribution $\mathcal{D}_G$ and a hypothesis h over Z is given by

$$\epsilon_{\mathcal{D}_G}(h) = \mathbb{E}_{\mathbf{z} \sim \mathcal{D}_G}\left|h(z) - \tilde{f}(x)\right|$$

The empirical hypothesis risk for a sample is defined similar to Definition 4.4.

To learn the distance reducing transformation G we use adversarial domain adaptation. Adversarial domain adaptation is a group of approaches that utilizes a GAN methodology for finding transformation G and hypothesis h that minimize the distance between distributions and the empirical risk by approximating them with neural networks G_θ (a "generator") and C_ψ (a "a classifier"). These networks are trained on the labelled source and the unlabelled target samples. A third neural network D_ϕ (a discriminator) empirically approximates the distance between the transformed source and target samples (see Remark 4.2). These networks are trained using a shared training objective to simultaneously maximize classification accuracy while minimizing distance between transformed distributions.

After learning G the transformed validation set $G(V)$ is used for calibration and thresholds setting.

6 Experiments

We compare performance of the baseline classifier trained on the original source data by minimising standard cross entropy loss with the performance of the classifier trained on the transformed source data using the training objective 3. We also compare to the UPA method. We consider performance at the threshold maximising the Youden's index Youden (1950)

$$\mathcal{J} = \text{Sensitivity} + \text{Specificity} - 1$$

We find an optimal threshold on an unseen validation subset of $\mathcal{S}$ and compare performance on the target domain $\mathcal{T}$ (Figure 3).

We conduct 5 experiments with random sampling of training and validation datasets for each target site and verify significance of the results by computing t-values for each site.

We report results for additional thresholds and results on calibration in the Appendix.

eICU is a multi-site database comprising de-identified health data for patients admitted to ICUs across different sites in the US (Pollard et al., 2018), (Pollard et al., 2019) that is a part of PhysioNet (Goldberger et al., 2000). The feature representation used in the experiments was created using the feature extraction method FIDDLE (FlexIble Data-Driven pipeLinE), an open-source preprocessing pipeline for structured clinical data (Tang et al., 2020), (Tang et al., 2021). For our experimental evaluation we

considered the "Mortality within 48 hours" and Shock within 4 hours binary prediction tasks. Each task corresponds to a time series dataset.

Outcome distributions, population characteristics and data collection methods vary widely across different clinical sites leading to domain shifts between different sites (Zhang et al., 2022). To increase robustness and transferability a model is often trained on a diverse dataset collected from multiple hospitals rather than training and applying individual models. In keeping with this approach, our source dataset was comprised of labelled data pooled across a subset of several hospitals. We considered 3 target datasets for each of the tasks, containing data from hospitals not belonging to the above subset. This setup mimics the plausible real-world scenario where a diagnostic system is trained on a body of data across several sites and deployed on a new unseen site. Class imbalance is present in both source and target datasets. Positive class comprised 20% of the source data for Mortality and Shock tasks. For target sites the percentage of positives varied between 5% and 16%.

The dataset used for the creation of the CURIAL model Soltan et al. (2022) contains EHR data from several UK National Healthcare Service trusts. In our experiments we have used the data from the Oxford University Hospitals (OUH) trust collected during the first wave of the Covid pandemic as described in Soltan et al. (2022) as the source training data for the creation of our baseline model. The source dataset contains approximately 5% positive cases. The data from OUH second wave, as well as the data from the University Hospitals Birmingham (UHB) NHS Foundation Trust was used as target data. The target sites contain 10% and 6% positive cases respectively. [1] For details on models and training see Appendix.

7 Limitations and further directions

Our theoretical analysis is necessarily restricted to the case of covariant shift, since we assume no knowledge of labels for the target domain. However it can be modified by combining adversarial domain adaptation with methods that tackle label shift (Garg et al., 2023).

The literature on domain adaptation for time series is relatively scarce (Purushotham et al., 2016; He et al., 2023). In the present work we opted to use a classical approach of DANN Ganin et al. (2016) combined with an LSTM network as feature extractor as in Purushotham et al. (2016). Newer variations of this algorithm have potential to improve performance even further.

Our experiments were limited to EHR datasets. The choice of modality was motivated by the fact that EHR models behavior in the setting of distributional shift is understudied compared to other data modalities (Avati et al., 2021). We believe that applications of our method to calibration on other types of datasets, such as e.g. medical imaging data, merits separate investigation.

Finally we note that we did not discuss potential privacy concerns related to the use of unlabelled target data. Various federated learning approaches exist to overcome this constraint, and we note that adversarial domain adaptation methods can be performed in federated learning mode (Peng et al., 2019), therefore it should be possible to learn distance -reducing transformation G without direct access to target data. We postpone detailed work up of this exciting direction to future work.

Robustness to distributional shifts is one of the key challenges for successful deployment of machine learning models based on electronic health record (EHR) data. To our knowledge our work is the first to propose a systematic approach to threshold setting on the EHR data in the presence of distributional shift. Our method requires only unlabelled target samples, thus avoiding the ethical concerns and significant costs associated with obtaining labelled healthcare data. As such it has potential to improve performance of machine learning models and facilitate their incorporation into clinical practice.

Acknowledgments

DAC is supported by the Pandemic Sciences Institute at the University of Oxford; the National Institute for Health Research (NIHR) Oxford Biomedical Research Center (BRC); an NIHR Research Professorship; a Royal Academy of Engineering Research Chair; the Wellcome Trust; the UKRI; and the InnoHK Hong Kong Center for Center for Cerebro-cardiovascular Engineering (COCHE).

[1] The OUH datasets are available from the Infections in Oxfordshire Research Database, subject to an application meeting the ethical and governance requirements of the database. Data from UHB are available on reasonable request to the trust, subject to NHS Health Research Authority requirements.

REFERENCES

Anand Avati, Martin Seneviratne, Emily Xue, Zhen Xu, Balaji Lakshminarayanan, and Andrew M Dai. Beds-bench: Behavior of ehr-models under distributional shift–a benchmark. *arXiv preprint arXiv:2107.08189*, 2021.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL `https://proceedings.neurips.cc/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf`.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010. URL `http://www.springerlink.com/content/q6qk230685577n52/`.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, jan 2016. ISSN 1532-4435.

Saurabh Garg, Nick Erickson, James Sharpnack, Alex Smola, Sivaraman Balakrishnan, and Zachary Chase Lipton. Rlsbench: Domain adaptation under relaxed label shift. In *International Conference on Machine Learning*, pp. 10879–10928. PMLR, 2023.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Chirag Gupta and Aaditya Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International conference on machine learning*, pp. 3942–3952. PMLR, 2021.

Huan He, Owen Queen, Teddy Koker, Consuelo Cuevas, Theodoros Tsiligkaridis, and Marinka Zitnik. Domain adaptation for time series under feature and label shifts. *arXiv preprint arXiv:2302.03133*, 2023.

José Hernández-Orallo, Peter Flach, and César Ferri Ramírez. A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 13:2813–2869, 2012.

Justin M Johnson and Taghi M Khoshgoftaar. Thresholding strategies for deep learning with highly imbalanced big data. *Deep Learning Applications, Volume 2*, pp. 199–227, 2021.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *VLDB*, volume 4, pp. 180–191. Toronto, Canada, 2004.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Bret Nestor, Matthew BA McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In *Machine Learning for Healthcare Conference*, pp. 381–405. PMLR, 2019.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Anusri Pampari and Stefano Ermon. Unsupervised calibration under covariate shift. *arXiv preprint arXiv:2006.16405*, 2020.

Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3219–3229. PMLR, 2020.

Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. *arXiv preprint arXiv:1911.02054*, 2019.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

T. Pollard, A. Johnson, J. Raffa, L. A. Celi, O. Badawi, and R. Mark. eicu collaborative research database (version 2.0), 2019. URL `https://doi.org/10.13026/C2WM1R`.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Sanjay Purushotham, Wilka Carvalho, Tanachat Nilanon, and Yan Liu. Variational recurrent adversarial deep domain adaptation. In *International Conference on Learning Representations*, 2016.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 4036–4054. PMLR, 2022.

Mélanie Roschewitz, Galvin Khara, Joe Yearsley, Nisha Sharma, Jonathan J James, Éva Ambrózay, Adam Heroux, Peter Kecskemethy, Tobias Rijken, and Ben Glocker. Automatic correction of performance drift under acquisition shift in medical image classification. *Nature Communications*, 14(1):6608, 2023.

Andrew AS Soltan, Jenny Yang, Ravi Pattanshetty, Alex Novak, Yang Yang, Omid Rohanian, Sally Beer, Marina A Soltan, David R Thickett, Rory Fairhead, et al. Real-world evaluation of rapid and laboratory-free covid-19 triage for emergency care: external validation and pilot deployment of artificial intelligence driven screening. *The Lancet Digital Health*, 4(4):e266–e278, 2022.

S. Tang, P. Davarmanesh, D. Song, Y.and Koutra, M. Sjoding, and J. Wiens. Mimic-iii and eicu-crd: Feature representation by fiddle preprocessing (version 1.0.0), 2021. URL `https://doi.org/10.13026/2qtg-k467`.

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.

Ping Wang, Rick Mathieu, Jie Ke, and HJ Cai. Predicting criminal recidivism with support vector machine. In *2010 International Conference on Management and Service Science*, pp. 1–9. IEEE, 2010.

Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable calibration with lower bias and variance in domain adaptation. *Advances in Neural Information Processing Systems*, 33:19212–19223, 2020.

William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pp. 609–616, 2001.

Angela Zhang, Lei Xing, James Zou, and Joseph C Wu. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, 6(12): 1330–1345, 2022.

Quan Zou, Sifa Xie, Ziyu Lin, Meihong Wu, and Ying Ju. Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5:2–8, 2016.

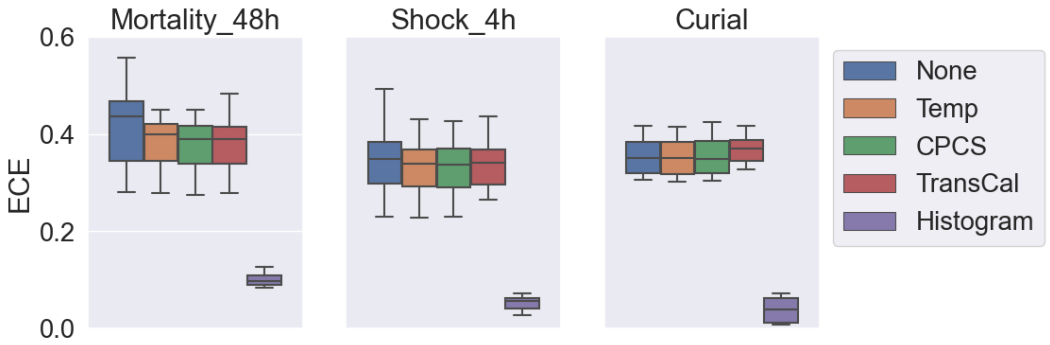

Figure 4: The comparison of ECE↓ distributions for the Temperature scaling, CPCS, TransCal and Histogram post-hoc calibration methods for the adapted model. Blue bar corresponds to the results with no calibration applied. We plot distributions across several *target* hospital sites on the eICU datasets for Mortality and Shock prediction tasks and on the CURIAL dataset. Histogram calibration method significantly outperforms other calibration methods in the setting of adversarial domain adaptation.

A Calibration

It is natural to utilize the transformed validation set $G_\theta(V)$ for calibration of the bins using the histogram binning procedure from Zadrozny & Elkan (2001). We compare the performance of this method to the three benchmark calibration methods that use domain adaptation: Temperature scaling (Guo et al., 2017), CPCS (Park et al., 2020) and TransCal (Wang et al., 2010), and also consider ablation case (performance on target domain without calibration). Figure 4 summarises ECE equation 1 distributions on $\mathcal{T}$ across all experiments and target domains. ECE is calculated by dividing the domain into 15 equal length bins as in Guo et al. (2017); Wang et al. (2020). We also compare Histogram binning with and without calibration in 5

We find that Histogram calibration method has the lowest ECE by a large margin. Interestingly the two benchmark methods developed specifically for the domain adaptation setting, CPCS and TransCal, perform on par or worse than Temperature scaling method in our experiments. A likely reason is that our datasets are highly imbalanced, which impacts both CPCS, which optimizes Brier score (Brier, 1950), known to perform poorly for minority classes, and TransCal, which uses accuracy as one of components of its method on top of Temperature scaling.

B Thresholds with given sensitivity value

Clinical tests are sometimes required to have a certain sensitivity value. For ML models required performance can be achieved by calibrating the threshold on the validation subset, but in the presence of distributional shift the sensitivity on the target domain can drop. We compare performance of baseline model and the models trained using shared feature representation and UPA approach for models trained and deployed on domains with distributional shift in Figure 6.

C Proofs

Theorem 3. *For any classifier C the absolute difference in the proportion of points contained in bins with same predicted probability range on $\mathcal{S}$ and $\mathcal{T}$ is bounded by the empirical $\mathcal{H}$-divergence between samples $\mathcal{S}$ and $\mathcal{T}$ for big enough hypothesis class $\mathcal{H}$.*

Proof. Given a classifier C, consider bins B_i and B'_i containing the points with predicted probabilities in the range $[p_1, p_2]$ of $\mathcal{S}$ and $\mathcal{T}$ respectively. Then clearly there exists a hypothesis $\hat{h}$ which maps the points belonging to B_i, B'_i to 1 and the points outside these to 0 on the samples $\mathcal{S}$ and $\mathcal{T}$ respectively. Hence for any $\mathcal{H}$ containing $\hat{h}$ we have

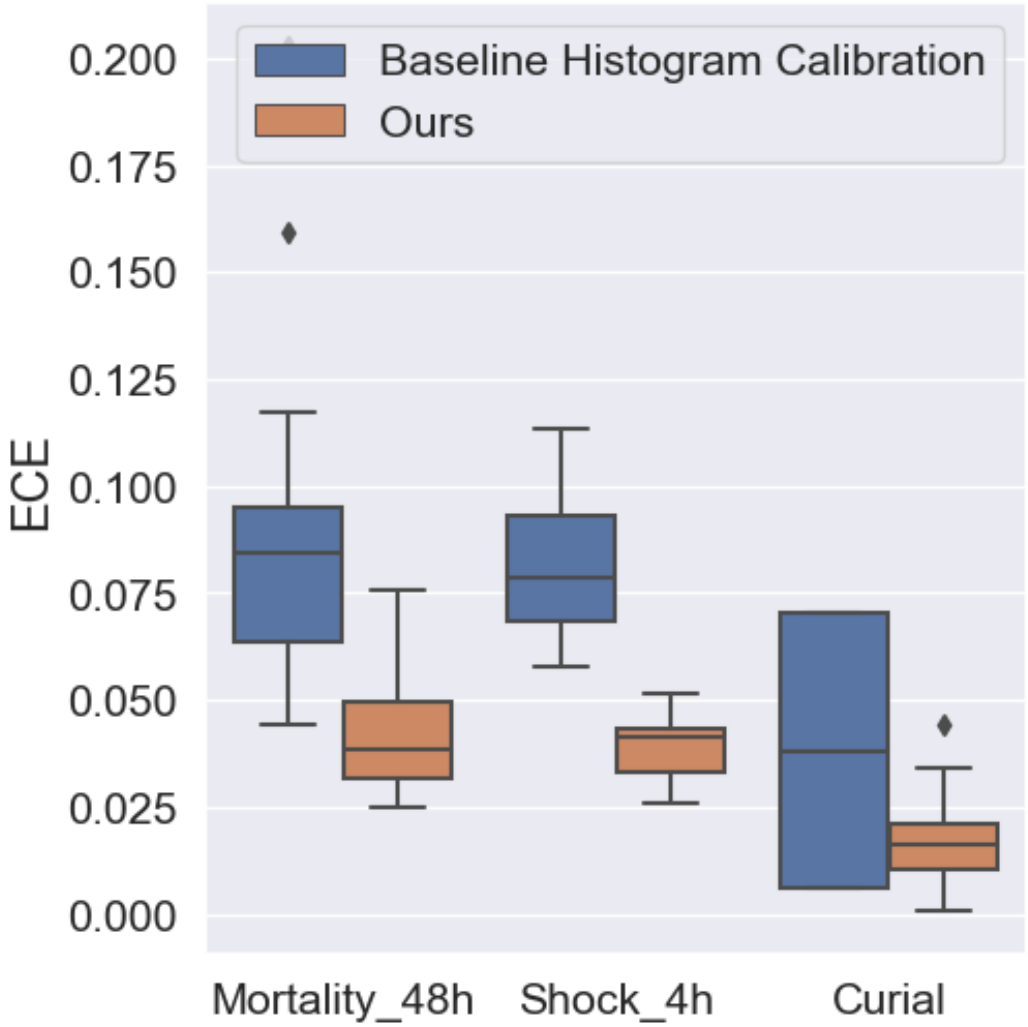

Figure 5: The comparison of ECE↓ distributions after application of histogram binning calibration method using transformed and original validation subset

$$\left|\frac{|B_i|}{|\mathcal{S}|} - \frac{|B_i'|}{|\mathcal{T}|}\right| = \left|\frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} \hat{h}(\mathbf{x}) - \frac{1}{|\mathcal{T}|}\sum_{\mathbf{x}\in\mathcal{T}} \hat{h}(\mathbf{x})\right| \leq$$
$$\sup_{h\in\mathcal{H}}\left|\frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} h(\mathbf{x}) - \frac{1}{|\mathcal{T}|}\sum_{\mathbf{x}\in\mathcal{T}} h(x)\right| = \frac{1}{2}\hat{d}_{\mathcal{H}}(\mathcal{S},\mathcal{T})$$

□

Theorem 4. *Given a classifier C, consider bins B_i and B_i' containing the points with predicted probabilities in the range $[p_1, p_2]$ of $\mathcal{S}$ and $\mathcal{T}$ respectively. The absolute difference between proportion of positively labelled instances in B_i and B_i' is bounded as follows*

$$\sum_{\mathbf{x}\in\mathcal{S}}\left|\frac{1}{|B_i|}f_{\mathcal{D}}(\mathbf{x}) - \frac{1}{|B_i|}f_{\mathcal{D}'}(\mathbf{x})\right| \leq \hat{d}_{\mathcal{H}} + \hat{\lambda}$$

where $\hat{d}_{\mathcal{H}}$ is the empirical $\mathcal{H}$-divergence between samples $\mathcal{S}$ and $\mathcal{T}$ and $\hat{\lambda}$ is the empirical combined error.

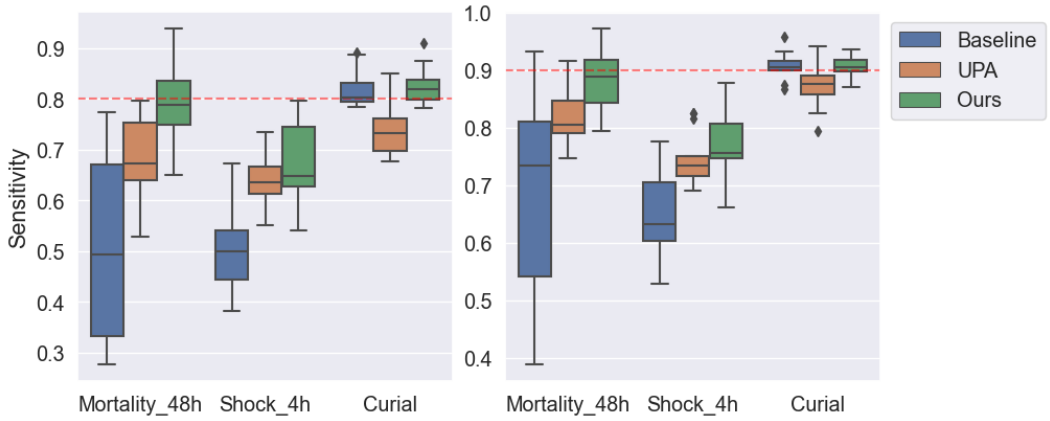

Figure 6: We calibrate the thresholds to achieve respectively 80% and 90% sensitivity on the validation subset of the source dataset. We plot distributions of the sensitivity values across several *target* hospital sites on the eICU datasets for Mortality within 48 hours and Shock within 4 hours prediction tasks as well as the distributions on the CURIAL dataset. Each box shows the quartiles summarizing the results across all target sites and the error bars indicate the minimal and maximal values across the experiments. Note that while the mean sensitivity of the models calibrated on the transformed validation subset for the Shock task is less than required, it is significantly improved compared to the sensitivity of the baseline and UPA models.

Proof. First note that the empirical H-divergence $\hat{d}_{\mathcal{H}}(B_i, B_i')$ between the bins $B_i \subset \mathcal{S}$ and $B_i' \subset \mathcal{T}$ is less or equal to the empirical $\mathcal{H}$-divergence between the whole source and target samples, $\hat{d}_{\mathcal{H}}(\mathcal{S}, \mathcal{T})$. This follows by considering the hypothesis $\hat{h}$ that returns 0 for points outside of B_i, B_i' and 1 otherwise. Then for any $h \in \mathcal{H}$

$$\left| \frac{1}{|B_i|} \sum_{\mathbf{x} \in B_i} h(\mathbf{x}) - \frac{1}{|B_i'|} \sum_{\mathbf{x} \in B_i'} h(\mathbf{x}) \right| =$$
$$\left| \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \hat{h} \circ h(\mathbf{x}) - \frac{1}{|\mathcal{T}|} \sum_{\mathbf{x} \in \mathcal{T}} \hat{h} \circ h(\mathbf{x}) \right| \leq$$
$$\sup_{h \in \mathcal{H}} \left| \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} h(\mathbf{x}) - \frac{1}{|\mathcal{T}|} \sum_{\mathbf{x} \in \mathcal{T}} h(\mathbf{x}) \right| = \frac{1}{2} \hat{d}_{\mathcal{H}}(\mathcal{S}, \mathcal{T})$$

and the claim follows. Then we have for the ideal joint hypothesis h^*:

$$\left| \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} h^*(\mathbf{x}) - \frac{1}{|\mathcal{T}|} \sum_{\mathbf{x} \in \mathcal{T}} h^*(\mathbf{x}) \right| \leq \hat{d}_{\mathcal{H}} \tag{2a}$$
$$\left| \frac{1}{|B_i|} \sum_{\mathbf{x} \in B_i} |h^*(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x})| - \frac{1}{|B_i'|} \sum_{\mathbf{x} \in B_i} |h^*(\mathbf{x}) - f_{\mathcal{D}'}(\mathbf{x})| \right|$$
$$\leq \hat{\lambda} \tag{2b}$$

The equations(2a) and (2b) imply

$$\left| \frac{1}{|B_i|} \sum_{\mathbf{x} \in B_i} f_{\mathcal{D}}(\mathbf{x}) - \frac{1}{|B_i'|} \sum_{\mathbf{x} \in B_i'} f_{\mathcal{D}'}(\mathbf{x}) \right| \leq \hat{d} + \hat{\lambda}$$

□

Algorithm 1 DANN

Input: Labelled batch $\{(x_L^{(i)}, y_L^{(i)}) : i \in (1, \ldots, n_L)\}$**, unlabelled batch** $\{(x_U^{(i)}, y_U^{(i)}) : i \in 1, \ldots, n_U\}$**, penalty** $\lambda \in \mathbb{R}$**, learning rates** η_G, η_C, η_D.
Compute loss for Discriminator D

$$L(D) = \frac{1}{n_L} \sum_{i=1}^{n_L} l(D \circ G(x_L^{(i)}), 1) + \frac{1}{n_U} \sum_{i=1}^{n_U} l(D \circ G(x_U^{(i)}), 0)$$

Compute loss for Classifier C

$$L(C) = \frac{1}{n_L} \sum_{i=1}^{n_L} l(C \circ G(x_L^{(i)}), y_L^{(i)})$$

Compute combined loss

$$L = L(C) - \lambda L(D)$$

Update C, D, G using learning rates η_G, η_C, η_D

D Models and Training

We use adversarial domain adaptation to learn the representation G that reduces the distance between distributions.

Adversarial domain adaptation is a group of approaches for training a triple of neural networks G_θ (a "generator"), C_ψ (a "a classifier") and D_ϕ (a discriminator). The training objective is given by $\min_\theta \min_\psi \max_\phi L_{\theta,\psi,\phi}$, where

$$\begin{aligned} L_{\theta,\psi,\phi} := & \mathop{\mathbb{E}}_{\mathbf{x} \in \mathcal{T}} \log(D_\phi(G_\theta(\mathbf{x})) \\ & + \mathop{\mathbb{E}}_{\mathbf{x} \in \mathcal{S}} \log(1 - D_\phi(G_\theta(\mathbf{x}))) \\ & + L^C_{x \in \mathcal{S}}(C_\psi(G_\theta(\mathbf{x})), y(\mathbf{x})) \end{aligned} \tag{3}$$

In the present work we opted to use a classical approach of DANN Ganin et al. (2016). For the eICU dataset G_θ is an LSTM network acting as a feature extractor Purushotham et al. (2016). On the CURIAL Dataset $G_\theta$is a DNN model with two hidden layers (see Algorithm 1).

We use weighted loss to address class imbalance and imbalance in the sizes of the source and target datasets for the adversarial training.

All experiments were run on Apple M2 Max CPU with 32 GB Memory. We used 5 different random data splits for each hospital site to obtain t-values and std. We used Adam optimizer for training and used hyperparameter tuning to find optimal penalty λ. Following the scheme of Ganin et al (2016) we tried 5 different values of lambda equally spaced on logarithmic scale between 10^-2 and 1. We note that in terms of the impact of hyperparameters on model performance in our experiments there was no statistically significant differences between Youden values for all values of λ tested.

E Efficiency Analysis

From wall-to-wall times analysis training with adversarial component took 1.5 longer per epoch than training baseline model. Explicitly, on eICU we had (in seconds)

- Mortality task: 9.3 ± 0.2 vs 6.2 ± 0.2
- Shock task: 1.34 ± 0.15 vs 0.9 ± 0.05

We note that retraining the models is expected to be needed relatively infrequently (e.g. at deployment and to account for time-related drift), which lowers the significance of the difference in training times in practice.