Label-free estimation of clinically relevant performance metrics under distribution shifts

Tim Flühmann[1,2,3(✉)], Alceu Bissoto[1,2,3], Trung-Dung Hoang[1,2,3], and Lisa M. Koch[1,2,3(✉)]

[1] University of Bern, Switzerland
[2] Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism, Inselspital, Bern University Hospital, University of Bern, Switzerland
[3] Diabetes Center Berne, Switzerland
{tim.fluehmann,lisa.koch}@unibe.ch

Abstract. Performance monitoring is essential for safe clinical deployment of image classification models. However, because ground-truth labels are typically unavailable in the target dataset, direct assessment of real-world model performance is infeasible. State-of-the-art performance estimation methods address this by leveraging confidence scores to estimate the target accuracy. Despite being a promising direction, the established methods mainly estimate the model's accuracy and are rarely evaluated in a clinical domain, where strong class imbalances and dataset shifts are common. Our contributions are twofold: First, we introduce generalisations of existing performance prediction methods that directly estimate the full confusion matrix. Then, we benchmark their performance on chest x-ray data in real-world distribution shifts as well as simulated covariate and prevalence shifts. The proposed confusion matrix estimation methods reliably predicted clinically relevant counting metrics on medical images under distribution shifts. However, our simulated shift scenarios exposed important failure modes of current performance estimation techniques, calling for a better understanding of real-world deployment contexts when implementing these performance monitoring techniques for postmarket surveillance of medical AI models.[1]

Keywords: performance estimation, label-free, postmarket surveillance

1 Introduction

Deep learning for medical image classification exhibits excellent performance in controlled settings [1], but distribution shifts in the target domain may cause silent failures in real-world applications [2,3]. For safe deployment in clinical domains, continuous performance monitoring is crucial. Several methods have been proposed to estimate model classification performance on unlabelled target datasets, enabling clinicians to anticipate model failures before they affect patients. Some performance estimation approaches estimate accuracy based on

[1] Code available at `https://github.com/mlm-lab-research/clin_perf_est`

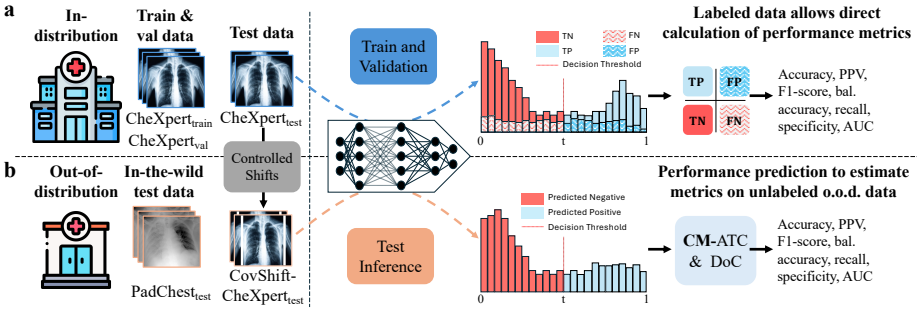

Fig. 1: Performance monitoring in a clinical setting. (**a**) An idealised, in-distribution scenario with labelled test data. (**b**) Our method for estimating metrics in the more realistic, out-of-distribution, unlabelled setting.

the distance between source and target distribution [4], or train a reverse model on the pseudo-labelled test data to evaluate reverse performance on the source distribution [5]. Here, we focus on performance estimation based on the model's confidence scores [4,6,7,8,9,10]. They have shown the best trade-off between accuracy and computational efficiency [7,8,11], as they require neither retraining, other labelled datasets, nor ensemble agreement [12]. Despite their potential, it remains unclear whether confidence-based performance-estimation methods apply in the clinical setting. Most were proposed and evaluated outside the clinical domain and focus on predicting only accuracy, which is often an inadequate metric for clinical tasks [13]. Instead, validating medical image classification models requires a suite of clinically relevant metrics depending on the domain of interest, such as precision, recall, etc. Currently, no benchmark exists for these metrics, and we are aware of only one first naive approach [10] to estimate these metrics without access to labelled test data.

In this paper, we propose a generalisation of two popular performance estimation techniques [4,7], which allows us to go beyond accuracy and reliably estimate the full confusion matrix (i.e., true positives (tp), false positives (fp), true negatives (tn), and false negatives (fn)), and thus any counting metric (e.g., recall, PPV) as well as the multi-threshold area under the ROC curve (AUC). We then present a comprehensive benchmark for confidence-based performance estimation in chest x-ray data, predicting a range of clinically relevant performance metrics (see overview in Fig. 1). We study real-world distribution shifts as well as the effects of covariate and prevalence shifts, which are common in medical data and can impair the clinical translation of medical classification models.

2 Background: performance estimation without labels

We define a binary classification problem with targets $y \in \{0, 1\}$, where models $f(x)$ are trained on inputs $\{(x_i, y_i)\}_{i=1}^{m} \sim \mathcal{D}_{\text{train}}$ and validated on $\mathcal{D}_{\text{val}}$, which

is assumed to be in-distribution (i.d.) with respect to $\mathcal{D}_{\text{train}}$. Predictions $\widehat{y}$ are obtained by applying a decision threshold $t \in [0, 1]$ to the model's sigmoid output $\tilde{s}_i = \sigma(f(x_i))$, which also represents the model's positive class confidence score. To obtain the confidence in the predicted class, we set $s_i = \mathbb{I}_{\{\tilde{s}_i \geq t\}}\, \tilde{s}_i + \mathbb{I}_{\{\tilde{s}_i < t\}}\, (1 - \tilde{s}_i)$. Our objective is to assess the model's performance through metrics like accuracy, recall, PPV, and AUC on a previously unseen, potentially out-of-distribution (o.o.d.) dataset $\{(x_i, y_i)\}_{i=1}^{n} \sim \mathcal{D}_{\text{test}}$, for which ground-truth labels are not available. Furthermore, we define the sets of positive and negative class predictions for distributions $d \in \{\text{val}, \text{test}\}$ as $I_d^+ = \{\tilde{s}_i \,|\, \tilde{s}_i \geq t,\, x_i \sim \mathcal{D}_d\}$ and $I_d^- = \{1 - \tilde{s}_i \,|\, \tilde{s}_i < t,\, x_i \sim \mathcal{D}_d\}$ with cardinalities n_d^+ and n_d^- respectively.

Confidence Based Performance Estimation (CBPE). A model is considered calibrated when its confidence score reflects the true probability of class 1; formally, $P(Y = 1 | \tilde{S} = \tilde{s}) = \tilde{s}, \forall\, \tilde{s} \in [0, 1]$ [14]. For calibrated models, accuracy can thus be estimated by the average confidence [9,15] $\widehat{\text{acc}}_{\text{CBPE}} = \frac{1}{n} \sum_{i=1}^{n} s_i$.

This approach has recently been generalised for other counting metrics based on the confusion matrix [10]. CBPE builds on the observation that scores with $\tilde{s}_i \geq t$ are either tp or fp, whereas scores with $\tilde{s}_i < t$ are either tn or fn. Averaging the subsets I_{test}^+ and I_{test}^- provides estimates for positive predictive value (PPV) and negative predictive value (NPV):

$$\widehat{\text{PPV}}_{\text{CBPE}} = \mathbb{E}_{s \sim I_{\text{test}}^+}[s]\,, \quad \widehat{\text{NPV}}_{\text{CBPE}} = \mathbb{E}_{s \sim I_{\text{test}}^-}[s]. \tag{1}$$

All counting metrics can then be estimated from the confusion matrix point estimates, which are derived from:

$$\begin{aligned} \hat{\text{tp}} &= n_{\text{test}}^+ \cdot \widehat{\text{PPV}}, \quad \hat{\text{fp}} = n_{\text{test}}^+ - \hat{\text{tp}}, \\ \hat{\text{tn}} &= n_{\text{test}}^- \cdot \widehat{\text{NPV}}, \quad \hat{\text{fn}} = n_{\text{test}}^- - \hat{\text{tn}}. \end{aligned} \tag{2}$$

Average Threshold Confidence (ATC). ATC [7] predicts the model accuracy on the test distribution $\mathcal{D}_{\text{test}}$ by computing the proportion of samples with scores s exceeding a learned threshold t_{ATC}. This threshold is determined on the validation set $\mathcal{D}_{\text{val}}$ such that the proportion of scores above t_{ATC} matches the empirical accuracy on $\mathcal{D}_{\text{val}}$: $\mathbb{E}_{x \sim \mathcal{D}_{\text{val}}}[\mathbb{I}[s > t_{\text{ATC}}]] = \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{val}}}[\mathbb{I}[\widehat{y} = y]]$.

The accuracy on $\mathcal{D}_{\text{test}}$ is then estimated as the fraction of test samples with confidence scores above the learned threshold t_{ATC}.

Difference of Confidences (DoC). DoC [4] can be used as a confidence-based performance estimation method that estimates test accuracy via:

$$\widehat{\text{acc}}_{\text{DoC}} = \text{acc}_{\text{val}} - \Delta; \quad \Delta = \mathbb{E}_{x \sim \mathcal{D}_{\text{val}}}[s] - \mathbb{E}_{x \sim \mathcal{D}_{\text{test}}}[s] \tag{3}$$

where acc_{val} is the observed accuracy in $\mathcal{D}_{\text{val}}$. In other words, the scores on the test distribution are re-calibrated by the amount that the average validation confidence deviates from the validation accuracy.

3 Performance estimation beyond accuracy

So far, estimating clinically relevant metrics beyond accuracy was only possible with the naive confidence-based CBPE. Here, we propose a method to extend ATC and DoC to estimate PPV (precision) and NPV. From these estimates, we can compute individual entries of the confusion matrix through Eq. (2), allowing us to then predict any counting metric from the confusion matrix estimators; for example, recall is estimated as $\hat{\text{tp}}/(\hat{\text{tp}} + \hat{\text{fn}})$. To approximate the AUC, we evaluate the true-positive rate and false-positive rate at 100 decision thresholds based on score quantiles and then numerically integrate the resulting ROC curve.

Confusion Matrix Estimation via Average Threshold Confidence (CM-ATC). To extend the ATC framework toward estimating elements of the confusion matrix, we build on an idea from a previously proposed variant of ATC [8]. They applied separate thresholds t^+_{ATC} to positive and t^-_{ATC} to negative predicted cases, learned on I^+_{val} and I^-_{val}, respectively. In [8], the positive and negative prediction sets were never used individually, but instead, accuracy was estimated by aggregating and counting all samples that exceeded their respective thresholds.

Here, we propose to use the positive I^+_{test} and negative I^-_{test} prediction sets in isolation along with their respective learned thresholds. The two class-specific thresholds are calculated such that the fraction of scores above t^+_{ATC} on I^+_{val} equals the validation PPV, and, analogously, the fraction of scores above t^-_{ATC} on I^-_{val} equals the validation NPV. We then define the following estimators for PPV and NPV, similarly to the original accuracy estimator described in Sec. 2:

$$\widehat{\text{PPV}}_{\text{CM-ATC}} = \mathbb{E}_{s \sim I^+_{\text{test}}}[\mathbb{I}[s > t^+_{\text{ATC}}]], \quad \widehat{\text{NPV}}_{\text{CM-ATC}} = \mathbb{E}_{s \sim I^-_{\text{test}}}[\mathbb{I}[s > t^-_{\text{ATC}}]]. \tag{4}$$

With estimates for PPV and NPV, we analogously estimate the confusion matrix through Eq. (2) to further estimate any counting metric of interest.

Confusion Matrix Estimation via Difference of Confidences (CM-DoC) Similarly to the ATC variant, [8] have extended DoC to re-calibrate the scores in I^+_{test} and I^-_{test} separately before estimating test accuracy. On I^+_{test}, the scores are offset by the gap between validation PPV and the mean confidence on I^+_{val}. Analogously, on I^-_{test} by the gap between validation NPV and the mean confidence on I^-_{val}.

Following the approach of CM-ATC, we now also extend the DoC method to estimate the confusion matrix elements. After calculating the offsets $\Delta^c := \mathbb{E}_{s \sim I^c_{\text{val}}}[s] - \mathbb{E}_{s \sim I^c_{\text{test}}}[s]$ for $c \in \{+, -\}$ and the realized PPV_{val} and NPV_{val} on validation, we can get estimates on the test set through:

$$\widehat{\text{PPV}}_{\text{CM-DoC}} = \text{PPV}_{\text{val}} - \Delta^+; \quad \widehat{\text{NPV}}_{\text{CM-DoC}} = \text{NPV}_{\text{val}} - \Delta^-. \tag{5}$$

From here, we again get the confusion matrix estimates from Eq. (2) and calculate estimates for the metrics.

4 Benchmark setup

With the methods for estimating performance metrics in place, we now set up their comparison on real-world and controlled distribution shifts on medical images. We compare CBPE and our proposed CM-ATC and CM-DoC. In addition, we include a naive baseline for ATC and DoC: we take the original ATC and DoC formulations and substitute accuracy with the metric of interest. In preliminary experiments, we have also analysed the impact of calibration (using temperature scaling [14] and class-wise temperature scaling [8]). As we had found no conclusive improvements (see Supplementary Fig. S1), we excluded an analysis of calibration techniques from the scope of this paper. We estimate a wide variety of metrics, covering prevalence-dependent (accuracy, PPV, F1-score), prevalence-independent counting metrics (balanced accuracy, recall, specificity), and a multi-threshold metric (AUC). Furthermore, we monitor the model's calibration using the Root Brier Score (RBS) and Adaptive Calibration Error (ACE).

4.1 Chest x-ray distribution shifts in the wild

First, we benchmark the performance estimators on real-world distribution shifts in chest x-ray data using three publicly available datasets from different cohorts: **CheXpertPlus** [16], **PadChest** [17], and **ChestX-Ray8** (NIH) [18]. They consist of 223,228, 160,861, and 112,120 chest radiographs, respectively. For each dataset, we set aside roughly 22,000 images for validation and testing (CheXpertPlus: 90/10/10; PadChest and NIH: 60/20/20) and use the remainder to train separate binary classifiers for three target conditions: Pleural Effusion, Cardiomegaly, and Pneumothorax. For each model, we estimate the performance metrics on the held-out i.d. test set, as well as on the two o.o.d. test sets. To evaluate the generalization capabilities of the performance estimation methods, we compute the difference between the estimated and realized performance metrics and report the mean absolute error (MAE).

4.2 Controlled distribution shifts in chest x-rays

Next, we investigate the behaviour of performance prediction methods under controlled distribution shifts. For this, we focus on the task of detecting Pleural Effusion in the CheXpertPlus dataset and simulate covariate shifts and prevalence shifts in the held-out test set.

To introduce **covariate shift**, we artificially modify the images by adding a visual artefact (two lateral vertical white bars) that is positively correlated with the positive class [19]. The model is then trained on this modified data, allowing it to learn a spurious correlation between the artefact and the target label. As a consequence, the model produces high-confidence predictions for the majority groups (i.e., label 1 with artefact present and label 0 without artefact), while generating lower-confidence predictions for the minority groups. For the training and validation sets, we set the proportion of majority samples to 80%,

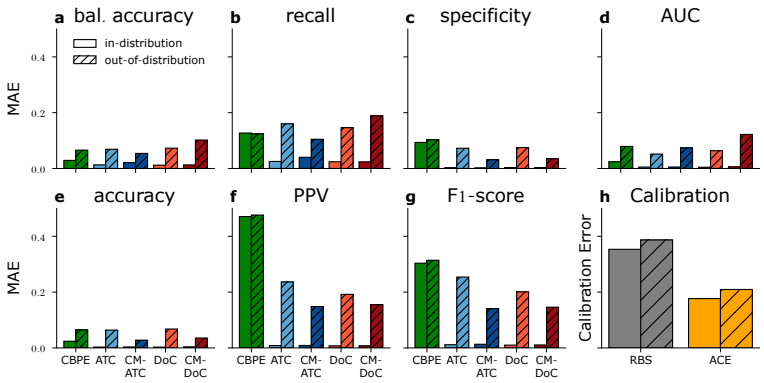

Fig. 2: (**a-g**) MAE evaluated on i.d. and o.o.d. data represented by solid and lined bars, respectively. (**h**) shows the mean RBS and ACE over all models. On average, CM methods outperformed the other estimators. Exact numerical values are reported in the Supplementary Table S2.

encouraging the model to rely on the spurious feature. Test sets are constructed by sampling 1000 images with varying proportions of majority samples from 0% to 100% and consequently minority samples from 100% to 0%. We keep the class distribution consistent between the validation and test sets to prevent additional prevalence shifts from confounding the analysis. We repeat the test set construction 50 times for each shift strength to reduce sampling variability, and the estimations are averaged over all repetitions.

To simulate **prevalence shift**, we repeatedly sample 1000 instances from the test set while targeting a positive class prevalence ranging from 5% to 95%. Since the overall prevalence of Pleural Effusion in the CheXpertPlus dataset is relatively high, at 38%, we can simulate label shifts in both directions, toward lower and higher prevalence levels without restricting the dataset size too much. We perform 50 resampling iterations at each prevalence level and compute averaged realised and estimated performance metrics [9].

5 Results

5.1 CM estimation methods perform best in the wild

The trained models performed comparably to the state-of-the-art in terms of classification performance [20], see Supplementary Table S1 for details. Overall, across both i.d. and o.o.d. scenarios and the different metrics in Fig. 2, our CM estimators outperformed the other methods, with CM-ATC performing best. Several prior studies, including CBPE, ATC, and DoC, have focused on accuracy (Fig. 2 **e**). Here, all methods worked well in-distribution (mean MAE of $(0.7 \pm 0.7) \cdot 10^{-2}$). Estimating accuracy in o.o.d. datasets was more difficult for all methods with a mean MAE of 0.05 ± 0.02.

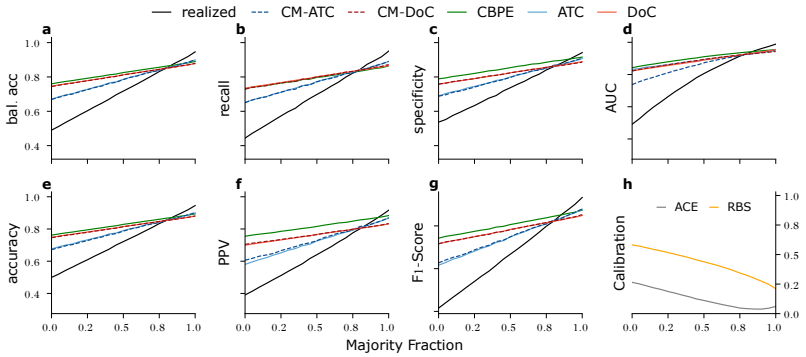

Fig. 3: (**a-g**) The estimated performance metrics capture the decline in performance under the simulated covariate shift, yet show overconfident performance towards minority groups in the test set. (**h**) Calibration error rises significantly for the minority group.

To estimate metrics beyond accuracy, only CBPE has been proposed before. While estimation of balanced accuracy, specificity, and AUC performed comparably to accuracy, performance dropped considerably for estimating recall, and failed dramatically for PPV and F1-score. CBPE's estimation error could be attributed to the significant calibration error (Fig. 2 **h**). In contrast, the ATC and DoC approaches led to consistently very low estimation error for all metrics in i.d. settings. O.o.d. performance varied, with CM-ATC generalising best overall. Since we have not explicitly quantified the distribution shift between the chest X-ray datasets, the observed drop is not straightforward to interpret.

5.2 Performance estimation methods can capture the impact of covariate shift

Next, we introduced covariate shifts using synthetic artefacts (see Sec. 4.2). By design, actual model performance deteriorated when increasing the proportion of minority groups (black curves in Fig. 3 **a-g**, original majority fraction $p = 0.8$). The metric estimators successfully captured this decline, yet still overestimated performance in these cases, and underestimated performance when the majority group was more prevalent than in the original distribution. Once again, (CM-)ATC methods performed best overall, while (CM-)DoC and CBPE likely suffered from model miscalibration on the minority samples (Fig. 3 **h**). The close agreement between the CM estimators and their respective naive counterparts can be attributed to this covariate shift affecting both negative and positive predictions similarly. As a result, the benefit of treating the two groups separately is limited.

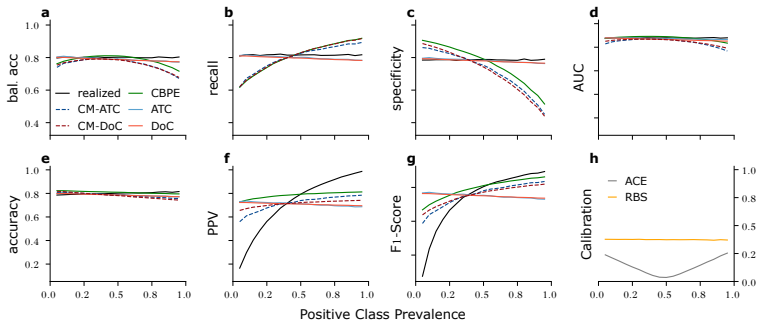

Fig. 4: (**a**–**d**) Prevalence-independent metrics remain constant under label shift; naive estimators perform best, whereas confusion-based ones remain prevalence-sensitive. (**e**–**g**) Prevalence-dependent metrics vary under label shifts; confusion-matrix-based methods capture these changes but still show high errors. (**h**) Prevalence shift directly affects model calibration.

5.3 Metric estimation degrades under prevalence shifts

All estimation methods struggled under prevalence shifts (Fig. 4), especially for prevalence-dependent metrics (Fig. 4 **e-g**). As CBPE is highly dependent on model calibration, it only accurately estimated performance when calibration error was low (Fig. 4 **h**). The ATC and DoC methods performed best when no shift was present (38% original prevalence, see Sec. 4.2). There, realised (black lines) and estimated performance (coloured) was very close, in line with the i.d. results in Sec. 5.1. When the prevalence shifted, the estimates diverged. The naive ATC and DoC implementations performed well on prevalence-independent metrics, but because they do not rely on class-specific calibration, they could not estimate prevalence-dependent metrics well (mainly visible in Fig. 4 **f,g**). In contrast, estimators derived from confusion matrix entries (CBPE, CM-ATC and CM-DoC) performed best for prevalence-dependent metrics, while they perform worse on the prevalence-independent metrics.

6 Discussion

In this paper, we proposed methods to estimate a wide range of classifier performance metrics without access to labelled test data. Our proposed estimators outperformed the only existing baseline (CBPE) for monitoring model performance in real-world distribution shifts and a simulated covariate shift. However, simulated prevalence shifts exposed systematic failures of all performance estimation techniques.

Our techniques for label-free performance estimation could be easily implemented in postmarket surveillance frameworks, as they can monitor deployed medical AI algorithms with clinically relevant performance estimators and little

computational overhead. However, as we have demonstrated, the accuracy of our estimators depends on the nature of the encountered distribution shifts. Therefore, we recommend that performance monitoring should be accompanied by distribution shift detection [21,22], identification [22], and mitigation [23]. Especially under prevalence shifts, domain adaptation techniques could help counter the systematic negative impact on model calibration.

Acknowledgments. This project was supported by the Diabetes Center Berne and strategic funding of the medical faculty of the University of Bern. Calculations were performed on UBELIX, the HPC cluster at the University of Bern.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J.R., Schmid, M.K., Balaskas, K., Topol, E.J., Bachmann, L.M., Keane, P.A., Denniston, A.K.: A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health **1**(6), e271–e297 (Oct 2019)
2. Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. npj Digital Medicine **5**(1), 1–8 (Apr 2022)
3. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: ACM conference on health, inference, and learning (CHIL). pp. 151–159 (2020)
4. Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., Schmidt, L.: Predicting with confidence on unseen distributions. In: International Conference on Computer Vision (ICCV). pp. 1134–1144 (2021)
5. Fan, W., Davidson, I.: Reverse testing: an efficient framework to select amongst classifiers under sample selection bias. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 147–156 (2006)
6. Białek, J., Kuberski, W., Perrakis, N., Bifet, A.: Estimating model performance under covariate shift without labels. arXiv preprint arXiv:2401.08348 (2024)
7. Garg, S., Balakrishnan, S., Lipton, Z.C., Neyshabur, B., Sedghi, H.: Leveraging unlabeled data to predict out-of-distribution performance. In: NeurIPS Workshop on Distribution Shifts: Connecting Methods and Applications (2021)
8. Li, Z., Kamnitsas, K., Islam, M., Chen, C., Glocker, B.: Estimating model performance under domain shifts with class-specific confidence scores. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 693–703. Springer (2022)
9. Kivimäki, J., Nurminen, J.K., Białek, J., Kuberski, W.: Confidence-based estimators for predictive performance in model monitoring. Journal of Artificial Intelligence Research **82**, 209–240 (2025)
10. Kivimäki, J., Białek, J., Kuberski, W., Nurminen, J.K.: Performance estimation in binary classification using calibrated confidence. arXiv preprint arXiv:2505.05295 (2025)

11. Elsahar, H., Gallé, M.: To annotate or not? predicting performance drop under domain shift. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2163–2173 (2019)
12. Baek, C., Jiang, Y., Raghunathan, A., Kolter, J.Z.: Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. Advances in Neural Information Processing Systems (NeurIPS) **35**, 19274–19289 (2022)
13. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., et al.: Metrics reloaded: recommendations for image analysis validation. Nature methods **21**(2), 195–212 (2024)
14. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning (ICML). pp. 1321–1330 (2017)
15. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (ICLR) (2017)
16. Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.: Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. arXiv preprint arXiv:2405.19538 (2024)
17. Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical Image Analysis **66**, 101797 (2020)
18. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2097–2106 (2017)
19. Sun, S., Koch, L.M., Baumgartner, C.F.: Right for the wrong reason: Can interpretable ml techniques detect spurious correlations? In: Medical Image Computing and Computer Assisted Interventions (MICCAI) (2023)
20. Cohen, J.P., Hashir, M., Brooks, R., Bertrand, H.: On the limits of cross-domain generalization in automated x-ray prediction. In: Medical Imaging with Deep Learning (MIDL). pp. 136–155 (2020)
21. Koch, L.M., Baumgartner, C.F., Berens, P.: Distribution shift detection for the postmarket surveillance of medical ai algorithms: A retrospective simulation study. npj Digital Medicine (2024)
22. Roschewitz, M., Mehta, R., Jones, C., Glocker, B.: Automatic dataset shift identification to support safe deployment of medical imaging ai. arXiv preprint arXiv:2411.07940 (2024)
23. Alexandari, A., Kundaje, A., Shrikumar, A.: Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In: International Conference on Machine Learning (ICML). pp. 222–232 (2020)