Revisiting Performance Claims for Chest X-Ray Models Using Clinical Context

Andrew Wang ANDREW_WANG3@BROWN.EDU
Brown University

Jiashuo Zhang JZHAN427@JHU.EDU
Johns Hopkins University

Michael Oberst MOBERST@JHU.EDU
Johns Hopkins University

Abstract

Public datasets of Chest X-Rays (CXRs) have long been a popular benchmark for developing machine learning (ML) computer vision models in healthcare. However, the reported strong average-case performance of these models do not necessarily reflect their actual utility when used in heterogeneous clinical settings, potentially masking weaker performance in medically significant scenarios. In this work we use clinical context to provide a more holistic evaluation of models for CXR diagnosis. In particular, we use discharge summaries, recorded prior to each CXR, to derive a "pre-CXR" probability of each CXR label, as a proxy for existing contextual knowledge available to clinicians when interpreting CXRs. We use this measure to probe model performance along two dimensions: First, using a stratified analysis, we show that models tend to have lower performance (as measured by AUROC and other metrics) among individuals with higher pre-CXR probability. Second, by controlling for pre-CXR probability via matching and re-weighting, we demonstrate that performance degrades when the correlation is broken between prior context and the current CXR label, suggesting that model performance is highly sensitive to the underlying distribution of clinical context. Specifically, cases with high pre-test probabilities present a fundamentally more difficult visual classification task, highlighting a gap in clinical utility when models are applied to high-risk cohorts.

Data and Code Availability We use data from the MIMIC-CXR (Johnson et al., 2019) and MIMIC-IV (Johnson et al., 2023) datasets, which are available on `physionet.org` via credentialed access under a data-use agreement. Code is publicly available at https://github.com/oberst-lab/revisiting-cxr-performance.

Institutional Review Board (IRB) This study uses de-identified, publicly available datasets under the PhysioNet Credentialed Health Data Use Agreement 1.5.0. Institutional Review Board (IRB) approval was not required for this study.

1. Introduction

Machine learning (ML) systems have shown impressive results in disease diagnosis from medical images, including on large public Chest X-Ray (CXR) datasets like MIMIC-CXR (Johnson et al., 2019). However, average-case evaluations do not fully capture the clinical utility of a model's predictions. First, these aggregate evaluations may mask differences in relevant subpopulations (Seyyed-Kalantari et al., 2020a; Oakden-Rayner et al., 2020), and clinical end-users of model predictions often need to understand the populations where a model should (or should not) be expected to perform well. Second, even on a well-defined subpopulation, standard "AI-alone" performance metrics may fail to capture the intrinsic diagnostic ability of models, and their ability to complement human decision-makers. Radiologists may not interpret CXRs in isolation, but also consider patient history and other contextual details. The clinical utility of a CXR model thus depends in part on its ability to rely on direct visual indicators of disease, rather than inferring prior clinical context already known to the clinician. For instance, Badgeley et al. (2019) found that the performance of a hip-fracture detection model was no better than random, once factors like scanner type and manufacturer were controlled for. Similarly, Oakden-Rayner et al. (2020) show that apparent high performance at

pneumothorax detection can be driven in part by high performance on cases where a chest drain is already present, a potentially less salient clinical population where the relevant condition is already being treated.

In this paper, we use prior clinical notes as a proxy for the "clinical context" of a diagnostic task, as captured by the "pre-test" or "pre-CXR" probability, before the CXR is actually taken, of each CXR disease label. We argue that this contextual information provides insights into the clinical utility of CXR models. Throughout, we use the MIMIC-CXR / MIMIC-IV datasets (Johnson et al., 2019, 2023) as a case study, though we expect that our approach could be broadly applied on any longitudinal health records dataset containing both medical images and prior clinical notes.

First, we show that prior discharge summaries (from prior hospital admissions) often contain sufficient information to predict disease labels in CXRs from future visits, even without access to the images themselves. We find evidence that this predictive signal can be derived (at least in part) from medically relevant terms in prior notes (e.g., the words "clavicle" and "rib" have high importance in predicting the label of "fracture" on future CXRs).

Second, we find that vision model performance varies significantly across different levels of the resulting pre-test probability derived from prior notes. The CXR models we study show significant degradation in performance (as measured by AUROC and other metrics) for those individuals with a higher prior probability of disease. A similar divergence in performance appears with a simple stratification of patients into those with and without an explicit prior mention of certain disease-relevant phrases. These results suggest that prior clinical text is a useful axis along which to understand variation in CXR model performance.

Finally, we consider performance of CXR models when controlling for pre-test probability. Our analysis is motivated by the existing work that has shown that vision models can recover surprising amounts of contextual information (e.g., scanner type (Badgeley et al., 2019), patient demographics (Gichoya et al., 2022), etc) from images, even if this information may not be readily distinguishable by humans. As such, it is not always clear whether strong performance arises from clinically relevant visual indicators of disease, or from information that is redundant between the context and the image itself. We probe this dependence in two ways: First, we demonstrate that model performance degrades substantially on a matched population subset where positive and negative examples are selected to have near-identical values of pre-test probability. Second, we demonstrate that performance degrades, though to a lesser degree, on a re-weighted population where the context-derived pre-test probability is rendered marginally independent of the label.

Taken together, these results suggest that model performance on standard CXR tasks is dependent on clinical context: Stratified analysis reveals underperformance in those cases where prior knowledge suggests higher risk of diagnosis, and matching / reweighting suggest that model performance may depend in part on inference of pre-CXR clinical context, although our analysis cannot rule out other causes, such as high intrinsic difficulty of distinguishing positive and negative high-risk cases. Our framework is summarized in Fig. 1.

To summarize, our contributions are as follows

- We demonstrate the ability to accurately predict future CXR disease labels using prior clinical notes, providing a "pre-test probability" that approximates the pre-imaging disease risk.
- We identify performance disparities of standard clinical vision models across subpopulations defined by the context-derived pre-test probability, and across subpopulations defined by prior mentions of medically relevant terms.
- We investigate the relationship between visual diagnostic signal and clinical context using a reweighted and context-matched evaluation framework, finding degradation in performance under both analyses, with degradation most severe in the matched evaluation.

2. Related Work

Recent studies have explored various approaches to improve medical imaging models and their evaluation using clinical context, though none address the specific challenge of characterizing model performance via prior medical history. Juodelyte et al. (2023) systematically analyzes dataset confounders in CXR, while Olesen et al. (2024) develops slice discovery methods to expose biased subpopulations, but neither incorporates clinical notes as stratification signals. Closest to our focus, Saab et al. (2024) demonstrates the value of workflow notes for EEG subgroup analysis, and Yang et al. (2025) reveals persistent demographic biases in vision-language models, but neither use longitudinal

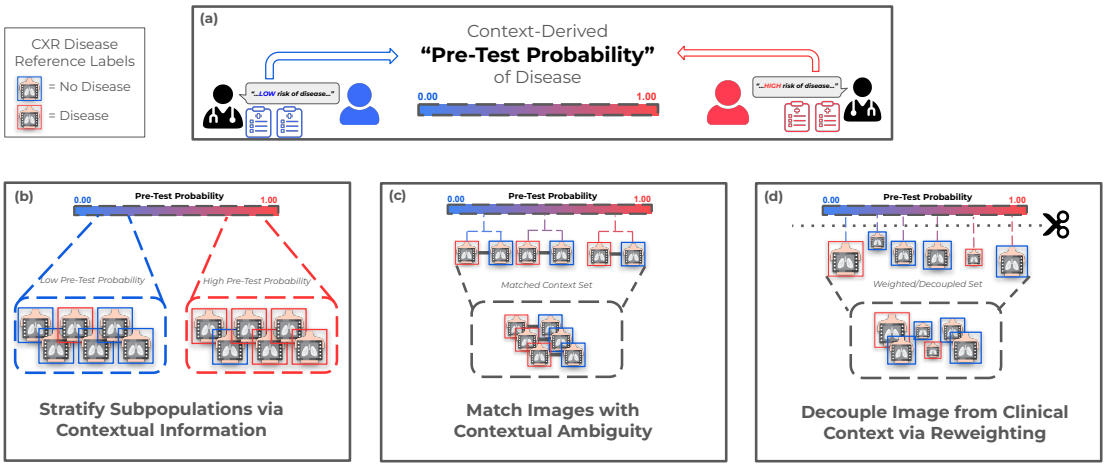

Figure 1: Overview of our evaluation framework (a): We use clinical context contained within discharge summaries to derive a pre-test probability estimate of disease risk, given knowledge obtained before a CXR is ordered. (b): We identify subpopulations of the evaluation using context from prior notes (such as pre-test probability, or prior mention of the disease label). We then evaluate performance on the resulting disjoint groups. (c): We create an evaluation set by matching positive/negative image pairs with similar context-derived pre-test probabilities. We then compare vision model performance on this balanced test set versus the original test set (d): We statistically decouple the image label from the contextual information via reweighting the evaluation set.

medical context to explain failures. Similarly, Subbaswamy et al. (2024) automates under-performance detection but requires manual feature engineering in their experiments, and does not make use of medical context as a stratifying feature in the evaluation set. On the model development side, several works incorporate radiological notes for label refinements (Syeda-Mahmood et al. (2021); Kim et al. (2022)) or model training (Bannur et al. (2023); Monajatipoor et al. (2021)), but treat these texts as auxiliary inputs rather than tools to interrogate performance. Closest to our perspective, Badgeley et al. (2019) showed that non-clinical correlates can inflate apparent performance, highlighting the need to disentangle true signal from contextual shortcuts when evaluating diagnostic models. Similarly, Gichoya et al. (2022) illustrates the ability of vision models to recover patient demographic information from images, while Makar et al. (2022) proposed enforcing conditional independence to reduce reliance on correlated features in ML classifiers. However, none of these works directly addresses the correlations between clinical context and imaging and its impact on vision model performance, which is the central focus of our study.

3. Data and Experimental Setup

3.1. Dataset Construction & Vision Model Training

To construct a cohort of CXR studies with prior context, we made use of the MIMIC-CXR database, which includes CXR studies with structured diagnosis labels from the CheXpert labeling tool for 14 different labels. We do not report the 'No Finding' label metrics due to lack of relevance in our downstream evaluation.[1] The MIMIC-CXR dataset provides linkages to the MIMIC-IV dataset, which we used to identify prior admissions for each patient, and pull the corresponding discharge summaries for those admissions. Further details are given in Appendix A.

We consider a single instance of the multilabel classification problem to be predicting labels for a single chest CXR study, where it is possible for multiple labels to be positive for a given CXR. As such, we defined prior context for a study as all of the associated ER visits and associated clinical notes whose discharge

1. We excluded the "No Finding" label from analysis as many of our experiments (e.g., stratifying based on prior mentions of a disease label) do not make sense in the context of the "No Finding" label.

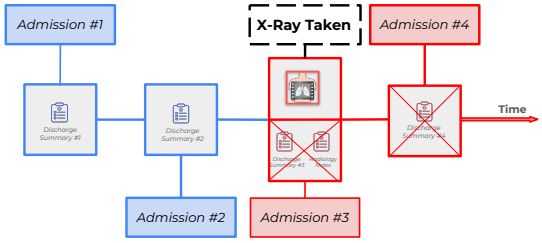

Figure 2: Visualization of "prior clinical context" used in this paper. For a given CXR study, we use all discharge summaries from admissions that occur before the current admission (blue). We do not use any information (discharge summaries, radiology reports, or otherwise) from the current or any future admission (red).

times and chart times, respectively, occur before the time of the CXR study (Fig. 2). We also restricted to individuals that had prior admissions and discharge summaries, in order to construct an evaluation set where all images had associated prior clinical context. We defined all CXR labels not explicitly corresponding to a positive diagnosis ("no label", "uncertain", among others) as a negative diagnosis in our label processing, following the procedure in Seyyed-Kalantari et al. (2020b). To prevent data leakage, we performed an 80-10-10 split by subject ID, ensuring that no patient's images or notes appeared across multiple splits. This yielded 167,291 images for training, 21,239 for validation, and 20,770 for testing. Full details of preprocessing, database joins, and methods to train the vision model used for experiments are provided in Appendix A. We broadly observed identical performance trends across both DenseNet121 and ResNet50 backbones, and all subsequent results in the main paper and Appendix are presented using the DenseNet121 backbone, unless otherwise specified.

3.2. Quantifying Pre-Test Probability from Prior Notes

Training Classifiers using Language Model (LM) Embeddings To capture the clinical context contained in prior notes, we evaluated a diverse set of text representations and classification heads. We first embedded all prior notes (after basic preprocessing) using six distinct open-weight language models: Mistral-7B-v0.1 (Jiang et al. (2023)), PubMedBERT (Gu et al. (2021)), BERT (Devlin et al. (2019)), ClinicalBERT (Huang et al. (2020)), BioLinkBERT (Yasunaga et al. (2022)), and RoBERTa (Liu et al. (2019)). Treating these embeddings as fixed feature representations, we trained classifiers to predict the image label from the text of prior discharge summaries alone. We performed a hyperparameter sweep using 5-fold stratified group cross-validation to select the optimal model class and hyperparameters from a set of 11 standard classification algorithms in `sklearn`. The "subject_id" was designated as the grouping variable to ensure that all data from a single patient resided within the same fold during cross-validation, preventing patient-level data leakage between folds. For each label, we selected the combination of language model encoder, classification architecture, and hyperparameters that yielded the highest mean cross-validation AUROC. Finally, the selected classifiers were calibrated using Platt scaling (sigmoid) prior to downstream analysis. See Appendix A for full details on text pre-processing, embedding extraction, the hyperparameter search space, and the best performing configurations.

Interpreting Predictive Signal with Bag of Words (BoW) Classifiers In addition to the models described above, which used language-model embeddings, we also constructed a bag-of-words (BoW) representation from these notes using `CountVectorizer` in `sklearn`, and evaluated seven standard classifiers (full training details in Table A.10, Table A.11), chosen for their interpretable features / coefficients, conducting hyperparameter and model selection as described above. Following the selection of the best classifier, we examined the top 10 most important features via feature importance measures specific to each model class (full list of feature importances in Appendix C).

Finally, we also trained XGBoost models on these BoW representations as well. Clinical text was pre-processed and tokenized in the same manner as in C. Model hyperparameters were optimized through a 60-iteration randomized search using 5-fold Stratified Group K-Fold cross-validation, preventing data leakage across splits. Finally, to ensure reliable pre-test probability estimates, the best-performing model from the cross-validation phase was post-hoc calibrated using Isotonic Regression.

3.3. Stratified Performance Analyses

Stratification via Pre-Test Probability Using the classification models trained on prior text, we obtained the predicted probabilities for each CXR label, for each example in our evaluation set. These label predictions were then binned by quantiles (bottom 25%, middle 50%, top 25%) to form disjoint subpopulations in the evaluation set.

Stratification via Prior Mentions As an alternative to using pre-test probability, we utilized the phrase list for each label from CheXpert (Irvin et al., 2019) to define prior mentions of the label in prior clinical notes. After adapting their phrase list to our context[2], we defined disjoint populations based on the presence of any relevant term, and the absence of all relevant terms.

3.4. Controlling for Pre-Test Probability

To assess the extent to which vision models depend on inference of pre-CXR context, rather than direct visual cues, we develop two complementary evaluation strategies to control for pre-test probability:

Matched Neighbor Analysis We constructed a context-balanced test set by selecting pairs of positive and negative examples (as determined by their ground truth labels) with similar pre-test probabilities. Positive (label = 1) and negative (label = 0) examples were separated, and the Hungarian algorithm applied (via `linear_sum_assignment` in SciPy) to perform 1:1 nearest-neighbor matching between positive and negative examples. The cost matrix was defined by the absolute difference between the pre-test probabilities of each positive–negative pair. This procedure yielded matched pairs with highly similar text-based probabilities but opposite ground truth labels. As illustrated in Fig. 3, this matched set represents a setting where the clinical context is balanced across class labels, making it impossible to predict the label by inferring prior clinical context alone. In practice, due to the limited number of positive cases, we found that this procedure tends to select all of the original positive examples, retaining only a subset (matched on pre-test probability) of the negative examples for each label. Hence, in contrast to the method described below, the resulting distribution of pre-test probability is unchanged among cases where $Y = 1$, and is substantially modified among cases where $Y = 0$.

Inverse Probability Weighting (IPW) To rigorously assess whether the vision model utilizes visual signals independent of clinical context, we estimate performance on a theoretical target distribution where diagnostic labels are statistically independent of the prior clinical context. We formalize this as follows:

Definition 1 (Reweighting Distributions) *Let $P(X, Y, C)$ denote the observed data distribution, where X is the image, $Y \in \{0, 1\}$ is the diagnostic label, and C is the pre-test probability derived from the prior clinical context. In the observational setting, Y and C are correlated. We define a target distribution $Q(X, Y, C)$ that satisfies two conditions:*

1. ***Independence of Context and Label:*** *$Q(Y, C) = Q(Y)Q(C) = P(Y)P(C)$. This represents a scenario where clinical history provides no information about the current diagnosis, but the marginal distributions of Y and C are unchanged.*
2. ***Invariance of Imaging Mechanism:*** *$Q(X \mid Y, C) = P(X \mid Y, C)$. Conditioned on context and the true label, the distribution of images is unchanged.*

We note that standard performance metrics evaluated on the target distribution Q can be estimated via weighted metrics on the observed distribution P (Sugiyama et al. (2007)). We provide the formal derivation for the importance weights and the estimators for the performance metrics in Appendix E, but to provide intuition here, we note the following basic result, which is an application of standard methods in importance sampling to our context.

Proposition 2 (Expected Equivalence) *For an observation (x, y, c), let the importance weight $w(y, c)$ be defined as the ratio of the marginal disease prevalence to the context-conditional probability:*

$$w(y, c) = \frac{P(Y = y)}{P(Y = y \mid C = c)} \tag{1}$$

Then, given a distribution P and a corresponding distribution Q (as defined in Def. 1), where $Q(X, Y, C) >$

2. Since the CheXpert labeler was specifically designed for radiology reports, some phrases in their list were not directly applicable to the clinical notes in our study. For example, for the "Pneumonia" label, the original corresponding phases included "infection" and "infectious", but these terms are commonly used in clinical notes to describe other infectious diseases and do not demonstrate a strong correlation with pneumonia specifically. We therefore adapted and refined selected terms to ensure precise identification of relevant mentions in our clinical context. Details of the specific phrases used are presented in Table A.15.

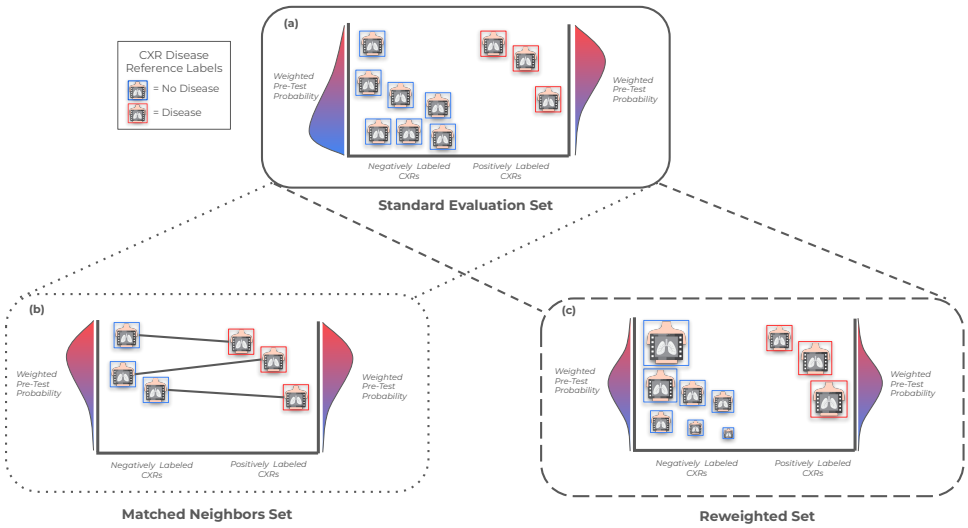

Figure 3: Details of Matched Neighbors vs Reweighted Evaluation Set (a): Original evaluation set: Negatively and positively labeled CXRs in the original evaluation set with their associated distributions of pre-test probabilities. All images and their associated pre-test probabilities are equally weighted. (b): Matched Neighbors set: Our procedure matches positive and negative examples 1-to-1 based on pre-test probability. In practice, given low prevalence of positive labels, this procedure tends to retain the subset of negatively labeled CXRs whose associated pre-test probability distribution more closely resembles that of the positively labeled CXRs, which are unchanged. (c): Reweighted set: After reweighting of images and their associated pre-test probabilities, the weighted distribution of pre-test probabilities become comparable across the positive and negative reweighted populations.

$0 \implies P(X,Y,C) > 0$*, then for any measurable function* $g(X,Y,C)$*:*

$$E_P[w(y,c)g(X,Y,C)] = E_Q[g(X,Y,C)]$$

In other words, for any metric that can be expressed as an expectation over the population Q, we can re-express that metric as a reweighted average over the population P. In Appendix E we illustrate how this result can be used to derive re-weighting estimators for our performance metrics of interest (Sensitivity, Specificity, AUROC, ACE, and BSS).

Implementation In practice, we estimate the weights using the training set and the text-only classifiers. The marginal prevalence $\hat{P}(Y=1)$ is estimated from the training set distribution. The conditional probability $\hat{P}(Y=1 \mid C=c_i)$ is obtained from the calibrated text-only classifier described in Section 3.2. To prevent numerical instability from extreme weights, text-derived probabilities are clipped to the range $[0.001, 0.999]$ prior to weight calculation. Finally, the weights are normalized such that their sum equals the total number of samples in the evaluation set ($\sum w_i = N$). All weighted metrics are computed using the `sample_weight` parameter in standard `scikit-learn` implementations.

3.5. Metrics and Statistical Uncertainty

We estimate performance on subpopulations with stratified bootstrapping. For each bootstrap iteration, we sample (with replacement) while maintaining the label distribution by stratifying on the ground truth labels. We calculate the area under the receiver operating characteristic curve (AUROC) for each bootstrap sample for each subgroup using the `MultilabelAUROC` metric across our 13 labels. The two metrics are subtracted to obtain our final difference in AUROC metric between the subgroups, and this process is repeated for a total of $10,000$ bootstrap iterations. Finally, we compute the mean difference and the corresponding 95% confidence intervals (CI) using the 2.5th and 97.5th percentiles of the bootstrapped metric difference distribution, adjusted via

Bonferroni correction to account for multiple comparisons across the 13 labels.

While AUROC provides an aggregate measure of diagnostic performance, it may obscure model behavior at specific operating points critical for clinical decision-making. We therefore further introduce sensitivity at 95% specificity: the true positive rate when the false positive rate is constrained to 5%.

To capture both the model's potential and its practical deployment risk, we evaluate this metric using both subgroup-specific thresholds, which reflect the model's intrinsic discriminative capacity, and a universal global threshold, which simulates clinical practice. For brevity, we denote these as **sensitivity at local 95% specificity** and **sensitivity at global 95% specificity**, respectively.

4. Results

4.1. CXR Labels are Predictable from Prior Clinical Notes Alone

We find that text classifiers trained on either LM or BoW representations are able to predict many of the MIMIC-CXR labels using only discharge summaries from prior to the CXR visit. Performance of models using the best LM embeddings is given in Table 1, with full results in Table A.6. Table A.13 gives performance of models using BoW representations. We make three general observations: First, we note that extensive hyperparameter optimization across both types of representations resulted in similar test set performance for most of the labels. Second, the predictive signal in the prior context can be at least partially traced to medically relevant terms in the discharged summaries: words like "clavicle" and "rib" are assigned high feature importance in the Fracture classifiers trained on BoW embeddings (full list in Table A.12). We reason that despite the presence of other noisier features such as "completing" and "18" that contribute to these predictions, the strong performance of both classes of text classifiers and relative importance of medically relevant terms provide orthogonal axes of evidence pointing towards the existence of medically relevant textual signal in the clinical context. Finally, we observe varying levels of text classifier calibration across different labels (refer to Fig. A.2 and Table A.7).

Table 1: Discrimination performance of text-only classifiers. We report the test set AUROC for the best performing language model and classifier combination for each label, trained solely on prior clinical notes. We observe that prior context is generally sufficient to build longitudinal ML classifiers of future CXR labels. Refer to table A.13 for the full results on the BoW representations and A.14 for the full results on the XGBoost models.

Label	AUROC
Atelectasis	0.623
Cardiomegaly	0.701
Consolidation	0.633
Edema	0.751
Enlarged Cardiomediastinum	0.566
Fracture	0.676
Lung Lesion	0.765
Lung Opacity	0.628
Pleural Effusion	0.745
Pleural Other	0.654
Pneumonia	0.592
Pneumothorax	0.712
Support Devices	0.682

4.2. Vision Model Performance Varies Across Context-Derived Subpopulations

We observe that when stratifying our dataset based on the pre-test probabilities (based on prior notes) produced by our text classifiers, statistically significant performance gaps emerge across our subgroups across several labels, particularly when comparing the "low risk" (bottom 25%) quantile against the "higher risk" (top 25%) quantile of pre-test probabilities (see Fig. 4 and Table A.16 for results using the text classifiers trained on LM embeddings). Notably, images within the lower quantiles of the text classifier predictions tend to yield higher vision model performance, which decreases as we move to higher quantiles. For example, the vision model's AUROC for Edema decreases from 0.9 in the bottom "low-risk" quantile to 0.812 in the top "high-risk" quantile, a statistically significant difference. Similarly, we observe that vision model performance for Atelectasis decreases from 0.831 in the bottom quantile to 0.738 in the top quantile. These examples indicate that vision models generally perform best on cases where the pre-test probability is low.

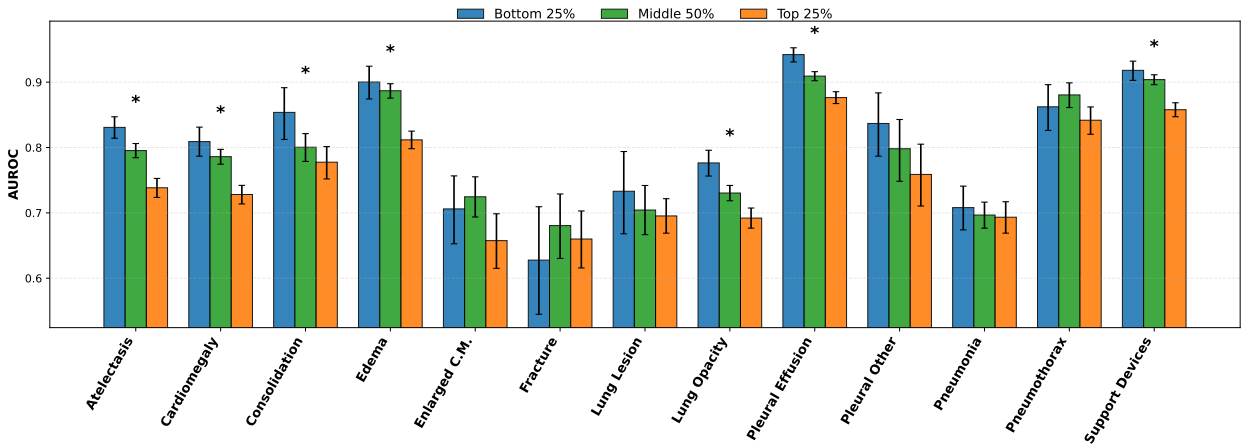

Figure 4: Held-out performance (in terms of **AUROC**) of CXR models across sub-populations stratified by pre-test probability (LM embeddings) of the CXR label. Vision model performance generally degrades across most labels as the prior probability of the label increases. Labels marked with an asterisk (*) indicate a statistically significant difference in AUROC between the Bottom 25% and Top 25% groups. The 95% confidence intervals in parentheses were calculated using percentile bootstrapping as described in section 3.3.

Complementing these findings, we observe a similar pattern of performance degradation in sensitivity at 95% local specificity, where thresholds are set independently in each subpopulation to maintain 95% specificity (see Table A.17). These results further reinforce the observation that vision model reliability at specific clinical operating points is significantly influenced by prior clinical context. However, this trend reverses when measuring sensitivity at 95% global specificity, where a single global threshold is chosen (see Table A.18), and where both sensitivity and specificty can vary across quantiles. In this setting, sensitivity is higher in the high-risk quantile, but as illustrated by the score distribution shifts for Edema and Atelectasis in Fig. A.3, this effect can be explained by an overall upward shift in model scores for high-risk patients (with a correspondingly lower specificity; see Table A.19), rather than improved discriminative performance. Consequently, applying a fixed global threshold can disproportionately limit sensitivity in low-risk cases: in clinical practice, the use of a single global threshold may systematically suppress performance in subgroups where the model's discriminative potential is higher. We also repeated the above analyses and using BoW representations and observed similar patterns, as shown in Tables A.23, A.25 and A.26. We also observe that the ResNet50 backbone provided similar performance trends, as indicated in Table A.24. Finally, we provide representative examples of images whose discharge notes yield high and low pre-test probabilties in Table A.8.

4.3. Vision Model Performance Varies Across Cases with and without Prior Mentions

As described in Section 3.3, we considered a simpler stratification of cases based on the presence or absence of label-relevant medical terms in previous discharge summaries. We observe a consistent pattern in which vision models perform systematically better on images from patients without prior clinical mentions of the corresponding label. Specifically, across all labels, the "no previous mentions" group demonstrates higher AUROC values than the "previous mentions" group, as shown in Table A.30. For example, the performance of the vision model for Atelectasis on the subset of images with a previous mention of the disease is 0.757, but the performance of the same model on the subset of images without a previous mention is 0.831. Similarly, the performance of the vision model for Cardiomegaly on the subset of images with a previous mention of the disease is 0.787, but the performance is 0.819 when evaluated on the images without this mention.

These patterns persist for sensitivity at local 95% specificity (Table A.31), which remains consistently

higher for cases without prior mentions. However, this advantage is obscured when measured via sensitivity at global 95% specificity (Table A.32). This contrast is similar with the pattern observed in Section 4.1: in clinical practice, the use of a single global threshold may systematically suppress performance in subgroups where the model's discriminative potential is higher.

4.4. Vision Model Performance Degrades when Controlling for Pre-Test probability

We next employ two complementary strategies to control for the predictive signal already present in clinical notes: Inverse Probability Weighting (IPW) and Matched Neighbor analysis.

First, using IPW (Section 3.4), we simulate a target population where diagnostic labels are statistically independent of clinical context, where "surprising" cases (e.g., low context-derived risk but positive diagnosis) are upweighted. Here, **we find that vision model performance degrades across all labels, though not all differences are statistically significant**, with six out of 13 labels exhibiting statistically significant declines (Table A.34). We note that the labels that do exhibit these difference correpsond to the labels whose associated text classifier predictions are comparably well calibrated. Given the reliance of these experiments on the text-derived pre-test probabilties, we posit that the relative levels of calibration in the text classifiers may be a factor in some labels exhibiting larger differences than others. For instance, while certain labels such as "Atelectasis" and "Cardiomegaly" display statistically significant differences between this reweighted set compared to the original, other labels such as "Fracture" and "Lung Lesion" do not appear to hold the same pattern.

Following the Matched Neighbor procedure (Section 3.4), we evaluated the model on a subset of paired positive and negative examples with effectively identical pre-test probabilities. **In this setting, we find that vision model performance drops across all labels in a statistically significant fashion.** For instance, the AUROC for Pleural Effusion falls from 0.92 in the standard set to 0.85 in the matched set, and the AUROC for Pneumothorax falls from 0.879 in the standard to 0.722 in the matched set (Fig. 5 and Table A.34). We observe a parallel trend in our threshold-based and calibration metrics. Sensitivity at 95% specificity (using subgroup-optimized thresholds) decreases notably in the matched evaluation set across all labels (see Table A.36 in the Appendix), and calibration metrics appear to follow the same trend (see Tables A.38 and A.39). We hypothesize that this result is due to the negative instances with high pre-test probability representing a particularly difficult subset of images, in that the contextual information does not support the actual diagnosis as closely. As a result, the performance drop is more pronounced as compared to our previous evaluation methods. We broadly observe similar performance trends for the ResNet50 backbone as well. For full results on the ResNet50 backbone, refer to Table A.35.

5. Discussion

In this work, we proposed clinical context as a foundation for evaluating state-of-the-art CXR vision models. Rather than relying solely on aggregate performance metrics, we introduced context-derived measures, such as pre-test probability and prior textual mentions of the label (e.g., disease), that give us a proxy for the information available to clinicians in practice. This framework allows us to assess not only whether models perform well on average, but also whether they perform reliably across clinically meaningful subpopulations and scenarios.

We first find that clinical context alone (in this case, the content of prior discharge summaries) is capable of accurately predicting future CXR labels across several different classes of embedding representations. In addition, we trace this predictive signal to discrete, clinically relevant terms in the associated discharge summaries. Next, by stratifying the images into subpopulations based on the pre-test probability of the CXR label, we find that vision models often show stronger performance on lower quantiles of risk, and worse performance in higher risk strata. This indicates that vision models may be most useful in triaging low-risk cases, such as helping to surface unexpected positives, while being less reliable in cases where the relevant label (e.g., a particular disease) is already suspected. We also find that performance gaps persist when stratifying by discrete terms in the clinical text, suggesting that clinical context can be used to inform evaluations of performance across several "resolutions" of medical history. More broadly, clinical context can be used to define subpopulations that are readily interpretable by clinical end-users, who may be expected to have a sense of the relevant medical context for a given patient. Hence, information on performance variation across these subpopulations (e.g., low risk

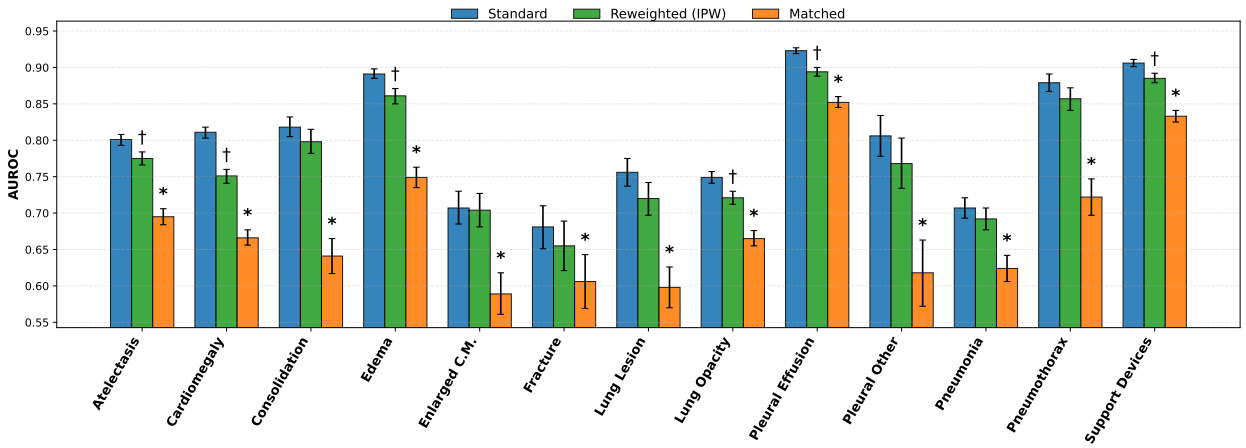

Figure 5: Comparison of **AUROC** across Standard, Reweighted (IPW), and Matched Neighbor settings. Metrics reported are mean (95% CI). Labels marked with a dagger (†) and an asterisk (*) indicate statistically significant differences in the Reweighted (IPW) and Matched settings, respectively, compared to the Standard setting. We observe that performance consistently drops across the Matched evaluation settings, and that vision model performance degrades across all labels in the Reweighted setting, though not all differences are statistically significant.

vs. high risk cases) may be useful for deciding when (and to what extent) to rely on the outputs of medical imaging models in clinical diagnostic tasks.

Using this contextual information, we also evaluate the extent to which model performance is driven by inference of pre-CXR context, as opposed to direct visual diagnostic signal. We probe this dependence in two ways, by employing both Inverse Probability Weighting (IPW) and Matched Neighbor analysis to decouple the visual signal from the clinical prior. The inconsistent stability of model performance under IPW suggests that while performance is not entirely driven by recovery of our particular notion of pre-CXR probability, there exists some dependence on performance on that underlying signal. Furthermore, the sharp performance drop observed in the Matched Neighbor analysis reveals a limitation in that for patients with similar clinical risk profiles, models struggle to use direct visual features to resolve diagnostic uncertainty.

While our Matched Neighbor and IPW analyses demonstrate a clear degradation in performance when the statistical link between context and label is severed, the exact mechanism driving this drop warrants careful interpretation. One hypothesis is that the vision model over-relies on non-clinically relevant visual confounders that happen to correlate with the pre-test probability. However, another explanation is that high pre-test probability negative cases simply possess clinically relevant visual features that make the classification task intrinsically harder.

Taken together, our results argue for the integration of contextual information into future evaluation pipelines. Standard metrics like AUROC reported in aggregate may overstate the true diagnostic contribution of vision models by conflating context-inference with visual diagnosis. By grounding evaluation in clinical context, we obtain a more complete, nuanced picture of model utility. This helps us distinguish between cases where the model adds genuine diagnostic value to human decision making and those where it merely reiterates the patient's history.

Looking forward, context-based evaluation opens several directions for future work. Beyond chest X-rays, many clinical imaging tasks are embedded within rich longitudinal records that could provide analogous context signals. Benchmarking frameworks that use these signals appropriately in training and evaluation may sharpen our understanding of precisely when and how models complement clinical decision making.

Limitations In this study, we restrict ourselves to MIMIC-CXR and MIMIC-IV, leveraging the unique longitudinal nature of the combined dataset, including the linkage of CXR images and prior discharge summaries. However, as a result, the generalization of our results to other clinical datasets from other institutions in not guaranteed.

We were also restricted to individuals with prior discharge summaries, which may not represent the general population (e.g., our analysis does not cover new patients who present without any prior context).

Our methodology establishes a robust statistical relationship between text-derived pre-test probabilities and vision model performance, but it does not empirically map these probabilities to specific visual artifacts within the image. Future work utilizing fine-grained multimodal interpretability techniques or bounding-box annotations is required to explicitly link a patient's textual clinical context to the specific visual features driving a model's prediction.

Finally, we choose to only use textual discharge summaries to represent our prior clinical context. However, many CXR datasets such as MIMIC-IV contain modalities other than text, such as time-series data with ICU lab results. Expanding the scope to modalities other than text may yield more complex, realistic contextual "stories" beyond what is captured in the discharge texts.

References

Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin S. Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder, and Joel T. Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2, 4 2019. doi: 10.1038/s41746-019-0105-1. URL http://dx.doi.org/10.1038/s41746-019-0105-1.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision–language processing, 2023. URL https://arxiv.org/abs/2301.04558.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, and Haoran Zhang. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022. ISSN 2589-7500. doi: https://doi.org/10.1016/S2589-7500(22)00063-2. URL https://www.sciencedirect.com/science/article/pii/S2589750022000632.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, October 2021. ISSN 2637-8051. doi: 10.1145/3458754. URL http://dx.doi.org/10.1145/3458754.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2020. URL https://arxiv.org/abs/1904.05342.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10, 1 2023. doi: 10.1038/s41597-022-01899-x. URL http://dx.doi.org/10.1038/s41597-022-01899-x.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6, 2019.

Dovile Juodelyte, Yucheng Lu, Amelia Jiménez-Sánchez, Sabrina Bottazzi, Enzo Ferrante, and Veronika Cheplygina. Source matters: Source dataset impact on model robustness in medical imaging. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, Lecture Notes in Computer Science, pages 163–173. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-82007-6_11. URL https://link.springer.com/chapter/10.1007/978-3-031-82007-6_11. Part of the MICCAI 2023 proceedings.

Doyun Kim, Joowon Chung, Jongmun Choi, Marc D. Succi, John Conklin, Maria Gabriela Longo,

Jeanne B. Ackman, Brent P. Little, Milena Petranovic, Mannudeep K. Kalra, et al. Accurate auto-labeling of chest x-ray images based on quantitative similarity to an explainable ai model. *Nature Communications*, 13(1), Apr 2022. doi: 10.1038/s41467-022-29437-8. URL https://doi.org/10.1038/s41467-022-29437-8.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels, 2022. URL https://arxiv.org/abs/2105.06422.

Masoud Monajatipoor, Mozhdeh Rouhsedaghat, Liunian Harold Li, Aichi Chien, C.-C. Jay Kuo, Fabien Scalzo, and Kai-Wei Chang. Berthop: An effective vision-and-language model for chest x-ray disease diagnosis. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3327–3336. IEEE, Oct 2021. doi: 10.1109/ICCVW54120.2021.00372. URL https://doi.org/10.1109/ICCVW54120.2021.00372. Workshop on Computer Vision for Automated Medical Diagnosis (CVAMD).

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *ACM CHIL '20: ACM Conference on Health, Inference, and Learning*. ACM, 4 2020. doi: 10.1145/3368555.3384468. URL http://dx.doi.org/10.1145/3368555.3384468.

Vincent Olesen, Nina Weng, Aasa Feragen, and Eike Petersen. Slicing through bias: Explaining performance gaps in medical image analysis using slice discovery methods. *arXiv preprint*, Oct 2024. URL https://arxiv.org/abs/2406.12142. Preprint submitted on 17 Jun 2024.

Khaled Saab, Siyi Tang, Mohamed Taha, Christopher Lee-Messer, Christopher Ré, and Daniel L. Rubin. Towards trustworthy seizure onset detection using workflow notes. *npj Digital Medicine*, 7(1), Feb 2024. doi: 10.1038/s41746-024-01008-9. URL https://doi.org/10.1038/s41746-024-01008-9.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. 2020a. URL https://arxiv.org/abs/2003.00827.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. *arXiv preprint (2003.00827v2)*, 2 2020b. URL http://arxiv.org/abs/2003.00827v2.

Adarsh Subbaswamy, Berkman Sahiner, Nicholas Petrick, Vinay Pai, Roy Adams, Matthew C. Diamond, and Suchi Saria. A data-driven framework for identifying patient subgroups on which an AI/-machine learning model may underperform. *npj Digital Medicine*, 7(1), Nov 2024. doi: 10.1038/s41746-024-01275-6. URL https://doi.org/10.1038/s41746-024-01275-6.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper.pdf.

Tanveer Syeda-Mahmood, K. C. L. Wong, Joy T. Wu, Ashutosh Jadhav, and Orest Boyko. Extracting and learning fine-grained labels from chest radiographs. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, Jan 2021. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8075457/. PMCID: PMC8075457.

Yuzhe Yang, Yujia Liu, Xin Liu, Avanti Gulhane, Domenico Mastrodicasa, Wei Wu, Edward J. Wang, Dushyant Sahani, and Shwetak Patel. Demographic bias of expert-level vision-language foundation models in medical imaging. *Science Advances*, 11(13):eadq0305, Mar 2025. doi: 10.1126/sciadv.adq0305. URL https://doi.org/10.1126/sciadv.adq0305. Article number eadq0305.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*, 2022.

Appendix A. Experimental Details

Dataset Version The study used data from the MIMIC-IV and its extensions, specifically MIMIC-IV v3.0, MIMIC-CXR-JPG v2.0.0, and MIMIC-IV-Note v2.2.

Dataset Processing The admission data was joined to the CXR studies and labels via the "subject_id" columns, which was then matched with the metadata via the "subject_id" and "study_id" columns. Finally, the notes was added to this output by joining the dataframes on the "subject_id" and "hadm_id" columns.

While there exists a standardized train/validation/test split for the MIMIC-CXR database, we chose to construct our data split by randomizing the subjects and performing an 80-10-10 split based on the unique subject IDs. Based on our definition of prior context and due to the nature of each subject having a variable number of admissions, CXR images, and associated clinical notes, our splitting by subject ID ensures that there is no overlap of subjects or their associated images and clinical notes between training, validation, and test data. Following this pipeline, we have 167,291 images in our training set, 21,239 images in our validation set, and 20,770 images in our test set. For distribution of positive and negative image labels in our dataset, refer to Table A.3.

Vision Model Training In order to evaluate performance impacts on state-of-the-art models for this dataset, we developed our vision models following a standard approach, as pre-trained weights are generally unavailable. Specifically, we based our design choices on the methodology outlined in Seyyed-Kalantari et al. (2020b), using both DenseNet121 and ResNet50 backbones. Using our custom train/validation/test split, we trained models on NVIDIA L40S GPUs, applying multi-GPU model and data parallelism to accelerate the training process. Our training procedure closely follows that in Seyyed-Kalantari et al. (2020b), including a batch size of 256. This approach yielded models with state-of-the-art performance (see Table A.1 for details), which enables us to perform our evaluations on state-of-the-art vision models.

We also attempted to evaluate BioMedCLIP in a zero-shot manner to avoid data leakage and use larger Vision-Language Models (VLMs), but it achieved a macro AUROC of only 0.585 (vs. 0.803 for DenseNet), making downstream analysis uninformative. See Table A.2 for details.

Preprocessing and Embedding of Prior Clinical Notes We first preprocessed all of the clinical notes by implementing punctuation and stopword removal, replacement of repetitive words, and lower-casing. To embed a summarized version of the prior clinical context of a CXR, we profiled six pre-trained language model encoders publicly available on HuggingFace: Mistral, BERT, BioLinkBERT, PubMedBERT, ClinicalBERT, and RoBERTa. We processed the clinical notes using a sliding window chunking strategy to handle notes exceeding the models' context limit. Each chunk was independently encoded, mean-pooled, and the resulting embeddings were aggregated via mean pooling to produce a final note representation. To ensure consistency and improve downstream stability, the pooled embeddings were normalized to have a unit L2 norm. Finally, in cases where multiple notes occurred prior to a CXR, their embeddings were averaged to produce a single summarized representation of the patient's prior context.

Hyperparameter and Model Class Selection We evaluated ten distinct classification models provided in sklearn: `Perceptron`, `RidgeClassifierCV`, `PassiveAggressiveClassifier`, `GaussianNB`, `LinearDiscriminantAnalysis`, `RandomForestClassifier`, `KNeighborsClassifier`, `MLPClassifier`, `LinearSVC`, and `SGDClassifier`.

We performed an automated search using `GridSearchCV` over predefined parameter grids (see Supplementary Tables Table A.4 and Table A.5 for specific grids). To select the optimal hyperparameters for each classifier type, we employed a 5-fold stratified group cross-validation (`StratifiedGroupKFold`) on the combined training-validation data. The "subject_id" were designated as the grouping variable to ensure that all data from a single patient resided within the same fold during cross-validation, preventing patient-level data leakage between folds. The specific combination of language model embedding, classifier architecture, and hyperparameters yielding the highest mean AUROC for each label was selected as the best predictor. These final models were calibrated using `CalibratedClassifierCV` with sigmoid calibration (Platt scaling). See Table A.6 for details on the best performing configurations.

BoW A bag-of-words (BoW) representation was constructed on these notes using the "CountVectorizer" in scikit-learn, restricted to a vocabulary of the top 8,192 most frequent tokens (after removing English stop words), with "min_df=5" and "max_df=0.90" to filter rare and overly common tokens/words.

We evaluated seven distinct classification models provided in sklearn: `Perceptron, RidgeClassifierCV, PassiveAggressiveClassifier, RandomForestClassifier, DecisionTreeClassifier, LinearSVC, and SGDClassifier`. These were chosen because of their interpretable features/coefficients.

We performed an automated search over the interpretable model types and hyperparameters in the same manner as Section 3.2. See Supplementary Tables Table A.10 and Table A.11 for specific grids), and Table A.13 for details on the best performing classifiers. Following the selection of the best classifier, we examined the most important features by extracting the feature importances/coefficients and subsetting to the top ten contributing terms in the clinical text (refer to Table A.12 for results).

Calibration Metrics To evaluate the reliability of the vision and text models' probability estimates, we computed the Adaptive Calibration Error (ACE). Unlike the static Expected Calibration Error (ECE), which uses bins of equal width (e.g., 0-0.1, 0.1-0.2), ACE utilizes an adaptive binning scheme where bins are constructed to contain an equal number of samples. This approach was chosen for its robustness in imbalanced datasets where model predictions may be heavily skewed toward low probabilities.

We partition the samples into disjoint bins such that the number of samples in each bin is approximately equal:

$$|B_k| \approx \frac{N}{K} \tag{2}$$

For each bin , we compute the average predicted probability and the observed label proportion :

$$\bar{p}_k = \frac{1}{|B_k|} \sum_{i \in B_k} p_i, \quad \bar{y}_k = \frac{1}{|B_k|} \sum_{i \in B_k} y_i \tag{3}$$

The ACE is then calculated as the mean absolute calibration error across these equal-frequency bins:

$$\text{ACE} = \frac{1}{K} \sum_{k=1}^{K} |\bar{y}_k - \bar{p}_k| \tag{4}$$

Additionally, we reported the Brier Skill Score (BSS), which measures the relative improvement in mean squared error compared to a baseline. We first compute the Brier Score (BS) as:

$$BS = \frac{1}{N} \sum_{i=1}^{N} (y_i - p_i)^2 \tag{5}$$

The Brier Skill Score (BSS) normalizes this metric against a baseline reference classifier that has no discriminative ability and simply predicts the sample prevalence. We define the reference Brier Score, as the Brier Score obtained if the predicted probability for every instance was fixed to the global disease prevalence. The Brier Skill Score is defined as:

$$BSS = 1 - \frac{BS}{BS_{\text{ref}}} \tag{6}$$

A BSS of 0 indicates the model performs no better than the prevalence baseline, while a BSS of 1 indicates perfect prediction. Negative values indicate performance worse than the baseline.

Appendix B. Additional Vision Model Results

	Our Architectures (AUROC)		Comparison (AUROC)
Label	DenseNet121	ResNet50	Chexclusion
Atelectasis	0.8008	0.7379	0.837
Cardiomegaly	0.8108	0.7355	0.828
Consolidation	0.8182	0.7246	0.844
Edema	0.8914	0.8398	0.904
Enlarged Cardiomediastinum	0.7074	0.6642	0.757
Fracture	0.681	0.6524	0.718
Lung Lesion	0.7565	0.6326	0.772
Lung Opacity	0.7491	0.6718	N/A
Pleural Effusion	0.9233	0.8322	N/A
Pleural Other	0.8061	0.7233	0.848
Pneumonia	0.7073	0.6173	0.748
Pneumothorax	0.8790	0.7818	0.903
Support Devices	0.906	0.8490	0.927

Table A.1: Comparisons of our trained DenseNet121 and ResNet50 vision models against the state-of-the-art Chexclusion model from Seyyed-Kalantari et al. (2020b).

Label	BioMedClip AUROC
Atelectasis	0.507
Cardiomegaly	0.646
Consolidation	0.587
Edema	0.698
Enlarged Cardiomediastinum	0.551
Fracture	0.499
Lung Lesion	0.545
Lung Opacity	0.563
Pleural Effusion	0.699
Pleural Other	0.444
Pneumonia	0.564
Pneumothorax	0.576
Support Devices	0.727

Table A.2: Performance of the BioMedClip Vision Model on Test Data

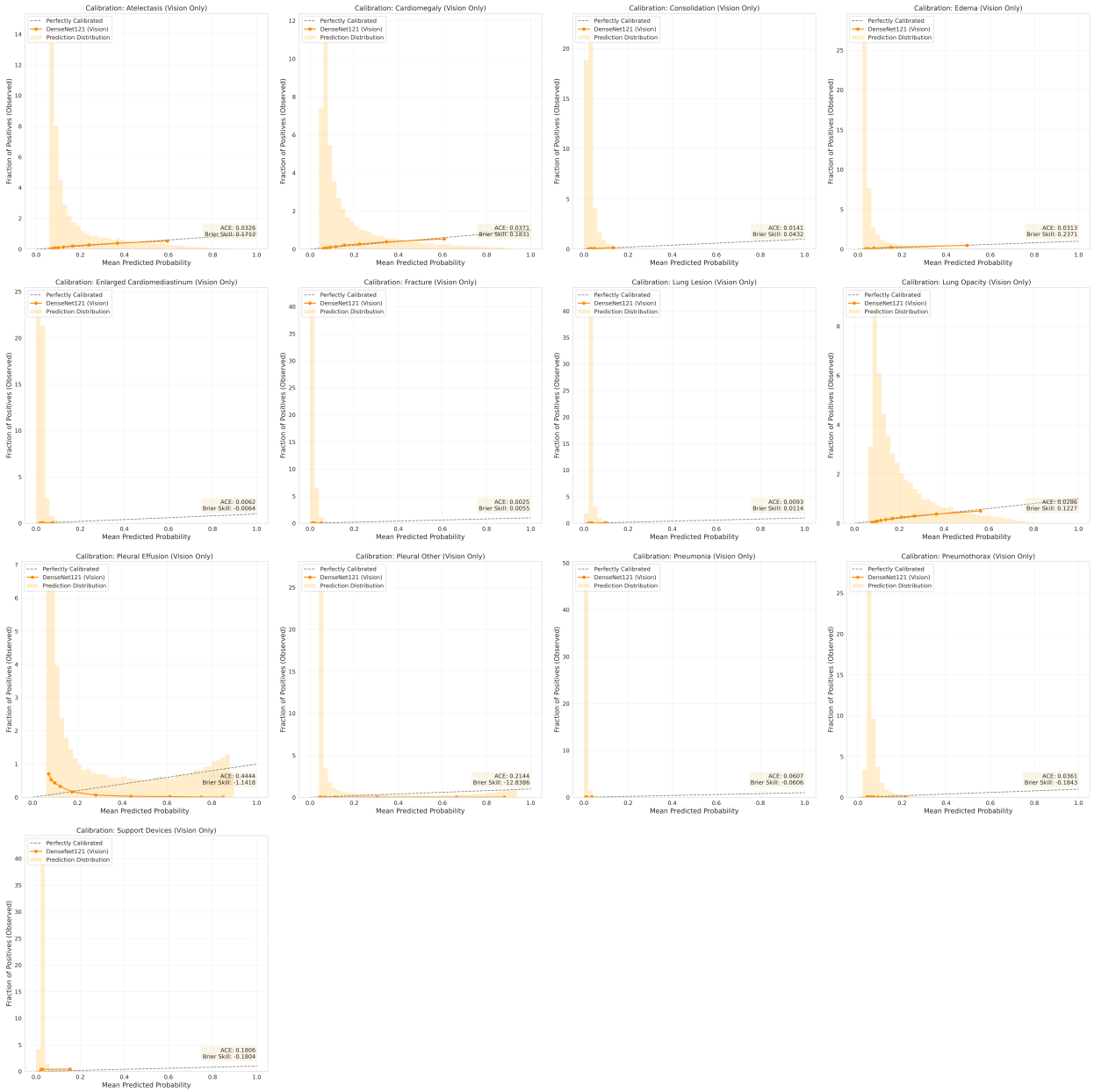

Figure A.1: Calibration Plots for Vision Model

Label	Negatives	Positives	Proportion of Positives
Atelectasis	171488	37812	0.22
Cardiomegaly	169585	39715	0.234
Consolidation	200875	8425	0.042
Edema	185961	23339	0.126
Enlarged Cardiomediastinum	203854	5446	0.027
Fracture	205555	3745	0.018
Lung Lesion	202778	6522	0.032
Lung Opacity	164244	45056	0.274
Pleural Effusion	161091	48209	0.299
Pleural Other	207075	2225	0.011
Pneumonia	193869	15431	0.079
Pneumothorax	202296	7004	0.035
Support Devices	163275	46025	0.282

Table A.3: Distribution of Positive and Negative CXR Labels in Our Dataset

Appendix C. Additional Text Model Results

Table A.4: Overview of evaluated LM text classifiers and their base configurations

Classifier	**Base Configuration**
Perceptron	`penalty='l2'`, `random_state=42`
RidgeClassifierCV	`alphas=np.logspace(-3, 3, 7)`
PassiveAggressiveClassifier	`random_state=42`
GaussianNB	default parameters
LinearDiscriminantAnalysis	`shrinkage='auto'`, `solver='lsqr'`
RandomForestClassifier	`random_state=42`
KNeighborsClassifier	default parameters
MLPClassifier	`early_stopping=True`, `validation_fraction=0.1`, `max_iter=300`, `random_state=42`
LinearSVC	`max_iter=1000`, `random_state=42`
SGDClassifier	`loss='log_loss'`, `penalty='l2'`, `random_state=42`

Table A.5: Hyperparameter grids used for each LM classifier during cross-validation

Classifier	**Hyperparameter Grid**
Perceptron	`alpha: [1e-5, 1e-4, 1e-3]`
RidgeClassifierCV	`alphas: [np.logspace(-4, 4, 9)]`
PassiveAggressiveClassifier	`C: [0.01, 0.1, 0.5, 1.0]`
GaussianNB	*(No hyperparameters tuned)*
LinearDiscriminantAnalysis	`solver: ['lsqr', 'eigen']`, `shrinkage: [None, 'auto']`
RandomForestClassifier	`n_estimators: [50, 100, 150]`, `max_depth: [None, 5, 10]`, `min_samples_leaf: [1, 5, 10]`
KNeighborsClassifier	`n_neighbors: [3, 5, 7, 9]`
MLPClassifier	`alpha: [1e-5, 1e-4, 1e-3]`, `hidden_layer_sizes: [(64,), (128,), (64,32)]`
LinearSVC	`C: [0.1, 1.0, 10.0]`
SGDClassifier	`alpha: [1e-5, 1e-4, 1e-3]`, `penalty: ['l1', 'l2']`

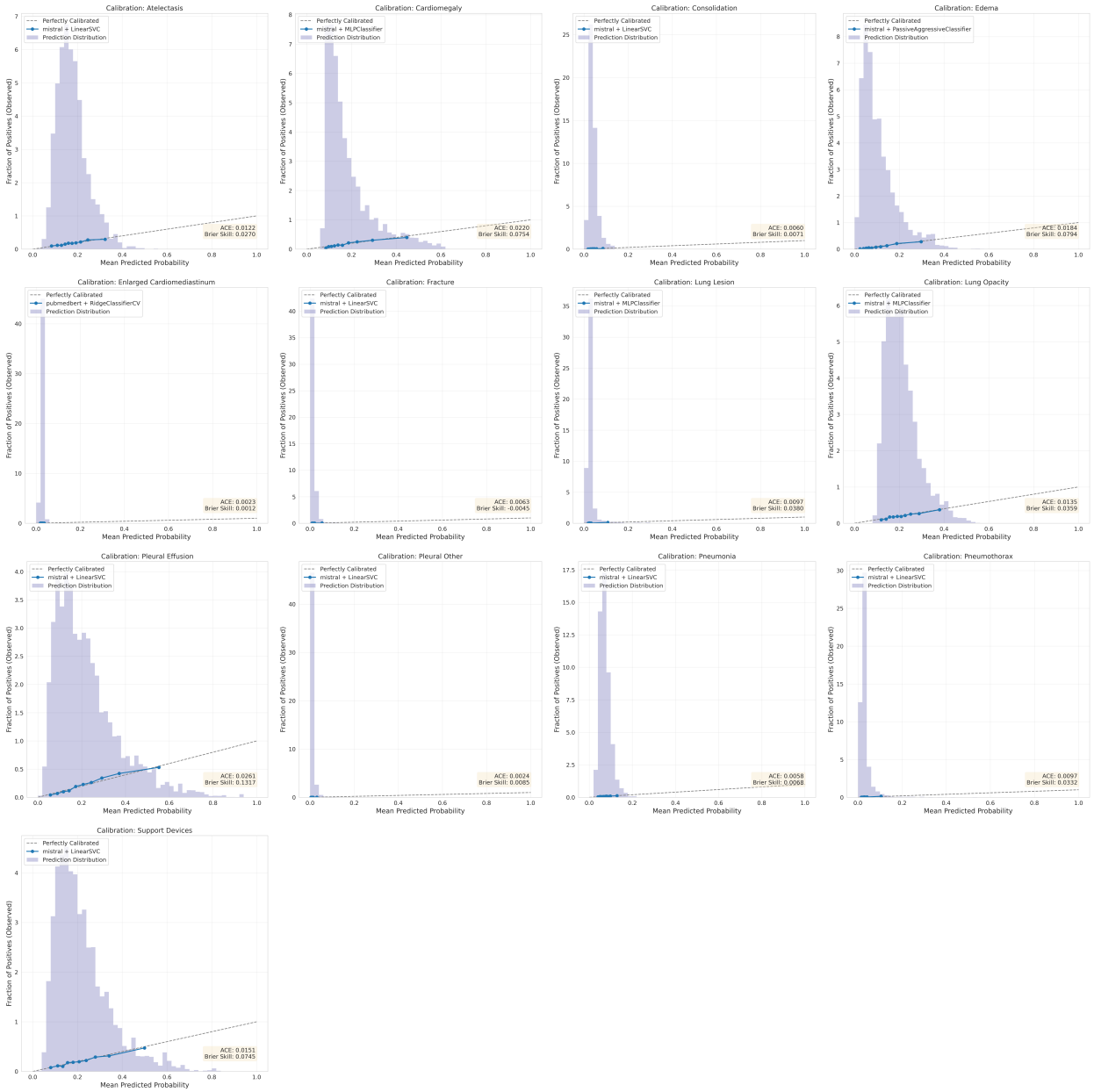

Figure A.2: Calibration Plots for Text Classifiers Trained on Prior Context

Table A.6: Best performing classifiers per label and Language Model

Label	BERT	BioLinkBERT	ClinicalBERT	Mistral	PubMedBERT	RoBERTa
Atelectasis	0.589 (SVC)	0.606 (SVC)	0.608 (SVC)	0.623 (SVC)	0.609 (SVC)	0.603 (SVC)
Cardiomegaly	0.684 (SVC)	0.695 (SVC)	0.683 (LinearDisc)	0.701 (MLP)	0.693 (SVC)	0.683 (Ridge)
Consolidation	0.613 (SVC)	0.624 (SVC)	0.623 (Ridge)	0.633 (SVC)	0.621 (SVC)	0.610 (SVC)
Edema	0.730 (PassAgg)	0.746 (SVC)	0.738 (SVC)	0.751 (PassAgg)	0.741 (SVC)	0.730 (SVC)
Enlarged Cardiomediastinum	0.558 (RF)	0.560 (Ridge)	0.569 (SGD)	0.569 (SVC)	0.566 (Ridge)	0.557 (SVC)
Fracture	0.620 (Ridge)	0.655 (SVC)	0.617 (SGD)	0.676 (SVC)	0.628 (SVC)	0.614 (SVC)
Lung Lesion	0.691 (SVC)	0.768 (SVC)	0.742 (SVC)	0.765 (MLP)	0.759 (SVC)	0.737 (SVC)
Lung Opacity	0.604 (Ridge)	0.625 (PassAgg)	0.614 (SVC)	0.628 (MLP)	0.620 (SVC)	0.607 (SVC)
Pleural Effusion	0.696 (MLP)	0.724 (Ridge)	0.711 (Percep)	0.745 (SVC)	0.728 (MLP)	0.721 (MLP)
Pleural Other	0.611 (SVC)	0.652 (RF)	0.682 (Ridge)	0.654 (SVC)	0.659 (SVC)	0.622 (SVC)
Pneumonia	0.561 (SVC)	0.577 (SVC)	0.555 (LinearDisc)	0.592 (SVC)	0.580 (Ridge)	0.581 (SVC)
Pneumothorax	0.689 (SVC)	0.695 (PassAgg)	0.705 (SVC)	0.712 (SVC)	0.683 (PassAgg)	0.694 (PassAgg)
Support Devices	0.663 (SVC)	0.672 (MLP)	0.623 (GaussianNB)	0.682 (SVC)	0.673 (SVC)	0.670 (SVC)

Table A.7: Calibration performance of text-only classifiers. We report the Adaptive Calibration Error (ACE), Brier Score, and Brier Skill Score (BSS) for the selected models trained on prior clinical notes.

Label	**Language Model**	**Classifier**	**ACE**	**Brier Score**	**Brier Skill Score**
Atelectasis	Mistral	LinearSVC	0.012	0.144	0.027
Cardiomegaly	Mistral	MLPClassifier	0.022	0.135	0.075
Consolidation	Mistral	LinearSVC	0.006	0.034	0.007
Edema	Mistral	PassiveAggressiveClassifier	0.018	0.080	0.079
Enlarged Cardiomediastinum	PubMedBERT	RidgeClassifierCV	0.002	0.024	0.001
Fracture	Mistral	LinearSVC	0.006	0.017	-0.005
Lung Lesion	Mistral	MLPClassifier	0.010	0.028	0.038
Lung Opacity	Mistral	MLPClassifier	0.014	0.156	0.036
Pleural Effusion	Mistral	LinearSVC	0.026	0.154	0.132
Pleural Other	Mistral	LinearSVC	0.002	0.010	0.008
Pneumonia	Mistral	LinearSVC	0.006	0.066	0.007
Pneumothorax	Mistral	LinearSVC	0.010	0.034	0.033
Support Devices	Mistral	LinearSVC	0.015	0.156	0.074

Table A.12: Top Features Ranked by Importance by the BoW Text Classifiers. Each feature corresponds to a token from the bag-of-words vocabulary. Importance scores correspond to the magnitude of the coefficient assigned to the feature by the classifier. Across all image labels, clinically meaningful words consistently appeared in the top ten features, which suggests the importance of specific language features within the discharge notes as strong indicators of disease status.

Feature	Rank	Label Name	Classifier Name
metastatic	1	Lung Lesion	RandomForestClassifier
metastasis	2	Lung Lesion	RandomForestClassifier
metastases	3	Lung Lesion	RandomForestClassifier
lung	4	Lung Lesion	RandomForestClassifier
nsclc	5	Lung Lesion	RandomForestClassifier
oncology	6	Lung Lesion	RandomForestClassifier
il2	7	Lung Lesion	RandomForestClassifier
tumor	8	Lung Lesion	RandomForestClassifier

Continued on next page

Table A.12: Top Features Ranked by Importance by the BoW Text Classifiers.

Feature	Rank	Label Name	Classifier Name
nodules	9	Lung Lesion	RandomForestClassifier
nonsmall	10	Lung Lesion	RandomForestClassifier
pleural	1	Pleural Effusion	RandomForestClassifier
thoracentesis	2	Pleural Effusion	RandomForestClassifier
effusion	3	Pleural Effusion	RandomForestClassifier
effusions	4	Pleural Effusion	RandomForestClassifier
atelectasis	5	Pleural Effusion	RandomForestClassifier
lasix	6	Pleural Effusion	RandomForestClassifier
compressive	7	Pleural Effusion	RandomForestClassifier
tube	8	Pleural Effusion	RandomForestClassifier
pleurx	9	Pleural Effusion	RandomForestClassifier
cytology	10	Pleural Effusion	RandomForestClassifier
pneumothorax	1	Pneumothorax	RandomForestClassifier
cardiothoracic	2	Pneumothorax	RandomForestClassifier
ptx	3	Pneumothorax	RandomForestClassifier
apical	4	Pneumothorax	RandomForestClassifier
tube	5	Pneumothorax	RandomForestClassifier
nl	6	Pneumothorax	RandomForestClassifier
spontaneous	7	Pneumothorax	RandomForestClassifier
pigtail	8	Pneumothorax	RandomForestClassifier
abnormal	9	Pneumothorax	RandomForestClassifier
vats	10	Pneumothorax	RandomForestClassifier
lasix	1	Edema	RandomForestClassifier
diuresed	2	Edema	RandomForestClassifier
cardiomegaly	3	Edema	RandomForestClassifier
furosemide	4	Edema	RandomForestClassifier
failure	5	Edema	RandomForestClassifier
ef	6	Edema	RandomForestClassifier
congestive	7	Edema	RandomForestClassifier
chf	8	Edema	RandomForestClassifier
diuresis	9	Edema	RandomForestClassifier
diastolic	10	Edema	RandomForestClassifier
cardiomegaly	1	Cardiomegaly	RandomForestClassifier
lasix	2	Cardiomegaly	RandomForestClassifier
furosemide	3	Cardiomegaly	RandomForestClassifier
diuresed	4	Cardiomegaly	RandomForestClassifier
fibrillation	5	Cardiomegaly	RandomForestClassifier
congestive	6	Cardiomegaly	RandomForestClassifier
ef	7	Cardiomegaly	RandomForestClassifier
diastolic	8	Cardiomegaly	RandomForestClassifier
chf	9	Cardiomegaly	RandomForestClassifier
atrial	10	Cardiomegaly	RandomForestClassifier
prior	1	Fracture	DecisionTreeClassifier
rib	2	Fracture	DecisionTreeClassifier
clavicle	3	Fracture	DecisionTreeClassifier

Continued on next page

Table A.12: Top Features Ranked by Importance by the BoW Text Classifiers.

Feature	Rank	Label Name	Classifier Name
velcade	4	Fracture	DecisionTreeClassifier
0945pm	5	Fracture	DecisionTreeClassifier
acute	6	Fracture	DecisionTreeClassifier
completing	7	Fracture	DecisionTreeClassifier
abdomen	8	Fracture	DecisionTreeClassifier
0220pm	9	Fracture	DecisionTreeClassifier
18	10	Fracture	DecisionTreeClassifier
picc	1	Support Devices	RandomForestClassifier
placement	2	Support Devices	RandomForestClassifier
tube	3	Support Devices	RandomForestClassifier
transferred	4	Support Devices	RandomForestClassifier
osh	5	Support Devices	RandomForestClassifier
line	6	Support Devices	RandomForestClassifier
flush	7	Support Devices	RandomForestClassifier
pacemaker	8	Support Devices	RandomForestClassifier
tip	9	Support Devices	RandomForestClassifier
svc	10	Support Devices	RandomForestClassifier
pneumonia	1	Consolidation	RandomForestClassifier
bronchoscopy	2	Consolidation	RandomForestClassifier
lobe	3	Consolidation	RandomForestClassifier
pleural	4	Consolidation	RandomForestClassifier
bronchiectasis	5	Consolidation	RandomForestClassifier
lung	6	Consolidation	RandomForestClassifier
tube	7	Consolidation	RandomForestClassifier
opacities	8	Consolidation	RandomForestClassifier
hcap	9	Consolidation	RandomForestClassifier
effusion	10	Consolidation	RandomForestClassifier
atelectasis	1	Atelectasis	RandomForestClassifier
pleural	2	Atelectasis	RandomForestClassifier
bibasilar	3	Atelectasis	RandomForestClassifier
collapse	4	Atelectasis	RandomForestClassifier
thoracentesis	5	Atelectasis	RandomForestClassifier
tube	6	Atelectasis	RandomForestClassifier
compressive	7	Atelectasis	RandomForestClassifier
effusions	8	Atelectasis	RandomForestClassifier
furosemide	9	Atelectasis	RandomForestClassifier
facility	10	Atelectasis	RandomForestClassifier
pneumonia	1	Lung Opacity	RandomForestClassifier
opacities	2	Lung Opacity	RandomForestClassifier
lung	3	Lung Opacity	RandomForestClassifier
bronchoscopy	4	Lung Opacity	RandomForestClassifier
opacity	5	Lung Opacity	RandomForestClassifier
aspiration	6	Lung Opacity	RandomForestClassifier
lobe	7	Lung Opacity	RandomForestClassifier
bronchiectasis	8	Lung Opacity	RandomForestClassifier

Continued on next page

Table A.12: Top Features Ranked by Importance by the BoW Text Classifiers.

Feature	Rank	Label Name	Classifier Name
sputum	9	Lung Opacity	RandomForestClassifier
copd	10	Lung Opacity	RandomForestClassifier
icd	1	Pleural Other	SGDClassifier
whipple	2	Pleural Other	SGDClassifier
gastropathy	3	Pleural Other	SGDClassifier
crohns	4	Pleural Other	SGDClassifier
aml	5	Pleural Other	SGDClassifier
cerebral	6	Pleural Other	SGDClassifier
hearing	7	Pleural Other	SGDClassifier
t12	8	Pleural Other	SGDClassifier
bacteremia	9	Pleural Other	SGDClassifier
paroxetine	10	Pleural Other	SGDClassifier
pneumonia	1	Pneumonia	DecisionTreeClassifier
bronchiectasis	2	Pneumonia	DecisionTreeClassifier
congestion	3	Pneumonia	DecisionTreeClassifier
ran	4	Pneumonia	DecisionTreeClassifier
inhibitor	5	Pneumonia	DecisionTreeClassifier
pump	6	Pneumonia	DecisionTreeClassifier
multifocal	7	Pneumonia	DecisionTreeClassifier
cefepime	8	Pneumonia	DecisionTreeClassifier
poorly	9	Pneumonia	DecisionTreeClassifier
emphysema	10	Pneumonia	DecisionTreeClassifier
tube	1	Enlarged Cardiomediastinum	RandomForestClassifier
cardiothoracic	2	Enlarged Cardiomediastinum	RandomForestClassifier
pleurex	3	Enlarged Cardiomediastinum	RandomForestClassifier
dissection	4	Enlarged Cardiomediastinum	RandomForestClassifier
wout	5	Enlarged Cardiomediastinum	RandomForestClassifier
jtube	6	Enlarged Cardiomediastinum	RandomForestClassifier
pox	7	Enlarged Cardiomediastinum	RandomForestClassifier
esophagectomy	8	Enlarged Cardiomediastinum	RandomForestClassifier
wk	9	Enlarged Cardiomediastinum	RandomForestClassifier
oxygenation	10	Enlarged Cardiomediastinum	RandomForestClassifier

Table A.8: Representative examples of clinical notes with high and low predicted probabilities for Pleural Effusion. We truncate the raw notes to the relevant portions for clarity, but the full set of notes may be obtained using the Subject and Study ID, which we provide for reference.

High Predicted Probability (True Positive)
Subject ID: 14965197 **Study ID:** 53934290
Model Probability: 0.9237 **Reference Label:** 1 (Positive)

```
name:  ___                    unit no:   ___
service: medicine
...
chief complaint:
massive pleural effusion

major surgical or invasive procedure:
chest tube placed ___ by ip
thoracentesis ___ by ip

history of present illness:
... found to have massive pleural effusion.
thoracentesis on ___ removed 1.5l of fluid.
```

Low Predicted Probability (True Negative)
Subject ID: 10486638 **Study ID:** 57867200
Model Probability: 0.0003 **Reference Label:** 0 (Negative)

```
name:  ___                     unit no:   ___
service: medicine
...
chief complaint:
chest pain and vertigo
...
physical examination:
lungs: clear to auscultation bilaterally. no wheezes, rales, or rhonchi.
...
brief hospital course:
chest x-ray showed no pleural effusion or pneumothorax.
```

Table A.9: Calibration performance of BoW classifiers. We report the Adaptive Calibration Error (ACE), Brier Score, and Brier Skill Score (BSS) for the selected models trained on prior clinical notes.

Label	Classifier	ACE	Brier Score	Brier Skill Score
Atelectasis	RandomForestClassifier	0.014	0.144	0.028
Cardiomegaly	RandomForestClassifier	0.015	0.136	0.068
Consolidation	RandomForestClassifier	0.009	0.034	0.003
Edema	RandomForestClassifier	0.016	0.082	0.053
Enlarged Cardiomediastinum	RandomForestClassifier	0.004	0.024	0.001
Fracture	DecisionTreeClassifier	0.003	0.017	0.021
Lung Lesion	RandomForestClassifier	0.004	0.027	0.056
Lung Opacity	RandomForestClassifier	0.012	0.157	0.030
No Finding	RandomForestClassifier	0.017	0.200	0.122
Pleural Effusion	RandomForestClassifier	0.016	0.154	0.132
Pleural Other	SGDClassifier	0.003	0.010	0.001
Pneumonia	DecisionTreeClassifier	0.006	0.067	0.003
Pneumothorax	RandomForestClassifier	0.006	0.033	0.043
Support Devices	RandomForestClassifier	0.018	0.157	0.070

Table A.10: Overview of BoW-based classifiers and their base configurations

Classifier	Base Configuration
Perceptron	`penalty='l2'`, `random_state=42`, `max_iter=1000`, `tol=1e-3`
RidgeClassifierCV	`alphas=np.logspace(-3, 3, 7)`
PassiveAggressiveClassifier	`random_state=42`, `max_iter=1000`, `tol=1e-3`
RandomForestClassifier	`random_state=42`
LinearSVC	`max_iter=2000`, `random_state=42`, `dual='auto'`
SGDClassifier	`loss='log_loss'`, `penalty='l2'`, `random_state=42`, `max_iter=1000`, `tol=1e-3`
DecisionTreeClassifier	`random_state=42`

Table A.11: Hyperparameter grids used for BoW classifiers during cross-validation

Classifier	Hyperparameter Grid
Perceptron	`alpha: [1e-6, 1e-5, 1e-4]`
RidgeClassifierCV	`alphas: [np.logspace(-4, 4, 9)]`
PassiveAggressiveClassifier	`C: [0.001, 0.01, 0.1, 1.0]`
RandomForestClassifier	`n_estimators: [50, 100, 200]`, `max_depth: [10, 20, None]`, `min_samples_leaf: [1, 5, 10]`
LinearSVC	`C: [0.01, 0.1, 1.0, 10.0]`
SGDClassifier	`alpha: [1e-6, 1e-5, 1e-4]`, `penalty: ['l1', 'l2', 'elasticnet']`
DecisionTreeClassifier	`max_depth: [5, 10, 20, None]`, `min_samples_leaf: [1, 5, 10, 20]`

Table A.13: Best Performing BoW Text Classifiers. We train classifiers to predict the disease labels of CXR images based on prior clinical notes without looking at the image itself. All models use an BoW embedding of prior clinical notes, and are trained on top of that representation. We note the specific label, classifier type, and best test set AUROC achieved.

Label	Classifier	AUROC
Lung Lesion	RandomForestClassifier	0.759
Pleural Effusion	RandomForestClassifier	0.734
Edema	RandomForestClassifier	0.717
Pneumothorax	RandomForestClassifier	0.703
Cardiomegaly	RandomForestClassifier	0.688
Support Devices	RandomForestClassifier	0.681
Fracture	DecisionTreeClassifier	0.657
Atelectasis	RandomForestClassifier	0.623
Lung Opacity	RandomForestClassifier	0.617
Consolidation	RandomForestClassifier	0.611
Pleural Other	SGDClassifier	0.577
Enlarged Cardiomediastinum	RandomForestClassifier	0.573
Pneumonia	DecisionTreeClassifier	0.550

Table A.14: XGBoost Classifier Performance. Predicting disease labels of CXR images using XGBoost. We note the specific label, classifier type, and best test set AUROC achieved.

Label	Classifier	AUROC
Atelectasis	XGBoost	0.640
Cardiomegaly	XGBoost	0.712
Consolidation	XGBoost	0.627
Edema	XGBoost	0.741
Enlarged Cardiomediastinum	XGBoost	0.574
Fracture	XGBoost	0.737
Lung Lesion	XGBoost	0.800
Lung Opacity	XGBoost	0.626
Pleural Effusion	XGBoost	0.751
Pleural Other	XGBoost	0.704
Pneumonia	XGBoost	0.581
Pneumothorax	XGBoost	0.736
Support Devices	XGBoost	0.675

Table A.15: Adapted CheXpert phrases used for prior mention detection. Since the original CheXpert phrases were designed for radiology reports and were not fully applicable to the clinical notes we used, we adapted the original phrase list primarily by removing terms to better align with our clinical context, where ~~strikethrough~~ indicates a term we removed. Due to excessive processing time, we did not perform classification of mentions (e.g., into positive or negative mentions) since the Chexpert labeler relies on a syntactic parser (BLLIP parser) to process the entire note, which generates an excessive number of possible parse tree structures. This process was not feasible to run on discharge summaries, even when broken down into individual sentences. Instead, we performed direct substring matching. To make the provided terms suitable for substring matching, we also removed underscores (e.g., in "drain_").

Label	**Adapted Phrases**
Atelectasis	atelecta, ~~collapse~~
Cardiomegaly	cardiomegaly, ~~the heart~~, heart size, cardiac enlargement, cardiac size, cardiac shadow, cardiac contour, cardiac silhouette, enlarged heart
Consolidation	consolidat
Edema	edema, heart failure, chf, vascular congestion, pulmonary congestion, ~~indistinctness~~, vascular prominence
Enlarged Cardiomediastinum	~~_~~mediastinum, cardiomediastinum, ~~contour~~, mediastinal configuration, mediastinal silhouette, pericardial silhouette, cardiac silhouette and vascularity
Fracture	fracture
Lung Lesion	~~mass~~, nodular density, nodular densities, nodular opacity, nodular opacities, nodular opacification, nodule, ~~lump~~, cavitary lesion, carcinoma, neoplasm, tumor
Lung Opacity	opaci, decreased translucency, ~~increased density~~, airspace disease, air-space disease, air space disease, infiltrate, infiltration, interstitial marking, interstitial pattern, interstitial lung, reticular pattern, reticular marking, reticulation, parenchymal scarring, peribronchial thickening, ~~wall thickening~~, ~~scar~~
Pleural Effusion	pleural fluid, effusion
Pleural Other	pleural thickening, ~~fibrosis~~, fibrothorax, pleural scar, pleural parenchymal scar, pleuro-parenchymal scar, pleuro-pericardial scar
Pneumonia	pneumonia, ~~infection~~, ~~infectious process~~, ~~infectious~~
Pneumothorax	pneumothorax, pneumothoraces
Support Devices	pacer, ~~_line_~~, ~~lines~~, picc, tube, valve, catheter, pacemaker, hardware, arthroplast, marker, icd, defib, device, drain~~_~~, plate, screw, cannula, apparatus, coil, ~~support~~, equipment, mediport

Appendix D. Additional Experimental Results

Table A.16: Held-out performance (in terms of **AUROC**) of CXR models across sub-populations stratified by pre-test probability (LM embeddings) of the CXR label. Labels marked with an asterisk (*) indicate a statistically significant difference between the Bottom 25% and Top 25% groups.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis*	0.831 (0.814–0.847)	0.795 (0.784–0.806)	0.738 (0.724–0.753)
Cardiomegaly*	0.809 (0.787–0.831)	0.786 (0.774–0.797)	0.728 (0.714–0.742)
Consolidation*	0.854 (0.812–0.892)	0.800 (0.779–0.821)	0.778 (0.752–0.801)
Edema*	0.900 (0.874–0.924)	0.887 (0.876–0.898)	0.812 (0.798–0.825)
Enlarged Cardiomediastinum	0.706 (0.653–0.757)	0.725 (0.694–0.755)	0.658 (0.615–0.699)
Fracture	0.628 (0.545–0.709)	0.681 (0.630–0.729)	0.660 (0.616–0.703)
Lung Lesion	0.733 (0.668–0.794)	0.704 (0.667–0.742)	0.695 (0.669–0.722)
Lung Opacity*	0.776 (0.756–0.796)	0.730 (0.719–0.742)	0.692 (0.677–0.707)
Pleural Effusion*	0.942 (0.931–0.953)	0.909 (0.902–0.916)	0.876 (0.867–0.885)
Pleural Other	0.837 (0.787–0.884)	0.798 (0.748–0.843)	0.759 (0.710–0.805)
Pneumonia	0.708 (0.674–0.741)	0.697 (0.677–0.716)	0.693 (0.669–0.717)
Pneumothorax	0.862 (0.826–0.896)	0.880 (0.861–0.899)	0.842 (0.820–0.862)
Support Devices*	0.918 (0.903–0.932)	0.904 (0.896–0.911)	0.858 (0.847–0.868)

Table A.17: Held-out performance (in terms of **sensitivity at local 95% specificity**) of CXR models across sub-populations stratified by pre-test probability (LM embeddings) of the CXR label. Labels marked with an asterisk (*) indicate a statistically significant difference between the Bottom 25% and Top 25% groups.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis*	0.330 (0.261–0.402)	0.276 (0.241–0.313)	0.184 (0.143–0.226)
Cardiomegaly	0.293 (0.216–0.379)	0.268 (0.225–0.315)	0.186 (0.147–0.224)
Consolidation	0.440 (0.286–0.615)	0.247 (0.177–0.327)	0.232 (0.154–0.314)
Edema*	0.595 (0.454–0.731)	0.447 (0.390–0.506)	0.299 (0.245–0.354)
Enlarged Cardiomediastinum	0.179 (0.073–0.292)	0.201 (0.129–0.284)	0.107 (0.044–0.189)
Fracture	0.188 (0.027–0.378)	0.187 (0.089–0.298)	0.182 (0.101–0.275)
Lung Lesion	0.120 (0.000–0.295)	0.178 (0.098–0.274)	0.177 (0.119–0.240)
Lung Opacity*	0.284 (0.223–0.347)	0.192 (0.158–0.227)	0.167 (0.127–0.208)
Pleural Effusion*	0.651 (0.561–0.747)	0.588 (0.542–0.635)	0.442 (0.366–0.519)
Pleural Other	0.286 (0.086–0.514)	0.306 (0.153–0.472)	0.277 (0.143–0.419)
Pneumonia	0.219 (0.148–0.297)	0.174 (0.129–0.221)	0.189 (0.138–0.242)
Pneumothorax	0.426 (0.263–0.592)	0.509 (0.415–0.596)	0.448 (0.371–0.532)
Support Devices*	0.714 (0.647–0.780)	0.595 (0.545–0.639)	0.412 (0.342–0.489)

Table A.18: Held-out performance (in terms of **sensitivity at global 95% specificity**) of CXR models across sub-populations stratified by pre-test probability (LM embeddings) of the CXR label. Labels marked with an asterisk (*) indicate a statistically significant difference between the Bottom 25% and Top 25% groups.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis*	0.192 (0.144–0.245)	0.235 (0.201–0.270)	0.329 (0.287–0.372)
Cardiomegaly*	0.047 (0.017–0.084)	0.193 (0.162–0.225)	0.403 (0.366–0.443)
Consolidation	0.313 (0.176–0.462)	0.207 (0.142–0.277)	0.396 (0.311–0.480)
Edema*	0.246 (0.130–0.378)	0.391 (0.333–0.449)	0.521 (0.474–0.565)
Enlarged Cardiomediastinum	0.141 (0.052–0.250)	0.194 (0.124–0.268)	0.143 (0.069–0.233)
Fracture	0.154 (0.022–0.333)	0.177 (0.089–0.274)	0.228 (0.143–0.317)
Lung Lesion*	0.045 (0.000–0.159)	0.151 (0.073–0.238)	0.348 (0.280–0.417)
Lung Opacity*	0.128 (0.091–0.172)	0.166 (0.141–0.194)	0.286 (0.247–0.327)
Pleural Effusion*	0.391 (0.311–0.470)	0.547 (0.509–0.587)	0.690 (0.655–0.726)
Pleural Other	0.172 (0.029–0.371)	0.223 (0.097–0.375)	0.401 (0.267–0.541)
Pneumonia	0.165 (0.103–0.236)	0.161 (0.119–0.206)	0.258 (0.205–0.315)
Pneumothorax*	0.268 (0.118–0.434)	0.434 (0.348–0.522)	0.601 (0.530–0.673)
Support Devices*	0.482 (0.416–0.549)	0.549 (0.510–0.590)	0.642 (0.602–0.684)

Table A.19: Held-out performance (in terms of **sensitivity at local 95% specificity**) of CXR models across sub-populations stratified by pre-test probability (LM embeddings) of the CXR label. Labels marked with an asterisk (*) indicate a statistically significant difference between the Bottom 25% and Top 25% groups.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis*	0.330 (0.261–0.402)	0.276 (0.241–0.313)	0.184 (0.143–0.226)
Cardiomegaly	0.293 (0.216–0.379)	0.268 (0.225–0.315)	0.186 (0.147–0.224)
Consolidation	0.440 (0.286–0.615)	0.247 (0.177–0.327)	0.232 (0.154–0.314)
Edema*	0.595 (0.454–0.731)	0.447 (0.390–0.506)	0.299 (0.245–0.354)
Enlarged Cardiomediastinum	0.179 (0.073–0.292)	0.201 (0.129–0.284)	0.107 (0.044–0.189)
Fracture	0.188 (0.027–0.378)	0.187 (0.089–0.298)	0.182 (0.101–0.275)
Lung Lesion	0.120 (0.000–0.295)	0.178 (0.098–0.274)	0.177 (0.119–0.240)
Lung Opacity*	0.284 (0.223–0.347)	0.192 (0.158–0.227)	0.167 (0.127–0.208)
Pleural Effusion*	0.651 (0.561–0.747)	0.588 (0.542–0.635)	0.442 (0.366–0.519)
Pleural Other	0.286 (0.086–0.514)	0.306 (0.153–0.472)	0.277 (0.143–0.419)
Pneumonia	0.219 (0.148–0.297)	0.174 (0.129–0.221)	0.189 (0.138–0.242)
Pneumothorax	0.426 (0.263–0.592)	0.509 (0.415–0.596)	0.448 (0.371–0.532)
Support Devices*	0.714 (0.647–0.780)	0.595 (0.545–0.639)	0.412 (0.342–0.489)

Table A.20: Held-out performance (in terms of **specificity at global 95% specificity**) of CXR models across sub-populations stratified by pre-test probability (LM embeddings) of the CXR label. Labels marked with an asterisk (*) indicate a statistically significant difference between the Bottom 25% and Top 25% groups.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis*	0.978 (0.972–0.984)	0.962 (0.957–0.967)	0.889 (0.878–0.899)
Cardiomegaly*	0.993 (0.989–0.996)	0.967 (0.963–0.972)	0.846 (0.834–0.857)
Consolidation*	0.977 (0.971–0.983)	0.959 (0.955–0.964)	0.902 (0.893–0.912)
Edema*	0.991 (0.987–0.995)	0.961 (0.957–0.966)	0.871 (0.860–0.882)
Enlarged Cardiomediastinum*	0.968 (0.961–0.974)	0.953 (0.948–0.957)	0.927 (0.919–0.935)
Fracture*	0.965 (0.958–0.971)	0.958 (0.954–0.962)	0.919 (0.909–0.927)
Lung Lesion*	0.981 (0.975–0.986)	0.969 (0.964–0.973)	0.877 (0.868–0.887)
Lung Opacity*	0.980 (0.974–0.986)	0.958 (0.953–0.962)	0.894 (0.883–0.905)
Pleural Effusion*	0.984 (0.979–0.989)	0.957 (0.953–0.962)	0.869 (0.855–0.883)
Pleural Other*	0.982 (0.977–0.987)	0.968 (0.964–0.972)	0.881 (0.872–0.890)
Pneumonia*	0.966 (0.959–0.972)	0.954 (0.949–0.958)	0.925 (0.916–0.934)
Pneumothorax*	0.974 (0.968–0.980)	0.968 (0.964–0.972)	0.886 (0.877–0.895)
Support Devices*	0.983 (0.977–0.988)	0.958 (0.953–0.963)	0.882 (0.870–0.894)

Table A.21: Held-out performance (in terms of **Adaptive Calibration Error (ACE)**) of CXR models across sub-populations stratified by pre-test probability (LM embeddings) of the CXR label.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis	0.039 (0.032–0.047)	0.030 (0.024–0.036)	0.046 (0.036–0.056)
Cardiomegaly	0.031 (0.027–0.037)	0.027 (0.021–0.033)	0.060 (0.049–0.070)
Consolidation	0.013 (0.011–0.016)	0.014 (0.012–0.017)	0.018 (0.015–0.021)
Edema	0.024 (0.023–0.026)	0.029 (0.024–0.033)	0.065 (0.056–0.073)
Enlarged Cardiomediastinum	0.008 (0.006–0.010)	0.007 (0.005–0.009)	0.009 (0.005–0.013)
Fracture	0.008 (0.007–0.010)	0.005 (0.004–0.006)	0.017 (0.016–0.018)
Lung Lesion	0.018 (0.017–0.018)	0.013 (0.012–0.014)	0.033 (0.031–0.036)
Lung Opacity	0.033 (0.028–0.039)	0.027 (0.021–0.034)	0.048 (0.038–0.058)
Pleural Effusion	0.039 (0.033–0.045)	0.037 (0.032–0.043)	0.057 (0.049–0.066)
Pleural Other	0.007 (0.006–0.008)	0.005 (0.004–0.006)	0.009 (0.007–0.011)
Pneumonia	0.015 (0.012–0.020)	0.016 (0.012–0.020)	0.024 (0.018–0.031)
Pneumothorax	0.014 (0.012–0.017)	0.017 (0.015–0.020)	0.032 (0.027–0.036)
Support Devices	0.030 (0.025–0.035)	0.032 (0.027–0.037)	0.044 (0.035–0.054)

Table A.22: Held-out performance (in terms of **Brier Skill Score (BSS)**) of CXR models across sub-populations stratified by pre-test probability (LM embeddings) of the CXR label.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis	0.157 (0.128–0.186)	0.167 (0.150–0.184)	0.123 (0.101–0.146)
Cardiomegaly	0.084 (0.059–0.109)	0.142 (0.125–0.160)	0.122 (0.099–0.145)
Consolidation	0.036 (−0.019–0.091)	0.028 (0.010–0.047)	0.047 (0.020–0.075)
Edema	0.085 (0.027–0.146)	0.190 (0.161–0.220)	0.187 (0.159–0.216)
Enlarged Cardiomediastinum	−0.007 (−0.023–0.009)	0.005 (−0.014–0.025)	−0.027 (−0.055–−0.004)
Fracture	−0.007 (−0.034–0.027)	0.002 (−0.009–0.013)	−0.003 (−0.016–0.010)
Lung Lesion	−0.052 (−0.087–−0.011)	−0.012 (−0.026–0.002)	−0.015 (−0.037–0.008)
Lung Opacity	0.120 (0.096–0.143)	0.100 (0.085–0.114)	0.085 (0.066–0.105)
Pleural Effusion	0.293 (0.238–0.348)	0.436 (0.414–0.458)	0.405 (0.379–0.431)
Pleural Other	−0.007 (−0.018–0.002)	0.011 (−0.046–0.071)	−0.003 (−0.041–0.036)
Pneumonia	0.019 (−0.008–0.047)	0.028 (0.013–0.045)	0.035 (0.012–0.059)
Pneumothorax	−0.026 (−0.109–0.054)	0.069 (0.019–0.118)	0.149 (0.106–0.191)
Support Devices	0.427 (0.385–0.469)	0.434 (0.411–0.457)	0.373 (0.345–0.401)

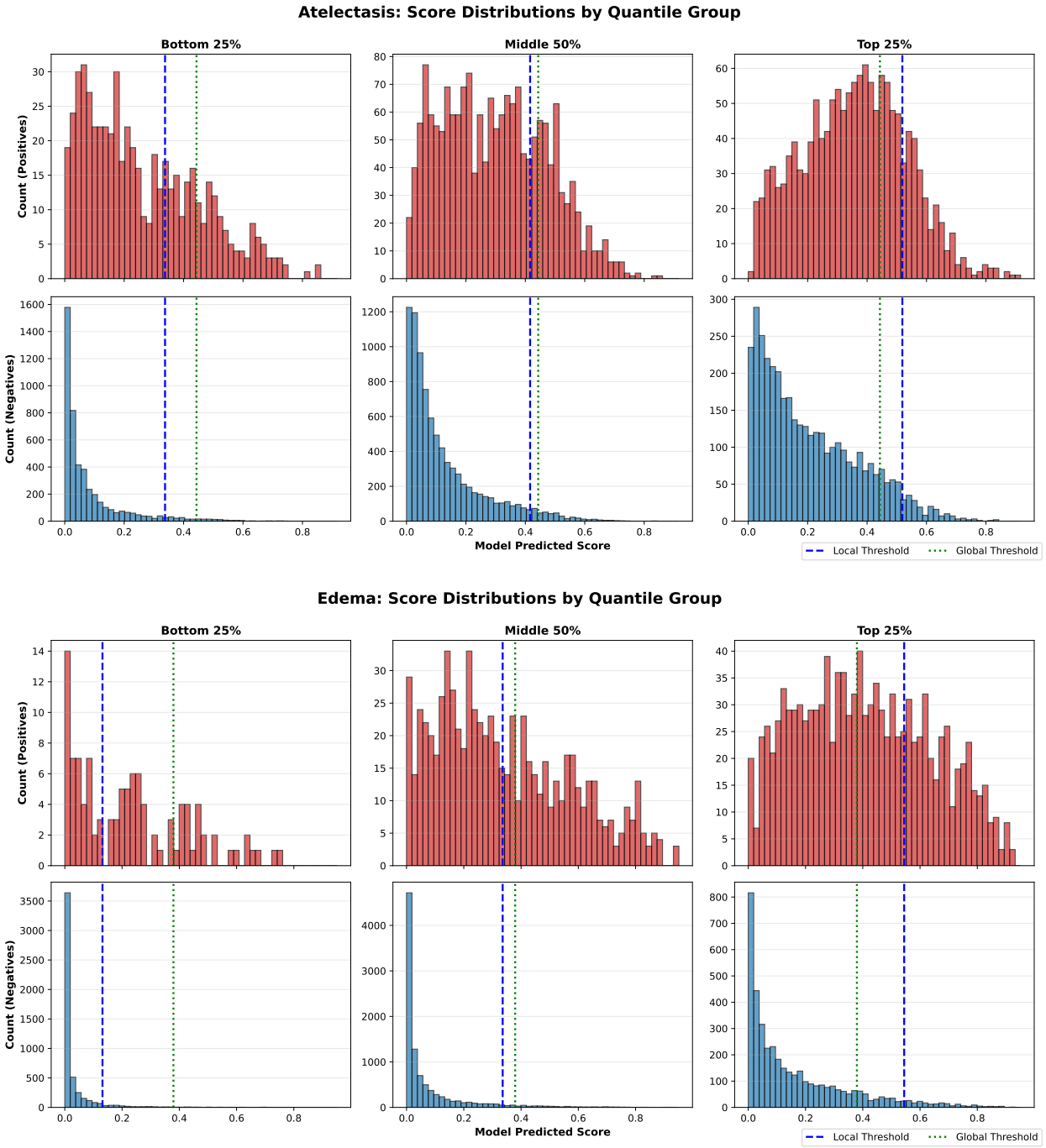

Figure A.3: Score distributions of the vision model predictions stratified by pre-test probability (LM embeddings) for two representative labels (Atelectasis and Edema). Histograms show the distribution of predicted scores for positive (top row) and negative (bottom row) cases across low-, medium-, and high-risk quantile groups defined by the text-based pre-test probability. Blue dashed lines indicate subgroup-specific (local) thresholds selected at 95% specificity within each quantile, while green dotted lines indicate a universal global threshold computed using all samples.

Table A.23: Held-out performance (in terms of **AUROC**) of CXR DenseNet121 models across sub-populations stratified by pre-test probability (BoW Representations) of the CXR label. Labels marked with an asterisk (*) indicate a statistically significant difference between the Bottom 25% and Top 25% groups.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis*	0.848 (0.824–0.872)	0.795 (0.779–0.811)	0.725 (0.703–0.746)
Cardiomegaly*	0.820 (0.788–0.850)	0.783 (0.766–0.799)	0.742 (0.721–0.762)
Consolidation*	0.862 (0.818–0.900)	0.825 (0.791–0.856)	0.751 (0.713–0.787)
Edema*	0.908 (0.873–0.938)	0.889 (0.873–0.903)	0.826 (0.805–0.845)
Enlarged Cardiomediastinum	0.775 (0.710–0.836)	0.696 (0.644–0.746)	0.654 (0.595–0.711)
Fracture	0.659 (0.595–0.719)	(–)	0.681 (0.637–0.723)
Lung Lesion	0.676 (0.558–0.782)	0.714 (0.659–0.768)	0.707 (0.667–0.744)
Lung Opacity*	0.778 (0.750–0.805)	0.733 (0.716–0.750)	0.697 (0.675–0.719)
Pleural Effusion*	0.951 (0.936–0.964)	0.914 (0.904–0.923)	0.870 (0.855–0.883)
Pleural Other	0.738 (0.617–0.836)	0.853 (0.796–0.903)	0.783 (0.715–0.843)
Pneumonia	0.721 (0.690–0.753)	0.683 (0.646–0.720)	0.703 (0.662–0.741)
Pneumothorax	0.828 (0.764–0.886)	0.879 (0.851–0.905)	0.852 (0.820–0.880)
Support Devices*	0.936 (0.916–0.953)	0.901 (0.890–0.912)	0.854 (0.838–0.870)

Table A.24: Held-out performance (in terms of **AUROC**) of CXR ResNet50 models across sub-populations stratified by pre-test probability (LM embeddings) of the CXR label. Labels marked with an asterisk (*) indicate a statistically significant difference between the Bottom 25% and Top 25% groups.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis*	0.776 (0.745–0.803)	0.733 (0.715–0.751)	0.652 (0.628–0.675)
Cardiomegaly*	0.757 (0.722–0.793)	0.702 (0.681–0.721)	0.628 (0.604–0.651)
Consolidation	0.772 (0.698–0.838)	0.697 (0.659–0.735)	0.685 (0.640–0.727)
Edema*	0.860 (0.810–0.907)	0.823 (0.801–0.844)	0.756 (0.734–0.778)
Fracture	0.669 (0.559–0.766)	0.620 (0.550–0.690)	0.640 (0.578–0.702)
Lung Lesion	0.604 (0.469–0.734)	0.590 (0.520–0.655)	0.566 (0.524–0.607)
Lung Opacity*	0.710 (0.680–0.740)	0.644 (0.624–0.663)	0.608 (0.584–0.632)
Pleural Effusion*	0.860 (0.830–0.887)	0.814 (0.799–0.828)	0.747 (0.728–0.766)
Pleural Other	0.749 (0.629–0.851)	0.701 (0.616–0.782)	0.685 (0.606–0.760)
Pneumonia	0.607 (0.554–0.662)	0.612 (0.579–0.644)	0.602 (0.562–0.639)
Pneumothorax	0.765 (0.696–0.826)	0.779 (0.740–0.817)	0.735 (0.698–0.769)
Support Devices*	0.879 (0.853–0.902)	0.841 (0.827–0.855)	0.785 (0.765–0.803)

Table A.25: Held-out performance (in terms of **sensitivity at local 95% specificity**) of CXR models across sub-populations stratified by pre-test probability (BoW Representations) of the CXR label. Labels marked with an asterisk (*) indicate a statistically significant difference between the Bottom 25% and Top 25% groups.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis*	0.392 (0.320–0.465)	0.287 (0.247–0.325)	0.165 (0.122–0.204)
Cardiomegaly	0.293 (0.219–0.369)	0.258 (0.212–0.302)	0.200 (0.159–0.240)
Consolidation	0.372 (0.239–0.519)	0.364 (0.282–0.452)	0.183 (0.118–0.260)
Edema*	0.615 (0.480–0.748)	0.457 (0.394–0.517)	0.323 (0.272–0.378)
Enlarged Cardiomediastinum	0.195 (0.085–0.330)	0.175 (0.103–0.253)	0.118 (0.056–0.197)
Fracture	0.191 (0.113–0.285)	(–)	0.201 (0.142–0.263)
Lung Lesion	0.077 (0.000–0.244)	0.190 (0.106–0.279)	0.172 (0.112–0.234)
Lung Opacity*	0.283 (0.218–0.345)	0.191 (0.161–0.222)	0.153 (0.116–0.192)
Pleural Effusion*	0.710 (0.624–0.797)	0.604 (0.559–0.651)	0.441 (0.375–0.519)
Pleural Other	0.215 (0.045–0.409)	0.385 (0.232–0.549)	0.235 (0.116–0.372)
Pneumonia	0.219 (0.165–0.278)	0.120 (0.069–0.175)	0.242 (0.181–0.304)
Pneumothorax	0.342 (0.188–0.522)	0.496 (0.407–0.586)	0.453 (0.374–0.536)
Support Devices*	0.783 (0.725–0.838)	0.576 (0.527–0.623)	0.417 (0.353–0.480)

Table A.26: Held-out performance (in terms of **sensitivity at global 95% specificity**) of CXR models across sub-populations stratified by pre-test probability (BoW Representations) of the CXR label. Labels marked with an asterisk (*) indicate a statistically significant difference between the Bottom 25% and Top 25% groups.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis	0.237 (0.181–0.295)	0.255 (0.220–0.290)	0.287 (0.244–0.329)
Cardiomegaly*	0.070 (0.034–0.116)	0.198 (0.167–0.229)	0.401 (0.362–0.445)
Consolidation	0.243 (0.128–0.359)	0.337 (0.261–0.417)	0.278 (0.201–0.360)
Edema*	0.267 (0.162–0.382)	0.400 (0.346–0.456)	0.529 (0.479–0.576)
Enlarged Cardiomediastinum	0.157 (0.056–0.266)	0.169 (0.103–0.245)	0.173 (0.090–0.258)
Fracture	0.181 (0.102–0.269)	(–)	0.204 (0.145–0.268)
Lung Lesion*	0.048 (0.000–0.171)	0.159 (0.084–0.246)	0.349 (0.280–0.422)
Lung Opacity*	0.135 (0.095–0.175)	0.182 (0.155–0.210)	0.265 (0.228–0.307)
Pleural Effusion*	0.453 (0.376–0.531)	0.547 (0.509–0.587)	0.685 (0.647–0.722)
Pleural Other	0.118 (0.000–0.273)	0.374 (0.220–0.524)	0.330 (0.186–0.477)
Pneumonia	0.192 (0.140–0.244)	0.097 (0.058–0.142)	0.295 (0.237–0.356)
Pneumothorax*	0.194 (0.058–0.348)	0.431 (0.346–0.521)	0.613 (0.541–0.691)
Support Devices	0.536 (0.467–0.603)	0.556 (0.517–0.595)	0.623 (0.582–0.666)

Table A.27: Held-out performance (in terms of **specificity at global 95% specificity**) of CXR models across sub-populations stratified by pre-test probability (BoW Representations) of the CXR label. Labels marked with an asterisk (*) indicate a statistically significant difference between the Bottom 25% and Top 25% groups.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis*	0.979 (0.973–0.984)	0.958 (0.953–0.963)	0.898 (0.887–0.908)
Cardiomegaly*	0.993 (0.989–0.996)	0.964 (0.959–0.968)	0.857 (0.846–0.868)
Consolidation*	0.971 (0.965–0.978)	0.957 (0.953–0.961)	0.914 (0.905–0.923)
Edema*	0.987 (0.982–0.991)	0.959 (0.955–0.964)	0.884 (0.873–0.893)
Enlarged Cardiomediastinum*	0.970 (0.964–0.976)	0.952 (0.947–0.956)	0.926 (0.917–0.934)
Fracture*	0.954 (0.950–0.957)	(–)	0.947 (0.944–0.950)
Lung Lesion*	0.986 (0.981–0.990)	0.964 (0.960–0.968)	0.882 (0.873–0.891)
Lung Opacity*	0.979 (0.974–0.985)	0.954 (0.949–0.959)	0.904 (0.894–0.915)
Pleural Effusion*	0.982 (0.976–0.987)	0.961 (0.956–0.966)	0.864 (0.850–0.878)
Pleural Other*	0.973 (0.967–0.979)	0.955 (0.951–0.960)	0.917 (0.908–0.926)
Pneumonia*	0.958 (0.953–0.963)	0.956 (0.949–0.962)	0.929 (0.920–0.938)
Pneumothorax*	0.982 (0.976–0.986)	0.964 (0.960–0.969)	0.886 (0.876–0.895)
Support Devices*	0.985 (0.980–0.989)	0.954 (0.949–0.959)	0.891 (0.879–0.902)

Table A.28: Held-out performance (in terms of **Adaptive Calibration Error (ACE)**) of CXR models across sub-populations stratified by pre-test probability (BoW representations) of the CXR label.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis	0.011 (0.005–0.019)	0.018 (0.014–0.023)	0.036 (0.024–0.051)
Cardiomegaly	0.013 (0.007–0.021)	0.012 (0.006–0.021)	0.030 (0.019–0.044)
Consolidation	0.005 (0.002–0.009)	0.006 (0.002–0.010)	0.013 (0.008–0.021)
Edema	0.014 (0.010–0.018)	0.017 (0.013–0.021)	0.019 (0.009–0.031)
Enlarged Cardiomediastinum	0.005 (0.003–0.009)	0.006 (0.003–0.009)	0.009 (0.004–0.016)
Fracture	0.006 (0.005–0.008)	–	0.003 (0.001–0.006)
Lung Lesion	0.016 (0.015–0.018)	0.013 (0.012–0.015)	0.022 (0.019–0.029)
Lung Opacity	0.014 (0.006–0.023)	0.017 (0.009–0.026)	0.055 (0.048–0.065)
Pleural Effusion	0.013 (0.006–0.022)	0.011 (0.006–0.019)	0.037 (0.027–0.048)
Pleural Other	0.004 (0.002–0.008)	0.004 (0.002–0.006)	0.007 (0.002–0.011)
Pneumonia	0.008 (0.003–0.014)	0.013 (0.006–0.021)	0.015 (0.008–0.024)
Pneumothorax	0.005 (0.002–0.009)	0.006 (0.005–0.008)	0.020 (0.016–0.025)
Support Devices	0.011 (0.006–0.017)	0.015 (0.008–0.022)	0.035 (0.022–0.049)

Table A.29: Held-out performance (in terms of **Brier Skill Score (BSS)**) of CXR models across sub-populations stratified by pre-test probability (BoW representations) of the CXR label.

Label	Bottom 25%	Middle 50%	Top 25%
Atelectasis	0.209 (0.166–0.250)	0.178 (0.155–0.201)	0.113 (0.084–0.140)
Cardiomegaly	0.110 (0.067–0.154)	0.150 (0.126–0.174)	0.153 (0.125–0.183)
Consolidation	0.061 (0.023–0.101)	0.072 (0.050–0.095)	0.045 (0.023–0.069)
Edema	0.126 (0.036–0.216)	0.220 (0.182–0.260)	0.234 (0.194–0.272)
Enlarged Cardiomediastinum	0.019 (−0.002–0.042)	0.013 (−0.002–0.028)	0.005 (−0.013–0.025)
Fracture	0.004 (−0.004–0.013)	–	0.012 (0.006–0.019)
Lung Lesion	−0.091 (−0.124–−0.059)	−0.014 (−0.042–0.015)	0.027 (0.002–0.051)
Lung Opacity	0.130 (0.097–0.164)	0.111 (0.090–0.131)	0.086 (0.059–0.111)
Pleural Effusion	0.387 (0.312–0.458)	0.457 (0.426–0.487)	0.406 (0.370–0.441)
Pleural Other	−0.006 (−0.031–0.029)	0.039 (0.003–0.086)	0.019 (−0.011–0.053)
Pneumonia	0.045 (0.021–0.069)	0.020 (−0.001–0.041)	0.072 (0.041–0.103)
Pneumothorax	0.010 (−0.055–0.081)	0.133 (0.078–0.188)	0.185 (0.128–0.243)
Support Devices	0.502 (0.445–0.560)	0.439 (0.407–0.471)	0.372 (0.331–0.409)

Table A.30: Held-out performance (in terms of **AUROC**) of CXR models across sub-populations stratified by the presence of label-relevant terms in prior discharge summaries. Statistically significant differences (as evaluated by bootstrapping) are highlighted with an asterisk (*) next to the label.

Label	Previous Mentions	No Previous Mentions
Atelectasis*	0.757 (0.740–0.773)	0.831 (0.816–0.846)
Cardiomegaly*	0.787 (0.772–0.802)	0.819 (0.804–0.834)
Consolidation	0.802 (0.770–0.832)	0.828 (0.801–0.854)
Edema	0.889 (0.879–0.898)	0.900 (0.838–0.945)
Enlarged Cardiomediastinum	0.699 (0.621–0.770)	0.710 (0.674–0.745)
Fracture	0.651 (0.595–0.705)	0.702 (0.630–0.771)
Lung Lesion	0.737 (0.703–0.769)	0.750 (0.691–0.803)
Lung Opacity*	0.724 (0.709–0.740)	0.768 (0.750–0.785)
Pleural Effusion	0.918 (0.911–0.925)	0.930 (0.918–0.940)
Pleural Other	0.704 (0.582–0.813)	0.818 (0.774–0.860)
Pneumonia	0.696 (0.670–0.723)	0.714 (0.682–0.744)
Pneumothorax	0.867 (0.840–0.891)	0.892 (0.866–0.915)
Support Devices	0.901 (0.893–0.909)	0.932 (0.901–0.959)

Table A.31: Held-out performance (in terms of **sensitivity at local 95% specificity**) of CXR models across sub-populations stratified by the presence of label-relevant terms in prior discharge summaries. Statistically significant differences (as evaluated by bootstrapping) are highlighted with an asterisk (*) next to the label.

Label	Previous Mentions	No Previous Mentions
Atelectasis*	0.201 (0.164–0.239)	0.333 (0.289–0.379)
Cardiomegaly	0.264 (0.225–0.303)	0.298 (0.254–0.345)
Consolidation	0.275 (0.206–0.349)	0.315 (0.249–0.391)
Edema	0.452 (0.413–0.491)	0.476 (0.255–0.697)
Enlarged Cardiomediastinum	0.165 (0.063–0.285)	0.170 (0.116–0.230)
Fracture	0.184 (0.113–0.264)	0.231 (0.130–0.343)
Lung Lesion	0.253 (0.193–0.314)	0.206 (0.114–0.312)
Lung Opacity	0.177 (0.149–0.206)	0.238 (0.200–0.275)
Pleural Effusion	0.588 (0.552–0.624)	0.623 (0.555–0.678)
Pleural Other	0.115 (0.000–0.275)	0.347 (0.249–0.459)
Pneumonia	0.195 (0.153–0.236)	0.195 (0.147–0.247)
Pneumothorax	0.490 (0.419–0.567)	0.527 (0.437–0.616)
Support Devices*	0.562 (0.527–0.595)	0.793 (0.702–0.872)

Table A.32: Held-out performance (in terms of **sensitivity at global 95% specificity**) of CXR models across sub-populations stratified by the presence of label-relevant terms in prior discharge summaries. Statistically significant differences (as evaluated by bootstrapping) are highlighted with an asterisk (*) next to the label.

Label	Previous Mentions	No Previous Mentions
Atelectasis	0.264 (0.230–0.298)	0.264 (0.225–0.304)
Cardiomegaly*	0.336 (0.301–0.372)	0.198 (0.165–0.233)
Consolidation	0.297 (0.225–0.375)	0.300 (0.236–0.367)
Edema	0.463 (0.425–0.502)	0.315 (0.154–0.489)
Enlarged Cardiomediastinum	0.168 (0.068–0.284)	0.168 (0.115–0.225)
Fracture	0.194 (0.122–0.275)	0.213 (0.118–0.327)
Lung Lesion*	0.312 (0.251–0.377)	0.148 (0.072–0.244)
Lung Opacity	0.226 (0.197–0.259)	0.171 (0.142–0.200)
Pleural Effusion*	0.629 (0.596–0.661)	0.525 (0.473–0.576)
Pleural Other	0.151 (0.022–0.326)	0.344 (0.246–0.453)
Pneumonia	0.209 (0.167–0.252)	0.177 (0.132–0.226)
Pneumothorax	0.497 (0.422–0.573)	0.519 (0.435–0.602)
Support Devices	0.583 (0.550–0.618)	0.570 (0.472–0.664)

Table A.33: Held-out performance (in terms of **specificity at global 95% specificity**) of CXR models across sub-populations stratified by the presence of label-relevant terms in prior discharge summaries. Statistically significant differences (as evaluated by bootstrapping) are highlighted with an asterisk (*) next to the label.

Label	Previous Mentions	No Previous Mentions
Atelectasis*	0.931 (0.926–0.937)	0.964 (0.960–0.968)
Cardiomegaly*	0.924 (0.918–0.929)	0.968 (0.964–0.972)
Consolidation*	0.942 (0.936–0.948)	0.955 (0.952–0.959)
Edema*	0.948 (0.946–0.949)	0.975 (0.963–0.985)
Enlarged Cardiomediastinum	0.948 (0.941–0.955)	0.951 (0.948–0.954)
Fracture*	0.936 (0.929–0.942)	0.958 (0.954–0.961)
Lung Lesion*	0.929 (0.924–0.933)	0.969 (0.965–0.973)
Lung Opacity*	0.931 (0.927–0.936)	0.969 (0.964–0.974)
Pleural Effusion*	0.941 (0.938–0.944)	0.968 (0.962–0.974)
Pleural Other*	0.908 (0.883–0.932)	0.952 (0.951–0.954)
Pneumonia*	0.943 (0.938–0.947)	0.958 (0.953–0.963)
Pneumothorax	0.948 (0.944–0.953)	0.952 (0.947–0.956)
Support Devices*	0.945 (0.944–0.947)	0.981 (0.972–0.989)

Table A.34: Comparison of DenseNet121 AUROC across Standard, Reweighted (IPW), and Matched Neighbor settings. Metrics reported are mean (95% CI). Statistically significant differences in the reweighted/matched sets compared to the standard are bolded.

Label	Standard	Reweighted (IPW)	Matched
Atelectasis	0.801 (0.793–0.808)	**0.775 (0.766–0.784)**	**0.695 (0.684–0.706)**
Cardiomegaly	0.811 (0.803–0.818)	**0.751 (0.741–0.760)**	**0.666 (0.656–0.677)**
Consolidation	0.818 (0.805–0.832)	0.798 (0.782–0.815)	**0.641 (0.617–0.665)**
Edema	0.891 (0.885–0.898)	**0.861 (0.850–0.871)**	**0.749 (0.735–0.763)**
Enlarged Cardiomediastinum	0.707 (0.685–0.730)	0.704 (0.681–0.727)	**0.589 (0.561–0.618)**
Fracture	0.681 (0.651–0.710)	0.655 (0.621–0.689)	**0.606 (0.569–0.643)**
Lung Lesion	0.756 (0.737–0.775)	0.720 (0.697–0.742)	**0.598 (0.570–0.626)**
Lung Opacity	0.749 (0.741–0.757)	**0.721 (0.712–0.730)**	**0.665 (0.655–0.676)**
Pleural Effusion	0.923 (0.919–0.927)	**0.894 (0.888–0.900)**	**0.852 (0.845–0.860)**
Pleural Other	0.806 (0.778–0.834)	0.768 (0.734–0.803)	**0.618 (0.572–0.663)**
Pneumonia	0.707 (0.693–0.721)	0.692 (0.677–0.707)	**0.624 (0.606–0.642)**
Pneumothorax	0.879 (0.867–0.891)	0.857 (0.841–0.872)	**0.722 (0.697–0.747)**
Support Devices	0.906 (0.901–0.911)	**0.885 (0.879–0.892)**	**0.833 (0.825–0.841)**

Table A.35: Comparison of ResNet50 AUROC across Standard, Reweighted (IPW), and Matched Neighbor settings. Metrics reported are mean (95% CI). Statistically significant differences in the reweighted/matched sets compared to the standard are bolded.

Label	Standard	Reweighted (IPW)	Matched
Atelectasis	0.801 (0.793–0.808)	**0.712 (0.698–0.725)**	**0.632 (0.615–0.650)**
Cardiomegaly	0.811 (0.803–0.818)	**0.689 (0.674–0.704)**	**0.597 (0.580–0.614)**
Consolidation	0.818 (0.805–0.832)	**0.704 (0.674–0.733)**	**0.574 (0.537–0.613)**
Edema	0.891 (0.885–0.898)	**0.810 (0.788–0.829)**	**0.694 (0.672–0.717)**
Enlarged Cardiomediastinum	0.707 (0.685–0.730)	0.660 (0.624–0.696)	**0.565 (0.519–0.612)**
Fracture	0.681 (0.651–0.710)	0.635 (0.589–0.682)	**0.570 (0.514–0.626)**
Lung Lesion	0.756 (0.737–0.775)	**0.602 (0.561–0.639)**	**0.536 (0.488–0.580)**
Lung Opacity	0.749 (0.741–0.757)	**0.644 (0.631–0.658)**	**0.593 (0.576–0.609)**
Pleural Effusion	0.923 (0.919–0.927)	**0.798 (0.785–0.810)**	**0.739 (0.725–0.753)**
Pleural Other	0.806 (0.778–0.834)	**0.689 (0.631–0.748)**	**0.597 (0.519–0.674)**
Pneumonia	0.707 (0.693–0.721)	**0.603 (0.579–0.626)**	**0.554 (0.526–0.582)**
Pneumothorax	0.879 (0.867–0.891)	**0.764 (0.737–0.790)**	**0.652 (0.613–0.691)**
Support Devices	0.906 (0.901–0.911)	**0.825 (0.813–0.837)**	**0.763 (0.749–0.777)**

Table A.36: Comparison of Sensitivity (@Global 95% Spec.) across Standard, Reweighted (IPW), and Matched Neighbor settings. Metrics reported are mean (95% CI). Statistically significant differences as compared to the standard set are highlighted in bold.

Label	Standard	Reweighted (IPW)	Matched
Atelectasis	0.264 (0.235–0.292)	0.232 (0.204–0.260)	**0.147 (0.118–0.178)**
Cardiomegaly	0.638 (0.609–0.665)	**0.444 (0.413–0.476)**	**0.380 (0.345–0.413)**
Consolidation	0.167 (0.133–0.203)	0.158 (0.125–0.193)	**0.046 (0.024–0.071)**
Edema	0.505 (0.472–0.537)	**0.382 (0.350–0.414)**	**0.258 (0.222–0.294)**
Enlarged Cardiomediastinum	0.052 (0.038–0.068)	0.055 (0.040–0.071)	**0.016 (0.007–0.027)**
Fracture	0.049 (0.034–0.065)	0.066 (0.049–0.084)	**0.008 (0.001–0.018)**
Lung Lesion	0.063 (0.045–0.083)	0.040 (0.025–0.057)	**0.005 (0.000–0.014)**
Lung Opacity	0.264 (0.244–0.284)	0.237 (0.217–0.257)	**0.149 (0.130–0.168)**
No Finding	0.306 (0.288–0.324)	0.323 (0.305–0.341)	**0.210 (0.187–0.234)**
Pleural Effusion	0.608 (0.576–0.639)	**0.488 (0.443–0.531)**	**0.393 (0.345–0.442)**
Pleural Other	0.303 (0.213–0.401)	0.223 (0.136–0.323)	**0.096 (0.033–0.179)**
Pneumonia	0.136 (0.111–0.162)	0.118 (0.093–0.142)	**0.040 (0.023–0.059)**
Pneumothorax	0.213 (0.171–0.257)	0.147 (0.110–0.187)	**0.058 (0.033–0.086)**
Support Devices	0.707 (0.687–0.727)	**0.627 (0.603–0.650)**	**0.490 (0.461–0.518)**

Table A.37: Comparison of Specificity across Standard, Reweighted (IPW), and Matched Neighbor settings. Metrics reported are mean (95% CI). Statistically significant differences as compared to the standard set are highlighted in bold.

Label	Standard	Reweighted (IPW)	Matched
Atelectasis	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.950 (0.950–0.951)
Cardiomegaly	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.950 (0.950–0.950)
Consolidation	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.950 (0.950–0.951)
Edema	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.950 (0.950–0.951)
Enlarged Cardiomediastinum	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.950 (0.949–0.951)
Fracture	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.950 (0.948–0.952)
Lung Lesion	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.950 (0.949–0.951)
Lung Opacity	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.950 (0.950–0.951)
No Finding	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.950 (0.950–0.951)
Pleural Effusion	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.950 (0.950–0.951)
Pleural Other	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.949 (0.939–0.962)
Pneumonia	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.950 (0.948–0.951)
Pneumothorax	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.950 (0.948–0.951)
Support Devices	0.950 (0.950–0.950)	0.950 (0.950–0.950)	0.950 (0.950–0.951)

Table A.38: Comparison of Adaptive Calibration Error (ACE) across Standard, Reweighted (IPW), and Matched Neighbor settings. Metrics reported are mean (95% CI).

Label	Standard	Reweighted (IPW)	Matched
Atelectasis	0.017 (0.014–0.021)	0.024 (0.018–0.030)	**0.245 (0.239–0.251)**
Cardiomegaly	0.017 (0.014–0.021)	**0.033 (0.026–0.039)**	**0.170 (0.165–0.175)**
Consolidation	0.009 (0.007–0.011)	**0.015 (0.012–0.017)**	**0.158 (0.154–0.163)**
Edema	0.012 (0.010–0.015)	**0.025 (0.021–0.029)**	**0.178 (0.174–0.183)**
Enlarged Cardiomediastinum	0.020 (0.015–0.025)	0.019 (0.014–0.024)	**0.354 (0.347–0.362)**
Fracture	0.010 (0.007–0.013)	0.015 (0.011–0.019)	**0.238 (0.233–0.244)**
Lung Lesion	0.011 (0.008–0.015)	0.010 (0.007–0.013)	**0.171 (0.167–0.176)**
Lung Opacity	0.020 (0.016–0.023)	0.022 (0.018–0.027)	**0.261 (0.257–0.266)**
No Finding	0.020 (0.014–0.026)	0.029 (0.020–0.038)	**0.444 (0.437–0.450)**
Pleural Effusion	0.011 (0.006–0.015)	**0.025 (0.018–0.033)**	**0.098 (0.090–0.106)**
Pleural Other	0.002 (0.001–0.003)	0.003 (0.002–0.004)	**0.062 (0.059–0.065)**
Pneumonia	0.015 (0.012–0.018)	0.016 (0.013–0.019)	**0.222 (0.218–0.226)**
Pneumothorax	0.011 (0.008–0.015)	0.009 (0.006–0.012)	**0.155 (0.151–0.159)**
Support Devices	0.014 (0.010–0.017)	0.023 (0.017–0.029)	**0.138 (0.134–0.143)**

Table A.39: Comparison of Brier Skill Score (BSS) across Standard, Reweighted (IPW), and Matched Neighbor settings. Metrics reported are mean (95% CI).

Label	Standard	Reweighted (IPW)	Matched
Atelectasis	0.180 (0.163–0.197)	0.147 (0.128–0.165)	**-0.137 (-0.161—0.114)**
Cardiomegaly	0.490 (0.468–0.510)	**0.292 (0.264–0.320)**	**0.003 (-0.030–0.035)**
Consolidation	0.156 (0.125–0.190)	0.131 (0.098–0.165)	**-0.279 (-0.315—0.245)**
Edema	0.359 (0.334–0.383)	**0.226 (0.198–0.254)**	**-0.049 (-0.081—0.016)**
Enlarged Cardiomediastinum	0.054 (0.029–0.078)	0.047 (0.023–0.071)	**-0.669 (-0.710—0.627)**
Fracture	0.094 (0.061–0.128)	0.081 (0.050–0.113)	**-0.366 (-0.395—0.335)**
Lung Lesion	0.047 (0.022–0.074)	0.033 (0.007–0.059)	**-0.229 (-0.258—0.200)**
Lung Opacity	0.210 (0.195–0.224)	0.183 (0.167–0.199)	**-0.216 (-0.244—0.189)**
No Finding	0.207 (0.187–0.226)	0.212 (0.192–0.231)	**-1.218 (-1.267—1.171)**
Pleural Effusion	0.492 (0.473–0.513)	**0.393 (0.363–0.421)**	**0.324 (0.296–0.353)**
Pleural Other	0.054 (0.019–0.098)	0.047 (0.009–0.094)	**-0.163 (-0.213—0.122)**
Pneumonia	0.111 (0.086–0.136)	0.099 (0.076–0.122)	**-0.400 (-0.436—0.365)**
Pneumothorax	0.191 (0.150–0.233)	0.146 (0.106–0.188)	**-0.220 (-0.264—0.178)**
Support Devices	0.584 (0.569–0.598)	**0.498 (0.479–0.517)**	**0.145 (0.121–0.169)**

Appendix E. Mathematical Proofs

To rigorously assess whether the vision model utilizes visual signals independent of clinical context, we define a target distribution Q where the diagnostic label Y is statistically independent of the prior clinical context C ($Y \perp C$). This is meant to represent a setting where the clinical history does not provide discriminative information about the label. We estimate the model's performance on this target distribution using Inverse Probability Weighting (IPW). By reweighting our observed evaluation set using the ratio of marginal disease prevalence to the context-conditional pre-test probability, we remove the correlation between context and label.

E.1. Formal Derivation of the Inverse Probability Metric Weights

Let $P(X, Y, C)$ denote the observed data distribution, where X is the image, $Y \in \{0, 1\}$ is the diagnostic label, and C is the prior clinical context. In the observational setting, Y and C are correlated. We define a target distribution $Q(X, Y, C)$ that satisfies two conditions:

1. **Independence of Context and Label:** $Q(Y, C) = Q(Y)Q(C) = P(Y)P(C)$. This represents a scenario where clinical history provides no information about the current diagnosis.
2. **Invariance of Imaging Mechanism:** $Q(X \mid Y, C) = P(X \mid Y, C)$. The physical process generating the X-ray from the patient's state remains unchanged.

We seek to evaluate performance metrics defined under Q using data sampled from P. The importance weight $w(x, y, c)$ is given by:

$$w(x, y, c) = \frac{Q(X \mid Y, C)Q(Y)Q(C)}{P(X \mid Y, C)P(Y \mid C)P(C)} = \frac{P(Y)P(C)}{P(Y \mid C)P(C)} = \frac{P(Y)}{P(Y \mid C)} \tag{7}$$

Under the assumption that $Q(X|Y, C) = P(X|Y, C)$. Thus, the weight is the ratio of the marginal label prevalence to the conditional pre-test probability.

Lemma 3 *Let P be defined as the observed distribution ,and Q refer to the target distribution. where $Q(X, Y, C) > 0 \implies P(X, Y, C) > 0$. Then for any measurable function $g(X, Y, C)$,*

$$E_P[w(X, Y, C)g(X, Y, C)] = E_Q[g(X, Y, C)]$$

Proof Using p, q to denote the PDF / PMF of P and Q respectively,

$$\begin{aligned} E_Q[g(X, Y, X)] &= \int_{x,y,c} g(x, y, c)q(x, y, c)dxdydc \\ &= \int_{x,y,c} g(x, y, c)p(x, y, c)\frac{q(x, y, c)}{p(x, y, c)}dxdydc \\ &= \int_{x,y,c} w(x, y, c)g(x, y, c)p(x, y, c)dxdydc \\ &= E_P[w(X, Y, C)g(X, Y, C)] \end{aligned}$$

■

E.2. Weighted Sensitivity and Specificity

We define the Sensitivity (True Positive Rate) and Specificity (True Negative Rate) of a classifier $f(X)$ at a decision threshold τ as conditional probabilities under the target distribution Q. We utilize the importance sampling identity $\mathbb{E}_Q[g(X, Y, C)] = \mathbb{E}_P[w(X, Y, C) \cdot g(X, Y, C)]$ to express these metrics in terms of the observed distribution P.

E.2.1. Sensitivity (True Positive Rate)

Proof Sensitivity under a distribution Q, denoted $\text{Sens}_Q(\tau)$, is defined as the conditional probability of a positive prediction given a positive label:

$$\text{Sens}_Q(\tau) = P_Q(f(X) > \tau \mid Y = 1) = \frac{\mathbb{E}_Q[\mathbb{I}(f(X) > \tau) \cdot \mathbb{I}(Y = 1)]}{\mathbb{E}_Q[\mathbb{I}(Y = 1)]} \tag{8}$$

We apply lemma Theorem 3 to both the numerator and the denominator. For the numerator, we set $g(X, Y, C) = \mathbb{I}(f(X) > \tau) \cdot \mathbb{I}(Y = 1)$, and for the denominator, we set $g(X, Y, C) = \mathbb{I}(Y = 1)$. This yields:

$$\text{Sens}_Q(\tau) = \frac{\mathbb{E}_P\left[w(X, Y, C) \cdot \mathbb{I}(f(X) > \tau) \cdot \mathbb{I}(Y = 1)\right]}{\mathbb{E}_P\left[w(X, Y, C) \cdot \mathbb{I}(Y = 1)\right]} \tag{9}$$

■

E.2.2. Specificity (True Negative Rate)

Similarly, the Specificity in the target distribution, denoted $\text{Spec}_Q(\tau)$, is defined as:

$$\text{Spec}_Q(\tau) = P_Q(f(X) \leq \tau \mid Y = 0) = \frac{\mathbb{E}_Q[\mathbb{I}(f(X) \leq \tau) \cdot \mathbb{I}(Y = 0)]}{\mathbb{E}_Q[\mathbb{I}(Y = 0)]} \tag{10}$$

Applying the same lemma Theorem 3 to the numerator (with $g = \mathbb{I}(f(X) \leq \tau) \cdot \mathbb{I}(Y = 0)$) and the denominator (with $g = \mathbb{I}(Y = 0)$), we express Specificity in terms of the observed distribution P:

$$\text{Spec}_Q(\tau) = \frac{\mathbb{E}_P\left[w(X, Y, C) \cdot \mathbb{I}(f(X) \leq \tau) \cdot \mathbb{I}(Y = 0)\right]}{\mathbb{E}_P\left[w(X, Y, C) \cdot \mathbb{I}(Y = 0)\right]} \tag{11}$$

E.3. Area Under the Curve (AUC)

We define the AUROC in the target distribution, denoted τ_Q, as the probability that the classifier f ranks a randomly sampled positive instance higher than a randomly sampled negative instance. We consider two independent samples (X, Y, C) and (X', Y', C') drawn from the target joint distribution Q.

Proof Formally, τ_Q is the conditional probability of a correct ranking given that the first sample is positive $(Y = 1)$ and the second is negative $(Y' = 0)$. We express this as a ratio of expectations over the independent joint distributions:

$$\tau_Q = P_Q(f(X) > f(X') \mid Y = 1, Y' = 0) = \frac{\mathbb{E}_{Q,Q'}\left[\mathbb{I}(f(X) > f(X')) \cdot \mathbb{I}(Y = 1) \cdot \mathbb{I}(Y' = 0)\right]}{\mathbb{E}_{Q,Q'}\left[\mathbb{I}(Y = 1) \cdot \mathbb{I}(Y' = 0)\right]} \tag{12}$$

We apply the reweighting identity from Theorem 3 to the independent samples in both the numerator and the denominator. For the numerator, we define the function g over the pair of samples as $g(\cdot) = \mathbb{I}(f(X) > f(X')) \cdot \mathbb{I}(Y = 1) \cdot \mathbb{I}(Y' = 0)$. The weight for the pair is the product of individual weights $w(X, Y, C) \cdot w(X', Y', C')$. This yields:

$$\tau_Q = \frac{\mathbb{E}_{P,P'}\left[w(X, Y, C)w(X', Y', C') \cdot \mathbb{I}(f(X) > f(X')) \cdot \mathbb{I}(Y = 1)\mathbb{I}(Y' = 0)\right]}{\mathbb{E}_{P,P'}\left[w(X, Y, C)w(X', Y', C') \cdot \mathbb{I}(Y = 1)\mathbb{I}(Y' = 0)\right]} \tag{13}$$

The denominator factors into the product of the weighted marginals for each class. Thus, we can estimate the target AUROC using pairs sampled from the observed distribution P:

$$\tau_Q = \frac{\mathbb{E}_{P,P'}\left[w(X, Y, C)w(X', Y', C') \cdot \mathbb{I}(f(X) > f(X')) \cdot \mathbb{I}(Y = 1)\mathbb{I}(Y' = 0)\right]}{\mathbb{E}_P\left[w(X, Y, C) \cdot \mathbb{I}(Y = 1)\right] \cdot \mathbb{E}_P\left[w(X', Y', C') \cdot \mathbb{I}(Y' = 0)\right]} \tag{14}$$

■

E.4. Adaptive Calibration Error (ACE)

We define calibration error under the target distribution Q. Let predictions be partitioned into K disjoint bins $\{B_1, \ldots, B_K\}$. In the adaptive setting, these bins are chosen such that each bin contains an equal amount of probability mass under Q:

$$Q(X \in B_k) = \frac{1}{K} \quad \forall k \in 1, \ldots, K \tag{15}$$

Applying importance sampling, this constraint implies that the bins must satisfy an equal weight condition in the observed distribution P:

$$\mathbb{E}_P[w(Y, C) \cdot \mathbb{I}(X \in B_k)] = \frac{1}{K}\mathbb{E}_P[w(Y, C)] \tag{16}$$

Within each bin B_k, the Calibration Error is the difference between the expected confidence and the expected label under Q:

$$\text{Error}_k = |\mathbb{E}_Q[Y \mid X \in B_k] - \mathbb{E}_Q[f(X) \mid X \in B_k]| \tag{17}$$

Expanding the conditional expectations using weighted estimators from P:

$$\mathbb{E}_Q[Y \mid X \in B_k] = \frac{\mathbb{E}_Q[Y \cdot \mathbb{I}(X \in B_k)]}{Q(X \in B_k)} = \frac{\mathbb{E}_P[w(Y, C) \cdot Y \cdot \mathbb{I}(X \in B_k)]}{\mathbb{E}_P[w(Y, C) \cdot \mathbb{I}(X \in B_k)]} \tag{18}$$

Similarly for the prediction $f(X)$:

$$\mathbb{E}_Q[f(X) \mid X \in B_k] = \frac{\mathbb{E}_P[w(Y, C) \cdot f(X) \cdot \mathbb{I}(X \in B_k)]}{\mathbb{E}_P[w(Y, C) \cdot \mathbb{I}(X \in B_k)]} \tag{19}$$

The overall Adaptive Calibration Error (ACE) under Q is the expected error across all bins:

$$\text{ACE}_Q = \sum_{k=1}^{K} Q(X \in B_k) \cdot \text{Error}_k = \frac{1}{K} \sum_{k=1}^{K} \left| \frac{\mathbb{E}_P[w \cdot (Y - f(X)) \cdot \mathbb{I}(X \in B_k)]}{\mathbb{E}_P[w \cdot \mathbb{I}(X \in B_k)]} \right| \tag{20}$$

E.5. Brier Skill Score (BSS)

We define the Brier Score (BS) under the target distribution Q as the expected mean squared error between the predicted probability $f(X)$ and the true label Y.

$$BS_Q = \mathbb{E}_Q\left[(Y - f(X))^2\right] \tag{21}$$

Applying the importance sampling identity $\mathbb{E}_Q[\cdot] = \mathbb{E}_P[w\cdot]$, we express this in terms of the observed distribution P:

$$BS_Q = \mathbb{E}_P\left[w(Y, C) \cdot (Y - f(X))^2\right] \tag{22}$$

The Brier Skill Score requires a reference baseline, BS_{ref}, corresponding to a "no-skill" classifier that we chose to simply predict the marginal prevalence of the target distribution, $Q(Y = 1)$. Since our target distribution preserves the marginal prevalence of the observed distribution ($Q(Y) = P(Y)$), this baseline prediction is simply $\pi = P(Y = 1)$.

$$BS_{\text{ref, Q}} = \mathbb{E}_Q\left[(Y - \pi)^2\right] = \mathbb{E}_P\left[w(Y, C) \cdot (Y - \pi)^2\right] \tag{23}$$

The Weighted Brier Skill Score is therefore defined as:

$$BSS_Q = 1 - \frac{BS_Q}{BS_{\text{ref},Q}} = 1 - \frac{\mathbb{E}_P\left[w(Y, C) \cdot (Y - f(X))^2\right]}{\mathbb{E}_P\left[w(Y, C) \cdot (Y - P(Y = 1))^2\right]} \tag{24}$$