Beyond Discrimination: Evaluating the Calibration of Computed Tomography Visual Encoders

Anonymous authors
Paper under double-blind review

Abstract

Recent progress in supervised multi-label abnormality classification for three-dimensional (3D) computed tomography (CT) has been driven by increasingly powerful visual encoders. However, current evaluation remains largely limited to discriminative and ranking-based metrics, overlooking the reliability of predicted probabilities. Calibration, the alignment between predicted confidence and true outcome likelihood, is critical for clinical deployment, where probabilities inform decision-making and risk assessment. In this work, we present an evaluation of calibration in multi-label abnormality classification from 3D CT scans. We benchmark representative architectures spanning convolutional, transformer-based, and hybrid CT visual encoders under a unified training protocol across two public datasets, and extend our analysis to CT foundation models spanning diverse supervision paradigms and training data domains. Our results show that architectures achieving the best performance on standard predictive metrics are not consistently the best calibrated, and that calibration is observed to further degrade under distribution shift. We further identify that miscalibration is structured and data-dependent, increasing with abnormality prevalence and volumetric size, varying across anatomical systems, and exacerbated by abnormality co-occurrence. These findings highlight fundamental limitations of current evaluation practices and emphasize the need for calibration-aware modeling in medical image analysis.

1 Introduction

Computed Tomography (CT) provides high-resolution volumetric imaging of human anatomy, enabling radiologists to detect a wide range of abnormalities with high diagnostic accuracy (Patel & De Jesus, 2024). However, the rapid increase in the number of CT examinations has created a growing demand for automated tools to assist clinicians in managing this workload (Bellolio et al., 2017; Broder & Warshauer, 2006). In this context, abnormality classification from 3D CT scans, consisting of the simultaneous detection of one or multiple pathologies within a volume, has emerged as a fundamental task, benefiting from pathology co-occurrence (Draelos et al., 2021), and enabling large-scale screening and triage applications (Najjar, 2023).

Recent progress in supervised abnormality classification has been driven by the availability of large annotated datasets (Hamamci et al., 2026; Blankemeier et al., 2026) and advances in visual encoders, including 3D convolutional networks (Carreira & Zisserman, 2018), vision transformers (Dosovitskiy et al., 2021), and CT-specific slice-wise (2.5D) approaches (Draelos et al., 2021; Di Piazza et al., 2025; 2026). While recent work has explored self-supervised (Pai et al., 2025; Hamamci et al., 2025; Wald et al., 2026a; Manna et al., 2024) and vision-language (Hamamci et al., 2026; Blankemeier et al., 2026; Wang et al., 2025; Shui et al., 2025; Kim et al., 2025; Lei et al., 2026; Zhao et al., 2026) pre-training paradigms, supervised learning with large-scale annotated datasets remains the dominant setting for multi-label abnormality classification from CT. Despite these advances, evaluation of such models remains largely limited to standard classification metrics (Sokolova & Lapalme, 2009), including threshold-dependent measures such as F1-score and accuracy (Provost et al., 1998), as well as threshold-free metrics such as AUROC (Bradley, 1997) and AUPRC (Everingham et al., 2010). While these metrics quantify discriminative performance, they do not assess the reliability of predicted probabilities. In particular, *calibration*, which measures the alignment between predicted confidence and true

outcome likelihood (DeGroot & Fienberg, 1983; Guo et al., 2017; Mukhoti et al., 2020; Berta et al., 2023), remains largely unexplored in the context of multi-label abnormality classification from CT scans.

Calibration is a critical property for clinical deployment (Flores et al., 2025). Medical decision-making is inherently risk-sensitive (Masic, 2022), and model outputs are often interpreted in terms of confidence to support diagnostic reasoning and downstream actions (Sanchez-Martinez et al., 2022). A well-calibrated model ensures that predicted probabilities accurately reflect the true likelihood of abnormality presence, enabling meaningful thresholding and uncertainty-aware decision-making. In contrast, poorly calibrated models may exhibit overconfidence, leading to false reassurance, or underconfidence, potentially delaying the detection of critical findings, both of which can adversely impact patient care. Prior work has explored calibration in medical imaging, primarily in 2D settings across modalities such as X-ray (Frenkel & Goldberger, 2022; Bani-Harouni et al., 2026), dermatoscopic (Shen et al., 2025), and ultrasound imaging (Geysels et al., 2025). In contrast, the calibration of 3D volumetric classifiers has received comparatively little attention.

In this work, we conduct a comprehensive study of probability calibration for multi-label abnormality classification from 3D CT scans. We address the following questions: (*i*) how well do existing CT visual encoders perform in terms of calibration?; and (*ii*) what factors influence calibration across abnormalities? To this end, we evaluate five representative visual encoders for CT imaging including 3D and 2.5D architectures, under a unified experimental protocol, and assess both discriminative and calibration performance across two public datasets with a focus on chest CT. Beyond this controlled benchmark, we further extend our analysis to four CT foundation models spanning a broad range of supervision paradigms and training data domains.

Our contributions and key findings are as follows:

- To the best of our knowledge, we provide the first evaluation of calibration in multi-label abnormality classification from 3D CT, benchmarking five representative visual encoders across two public datasets with six complementary calibration metrics. We further contextualize our findings by extending the analysis to four CT foundation models. Our results reveal that model rankings differ between discriminative and calibration metrics: the most accurate ones are not necessarily the best calibrated.
- We identify several data-dependent factors driving miscalibration. Calibration error tends to increase with abnormality prevalence and volumetric size, varies across anatomical systems, and is exacerbated by pathology co-occurrence, suggesting that miscalibration may be highly structured and that current models struggle to reliably quantify uncertainty in the presence of inter-label dependencies.
- We observe that attention-based architectures tend to be better calibrated in-domain than purely convolutional approaches, although these gains appear less robust under distribution shift. Post-hoc calibration methods can reduce this gap in-domain but do not reliably generalize on external cohorts.

2 Background and Related Work

2.1 Abnormality classification from 3D CT

Visual encoders for 3D CT analysis can be broadly categorized into three families. First, 3D convolutional neural networks operate directly on volumetric data using 3D convolutions (Carreira & Zisserman, 2018). By leveraging inductive biases such as locality and translation equivariance, they have been adopted for both dense and global prediction tasks, including abnormality classification, semantic segmentation, and automated report generation (Blankemeier et al., 2026). Second, 3D vision transformers (Dosovitskiy et al., 2021) represent CT volumes as sets of volumetric patches that interact through self-attention (Vaswani et al., 2023), enabling modeling of long-range dependencies. While powerful, these models typically require large-scale pretraining (Hamamci et al., 2026; Di Salvo et al., 2026) and have been explored in settings such as vision-language pretraining (Shui et al., 2025; Wang et al., 2025), synthetic volume generation (Hamamci et al., 2023) and report generation (Hamamci et al., 2024). Third, 2.5D approaches model CT scans as sequences of axial slices, combining 2D feature extraction with cross-slice aggregation. These methods have demonstrated strong performance for classification tasks, often matching or exceeding their 3D counterparts. Prior work has explored diverse aggregation strategies, including lightweight 3D convolutions (Draelos et al.,

2021), attention-based modules (Di Piazza et al., 2025), and spectral graph-based methods (Di Piazza et al., 2026). Across these architectures, models are predominantly trained with binary cross-entropy (BCE) objectives (Mao et al., 2023) yet evaluated primarily through discriminative and ranking metrics (Sokolova & Lapalme, 2009). Although recent work on natural images has explored strictly proper losses with label-pair regularization (Cheng & Vasconcelos, 2024) and joint copula-based calibration (Panda & Garain, 2025) to capture inter-label dependencies during training, we restrict our study to BCE supervision as a controlled and widely adopted baseline, evaluating 3D and 2.5D backbones across a comprehensive set of calibration metrics.

2.2 Calibration and probabilistic evaluation in classification

Predictive calibration (Gneiting et al., 2007; Guo et al., 2017; Li & Sur, 2025) refers to the agreement between predicted probabilities and observed outcome frequencies. It is an important topic in machine learning for many applications (Sanders et al., 2022; Zou et al., 2022) including survival analysis (Lee et al., 2024; Alberge et al., 2024), self-driving systems (Michelmore et al., 2018), time series forecasting (Adler et al., 2026), natural language processing (Nguyen & O'Connor, 2015; Yuan et al., 2025; Yang et al., 2025) and medical diagnosis (Liou et al., 2024; Barfoot et al., 2025; Vasconcelos et al., 2023; Sambyal et al., 2023).

In clinical risk-sensitive decision making systems, such as abnormality classification from medical images, classification networks should be accurate while indicating when they are likely to be incorrect. Specifically for CT scan analysis, confidence scores should faithfully reflect the likelihood that a prediction is correct, enabling uncertain or high-risk cases to be appropriately deferred to human radiologists (Jiang et al., 2012). Such networks must be able to predict a binary prediction, as well as a corresponding predicted class probability that should reflect its ground truth correctness likelihood (Gómez-Gálvez et al., 2026; Kiryo et al., 2025).

In this study, we evaluate the calibration of state-of-the-art visual encoders for multi-label abnormality classification from 3D CT scans, motivated by prior findings in natural images showing that deep models tend to be overconfident (Ye et al., 2023; Pinto et al., 2022). Recent work further shows that calibration improvements achieved in-domain may not transfer robustly under distribution shift (Roschewitz & Mehta, 2025), a setting particularly relevant to medical imaging due to variations in acquisition protocols, patient demographics, and pathology prevalence across institutions (David et al., 2018; Zhou et al., 2026).

Setup. We consider a multi-label classification problem with input $X \in \mathcal{X}$ and label vector $Y = (Y_1, \ldots, Y_K) \in \mathcal{Y} = \{0,1\}^K$, where $Y_k = 1$ indicates the presence of abnormality k. Let $f : \mathcal{X} \to [0,1]^K$ be a neural network that outputs a vector of predicted probabilities $\hat{P} = (\hat{P}_1, \ldots, \hat{P}_K) \in [0,1]^K$, where $\hat{P}_k$ estimates the probability that abnormality k is present.

For a fixed abnormality $k \in \{1, \ldots, K\}$, calibration is assessed by treating $(Y_k, \hat{P}_k)$ as a binary probabilistic prediction problem. A perfectly calibrated model for label k satisfies:

$$\mathbb{P}(Y_k = 1 \mid \hat{P}_k = p) = p, \qquad \forall\, p \in [0,1], \tag{1}$$

where $\mathbb{P}(\cdot)$ denotes the true probability measure over the joint distribution of (X, Y).

For a fixed label k, let $\{(\hat{p}_{ik}, y_{ik})\}_{i=1}^n$ denote the predicted probabilities and binary outcomes over n evaluation samples, with $\hat{p}_{ik} \in [0,1]$ and $y_{ik} \in \{0,1\}$. Reported calibration scores are computed for each label and then macro-averaged across labels.

Expected Calibration Error. For a fixed k, ECE (Naeini et al., 2015) is a plug-in estimator of the population calibration error $\mathbb{E}[|\mathbb{P}(Y_k = 1 \mid \hat{P}_k = p) - p|]$, approximated in practice via binning. Let $[0,1]$ be partitioned into $B \in \mathbb{N}^+$ disjoint bins $\{I_b\}_{b=1}^B$, and define $B_{b,k} = \{i : \hat{p}_{ik} \in I_b\}$ as the set of samples assigned to bin b for label k. The label-wise ECE is defined as:

$$\mathrm{ECE}_k = \sum_{b=1}^{B} \frac{|B_{b,k}|}{n} \left|\mathrm{acc}_k(B_{b,k}) - \mathrm{conf}_k(B_{b,k})\right|, \tag{2}$$

where:

$$\mathrm{acc}_k(B_{b,k}) = \frac{1}{|B_{b,k}|} \sum_{i \in B_{b,k}} y_{ik}, \qquad \mathrm{conf}_k(B_{b,k}) = \frac{1}{|B_{b,k}|} \sum_{i \in B_{b,k}} \hat{p}_{ik}. \tag{3}$$

Adaptive Calibration Error. ACE (Nixon et al., 2019) follows the same formulation as ECE but uses adaptive (quantile-based) binning, such that each bin contains approximately the same number of samples. This reduces sensitivity to uneven confidence distributions by weighting bins uniformly.

Kernel Calibration Error. Because ECE can be sensitive to the choice of binning strategy, we also report KCE, a non-parametric measure of miscalibration that avoids confidence discretization (Widmann et al., 2019). Unlike bin-based metrics, KCE evaluates calibration continuously over the prediction space, reducing sensitivity to binning choices. Formally, KCE is defined as follows:

$$\text{KCE}_k = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (y_{i,k} - \hat{p}_{i,k})\ (y_{j,k} - \hat{p}_{j,k})\ ker(\hat{p}_{i,k}, \hat{p}_{j,k})\,, \tag{4}$$

where $ker(\cdot,\cdot)$ denotes a positive-definite kernel. In our experiments, we use the gaussian kernel such that:

$$ker(\hat{p}_{i,k}, \hat{p}_{j,k}) = \exp\left(-\frac{(\hat{p}_{i,k}, \hat{p}_{j,k})^2}{2\sigma^2}\right)\,, \tag{5}$$

where σ is the bandwidth of the probability kernel, fixed to 0.1 in our experiments. KCE quantifies the dependence between prediction residuals $(y_{i,k} - \hat{p}_{i,k})$ and confidence estimates, yielding a non-parametric measure of miscalibration that avoids confidence binning. The inclusion of KCE ensures that the calibration trends observed across models, labels, and distribution shifts are not artifacts of a particular binning strategy.

Hosmer-Lemeshow test. The HL test (Hosmer & Lemeshow, 1980) is a goodness-of-fit test for probabilistic classifiers that evaluates calibration by comparing predicted and observed frequencies across subgroups of the model population. In the multi-label setting, it is computed independently for each abnormality $k \in \{1, \ldots, K\}$, treating it as a separate binary problem. As in ECE, predictions are partitioned into B bins, typically using quantiles. For each bin $b \in \{1, \ldots, B\}$, let $o_{b,k} = \sum_{i \in B_{b,k}} y_{ik}$ denote the observed number of positive outcomes and $e_{b,k} = \sum_{i \in B_{b,k}} \hat{p}_{ik}$ the expected number under the model. The HL statistic is defined as follows:

$$\text{HL}_k = \sum_{b=1}^{B} \frac{(o_{b,k} - e_{b,k})^2}{e_{b,k}\left(1 - \frac{e_{b,k}}{|B_{b,k}|}\right)}\,. \tag{6}$$

Under the null hypothesis of perfect calibration, HL_k asymptotically follows a χ^2 distribution with $B-2$ degrees of freedom, enabling a hypothesis test of calibration. A large value of the statistic leads to rejecting this hypothesis, indicating evidence of miscalibration.

Brier score. The Brier score (Brier, 1950) is a proper scoring rule that measures the accuracy of probabilistic predictions. It is defined as the mean square error between predicted confidences and outcomes:

$$\text{Brier}_k = \frac{1}{n} \sum_{i=1}^{n} (\hat{p}_{ik} - y_{ik})^2. \tag{7}$$

The Brier score jointly captures aspects of both calibration and discrimination. It admits a decomposition into uncertainty, resolution and reliability terms (Murphy, 1973), where the reliability component (the second term in Equation 8) reflects calibration. Formally, the decomposition is defined as follows:

$$\text{Brier}_k = \frac{1}{n} \sum_{i=1}^{n} (\hat{p}_{ik} - \bar{p}_k)^2 - \frac{1}{n} \sum_{i=1}^{n} (\hat{p}_{ik} - y_{ik})^2 + \bar{p}(1-\bar{p})\,, \quad \bar{p}_k = \frac{1}{n} \sum_{i=1}^{n} y_{ik}\,, \tag{8}$$

where $\bar{p}_k$ is the empirical label prevalence. Unlike ECE and ACE, which isolate calibration via binning, the Brier score provides a strictly proper objective that evaluates the overall quality of probabilistic predictions.

While the metrics above evaluate marginal calibration for each abnormality independently, we further analyze how inter-label dependencies affect calibration through co-occurrence in Section 3.3, and we provide a preliminary analysis of canonical calibration of the model's joint prediction vector in Section. A.2.9.

3 Experiments

3.1 Experimental setup

Datasets. Models are trained and evaluated using the publicly available CT-RATE dataset (Hamamci et al., 2026), consisting of chest CT scans from 21,304 unique patients, and 18 abnormalities extracted from radiology reports. To perform a 5-fold cross-validation, the initial training set is split with a 85/15 ratio performed at the patient-level, and the official test set is untouched. Consistent with prior works (Hamamci et al., 2026; Draelos et al., 2021), volumes are reformatted with SLP orientation, resampled to $1.5 \times 0.75 \times 0.75$ mm and cropped or padded to a fixed size of $240 \times 480 \times 480$. Hounsfield Units are clipped to $[-1000, +200]$ reflecting the diagnosis limits (DenOtter & Schubert, 2024), before being mapped to the range $[0, 1]$ and normalized with ImageNet (Deng et al., 2009) statistics (-0.449). To assess cross-dataset evaluation, we further evaluate models on the external RAD-ChestCT (Draelos et al., 2021), with the same data pre-processing steps. RAD-ChestCT differs from CT-RATE in terms of data distribution, including variations in acquisition protocols, patient demographics, and abnormality prevalence. Following the CT-RATE protocol, we restrict the evaluation on RAD-ChestCT to 16 abnormalities, grouping both *coronary artery wall calcification* and *arterial wall calcification* under *calcification*, and excluding *mosaic attenuation pattern* which is not annotated.

Visual encoders. We consider three families of architectures capturing complementary inductive biases for volumetric medical imaging. First, we include a 3D vision transformer (Dosovitskiy et al., 2021) (ViT-3D), which models CT scans as a set of volumetric patches interacting through global self-attention (Vaswani et al., 2023). Second, we consider a fully convolutional 3D visual encoder based on a ResNet backbone (He et al., 2015; Carreira & Zisserman, 2018) (I3D), which relies on localized kernels and hierarchical feature extraction. Third, we evaluate a family of 2.5D approaches that represent the CT volume as sequences of axial slice triplets, inducing different receptive fields along the craniocaudal axis. Specifically, CT-Net (Draelos et al., 2021) aggregates triplet-level feature maps using a lightweight 3D convolutional module, CT-Scroll (Di Piazza et al., 2025) models both global and local inter-slice dependencies via self-attention, while CT-SSG (Di Piazza et al., 2026) incorporates localized spectral graph-based aggregation.

Initialization. To ensure fair comparison across architectures, all visual encoders are initialized from ImageNet-21k pre-trained weights (Deng et al., 2009). For the 3D vision transformer (ViT-3D), the patch embedding layer is inflated from 2D to 3D, and transformer blocks are initialized directly from their 2D Base-16 pre-trained counterparts. For the 3D convolutional model (I3D), weights are obtained by inflating a 2D ResNet-18 into 3D (Carreira & Zisserman, 2018). For all 2.5D approaches (CT-Net, CT-Scroll, CT-SSG), the triplet-wise 2D visual encoder is a 2D ResNet-18 (He et al., 2015) initialized from pre-trained weights.

Metrics. All metrics are computed per abnormality and macro-averaged across abnormalities and folds. Calibration metrics use 10 equal-width bins following standard practice (Lee et al., 2024). See Table 11 for binning sensitivity, Table 10 for micro-level results, and Tables 12 & 13 for abnormality-wise performance.

3.2 Main quantitative results

Internal evaluation. Table 1 shows that all models exhibit consistent ranking across classification metrics on CT-RATE, with the 2.5D CT-Scroll and CT-SSG achieving the strongest performance overall. In contrast, calibration metrics reveal a different ordering: ViT-3D and CT-Scroll consistently obtain the lowest ECE, ACE, and KCE. Notably, both models incorporate global dependency modeling through self-attention, suggesting that attention-based architectures may produce better-calibrated predictions than purely convolutional approaches. Paired t-tests (Ross & Willson, 2017) on ECE across folds indicate that the calibration improvements achieved by ViT-3D and CT-Scroll over CT-Net and I3D are statistically significant ($p < 0.05$). Complete pairwise p-values for calibration metrics are in Table 8. However, when considering the Brier score, which couples calibration and discrimination, CT-Scroll and CT-SSG perform best, indicating that higher discriminative capacity partially compensates calibration error in this composite metric.

Table 1: Evaluation of visual encoders on the multi-label abnormality classification task on CT-RATE. **Best results** are in bold, and second best are underlined. Higher is better (↑). Lower is better (↓). Standard deviation (±) are computed over the 5 folds. Random classification predictions are sampled from a Bernoulli distribution with $p = 0.5$, while probabilistic ones are drawn uniformly from $[0, 1]$.

	Classification ↑			**Calibration ↓**			**Brier ↓**
Method	F1	AUROC	AUPRC	ECE	ACE	KCE (10^{-1})	Score
Random	0.278±0.005	0.499±0.006	0.209±0.001	0.359±0.002	0.359±0.002	0.605±0.010	0.334±0.001
ViT-3D	0.495±0.005	0.790±0.004	0.463±0.003	**0.029**±0.002	**0.032**±0.002	**0.008**±0.002	0.115±0.001
I3D	0.529±0.005	0.817±0.003	0.522±0.004	0.040±0.011	0.040±0.009	0.013±0.007	0.110±0.001
CT-Net	0.510±0.006	0.797±0.003	0.485±0.006	0.054±0.012	0.052±0.011	0.023±0.009	0.114±0.003
CT-Scroll	0.543±0.002	0.822±0.003	0.531±0.005	0.031±0.005	**0.032**±0.004	0.010±0.002	0.107±0.001
CT-SSG	**0.572**±0.002	**0.836**±0.002	**0.582**±0.005	0.041±0.006	0.040±0.005	0.014±0.005	**0.103**±0.002

External evaluation. Table 2 presents results on RAD-ChestCT. For the classification metrics, I3D, CT-Scroll and CT-SSG remain among the strongest baselines. However, calibration behavior becomes more heterogeneous under distribution shift. The clear advantage of attention-based models observed on CT-RATE diminishes: ViT-3D, I3D, and CT-Scroll achieve similar ECE and ACE, while CT-Net and CT-SSG exhibit higher calibration error. Importantly, ViT-3D maintains a consistent advantage over CT-SSG under distribution shift, with the only paired comparison achieving statistical significance (paired t-test across folds, $p < 0.05$). Notably, I3D yields the best Brier score, suggesting improved robustness of 3D convolutional representations under distribution shift. These results suggest that calibration rankings do not consistently align with discriminative rankings across architectures, particularly under distribution shift.

Table 2: Cross-dataset evaluation of CT-RATE trained backbones on the external RAD-ChestCT.

	Classification ↑			**Calibration ↓**			**Brier ↓**
Method	F1	AUROC	AUPRC	ECE	ACE	KCE (10^{-1})	Score
Random	0.359±0.004	0.497±0.006	0.327±0.008	0.336±0.001	0.336±0.001	0.532±0.011	0.333±0.001
ViT-3D	0.479±0.008	0.669±0.002	0.479±0.003	**0.133**±0.005	**0.134**±0.005	0.208±0.025	0.181±0.001
I3D	0.507±0.010	0.713±0.006	0.530±0.034	0.134±0.007	**0.134**±0.007	0.207±0.030	**0.171**±0.002
CT-Net	0.491±0.008	0.691±0.005	0.497±0.037	0.141±0.009	0.140±0.009	**0.187**±0.016	0.180±0.004
CT-Scroll	0.495±0.008	0.714±0.003	0.525±0.052	0.134±0.008	0.135±0.008	0.198±0.032	0.173±0.003
CT-SSG	**0.523**±0.009	**0.746**±0.004	**0.589**±0.028	0.147±0.004	0.148±0.005	0.228±0.024	0.174±0.003

Hosmer-Lemeshow test. Table 3 reports the Hosmer-Lemeshow test pass rate across abnormalities and cross-validation folds. On CT-RATE, ViT-3D and CT-Scroll achieve the highest rates, confirming the calibration advantage of attention-based models over 3D convolution-based (I3D) and other 2.5D baselines (CT-Net and CT-SSG) observed in Tables 1 and 2. In contrast, all models exhibit consistently low pass rates on RAD-ChestCT, indicating a systematic degradation of calibration under distribution shift regardless of architecture. Across all three evaluation settings (Tables 1, 2 and 3), two findings remain consistent in our study: i) attention-based visual encoders appear to exhibit better calibration in-domain, but ii) this advantage diminishes under distribution shift.

Table 3: Hosmer-Lemeshow test pass counts ($\alpha = 0.01$) across folds and labels, on CT-RATE (internal) and RAD-ChestCT (external). Higher values (↑) indicate fewer rejections of the null hypothesis.

Method	**Internal**	**External**
ViT-3D	**34** / 90	1 / 80
I3D	16 / 90	0 / 80
CT-Net	2 / 90	3 / 80
CT-Scroll	30 / 90	**4** / 80
CT-SSG	3 / 90	0 / 80

Reliability diagrams. Figure 1 shows reliability diagrams (DeGroot & Fienberg, 1983; Niculescu-Mizil & Caruana, 2005) highlighting calibration differences across datasets. For lung opacity, models appear well-calibrated on CT-RATE but exhibit over-confidence under distribution shift. In contrast, for lymphadenopathy,

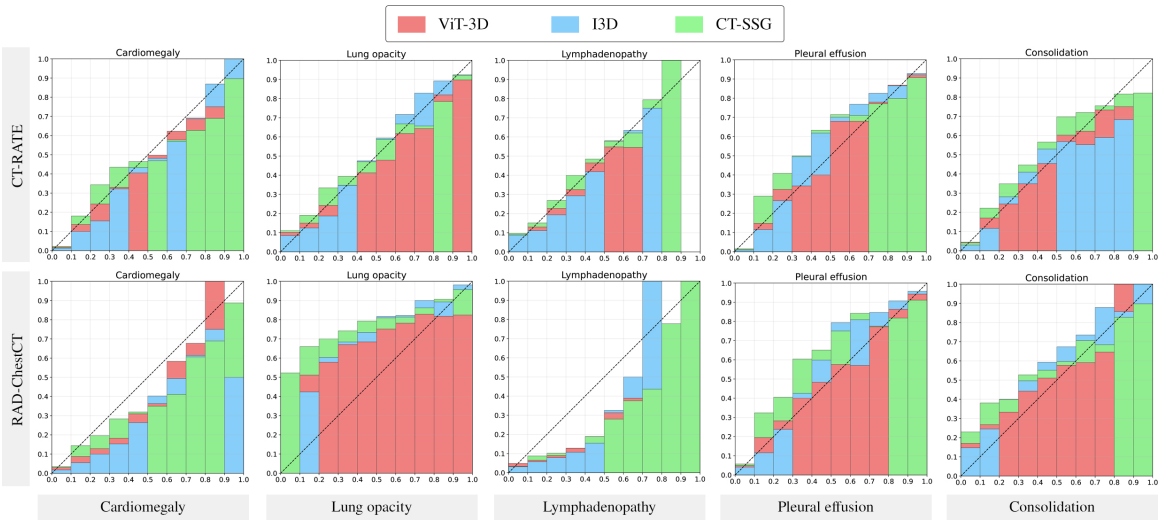


Figure 1: Reliability diagrams for three representative architectures (ViT-3D, I3D and CT-SSG) across datasets and a subset of abnormalities, illustrating contrasting over- and under-confidence behaviors.

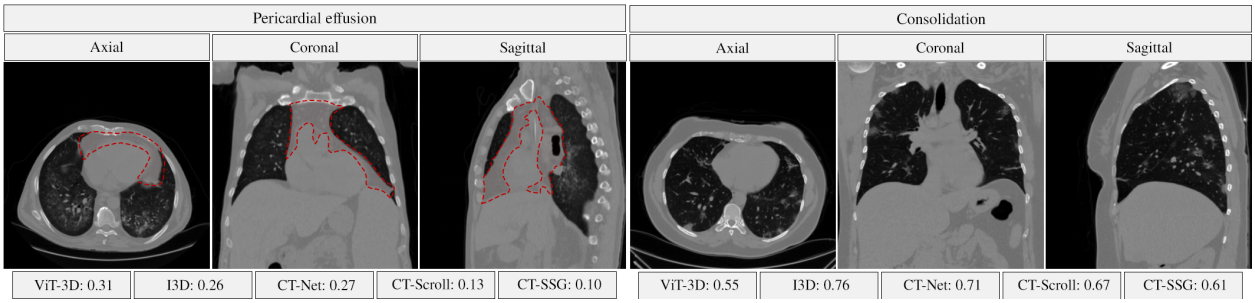


Figure 2: Axial, coronal and sagittal slices illustrating under- (left) and over-confidence (right). Manually drawn masks are shown for visualization. Predicted probabilities are reported below.

models tend to be under-confident on RAD-ChestCT, indicating that miscalibration manifests differently across pathologies. Abnormality-wise reliability diagrams are provided in Figure 16 for 3D CT visual encoders (ViT-3D, I3D), and in Figure 15 for 2.5D CT visual encoders (CT-Net, CT-Scroll, CT-SSG).

Figure 2 provides representative qualitative examples that illustrate calibration failures at the image level. For pericardial effusion, models produce low-confidence predictions despite clear radiological evidence, reflecting under-confidence. In contrast, for consolidation, where multifocal ground-glass opacities lead to elevated predicted probabilities despite the absence of clear consolidation patterns, highlighting calibration errors induced by overlapping visual features across abnormalities. Additional examples are provided in Figure 12.

Relationship between discrimination and calibration. We next analyze the abnormality-wise relationship between discrimination and calibration through AUROC-ECE associations. This analysis operates at a different granularity: whereas Tables 1 & 2 compare macro-averaged model rankings across architectures, Figure 3 examines label-wise AUROC-ECE associations within each model and dataset. Spearman correlations are consistently negative across all model-dataset pairs (correlation values ranging from -0.11 to -0.40 across abnormalities), suggesting a weak-to-moderate tendency for abnormalities with higher AUROC to also exhibit lower ECE. This finer-grained trend coexists with the architecture-level ranking differences observed earlier.

We further contextualize our findings by evaluating pre-trained CT foundation models in Section 3.5, extending our analysis beyond the controlled supervised setting. Furthermore, while our evaluation primarily considers label-wise marginal calibration, we include a preliminary evaluation of canonical calibration in Section A.2.9.

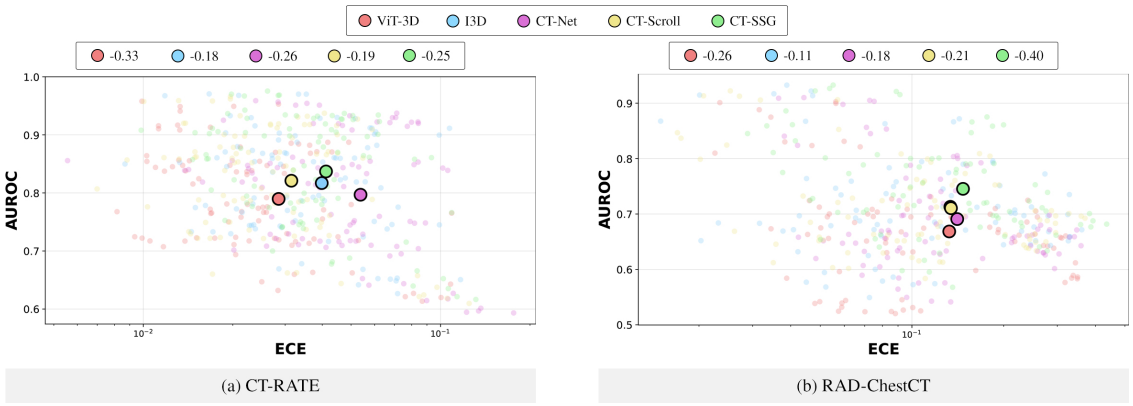


Figure 3: Relationship between AUROC and ECE, on (a) CT-RATE and (b) RAD-ChestCT. Each point represents a pair (ECE, AUROC) across abnormalities, folds and models, with colors indicating CT visual encoders. Spearman rank correlations are summarized on top of each plot.

3.3 Factors impacting calibration

We next investigate factors influencing calibration performance. Specifically, we analyze the impact of abnormality prevalence, volumetric size, anatomical system, and inter-abnormality co-occurrence.

Prevalence. Table 4 reports Spearman correlations between abnormality prevalence and performance metrics. AUROC exhibits a moderate negative correlation in-domain (-0.26 to -0.39) that does not replicate under distribution shift, where correlations cluster near zero. ECE shows consistent positive correlations across architectures on CT-RATE ($+0.58$ to $+0.81$), and remains moderately positive on RAD-ChestCT ($+0.49$ to $+0.67$), indicating that calibration error tends to scale with abnormality prevalence.

Table 4: Spearman rank correlation between abnormality prevalence and model performance. Positive (+) and negative (-) values indicate increasing or decreasing trends with prevalence, respectively.

	CT-RATE			**RAD-ChestCT**		
Method	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓
ViT-3D	-0.36±0.05	+0.74±0.08	+0.90±0.01	-0.13±0.05	+0.67±0.08	+0.91±0.01
I3D	-0.39±0.01	+0.71±0.09	+0.87±0.02	-0.03±0.07	+0.49±0.12	+0.85±0.04
CT-Net	-0.26±0.04	+0.81±0.05	+0.84±0.01	+0.07±0.08	+0.55±0.07	+0.86±0.04
CT-Scroll	-0.27±0.02	+0.58±0.20	+0.85±0.01	+0.03±0.03	+0.53±0.09	+0.81±0.03
CT-SSG	-0.36±0.03	+0.63±0.04	+0.84±0.01	-0.28±0.08	+0.65±0.10	+0.83±0.06

Volumetric size. We assess the effect of abnormality size using ReXGroundingCT (Baharoon et al., 2026), a CT-RATE subset providing segmentation masks for 12 pathologies, and restrict the analysis to the 231 test subjects shared between both datasets. To align with CT-RATE labels, we group pathologies into six meta-categories, detailed in Table 7, and define volumetric size as the union of corresponding masks. Model predictions are aggregated via max-pooling over associated classes. Table 5 reports Spearman correlations between abnormality size and performance. We observe a consistent positive correlation both for ECE and Brier, while AUROC shows no consistent trend with Spearman correlation ranging from -0.10 to $+0.13$. Since the convergent ECE/Brier signal coexists with a negligible

Table 5: Spearman correlation between abnormality size and performance on CT-RATE.

Method	**AUROC**	**ECE**	**Brier**
ViT-3D	-0.10±0.08	+0.57±0.22	+0.60±0.11
I3D	-0.09±0.01	+0.41±0.18	+0.59±0.09
CT-Net	-0.06±0.03	+0.53±0.20	+0.54±0.01
CT-Scroll	-0.03±0.14	+0.25±0.09	+0.53±0.02
CT-SSG	+0.13±0.11	+0.43±0.28	+0.54±0.01

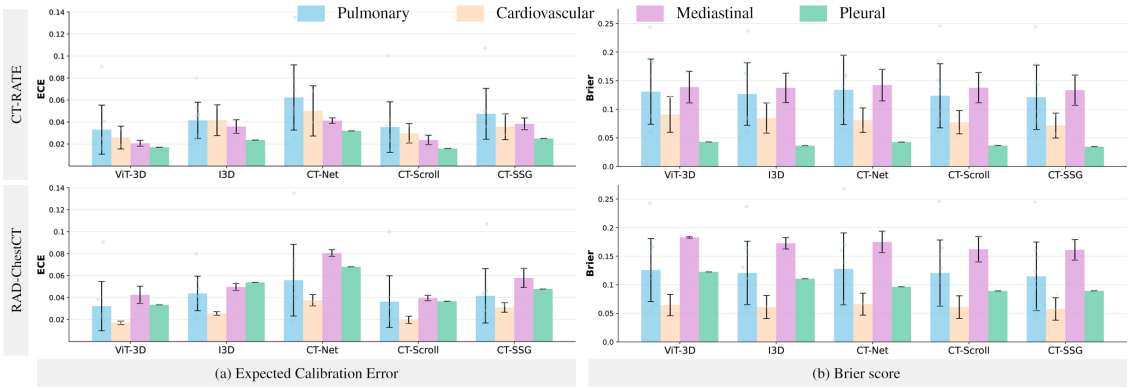


Figure 4: ECE (left) and Brier (right) performances across anatomical systems, both for CT-RATE (top) and RAD-ChestCT (bottom). Standard deviations are computed over abnormality performances.

AUROC trend, the Brier increase is attributable to its calibration component rather than to changes in discrimination, highlighting that volumetric extent primarily affects calibration rather than discrimination.

Anatomical system analysis. We examine calibration across anatomical systems by grouping abnormalities into four categories: *pulmonary*, *cardiovascular*, *mediastinal*, and *pleural*. Figure 4 shows that calibration varies substantially across systems. On CT-RATE, the pleural abnormality achieves the lowest calibration error. Meanwhile on RAD-ChestCT, mediastinal diseases exhibit the worst performance, indicating reduced robustness under distribution shift. In contrast, cardiovascular abnormalities remain comparatively stable across datasets, suggesting better generalization. Notably, these trends are consistent across models, indicating that calibration differences are primarily driven by data characteristics rather than architectural choices.

Abnormality co-occurrence. Prior work has shown that classification performance can benefit from abnormality co-occurrence (Draelos et al., 2021). We examine whether a similar effect holds for calibration. For each pair of abnormalities (i, j), with $i \neq j$, we measure the change in calibration of j conditioned on the presence of i, defined as $\Delta\text{ECE} = \text{ECE}(j \mid i = 1) - \text{ECE}(j \mid i = 0)$, and relate it to the conditional co-occurrence frequency $\mathbb{P}(j \mid i = 1)$. As shown in Figure 5, ΔECE increases with co-occurrence frequency, suggesting that calibration tends to deteriorate when abnormalities co-occur. This contrasts with gains in discriminative performance (Draelos et al., 2021), and suggests that models exploit inter-abnormality dependencies for prediction while failing to properly calibrate the associated uncertainty.

Furthermore, we extend our analysis to the spatial extent type (focal versus diffuse pathologies) in Figure 10, and observe only minor differences across models and datasets, suggesting a limited impact on calibration.

3.4 Towards post-hoc calibration

We evaluate whether standard post-hoc calibration methods, consisting of lightweight transformations applied to model outputs without retraining, can correct the observed miscalibration. We consider three representative approaches: (i) *temperature scaling* (Guo et al., 2017), which learns a single scalar to rescale logits and correct global over- or under-confidence; (ii) *Platt scaling* (Böken, 2021; Phelps et al., 2025), which fits a parametric sigmoid transformation of logits; and (iii) *isotonic regression* (Jiang et al., 2011), a non-parametric monotonic mapping with higher flexibility. As shown in Figure 6, temperature scaling has negligible impact, indicating that miscalibration is not driven by simple global confidence miscalibration. Isotonic regression substantially improves in-domain calibration on CT-RATE, but fails to generalize under distribution shift on RAD-ChestCT, suggesting overfitting of the calibration function. In contrast, Platt scaling provides more consistent improvements across datasets, indicating that calibration errors are structured but require regularized transformations to generalize across domains. As expected, post-hoc calibration has negligible impact on discriminative performance, indicating that these methods primarily adjust confidence calibration without altering overall prediction accuracy.

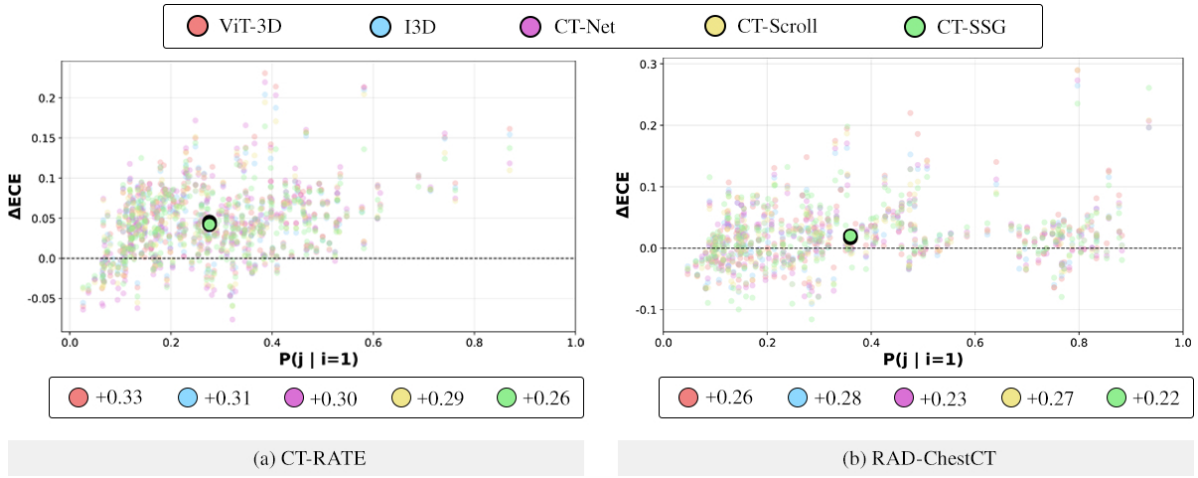


Figure 5: Relationship between conditional co-occurrence frequency $\mathbb{P}(j \mid i = 1)$ and change in calibration ΔECE, on CT-RATE (left) and RAD-ChestCT (right). Positive values indicate degraded calibration when abnormalities co-occur. Spearman rank correlations are reported below each plot, where colored markers correspond to models and are annotated with their respective coefficients.

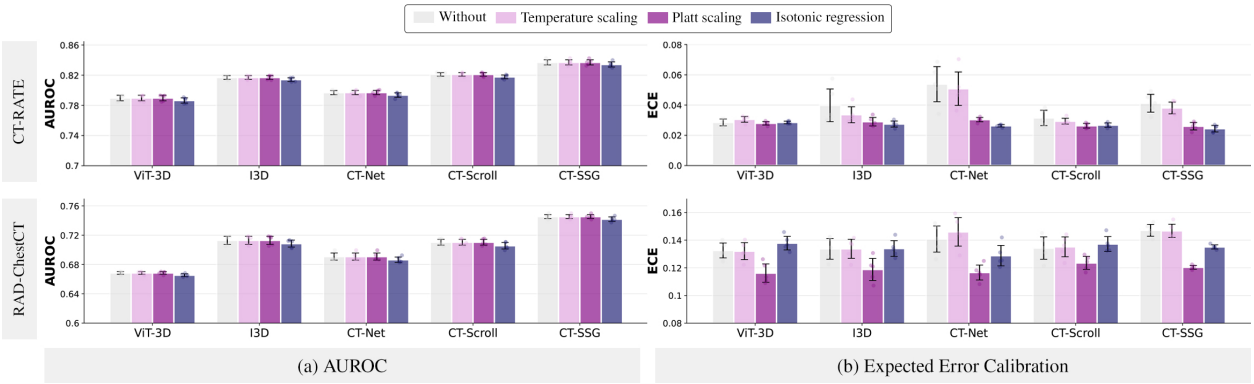


Figure 6: Comparison of three post-hoc calibration methods on AUROC (left, higher is better) and ECE (right, lower is better), evaluated both on CT-RATE (top) and RAD-ChestCT (bottom).

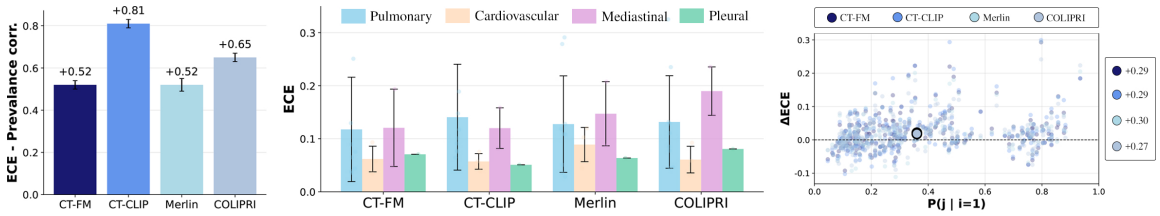
3.5 Contextualization with CT foundation models

Beyond our controlled benchmark in supervised settings, we further contextualize our findings by evaluating pre-trained computed tomography foundation models (CT FMs) under a linear probing protocol. We consider representative CT foundation models covering different pre-training paradigms. CT-FM (Pai et al., 2025) is trained with patch-level contrastive learning across heterogeneous CT volumes from multiple anatomical regions, emphasizing general visual representation learning. CT-CLIP (Hamamci et al., 2026) employs vision-language pre-training with a 3D-ViT encoder on chest CT scans, while Merlin (Blankemeier et al., 2026) follows a similar contrastive alignment paradigm using a convolutional I3D backbone on abdominal CT data. Finally, COLIPRI (Wald et al., 2026a), leveraging Primus (Wald et al., 2026b) for visual encoding, adopts a multi-stage objective combining vision-language alignment, report generation, and masked reconstruction on chest CT cohorts. Because CT-CLIP and COLIPRI pre-training data overlaps with our cross-validation sets, we restrict the evaluation to the held-out RAD-ChestCT. As CT FMs differ substantially from our controlled setting; spanning pre-training data, supervision, architectural design, and input pre-processing (*e.g.*, HU values, spacing); this comparison is intended for contextualization rather than direct benchmarking.

Table 6: Performance comparison on RAD-ChestCT, including pre-trained CT foundation models evaluated via linear probing on the abnormality classification task. *Data* refers to the body region(s) for model training, *Param.* denotes the number of visual encoder parameters (in millions, M), and *HL* is the number of Hosmer-Lemeshow test pass counts across labels and folds, with $\alpha = 0.05$. COLIPRI with (†) indicates sequential pre-training, including contrastive scan-report alignment (CLIP) and masked reconstruction objectives. Among CT foundation models, **best results** are in bold and second best are underlined.

Setup				Classification	Calibration			Brier
Method	Pre-training	Data	Param.	AUROC ↑	ECE ↓	KCE (10^{-1}) ↓	HL ↑	Score ↓
CT Visual Encoders								
ViT-3D	-	Chest	23 M	0.669±0.002	0.133±0.005	0.208±0.025	1 / 80	0.181±0.001
I3D	-	=	33 M	0.713±0.006	0.134±0.007	0.207±0.030	0 / 80	0.171±0.002
CT-Net	-	=	27 M	0.691±0.005	0.141±0.009	0.187±0.016	3 / 80	0.180±0.004
CT-Scroll	-	=	16 M	0.710±0.004	0.134±0.008	0.198±0.032	4 / 80	0.173±0.003
CT-SSG	-	=	12 M	0.745±0.003	0.147±0.004	0.228±0.024	0 / 80	0.174±0.003
CT Foundation Models								
CT-FM	SimCLR	Whole-body	77 M	0.633±0.001	0.129±0.001	0.285±0.004	0 / 80	0.187±0.001
CT-CLIP	CLIP	Chest	26 M	0.634±0.001	0.148±0.001	0.288±0.002	0 / 80	0.197±0.002
Merlin	CLIP	Abdomen	121 M	0.688±0.001	**0.128**±0.002	0.208±0.001	0 / 80	0.171±0.001
COLIPRI	Multi-stage†	Chest	147 M	**0.740**±0.001	0.132±0.003	**0.166**±0.002	**1** / 80	**0.164**±0.002

Figure 7: Comprehensive analysis of factors of miscalibration for CT foundation models. We include (a) correlation between ECE and abnormality prevalance; (b) per-organ calibration performance; and (c) relationship between conditional co-occurrence frequency and change in calibration.



(a) Prevalance (b) Organ type (c) Co-occurence

3.5.1 Quantitative results for CT foundation models

Table 6 reports classification and calibration performance alongside pre-training objectives and data, as well as parameter scales of visual encoders. Overall, CT FMs exhibit heterogeneous behavior. While CT-FM achieves competitive calibration despite weaker discriminative performance, others combine both competitive classification and calibration results (Merlin, COLIPRI), suggesting better probability estimation. Notably, COLIPRI achieves the best AUROC, KCE and Brier score among CT foundation models while also attaining competitive ECE, outperforming fully supervised CT visual encoders across most calibration metrics in our settings. In contrast, foundation models with lower AUROC, such as CT-FM or Merlin, can still yield competitive ECE, indicating a partial decoupling between discrimination and calibration. These results suggest that large-scale pre-training induces structured prediction behaviors that may benefit calibration, but remain inconsistent across architectures and objectives. As shown in Table 14, which details paired t-test p-values across calibration metrics, all tests yield $p < 0.05$, indicating statistically significant differences in calibration performance across CT FMs. Abnormality-wise ECE results are reported in Table 14.

3.5.2 Characterization analysis of CT foundation models

Prevalence. Figure 7 (a) reports Spearman correlations between abnormality prevalence and ECE across CT foundation models. We consistently observe strong positive correlations (from +0.52 to +0.81) across all models. As also shown for CT visual encoders (see Table 4), these results indicate that calibration error increases with abnormality prevalence under distribution shift, even for CT FMs.

Organ type. Figure 7 (b) further reports ECE stratified by organ type, revealing consistently lower calibration error for cardiovascular and pleural pathologies across models. The relative ranking of organ groups is stable across CT FMs and closely matches that observed for CT visual encoders.

Co-occurence. We analyze the relationship between calibration change for pathology j and the conditional occurrence of j given i. As illustrated in Figure 7 (c), we observe weak to moderate positive Spearman correlations across CT FMs (from +0.27 to +0.30), indicating that calibration tends to degrade when abnormalities co-occur. This pattern is consistent with observations for 3D CT visual encoders (see Figure 5).

Overall, the consistent patterns observed across both CT visual encoders and CT FMs suggest that calibration is primarily driven by data characteristics rather than architectural or supervision differences.

4 Discussion

Evaluating the calibration properties of visual encoders for 3D computed tomography addresses a critical yet underexplored aspect of medical image analysis, where reliable probability estimates are essential for clinical decision-making. While prior work has largely emphasized discriminative performance, our study shows that strong classification metrics do not systematically translate into well-calibrated predictions. In particular, we observe a consistent gap between in-domain and out-of-distribution calibration, indicating that current models lack robustness under distribution shift.

Our analyses further reveal that miscalibration is not uniform, but structured and data-dependent. Calibration error increases with abnormality prevalence and volumetric size, and varies across anatomical systems, highlighting heterogeneous behavior that is not captured by aggregate metrics. These findings suggest that calibration in multi-label CT classification is shaped by underlying data properties and inter-label dependencies rather than simple confidence mis-scaling.

Importantly, standard post-hoc calibration methods only partially mitigate these effects. While flexible approaches improve in-domain calibration, their limited generalization under distribution shift indicates that miscalibration cannot be reliably corrected after training. This points to the need for calibration-aware learning strategies that account for dataset structure and domain variability.

Our controlled benchmark of CT visual encoders and linear probing evaluation of CT foundation models are both limited to binary cross-entropy supervision, the dominant training objective in CT abnormality classification. We anticipate that future work should explore alternative supervision strategies, including strictly proper losses with label-pair regularization and copula-based calibration, that explicitly account for inter-label dependencies during training and may yield better-calibrated joint predictions.

Overall, our results highlight that calibration remains a fundamental challenge for 3D CT abnormality classification. Developing models that are both discriminative and reliably calibrated across pathologies, anatomical systems, and domains is essential for safe and effective clinical deployment. Additional discussion of these limitations and broader impacts are provided in Section A.1.

Acknowledgments

This work of academic research makes use of two publicly available 3D computed tomography datasets. We gratefully acknowledge the contributors of the CT-RATE (Hamamci et al., 2026), and RAD-ChestCT (Draelos et al., 2021) datasets for making these resources publicly available to the research community. We extend our gratitude to the contributors from CT-FM (Pai et al., 2025), CT-CLIP (Hamamci et al., 2026), Merlin (Blankemeier et al., 2026) and COLIPRI (Wald et al., 2026a) for releasing their CT foundation models.

References

Coen Adler, Yuxin Chang, Felix Draxler, Samar Abdi, and Padhraic Smyth. Beyond accuracy: Are time series foundation models well-calibrated? *ICLR*, 2026.

Julie Alberge, Vincent Maladière, Olivier Grisel, Judith Abécassis, and Gaël Varoquaux. Survival Models: Proper Scoring Rule and Stochastic Optimization with Competing Risks, October 2024. URL `http://arxiv.org/abs/2410.16765`. arXiv:2410.16765 [stat].

Mohammed Baharoon, Luyang Luo, Michael Moritz, Abhinav Kumar, Sung Eun Kim, Xiaoman Zhang, Miao Zhu, Mahmoud Hussain Alabbad, Maha Sbayel Alhazmi, Neel P. Mistry, Lucas Bijnens, Kent Ryan Kleinschmidt, Brady Chrisler, Sathvik Suryadevara, Sri Sai Dinesh Jaliparthi, Noah Michael Prudlo, Mark David Marino, Jeremy Palacio, Rithvik Akula, Di Zhou, Hong-Yu Zhou, Ibrahim Ethem Hamamci, Scott J. Adams, Hassan Rayhan AlOmaish, and Pranav Rajpurkar. ReXGroundingCT: A 3D Chest CT Dataset for Segmentation of Findings from Free-Text Reports. *NEJM AI*, 2026.

David Bani-Harouni, Chantal Pellegrini, Julian Lüers, Su Hwan Kim, Markus Baalmann, Benedikt Wiestler, Rickmer Braren, Nassir Navab, and Matthias Keicher. Calibrated confidence expression for radiology report generation. In *MICCAI*, 2026.

Theodore Barfoot, Luis C. Garcia-Peraza-Herrera, Samet Akcay, Ben Glocker, and Tom Vercauteren. Average Calibration Losses for Reliable Uncertainty in Medical Image Segmentation, June 2025. URL `http://arxiv.org/abs/2506.03942`. arXiv:2506.03942 [cs] version: 1.

M. Fernanda Bellolio, Herbert C. Heien, Lindsey R. Sangaralingham, Molly M. Jeffery, Ronna L. Campbell, Daniel Cabrera, Nilay D. Shah, and Erik P. Hess. Increased Computed Tomography Utilization in the Emergency Department and Its Association with Hospital Admission. *The Western Journal of Emergency Medicine*, 18(5):835–845, August 2017. ISSN 1936-9018. doi: 10.5811/westjem.2017.5.34152.

Emanual Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, 2021.

Eugene Berta, Francis Bach, and Michael Jordan. Classifier Calibration with ROC-Regularized Isotonic Regression, November 2023. URL `http://arxiv.org/abs/2311.12436`. arXiv:2311.12436 [cs].

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Robbie Holland, Cesar Truyts, Christian Bluethgen, Yufu Wu, Long Lian, Malte Engmann Kjeldskov Jensen, Sophie Ostmeier, Maya Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson Amaro Junior, Neera Ahuja, Jason Fries, Nigam H. Shah, Greg Zaharchuk, Marc Willis, Adam Yala, Andrew Johnston, Robert D. Boutin, Andrew Wentland, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, and Akshay S. Chaudhari. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, March 2026. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-026-10181-8. URL `https://www.nature.com/articles/s41586-026-10181-8`.

Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997. ISSN 0031-3203. doi: 10.1016/S0031-3203(96)00142-2. URL `https://www.sciencedirect.com/science/article/pii/S0031320396001422`.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1): 1–3, January 1950. ISSN 1520-0493, 0027-0644. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Joshua Broder and David M. Warshauer. Increasing utilization of computed tomography in the adult emergency department, 2000–2005. *Emergency Radiology*, 13(1):25–30, September 2006. ISSN 1438-1435. doi: 10.1007/s10140-006-0493-9. URL https://doi.org/10.1007/s10140-006-0493-9.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. In *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 2008.

Björn Böken. On the appropriateness of Platt scaling in classifier calibration. *Information Systems*, 95: 101641, January 2021. ISSN 0306-4379. doi: 10.1016/j.is.2020.101641. URL https://www.sciencedirect.com/science/article/pii/S0306437920301083.

Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, February 2018. URL http://arxiv.org/abs/1705.07750. arXiv:1705.07750 [cs].

Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Samit Hirawat, Venkat Sethuraman, Mariann Micsinai Balan, and Kevin Brown. Masked image modeling advances 3d medical image analysis. In *WACV*, 2023.

Jiacheng Cheng and Nuno Vasconcelos. Towards calibrated multi-label deep neural networks. In *CVPR*, 2024.

Anne T. David, Antony L. Palmer, Silvia Pani, and Andrew Nisbet. Assessment of the variation in ct scanner performance with scan parameters, for image optimisation in radiotherapy treatment planning. *Physica Medica*, 2018.

Morris H. DeGroot and Stephen E. Fienberg. The Comparison and Evaluation of Forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1):12–22, 1983. ISSN 0039-0526. doi: 10.2307/2987588. URL https://www.jstor.org/stable/2987588.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/document/5206848/authors.

Tami D. DenOtter and Johanna Schubert. Hounsfield Unit. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024. URL http://www.ncbi.nlm.nih.gov/books/NBK547721/.

Theo Di Piazza, Carole Lazarus, Olivier Nempont, and Loic Boussel. Imitating Radiological Scrolling: A Global-Local Attention Model for 3D Chest CT Volumes Multi-Label Anomaly Classification. *MIDL*, January 2025. URL https://openreview.net/forum?id=YPBbNAmGXW.

Theo Di Piazza, Carole Lazarus, Olivier Nempont, and Loic Boussel. Structured Spectral Graph Representation Learning for Multi-label Abnormality Analysis from 3D CT Scans, 2026. URL https://arxiv.org/abs/2510.10779v2.

Francesco Di Salvo, Sebastian Doerrich, Jonas Alle, and Christian Ledig. HypCBC: Domain-Invariant Hyperbolic Cross-Branch Con- sistency for Generalizable Medical Image Analysis. *TMLR*, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

Rachel Lea Draelos, David Dov, Maciej A. Mazurowski, Joseph Y. Lo, Ricardo Henao, Geoffrey D. Rubin, and Lawrence Carin. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical Image Analysis*, 67:101857, January 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101857. URL https://www.sciencedirect.com/science/article/pii/S1361841520302218.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. ISSN 1573-1405. doi: 10.1007/s11263-009-0275-4. URL https://doi.org/10.1007/s11263-009-0275-4.

Gerardo A. Flores, Alyssa H. Smith, Julia A. Fukuyama, and Ashia C. Wilson. Aligning Evaluation with Clinical Priorities: Calibration, Label Shift, and Error Costs, 2025. URL https://arxiv.org/abs/2506.14540. Version Number: 3.

Lior Frenkel and Jacob Goldberger. Calibration of Medical Imaging Classification Systems with Weight Scaling. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (eds.), *MICCAI*, volume 13438, pp. 642–651. Springer Nature Switzerland, Cham, 2022. ISBN 978-3-031-16451-4 978-3-031-16452-1. doi: 10.1007/978-3-031-16452-1_61. URL https://link.springer.com/10.1007/978-3-031-16452-1_61. Series Title: Lecture Notes in Computer Science.

Axel Geysels, Ben Van Calster, Bart De Moor, Wouter Froyman, and Dirk Timmerman. Calibration in Multiple Instance Learning: Evaluating Aggregation Methods for Ultrasound-Based Diagnosis. *MICCAI*, 2025.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. 2007.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks, June 2017. URL https://arxiv.org/abs/1706.04599v2.

Sandra Gómez-Gálvez, Tobias Olenyi, and Gillian Dobbie. Socrates Loss: Unifying Confidence Calibration and Classification by Leveraging the Unknown. *TMLR*, 2026.

Ibrahim Ethem Hamamci, Sezgin Er, Enis Simsar, Anjany Sekuboyina, Chinmay Prabhakar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, Hadrien Reynaud, Sarthak Pati, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. GenerateCT: Text-Conditional Generation of 3D Chest CT Volumes, November 2023. URL http://arxiv.org/abs/2305.16037. arXiv:2305.16037 [cs].

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging. March 2024. URL http://arxiv.org/abs/2403.06801. arXiv:2403.06801 [cs, eess].

Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Dong Yang, Pengfei Guo, Marc Edgar, Daguang Xu, Bernhard Kainz, and Bjoern Menze. Better Tokens for Better 3D: Advancing Vision-Language Modeling in 3D Medical Imaging, October 2025. URL http://arxiv.org/abs/2510.20639. arXiv:2510.20639 [cs].

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*, February 2026. ISSN 2157-846X. doi: 10.1038/s41551-025-01599-y. URL https://www.nature.com/articles/s41551-025-01599-y.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs].

David Hosmer and Stanley Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, 9(10):1043–1069, January 1980. ISSN 0361-0926, 1532-415X. doi: 10.1080/03610928008827941. URL https://www.tandfonline.com/doi/full/10.1080/03610928008827941.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Smooth Isotonic Regression: A New Method to Calibrate Predictive Models. *AMIA Summits on Translational Science Proceedings*, 2011:16–20, March 2011. ISSN 2153-4063. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3248752/.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association : JAMIA*, 19 (2):263–274, 2012. ISSN 1067-5027. doi: 10.1136/amiajnl-2011-000291. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3277613/.

SooYong Kim, Kyeonghun Kim, Taejin Kim, Yoonkyung Jeon, Jungmin Shin, Dayoon Lee, Hyunjun Kim, Won-jae Lee, Woo-kyung Jung, Hyuk-jae Lee, Pa Hong, and Namjoon Kim. Boundary Guidance for Efficient 3D CT Vision–Language Reasoning. *NeurIPS Workshop ER*, 2025.

Ryuichi Kiryo, Futoshi Futami, and Masashi Sugiyama. Estimating Expected Calibration Error for Positive-Unlabeled Learning. *TMLR*, 2025.

Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2015.

Dongjoon Lee, Hyeryn Park, and Changhee Lee. Toward a Well-Calibrated Discrimination via Survival Outcome-Aware Contrastive Learning. *NeurIPS*, 2024.

Jiayu Lei, Ziqing Fan, Yanyong Zhang, Weidi Xie, Ya Zhang, and Yanfeng Wang. Versatile Vision-Language Model for 3D Computed Tomography. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (8):5945–5954, March 2026. ISSN 2374-3468. doi: 10.1609/aaai.v40i8.37517. URL https://ojs.aaai.org/index.php/AAAI/article/view/37517.

Yufan Li and Pragya Sur. Optimal and Provable Calibration in High-Dimensional Binary Classification: Angular Calibration and Platt Scaling, 2025. URL https://arxiv.org/abs/2502.15131. Version Number: 3.

Lathan Liou, Erick Scott, Prathamesh Parchure, Yuxia Ouyang, Natalia Egorova, Robert Freeman, Ira S. Hofer, Girish N. Nadkarni, Prem Timsina, Arash Kia, and Matthew A. Levin. Assessing calibration and bias of a deployed machine learning malnutrition prediction model within a large healthcare system. *npj Digital Medicine*, 7(1):149, June 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01141-5. URL https://www.nature.com/articles/s41746-024-01141-5.

Siladittya Manna, Saumik Bhattacharya, and Umapada Pal. Self-Supervised Visual Representation Learning for Medical Image Analysis: A Comprehensive Survey. *TMLR*, 2024.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-Entropy Loss Functions: Theoretical Analysis and Applications, June 2023. URL http://arxiv.org/abs/2304.07288. arXiv:2304.07288 [cs].

Izet Masic. Medical Decision Making - an Overview. *Acta Informatica Medica*, 30(3):230–235, September 2022. ISSN 0353-8109. doi: 10.5455/aim.2022.30.230-235. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9560052/.

Rhiannon Michelmore, Marta Kwiatkowska, and Yarin Gal. Evaluating Uncertainty Quantification in End-to-End Autonomous Driving Control, November 2018. URL http://arxiv.org/abs/1811.06817. arXiv:1811.06817 [cs].

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. Calibrating Deep Neural Networks using Focal Loss, February 2020. URL https://arxiv.org/abs/2002.09437v2.

Allan H. Murphy. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, June 1973. ISSN 1520-0450. doi: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/apme/12/4/1520-0450_1973_012_0595_anvpot_2_0_co_2.xml.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), February 2015. ISSN 2374-3468. doi: 10.1609/aaai.v29i1.9602. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

Reabal Najjar. Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. *Diagnostics*, 13(17):2760, August 2023. ISSN 2075-4418. doi: 10.3390/diagnostics13172760. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10487271/.

Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models, September 2015. URL http://arxiv.org/abs/1508.05154. arXiv:1508.05154 [cs].

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning | Proceedings of the 22nd international conference on Machine learning, 2005. URL https://dl.acm.org/doi/10.1145/1102351.1102430.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring Calibration in Deep Learning. *CVPR Workshop*, 2019.

Suraj Pai, Ibrahim Hadzic, Dennis Bontempi, Keno Bressem, Benjamin H. Kann, Andriy Fedorov, Raymond H. Mak, and Hugo J. W. L. Aerts. Vision Foundation Models for Computed Tomography, February 2025. URL http://arxiv.org/abs/2501.09001. arXiv:2501.09001 [eess].

Arkapal Panda and Utpal Garain. Copula based trainable calibration error estimator of multi-label classification with label interdependencies. In *International Conference on Artificial Intelligence and Statistics*, 2025.

Paula R. Patel and Orlando De Jesus. CT Scan. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024. URL http://www.ncbi.nlm.nih.gov/books/NBK567796/.

Nathan Phelps, Daniel J Lizotte, and Douglas G Woolford. Using Platt's scaling for calibration after undersampling – limitations and how to address them. *TMLR*, 2025.

Francesco Pinto, Philip H. S. Torr, and Puneet K. Dokania. An Impartial Take to the CNN vs Transformer Robustness Contest, July 2022. URL http://arxiv.org/abs/2207.11347. arXiv:2207.11347 [cs].

Foster J. Provost, Tom Fawcett, and Ron Kohavi. The Case against Accuracy Estimation for Comparing Induction Algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pp. 445–453, San Francisco, CA, USA, July 1998. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-556-5.

Mélanie Roschewitz and Raghav Mehta. Where are we with calibration under dataset shift in image classification? *TMLR*, 2025.

Amanda Ross and Victor L. Willson. Paired Samples T-Test. In Amanda Ross and Victor L. Willson (eds.), *Basic and Advanced Statistical Tests: Writing Results Sections and Creating Tables and Figures*, pp. 17–19. SensePublishers, Rotterdam, 2017. ISBN 978-94-6351-086-8. doi: 10.1007/978-94-6351-086-8_4. URL https://doi.org/10.1007/978-94-6351-086-8_4.

Abhishek Singh Sambyal, Usma Niyaz, Narayanan C. Krishnan, and Deepti R. Bathula. Understanding calibration of deep neural networks for medical image classification. *Computer Methods and Programs in Biomedicine*, 2023.

Sergio Sanchez-Martinez, Oscar Camara, Gemma Piella, Maja Cikes, Miguel Ángel González-Ballester, Marius Miron, Alfredo Vellido, Emilia Gómez, Alan G. Fraser, and Bart Bijnens. Machine Learning for Clinical Decision-Making: Challenges and Opportunities in Cardiovascular Imaging. *Frontiers in Cardiovascular Medicine*, 8:765693, January 2022. ISSN 2297-055X. doi: 10.3389/fcvm.2021.765693. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8764455/.

Kate Sanders, Reno Kriz, Anqi Liu, and Benjamin Van Durme. Ambiguous Images With Human Judgments for Robust Visual Event Classification. *NeurIPS*, 2022.

Xing Shen, Justin Szeto, Mingyang Li, Hengguan Huang, and Tal Arbel. Exposing and Mitigating Calibration Biases and Demographic Unfairness in MLLM Few-Shot In-Context Learning for Medical Image Classification. *MICCAI*, 2025.

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and Ling Zhang. Large-scale and Fine-grained Vision-language Pre-training for Enhanced CT Image Understanding, January 2025. URL http://arxiv.org/abs/2501.14548. arXiv:2501.14548 [cs].

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, July 2009. ISSN 0306-4573. doi: 10.1016/j.ipm.2009.03.002. URL https://www.sciencedirect.com/science/article/pii/S0306457309000259.

Francisca Vasconcelos, Bobby He, Nalini Singh, and Yee Whye Teh. UncertaINR: Uncertainty Quantification of End-to-End Implicit Neural Representations for Computed Tomography. *TMLR*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. URL http://arxiv.org/abs/1706.03762.

Tassilo Wald, Constantin Ulrich, Stanislav Lukyanenko, Andrei Goncharov, Alberto Paderono, Maximilian Miller, Leander Maerkisch, Paul F. Jager, and Klaus Maier-Hein. Revisiting mae pre-training for 3d medical image segmentation. In *CVPR*, 2025.

Tassilo Wald, Ibrahim Ethem Hamamci, Yuan Gao, Sam Bond-Taylor, Harshita Sharma, Maximilian Ilse, Cynthia Lo, Olesya Melnichenko, Anton Schwaighofer, Noel C. F. Codella, Maria Teodora Wetscherek, Klaus H. Maier-Hein, Panagiotis Korfiatis, Valentina Salvatelli, Javier Alvarez-Valle, and Fernando Pérez-García. Comprehensive language-image pre-training for 3D medical image understanding, January 2026a. URL http://arxiv.org/abs/2510.15042. arXiv:2510.15042 [cs].

Tassilo Wald, Saikat Roy, Fabian Isensee, Constantin Ulrich, Sebastian Ziegler, Dasha Trofimova, Raphael Stock, Michael Baumgartner, Gregor Köhler, and Klaus Maier-Hein. Primus: Enforcing Attention Usage for 3D Medical Image Segmentation. *TMLR*, 2026b.

Rongsheng Wang, Fenghe Tang, Qingsong Yao, Rui Yan, Xu Zhang, Zhen Huang, Haoran Lai, Zhiyang He, Xiaodong Tao, Zihang Jiang, and S Kevin Zhou. SimCroP: Radiograph Representation Learning with Similarity-driven Cross-granularity Pre-training. *MICCAI*, 2025.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/1c336b8080f82bcc2cd2499b4c57261d-Abstract.html.

Yahan Yang, Soham Dan, Dan Roth, and Insup Lee. On Calibration of Multilingual Question Answering LLMs. *TMLR*, 2025.

Wenqian Ye, Yunsheng Ma, Xu Cao, and Kun Tang. Mitigating Transformer Overconfidence via Lipschitz Regularization, July 2023. URL http://arxiv.org/abs/2306.06849. arXiv:2306.06849 [cs].

Zhangdie Yuan, Zifeng Ding, and Andreas Vlachos. Introducing FORECAST: The Future Outcome Reasoning and Confidence Assessment Benchmark. *NeurIPS*, 2025.

Chenhui Zhao, Yiwei Lyu, Asadur Chowdury, Edward Harake, Akhil Kondepudi, Akshay Rao, Xinhai Hou, Honglak Lee, and Todd Hollon. Towards Scalable Language-Image Pre-training for 3D Medical Imaging. *TMLR*, 2026.

Qian Zhou, Chengting Lin, Jinyi Jiang, Yuwei Li, Yue Yu, Shiyang Huang, Chaokang Han, Liting Shi, and Lei Shi. Impact of ct acquisition settings on the stability of radiomic features and the performance of pulmonary nodule classification models. *Insights into Imaging*, 2026.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting Future World Events with Neural Networks. *NeurIPS*, 2022.

A Technical Appendices and Supplementary Material

In this appendix, we provide additional materials supporting the results presented in the main paper. We begin by discussing the limitations of our evaluation and the broader implications of calibration for multi-label abnormality classification in 3D CT (Section A.1). We then detail the experimental setup, including implementation specifics and evaluation protocols, followed by complementary quantitative and qualitative results that further substantiate our findings (Section A.2).

A.1 Limitations and broader impacts

Limitations and future work. While our study provides the first systematic evaluation of calibration for multi-label abnormality classification in 3D CT, several limitations remain.

First, our controlled analysis focuses on models trained with binary cross-entropy loss under a unified supervision protocol. We anticipate that a comprehensive and controlled study of other training objectives such as focal (Mukhoti et al., 2020) and asymmetric (Ben-Baruch et al., 2021) losses, label-pair regularization (Cheng & Vasconcelos, 2024) and copula-based calibration (Panda & Garain, 2025), as well as alternative learning paradigms including self-supervised (Chen et al., 2023; Wald et al., 2025) and fine-grained vision-language pre-training (Shui et al., 2025; Wang et al., 2025), remains an important direction for future work.

Second, although we consider two public datasets including an internal and an external evaluation, our study is limited to chest CT and a fixed set of abnormalities, primarily due to the scarcity of large-scale annotated datasets for other anatomical regions (*e.g.* head, abdominal, whole-body) and 3D medical imaging modalities (*e.g.* PET/MRI). Evaluating calibration across additional pathologies, regions, modalities, and more diverse clinical settings is necessary to assess generalization and remains an important avenue for future work.

Third, our evaluation focuses on standard post-hoc methods and does not explore training-time calibration strategies or uncertainty-aware modeling, which may provide more robust solutions under distribution shift.

Fourth, while we identify several factors influencing calibration (*e.g.*, prevalence, anatomical system, pathology size, co-occurrence), these analyses remain correlational and do not establish causal mechanisms, motivating future work on structured and causal approaches to calibration.

Finally, our evaluation is conducted offline and does not include clinical validation; assessing how calibration affects downstream decision-making in real-world workflows remains an open challenge.

Broader impacts. From an impact perspective, this work highlights that reliable probability estimates cannot be assumed even for the most accurate models, which is critical for risk-sensitive applications such as medical diagnosis and triage. Our findings suggest that deploying models based solely on discriminative metrics may lead to misinformed decisions due to miscalibrated confidence. This work of academic research is not intended for direct clinical deployment. Transitioning these findings into real-world medical practice requires extensive prospective evaluation alongside expert radiologists to assess human-AI interaction and clinical outcomes.

A.2 Additional Experimental Details

A.2.1 Implementation

All visual encoders are trained using Adam with a learning rate of $1e-4$, and $(B_1, B_2) = (0.9, 0.99)$. We use a batch size of 4 and training runs for 200,000 iterations. Depending on the architecture, a single training run takes approximately 24-36 hours on a single NVIDIA RTX A6000. Model selection is performed via early stopping based on the validation loss. For evaluation, classification thresholds are optimized independently for each abnormality, maximizing the F1-Score on the validation set.

A.2.2 Meta-categories for volumetric analysis

Due to inconsistencies in pathology definitions, Table 7 details the mapping between CT-RATE abnormalities and ReXGroundingCT annotations used to define meta-categories for the volumetric size analysis. Importantly, we focus on the 231 unique held-out test subjects with available abnormality segmentation masks.

Meta-category	**CT-RATE**	**ReXGroundingCT**	**Prevalence**	**Average size (in pixels)**
Airspace / Opacification	Lung opacity Consolidation Atelectasis	Groundglass opacity Atelectasis, consolidation	136/231	174,370
Fibrotic / Interstitial	Pulmonary fibrotic sequela Interlobular septal thickening	Linear Septhal thickening Honeycombing	70/231	21,023
Airway disease	Peribronchial thickening Bronchiectasis	Bronchial wall thickening Bronchiectasis	34/231	38,265
Nodular lesions	Lung nodule	Micronodules Pulmonary nodes	113/231	44,015
Pleural disease	Pleural effusion	Pleural effusion	12/231	7,498
Parenchymal destruction	Emphysema	Emphysema	28/231	28,877

Table 7: Meta-category mapping between CT-RATE and ReXGroundingCT.

A.2.3 Spatial extent type

Another natural question is whether calibration performance is affected by the spatial extent of abnormalities. We distinguish between *focal* abnormalities, characterized by localized visual patterns, and *non-focal* ones, which exhibit more diffuse spatial distributions across the CT volume. To study this, we group abnormalities accordingly and evaluate calibration performance across models. As shown in Figure 10, non-focal abnormalities on RAD-ChestCT are slightly better calibrated than focal ones across all models, although differences remain small relative to variability across abnormalities. In terms of Brier score, CT-RATE shows a marginal advantage for non-focal cases, while results on RAD-ChestCT are nearly indistinguishable between the two categories, suggesting that spatial extent has a limited and inconsistent effect on calibration.

A.2.4 Signed error calibration and reliability analysis

By replacing the absolute deviation in ECE with a signed difference, the Signed Calibration Error (SCE) captures the direction of miscalibration, as follows:

$$\mathrm{SCE}_k = \sum_{b=1}^{B} \frac{|B_{b,k}|}{n} \left(\mathrm{conf}_k(B_{b,k}) - \mathrm{acc}_k(B_{b,k})\right). \tag{9}$$

Positive values indicate over-confidence, whereas negative values indicate under-confidence. As shown in Figure 13 offering abnormality-wise SCE and Table 11 reporting performances per model, we observe distinct model-specific patterns on CT-RATE: both ViT-3D and I3D exhibit over-confidence across most abnormalities, while CT-Net and CT-SSG show predominantly negative SCE, indicating systematic under-confidence. CT-Scroll displays a more heterogeneous behavior, although certain abnormalities (*e.g.*, lung nodules, pulmonary fibrotic sequela) consistently induce strong over-confidence across all architectures, suggesting shared sources of bias. The results in Figure 15 show that these trends are corroborated by reliability diagrams, which show systematic deviations above or below the diagonal depending on the model and abnormality.

Under distribution shift on RAD-ChestCT, these model-specific patterns largely diminish. Both SCE and reliability diagrams, reported in Figure 16, indicate more uniform behavior across architectures, with several abnormalities shifting toward stronger under-confidence. This suggests that distribution shift affects not

only the magnitude of calibration error but also its direction, leading to more homogeneous miscalibration patterns across models.

A.2.5 Effect of binning on ECE estimation

We further analyze the sensitivity of ECE to the number of discretization bins, varying the bin count from 5 to 100. As shown in Figure 11, finer binning leads to higher estimated calibration error, reflecting increased resolution in detecting miscalibration. However, the relative ranking between models remains stable across all bin configurations, indicating that our comparative conclusions are robust to the choice of binning strategy.

A.2.6 Micro-level calibration analysis

Table 10 reports micro-averaged calibration metrics (ECE, ACE, and MCE) on CT-RATE. Results are consistent with the macro-level analysis: in-domain, attention-based architectures (ViT-3D and CT-Scroll) achieve the strongest calibration performance, particularly in terms of ECE and ACE. Under distribution shift, I3D becomes comparatively more robust, achieving the best calibration among models. This confirms that the observed trends are stable across aggregation levels and not driven by macro-averaging effects.

A.2.7 Qualitative calibration analysis

Figure 12 presents representative qualitative examples illustrating calibration behavior across datasets. On CT-RATE, models can produce well-calibrated predictions for visually salient abnormalities such as pleural effusion, while exhibiting under-confidence for more diffuse or subtle patterns such as interlobular septal thickening. A similar trend is observed on RAD-ChestCT, where well-calibrated predictions remain possible (e.g., lung opacity), but under-confidence persists for certain abnormalities such as atelectasis. These examples highlight that calibration is not uniform across pathologies and is influenced by visual characteristics, with more subtle or ambiguous patterns tending to induce under-confident predictions.

A.2.8 Abnormality-wise calibration performance

Tables 12 and 13 report abnormality-level ECE on CT-RATE and RAD-ChestCT, respectively. Across datasets, architectures incorporating attention mechanisms (*e.g.*, ViT-3D and CT-Scroll) consistently achieve lower calibration error on a majority of abnormalities, indicating more reliable probability estimates at the fine-grained level. This trend is particularly stable under distribution shift, where these models maintain comparatively better calibration on several abnormalities despite the overall degradation observed on RAD-ChestCT. In contrast, convolutional and graph-based models exhibit larger variability across abnormalities, reflecting less consistent calibration behavior. These results highlight that calibration differences between architectures are not driven by a small subset of findings, but instead persist across a broad range of abnormalities, reinforcing the importance of modeling long-range dependencies for reliable uncertainty estimation.

A.2.9 Joint calibration evaluation

Standard marginal calibration metrics (ECE, ACE, KCE) estimate calibration error per label, treating each abnormality independently. In CT scans, however, abnormalities co-occur in structured ways, and a model can be well-calibrated marginally yet miscalibrated in its joint predictions, a critical limitation for multi-abnormality analysis where joint uncertainty matters (Draelos et al., 2021).

To address this, canonical calibration evaluates the full joint prediction vector (Bröcker, 2008; Kull & Flach, 2015). Formally, the L_p Canonical Calibration Error ($\mathrm{CCE_p}$) is defined as:

$$\mathrm{CCE_p} = \left(\mathbb{E}\left[||\mathbb{E}[y|h(x)] - h(x)||_p^p\right]\right)^{\frac{1}{p}}, \tag{10}$$

where $y \in \{0,1\}^M$ is the ground-truth vector, $h(x) : \mathbb{R}^{S\times H\times W} \to [0,1]^M$ is the model's joint prediction, and $M \in \mathbb{R}^+$ is the number of abnormalities.

The Copula factorization, which makes it joint rather than marginal, is defined such that:

$$F_{y,h(x)}\left(z_1,\ldots,z_M,p_1,\ldots,p_M\right) = C\Big(F_1(z_1),\ldots F_M(z_M),F_1(p_1),\ldots,F_M(p_M);\Sigma\Big)\,, \tag{11}$$

where $C:[0,1]^{2M}\rightarrow[0,1]$ is the Gaussian copula with correlation matrix $\Sigma\in[-1,+1]^{(2M,2M)}$ (symmetric positive definite), $\{F_j(z_j)\}_{j=1}^{M}$ are the smoothed label marginals, and $\{F_j(p_j)\}_{j=1}^{M}$ are the empirical marginal cumulative distributions functions of the predicted probabilities. Exploration of other copulas is left for future work. In our experiments, the L_2 empirical estimator denoted as CCE_2 is defined as follows:

$$\mathrm{CCE}_2 = \frac{1}{N}\sum_{l=1}^{N}||\hat{\mathbb{E}}[z|h(x^{(l)})]-h(x^{(l)})||_2^2\,, \tag{12}$$

where N is the number of samples, and $\hat{\mathbb{E}}[z|h(x^{(l)})]$ is estimated via the Gaussian copula density as:

$$\hat{\mathbb{E}}[z_j|h(x)] = \frac{\sum_{i=1}^{N} z_j^i\; c\Big(\hat{F}_j^{(i)}(z_j^{(i)}),\hat{F}_1(p_1),\ldots,\hat{F}_M(p_M);\hat{\Sigma}\Big)}{\sum_{i=1}^{N} c\Big(\hat{F}_j^{(i)}(z_j^{(i)}),\hat{F}_1(p_1),\ldots,\hat{F}_M(p_M);\hat{\Sigma}\Big)} \quad \forall\; j\in\{1,\ldots,M\}\,, \tag{13}$$

where $c\in[0,1]^{M+1}$ is a $(M+1)$-dimensional margin of C obtained by marginalizing over the M labels except z_j, keeping $(z_j,p_1,\ldots,p_M)$, and $\hat{\Sigma}$ is a $(M+1)\times(M+1)$ sub-matrix of Σ.

Table 8 reports the canonical calibration error (CCE_2) on the internal CT-RATE test set across CT visual encoders, alongside ECE for reference. Consistent with the marginal calibration analysis, ViT-3D achieves the lowest CCE_2, indicating the strongest canonical calibration in-domain. Paired t-tests across folds show that the improvements of ViT-3D are statistically significant compared with all other encoders ($p<0.05$), except for I3D ($p=0.506$). Relative to CT-SSG and CT-Net, ViT-3D reduces CCE_2 by 28.9% and 30.2%, respectively. These findings suggest that the superior calibration properties of ViT-3D extend beyond pathology-wise marginal estimates and remain evident when accounting for label dependencies.

Table 8: Comparison of ECE and CCE_2 across CT visual encoders on CT-RATE.

Method	**ECE**	**CCE_2**
Random	0.359±0.002	1.858±0.008
ViT-3D	**0.029**±0.002	**0.316**±0.022
I3D	0.040±0.009	<u>0.331</u>±0.035
CT-Net	0.054±0.012	0.453±0.070
CT-Scroll	<u>0.031</u>±0.005	0.355±0.014
CT-SSG	0.041±0.006	0.445±0.032

To assess cross-dataset generalization, Table 9 reports CCE_2 on the external RAD-ChestCT dataset for both CT visual encoders and CT foundation models. Under distribution shift, ViT-3D, I3D, and CT-Scroll remain the most competitive architectures, mirroring the trends observed with ECE. However, pairwise comparisons involving I3D yield p-values greater than 0.05 for all methods except CT-SSG, suggesting that the observed ranking differences should be interpreted as empirical trends rather than strong statistical evidence. Among CT foundation models, the ranking induced by CCE_2 is identical to that obtained with ECE, and all pairwise comparisons are statistically significant ($p<0.05$). In particular, Merlin achieves the lowest canonical calibration error, followed by CT-FM, whereas weakly supervised foundation models exhibit the highest miscalibration. These findings further indicate that large-scale pre-training alone

Table 9: ECE and CCE_2 results on RAD-ChestCT. For each setup, **best results** are in bold, and <u>second-best</u> are underlined.

Setup	Method	**ECE**	**CCE_2**
Random	-	0.336±0.001	1.642±0.005
CT Visual Encoders	ViT-3D	**0.133**±0.005	<u>0.765</u>±0.010
	I3D	<u>0.134</u>±0.007	**0.757**±0.020
	CT-Net	0.141±0.009	0.818±0.048
	CT-Scroll	<u>0.134</u>±0.008	0.769±0.030
	CT-SSG	0.147±0.004	0.914±0.021
CT Foundation Models	CT-FM	<u>0.129</u>±0.001	<u>0.724</u>±0.002
	CT-CLIP	0.148±0.001	0.868±0.002
	Merlin	**0.128**±0.002	**0.704**±0.002
	COLIPRI	0.132±0.003	0.777±0.003

does not guarantee reliable probability estimates and that the choice of pre-training paradigm substantially influences both marginal and canonical calibration.

Finally, ECE and CCE_2 exhibit strong positive correlations across models on both CT-RATE ($\rho = 0.77 \pm 0.11$) and RAD-ChestCT ($\rho = 0.89 \pm 0.07$). This finding suggests that, in our setting, pathology-wise marginal calibration provides a reasonable proxy for dependency-aware canonical calibration. Nevertheless, the small discrepancies in model rankings indicate that accounting for label interdependencies may reveal complementary aspects of predictive reliability that are not fully captured by marginal calibration metrics.

Figure 8: Paired t-test p-values across cross-validation folds for all pairs of CT vision encoders for (a) ECE, (b) ACE, (c) KCE, and (d) Brier score. Results are reported for in-domain evaluation (top, CT-RATE) and under distribution shift (bottom, RAD-ChestCT). Asterisks (*) denote $p < 0.05$.

Table 10: Micro-level calibration evaluation of CT visual encoders on the multi-label abnormality classification task on CT-RATE and RAD-ChestCT. **Best results** are in bold, and <u>second best</u> are underlined.

	CT-RATE			RAD-ChestCT		
Method	ECE	ACE	KCE (10^{-1})	ECE	ACE	KCE (10^{-1})
Random	0.348±0.003	0.348±0.003	0.570±0.007	0.282±0.002	0.282±0.002	0.352±0.008
ViT-3D	**0.012**±0.003	**0.012**±0.002	**0.001**±0.001	0.058±0.007	0.057±0.008	0.023±0.007
I3D	0.025±0.014	0.025±0.014	0.005±0.004	**0.043**±0.005	**0.043**±0.006	**0.010**±0.005
CT-Net	0.047±0.012	0.047±0.012	0.012±0.006	0.070±0.013	0.069±0.013	0.036±0.014
CT-Scroll	<u>0.016</u>±0.008	<u>0.017</u>±0.007	<u>0.002</u>±0.001	<u>0.050</u>±0.010	<u>0.050</u>±0.009	<u>0.016</u>±0.008
CT-SSG	0.027±0.006	0.026±0.006	0.004±0.002	0.076±0.006	0.076±0.006	0.057±0.013

Figure 9: Abnormality-wise calibration degradation under distribution shift, measured as the difference in Expected Calibration Error between RAD-ChestCT and CT-RATE ($\Delta\text{ECE} = \text{ECE}_{\text{RAD-ChestCT}} - \text{ECE}_{\text{CT-RATE}}$), across CT visual encoders. Results are reported for the 15 abnormalities shared between datasets, excluding *mosaic attenuation pattern* and *calcifications*. Positive values indicate increased miscalibration on the external RAD-ChestCT dataset. Abnormalities are ordered by increasing calibration degradation.

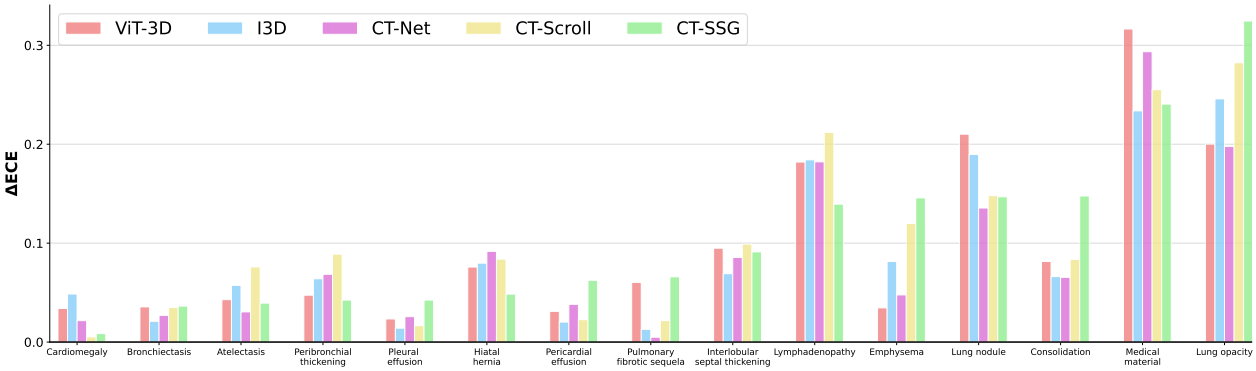


Figure 10: Comparison of ECE (left) and Brier (right) performances across spatial extent type. Calibration performance is largely consistent across focal and diffuse abnormalities, with only a slight improvement for non-focal findings on RAD-ChestCT in terms of ECE.

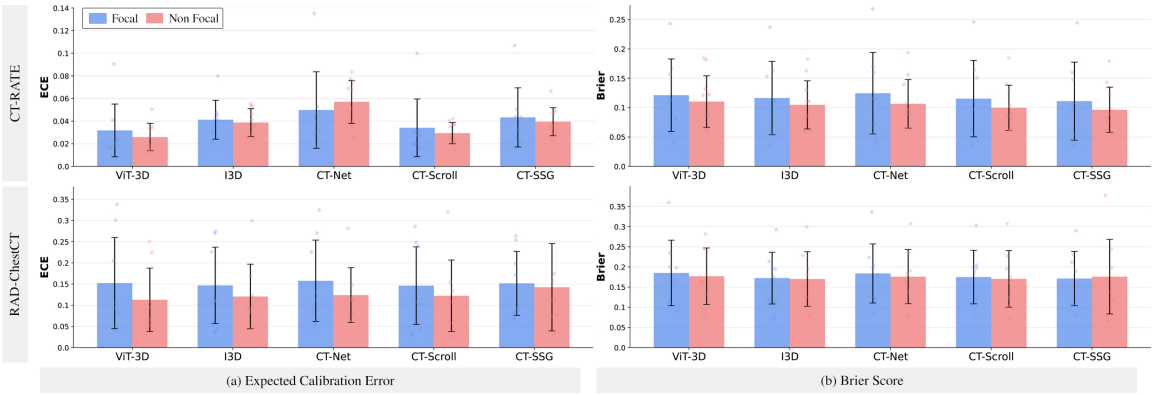


Figure 11: Sensitivity to the number of bins. While ECE increases with the number of bins, the relative ranking of models remains stable across all bin settings, both for CT-RATE and RAD-ChestCT.

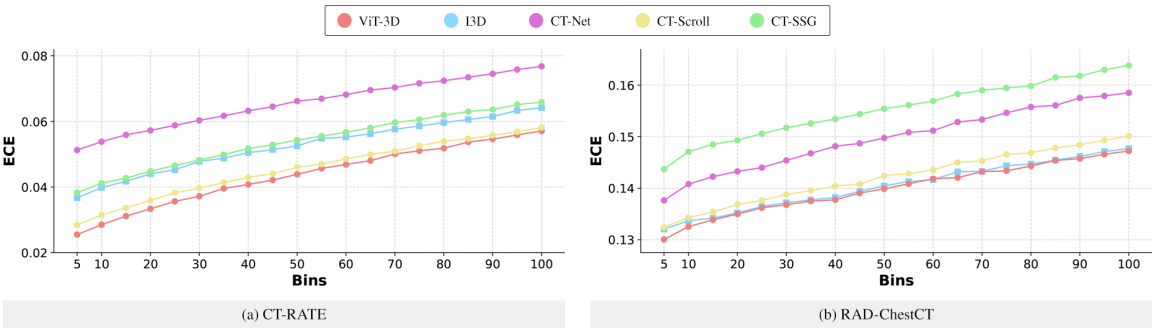


Figure 12: Representative axial, coronal and sagittal slices, with predicted probabilities. Manually drawn masks (in red) are shown for visualization. Both for CT-RATE (top) and RAD-ChestCT (bottom), examples illustrate well-calibrated (left) and under-confident (right) predictions.

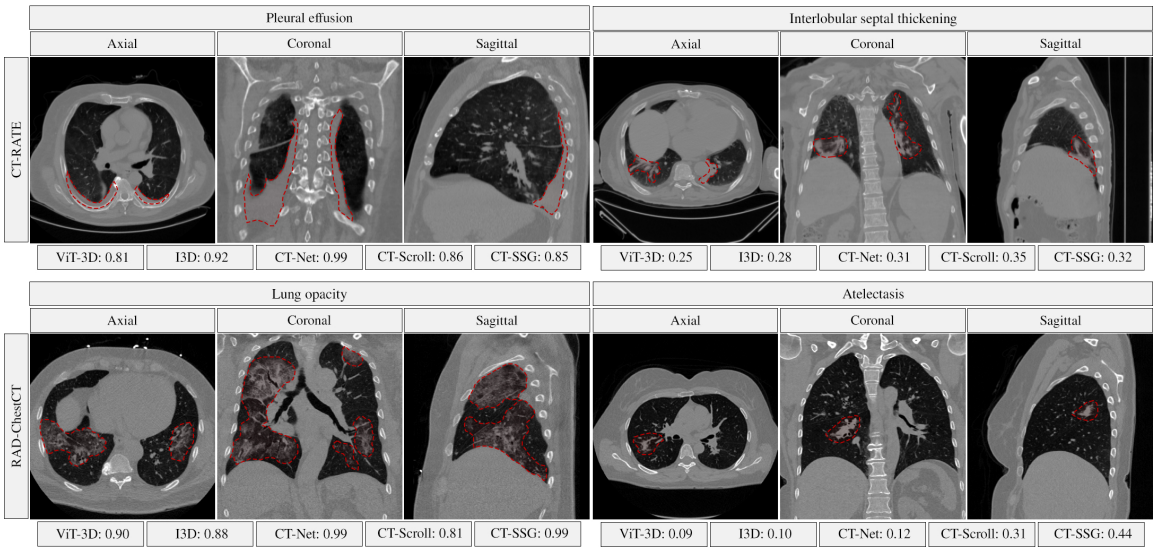
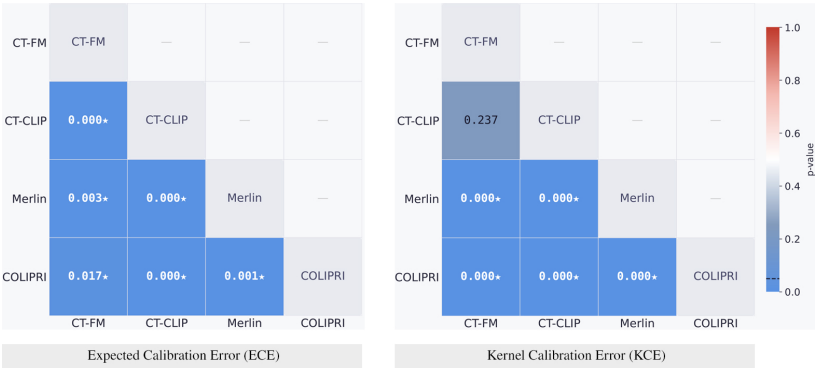


Figure 13: Abnormality-wise SCE results on CT-RATE (top) and RAD-ChestCT (bottom). Positive values indicate over-confident predictions, while negative ones are associated with under-confidence.

Table 11: Signed Error Calibration (SCE) evaluation of CT visual encoders on the abnormality classification task. *Under* and *Over* denote the counts across abnormalities and folds, exhibiting under- and over-confidence based on SCE. Underlined values indicate the dominant trend for each model.

	CT-RATE			**RAD-ChestCT**		
Method	SCE	Under	Over	SCE	Under	Over
Random	+0.313±0.002	0 / 90	90 / 90	+0.178±002	19 / 80	61 / 80
ViT-3D	+0.010±0.001	31 / 90	59 / 90	-0.055±0.010	45 / 80	35 / 80
I3D	+0.018±0.020	27 / 90	63 / 90	-0.032±0.015	43 / 80	37 / 80
CT-Net	-0.003±0.005	60 / 90	30 / 90	-0.036±0.009	40 / 80	40 / 80
CT-Scroll	+0.007±0.009	42 / 90	48 / 90	-0.028±0.01	40 / 80	40 / 80
CT-SSG	-0.005±0.005	58 / 90	32 / 90	-0.051±0.009	45 / 80	35 / 80

Figure 14: Paired t-test p-values across cross-validation folds for all pairs of CT foundation models for (a) ECE, (b) ACE, (c) KCE, and (d) Brier score. Results are reported for RAD-ChestCT only, as CT-CLIP and COLIPRI pre-training data partially overlap with our validation sets. Asterisks (*) denote $p < 0.05$.



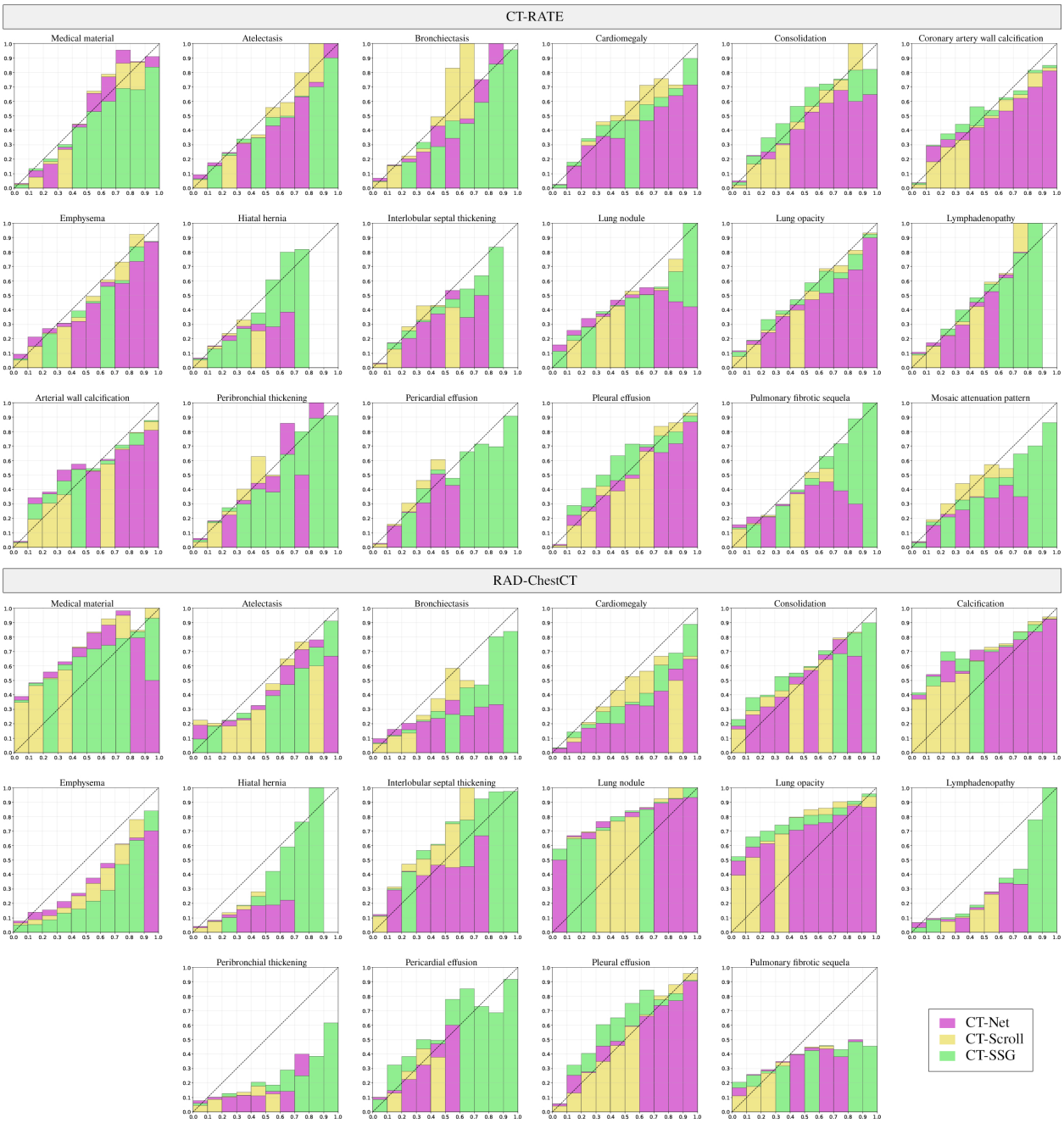


Figure 15: Abnormality-wise reliability diagrams for 2.5D methods across CT-RATE (top) and RAD-ChestCT (bottom). We observe calibration variability across i) abnormalities, with certain findings (*e.g.* lung opacity) consistently well-calibrated in-domain; ii) models certain encoders demonstrating stronger in-domain calibration on several pathologies (*e.g.* CT-SSG for pulmonary fibrotic sequela and peribronchial thickening); and iii) datasets, where well-calibrated in-domain predictions may shift toward over- or under-confidence under distribution shift (*e.g.*, lung nodule, emphysema).

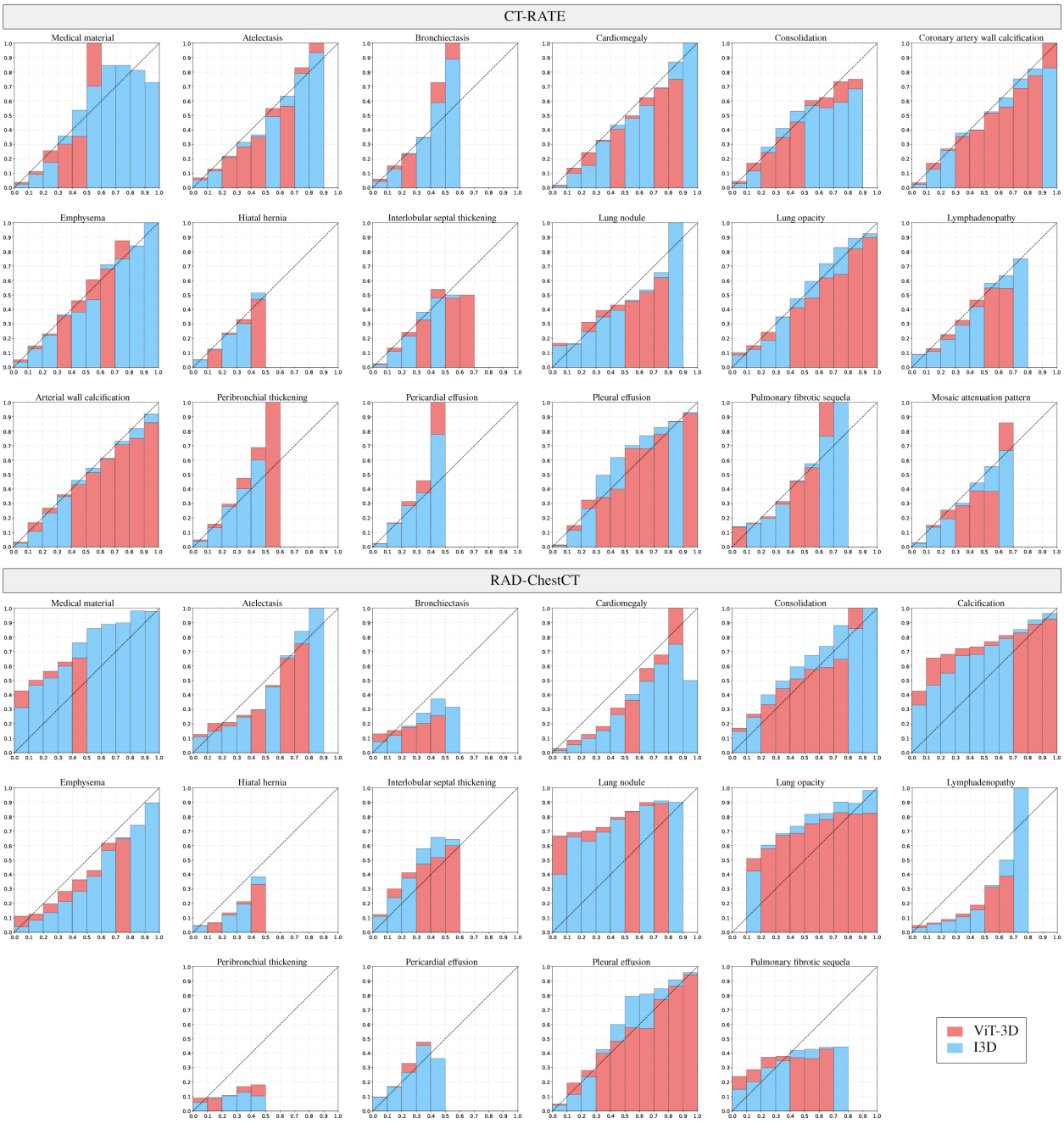


Figure 16: Abnormality-wise reliability diagrams for 3D methods across CT-RATE (top) and RAD-ChestCT (bottom). Calibration varies across (i) abnormalities, with arterial wall calcification showing consistently reliable calibration across encoders in-domain; (ii) models, where I3D demonstrates slightly better in-domain calibration than ViT-3D for emphysema; and (iii) datasets, where distribution shift induces changes in confidence, including over-confidence for consolidation and under-confidence for peribronchial thickening.

Table 12: Abnormality-wise ECE on the CT-RATE test set, averaged across folds. We also include the prevalence ratio of each abnormality in the test set, reported from Hamamci et al. Hamamci et al. (2025). ViT-3D and CT-Scroll consistently achieve the lowest ECE across most categories.

Abnormality	**Ratio**	**ViT-3D**	**I3D**	**CT-Net**	**CT-Scroll**	**CT-SSG**
Medical material	0.103	0.021	0.041	0.031	0.031	0.023
Arterial wall calcification	0.285	0.038	0.056	0.075	0.040	0.046
Cardiomegaly	0.107	0.013	0.036	0.040	0.022	0.028
Pericardial effusion	0.074	0.019	0.022	0.018	0.020	0.020
Coronary artery wall calc.	0.252	0.033	0.054	0.068	0.037	0.048
Hiatal hernia	0.137	0.018	0.030	0.068	0.019	0.044
Lymphadenopathy	0.260	0.023	0.042	0.044	0.028	0.033
Emphysema	0.197	0.025	0.041	0.069	0.035	0.029
Atelectasis	0.235	0.041	0.048	0.053	0.031	0.044
Lung nodule	0.448	0.091	0.080	0.135	0.100	0.107
Lung opacity	0.390	0.050	0.053	0.084	0.037	0.067
Pulmonary fibrotic sequela	0.273	0.035	0.046	0.077	0.042	0.049
Pleural effusion	0.124	0.017	0.024	0.032	0.016	0.025
Mosaic attenuation pattern	0.083	0.015	0.025	0.036	0.024	0.033
Peribronchial thick.	0.117	0.019	0.028	0.044	0.019	0.035
Consolidation	0.191	0.025	0.045	0.048	0.028	0.051
Bronchiectasis	0.109	0.017	0.024	0.051	0.023	0.036
Interlobular septal thickening	0.082	0.014	0.024	0.026	0.015	0.023
Mean	**0.193**	**0.029**	**0.040**	**0.053**	**0.031**	**0.041**

Table 13: Abnormality-wise ECE on RAD-ChestCT, averaged across folds. ViT-3D, I3D and CT-Scroll achieve the strongest generalization performance across most abnormalities.

Abnormality	**Ratio**	**ViT-3D**	**I3D**	**CT-Net**	**CT-Scroll**	**CT-SSG**
Medical material	0.327	0.338	0.274	0.324	0.286	0.264
Calcification	0.706	0.224	0.169	0.148	0.133	0.157
Cardiomegaly	0.109	0.047	0.084	0.061	0.027	0.037
Pericardial effusion	0.155	0.050	0.042	0.056	0.042	0.083
Hiatal hernia	0.117	0.094	0.109	0.130	0.103	0.092
Lymphadenopathy	0.165	0.205	0.226	0.226	0.240	0.172
Emphysema	0.273	0.060	0.123	0.116	0.155	0.175
Atelectasis	0.298	0.084	0.105	0.083	0.107	0.083
Lung nodule	0.802	0.301	0.269	0.271	0.248	0.254
Lung opacity	0.539	0.250	0.299	0.281	0.319	0.391
Pulmonary fibrotic sequela	0.132	0.095	0.059	0.082	0.064	0.115
Pleural effusion	0.200	0.040	0.038	0.058	0.032	0.067
Peribronchial thickening	0.085	0.066	0.092	0.113	0.108	0.078
Consolidation	0.139	0.106	0.111	0.113	0.111	0.198
Bronchiectasis	0.158	0.052	0.045	0.078	0.058	0.073
Interlobular septal thickening	0.069	0.108	0.093	0.112	0.114	0.114
Mean	**0.267**	**0.133**	**0.134**	**0.141**	**0.134**	**0.147**

Figure 17: Abnormality-wise reliability diagrams for CT foundation models on RAD-ChestCT. Calibration varies across pathologies, with lung nodule and calcification demonstrating consistently over-confidence across CT foundation models. Lymphadenopathy, however, illustrates an example of under-confidence under distribution shift for all models, highlighting similar observation patterns for CT visual encoders.

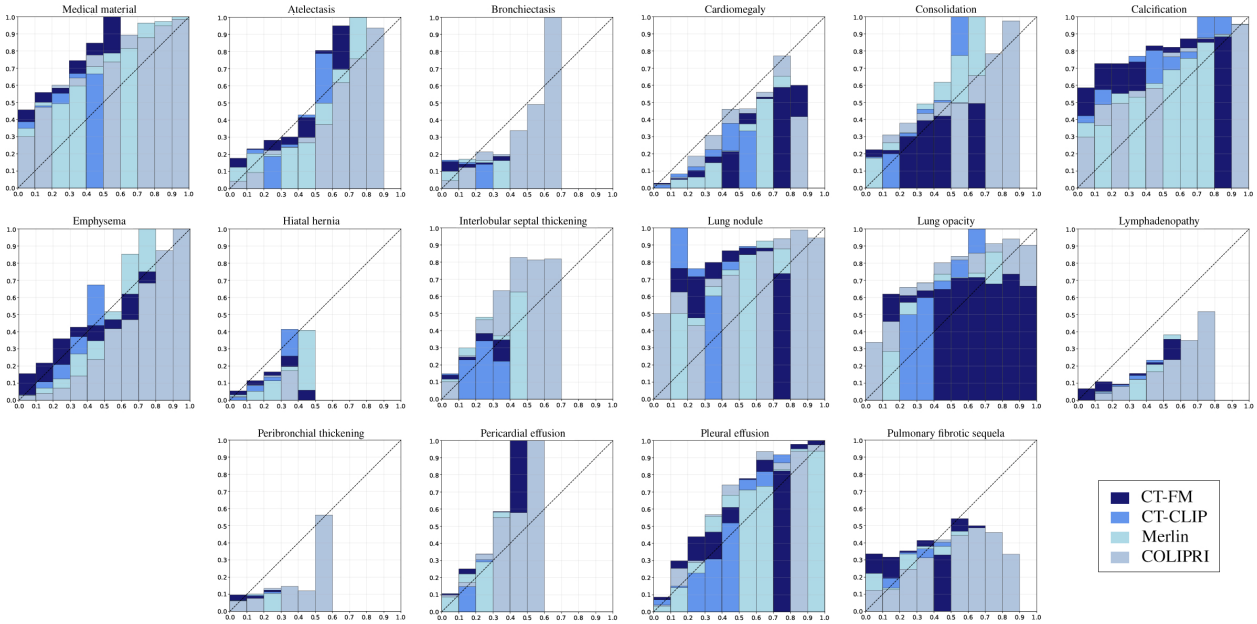


Table 14: Abnormality-wise ECE on RAD-ChestCT for CT foundation models, averaged across folds.

Abnormality	**Ratio**	**CT-FM**	**CT-CLIP**	**Merlin**	**COLIPRI**
Medical material	0.327	0.321	0.403	0.283	0.272
Calcification	0.706	0.324	0.380	0.119	0.138
Cardiomegaly	0.109	0.058	0.057	0.103	0.031
Pericardial effusion	0.155	0.035	0.075	0.045	0.058
Hiatal hernia	0.117	0.093	0.039	0.120	0.093
Lymphadenopathy	0.165	0.194	0.158	0.208	0.235
Emphysema	0.273	0.048	0.082	0.087	0.144
Atelectasis	0.298	0.065	0.064	0.108	0.085
Lung nodule	0.802	0.331	0.402	0.279	0.235
Lung opacity	0.539	0.251	0.189	0.291	0.325
Pulmonary fibrotic sequela	0.132	0.039	0.117	0.051	0.078
Pleural effusion	0.200	0.044	0.091	0.029	0.045
Peribronchial thickening	0.085	0.071	0.051	0.064	0.081
Consolidation	0.139	0.075	0.077	0.119	0.129
Bronchiectasis	0.158	0.036	0.076	0.036	0.046
Interlobular septal thickening	0.069	0.089	0.101	0.109	0.110
Mean	**0.267**	**0.129**	**0.148**	**0.128**	**0.132**