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

ANDREW WANG[1], JIASHUO ZHANG[2], MICHAEL OBERST[2]

[1]Brown University, [2]Johns Hopkins University

Introduction

Computer Vision models achieve impressive aggregate evaluation scores on public Chest X-Ray (CXR) datasets. However, the reported strong average-case performance of these models do not necessarily reflect their actual utility when used in heterogeneous clinical settings.

Clinicians do not interpret and diagnose images in isolation, but within the context of the patient information and histories. We model this contextual heterogeneity to simulate clinically relevant use cases by using prior discharge summaries occurring before each CXR, serving as the foundation of our evaluation framework.

Overview

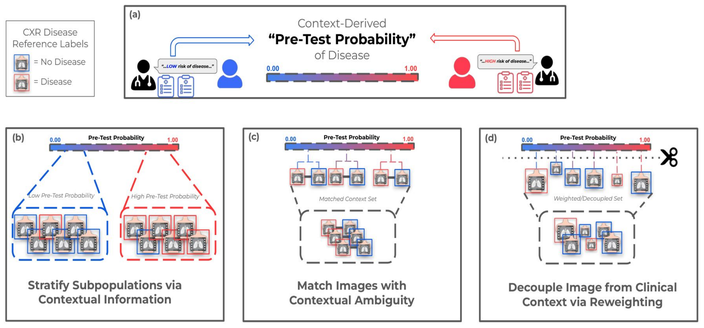

Figure 1: Overview of our evaluation framework

(a): We use clinical context contained within discharge summaries to derive a ***pre-test probability*** estimate of disease risk, given knowledge obtained before a CXR is ordered, and show that the binarized estimate accurately predicts future disease labels

(b): We identify subpopulations of the evaluation using this pre-test probability. We then evaluate performance on the resulting disjoint groups.

(c): We create an evaluation set by matching positive/negative image pairs with similar pre-test probabilities.

(d): We statistically decouple the image label from the contextual information via reweighting the evaluation set.

Results

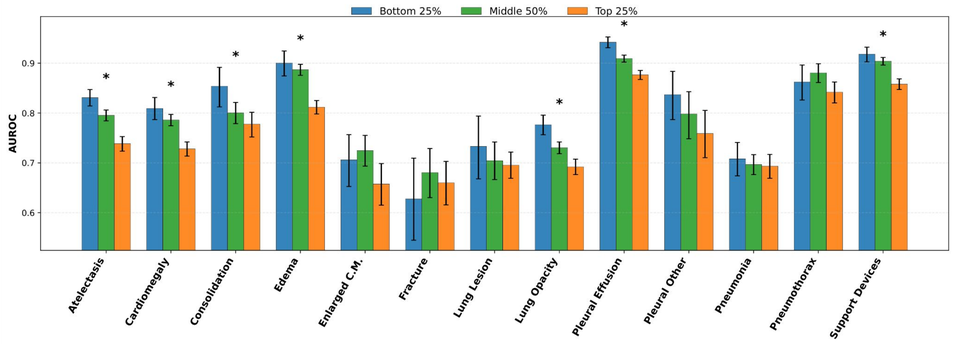

Figure 4: Comparison of AUROC of DenseNet121 model across subpopulations stratified by pre-test probability of the CXR label.

Labels marked with an asterisk indicate a statistically significant difference in AUROC between the Bottom 25% and Top 25% groups. We observe that vision model performance generally degrades across most labels as the prior probability of the label increases.

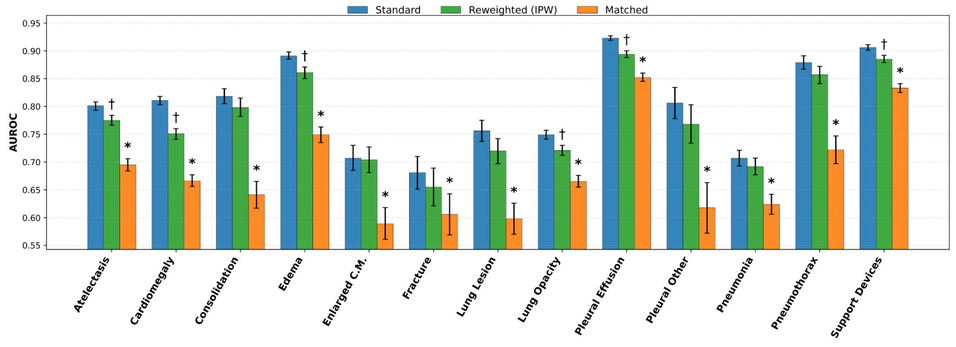

Figure 5: Comparison of AUROC of DenseNet121 model across Standard, Reweighted, and Matched Neighbor settings.

Labels marked with a dagger and an asterisk indicate statistically significant differences in the Reweighted (IPW) and Matched settings, respectively.. We observe that performance consistently drops across the Matched and Reweighted sets across several labels, illustrating vision model degradation in clinically significant use cases.

Methods

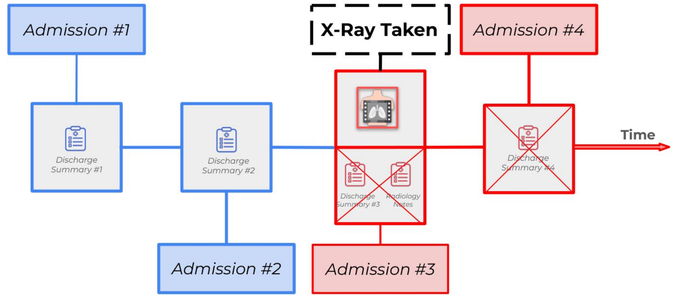

Figure 2: Visualization of "prior clinical context" used in this paper.

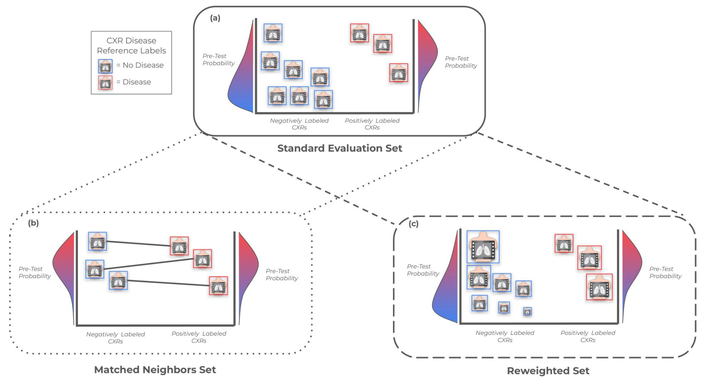

Figure 3: Evaluation Set Construction

a) Original evaluation set: Negatively and positively labeled CXRs in the original evaluation set with their associated distributions of pre-test probabilities. All images and their associated pre-test probabilities are equally weighted.

(b) Matched Neighbors set: We create a subset of the evaluation data by 1:1 matching negatively and positively labeled images with similar pre-test probabilities

(c) Reweighted set: After reweighting of images and their associated pre-test probabilities, the weighted distribution of pre-test probabilities become comparable across the positive and negative reweighted populations, creating an evaluation set where the disease label is statistically independent of the context

Acknowledgements

The authors thank the Oberst Lab at Johns Hopkins and the Approximately Correct Machine Intelligence (ACMI) lab at Carnegie Mellon University broadly for providing compute resources and mentorship in the development of the project.

Conclusions

Computer vision model performance is uneven across patient subpopulations defined by contextual risk: model excel at low-risk triage but degrade significantly on high-risk patients; exactly where clinicians need the most decision support.

Models struggle to resolve diagnostic uncertainty using only visual evidence: When diagnosing patients with identical clinical contextual risk profiles but differing disease statuses, performance drops significantly.

Contextual information is correlated with performance: Across several disease labels, performance drops when the statistical link between context and the label is severed. This suggests that diagnostic models do not strictly use visual signals for diagnosis, and that this process may be confounded by contextual history.