SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Discordance Analysis Between Automated Cardiothoracic Ratio Measurements and NLP-Extracted Cardiomegaly Labels

Frank Li, PhD, Emory University; Abdulhameed Dere, MBBS; Abdulquddus Ajibade, MBBS; Theodorus Dapamede, MD, PhD; Mohammadreza Chavoshi, MD; Janice M. Newsome, MD; Hari Trivedi, MD; Judy W. Gichoya, MD, MS, FSIIM

Introduction/Background
This study quantifies discordance between objective cardiothoracic ratio (CTR) measurements and cardiomegaly labels extracted from radiology reports using both a traditional NLP (CheXpert) and a large language model (Llama 3.1 8b). We aim to identify potential sources of label noise that could propagate algorithmic biases in downstream AI diagnostic models.

Methods/Intervention
We analyzed 246,904 PA chest radiographs from our institution using CheXmask for heart and lung segmentation, excluding images with unbalanced left and right lungs. CTR was calculated as the ratio of heart width to thoracic diameter, with CTR > 0.55 defining cardiomegaly. Cardiomegaly labels were independently extracted from corresponding reports using CheXpert labeler and Llama 3.1 8b, with concordance evaluated through confusion matrices.

Results/Outcome
CTR vs. CheXpert comparison showed 81% agreement on normal heart size and 4.2% agreement on cardiomegaly presence. However, 13% of cases with high CTR were not identified as cardiomegaly by CheXpert, while 1.4% were labeled as cardiomegaly despite normal CTR. The LLM showed better CTR alignment with just a 10% discordance in high CTR cases, and higher agreement on cardiomegaly presence at 7.4%. Both NLP tools demonstrated a significant tendency to miss cardiomegaly when CTR was high, though the LLM showed less propensity for this error pattern. Despite these differences, CheXpert and LLM showed strong overall concordance (94.5%), though LLM identified 5.2% of cases as cardiomegaly that CheXpert did not, while the reverse scenario was extremely rare (0.089%).

Conclusion
Significant discordance between CTR measurements and NLP-extracted labels highlights potential label noise in AI training data. While LLMs show improved alignment with objective measurements, systematic differences observed indicate researchers should exercise caution when using proxy labels from reports as reference standards.

Statement of Impact
AI systems trained on automatically extracted labels may perpetuate systematic biases, potentially leading to improper clinical decision-making. Improved concordance between LLM-derived labels and objective measurements represents a promising advancement for automated report interpretation, though label noise persists in both methods.

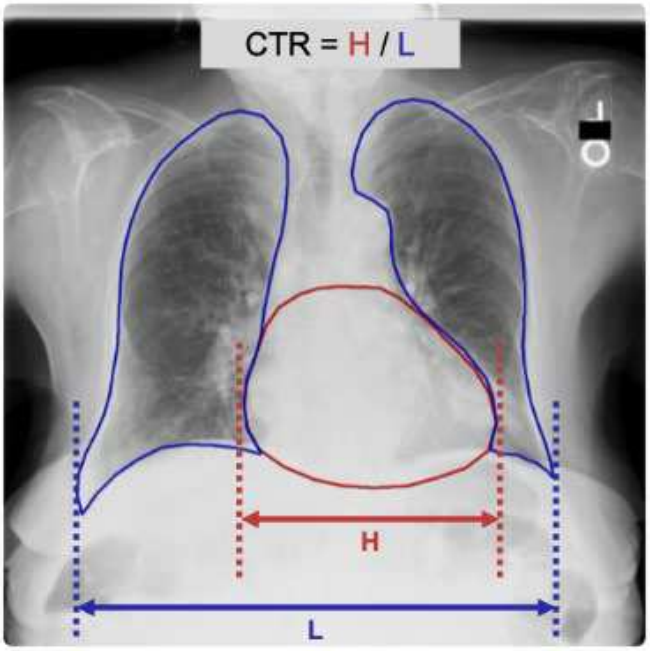

Figure 1. CTR was calculated as the ratio of the maximal horizontal width of the cardiac silhouette (H) to the maximal internal thoracic diameter (L) measured between the outer borders of the lungs, as derived from the heart and lung segmentation masks.

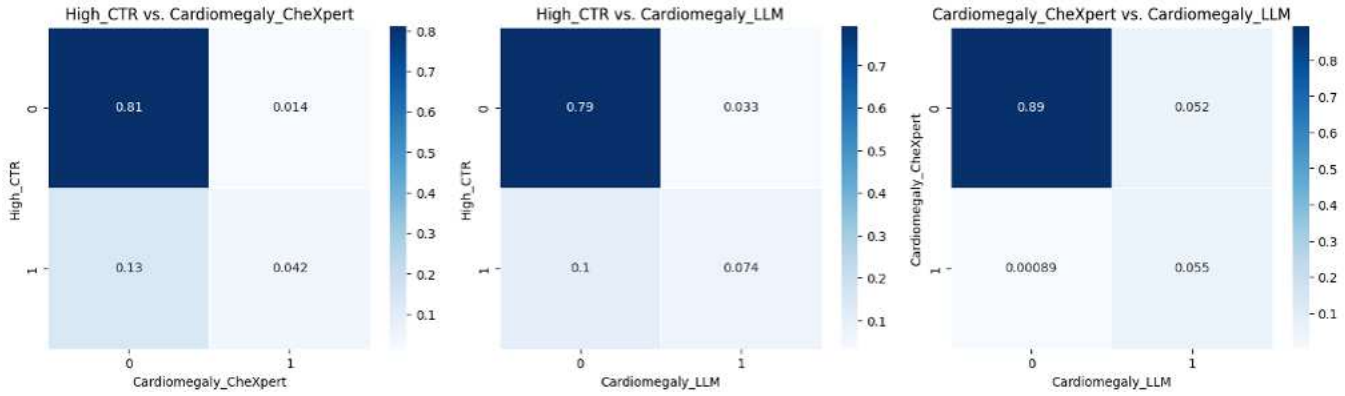

Figure 2. Confusion matrices compare three cardiomegaly labeling methods. Left: Automated high CTR (>0.55) vs. CheXpert NLP label, showing 81% negative agreement but 13% high CTR cases not classified as cardiomegaly by CheXpert. Center: Automated high CTR vs. LLM label, showing improved agreement with 10% discordance in high CTR cases and higher positive concordance (7.4%). Right: CheXpert vs. LLM label, showing strong overall agreement (94.5%) with LLM identifying 5.2% of cardiomegaly cases missed by CheXpert, while CheXpert missed only 0.089%.

Keywords

Natural Language Processing; Cardiothoracic Ratio; Label Noise