EPH6724

Evaluating A Deep Learning Model For Classifying Pediatric Pneumonia In Israel

Dong Wang, Ph.D.[1]; Meghan White, PharmD[1]; Sivan Gazit, MD, MA[2]; Boshu Ru, Ph.D.[1]; Jessica Weaver[1], Tal Patalon, MD, LLB,MBA[2]; Craig Roberts, MBA, PharmD[1]

[1]Merck & Co., Inc., Rahway, NJ, USA;
[2]Maccabi Healthcare Services, Tel Aviv, Israel

Introduction

Primary Endpoint Pneumonia (PEP) is a WHO-defined radiological measure used in vaccine studies for its high specificity for bacterial pneumonia, particularly *Streptococcus pneumoniae*. Retrospective studies using ICD codes for all-cause pneumonia are less time consuming and have high sensitivity for bacterial pneumonia but lack specificity. Although the WHO method has proven valuable in multiple studies, human interpretation remains time-consuming and subject to inter-reader disagreement, often necessitating adjudication panels.

Advances in deep learning have made automated chest X-ray (CXR) interpretation practical. Chen et al.[1] developed a transfer-learning–based model to classify pediatric CXRs using WHO criteria for PEP, demonstrating strong performance on both internal and independent test sets. Subsequent validation with a Hong Kong CXR dataset confirmed the model's real-world applicability [2].This study aims to further validate that model by assessing its ability to identify PEP in pediatric CXRs in Israel, thereby evaluating generalizability across healthcare systems and populations.

Methods

Study Design and Study Sample

This cross-sectional validation study utilized a deep learning model designed to classify CXRs for PEP based on WHO criteria. Pediatric CXR images were sourced from Maccabi healthcare system, with classifications determined by three board-certified radiologist. The dataset comprised 537 anonymized CXR images selected from children <5 years of age in Israel (2004-2018). The same dataset was then subjected to prediction by the deep learning model, which was pre-trained on adult CXR datasets and fine-tuned on pediatric data. The level of agreement among each rater and between each rater and the model was evaluated as well.

In a secondary analysis, we evaluated the effect of an autosegmentation algorithm on model performance, since autosegmentation previously improved results in our validation work in Hong Kong. The autosegmentation algorithm used a U-Net–inspired convolutional neural network. It retains the U-shaped encoder-decoder segmentation design of the original U-Net but reduces feature dimensionality. The model first segments the pulmonary region and computes a bounding rectangle from the mask (with an added margin) to crop the image[2].

Statistical analysis

The deep-learning model's classification performance was compared with the radiologists' consensus (the gold standard) using standard metrics: accuracy, F1 score, precision, sensitivity, specificity, area under the receiver operating characteristic curve(AUROC), and the area under the precision and recall curve (AUPRC). Performance was evaluated on the full dataset and within subgroups defined by age, sex, inter-observer agreement, ICU admission, and mortality. Inter-observer agreement was measured with Cohen's kappa, calculated between individual radiologists and between each radiologist and the model. The metrics were computed on both the original images and the auto-cropped images.

Results

Among the 537 CXRs, radiologists classified 78 as PEP (14.5%), while the model classified 25 as PEP (4.7%). The model achieved an accuracy of 89.4%, sensitivity of 29.5%, specificity of 99.6%, precision of 92.0% and AUROC of 91.8%. Performance varied with inter-observer agreement with AUROC at 95.2% when radiologists agreed (n=443; 82.5%) and 74.5% when they disagreed (n=94; 17.5%). Other evaluation results stratified by age group and gender and inter-observer agreement are shown in **Table 1**. The ROC and Precision-Recall curve are shown in **Figure 1**.

Model classification following the auto-cropping algorithm performed comparably to that with the original images, because the Maccabi dataset is generally high quality and contains minimal extraneous area.

Kappa statistics for inter-observer agreement are presented in **Table 2**. Radiologists showed moderate to substantial agreement with one another (K = 0.540–0.641). Agreement between each radiologist and the model was fair (K = 0.290–0.398).

Table 1 Primary performance metrics

	No. images	% of positive	Accuracy	Precision	Sensitivity	Specificity	F1	AUROC	AUPRC
All images	537	14.5%	89.4%	92.0%	29.5%	99.6%	44.7%	91.8%	72.3%
Age									
< 1 year	40	10.0%	90.0%	-	0.0%	100.0%	0.0%	67.4%	15.0%
1 year	167	12.0%	89.2%	75.0%	15.0%	99.3%	25.0%	88.6%	52.5%
2 year	125	12.8%	89.6%	80.0%	25.0%	99.1%	38.1%	88.9%	61.7%
3 year	109	20.2%	88.1%	100.0%	40.9%	100.0%	58.1%	98.1%	92.6%
4 year	62	16.1%	88.7%	100.0%	30.0%	100.0%	46.2%	93.5%	88.0%
5 year	34	17.6%	94.1%	100.0%	66.7%	100.0%	80.0%	100.0%	100.0%
Sex									
Male	231	19.4%	90.9%	92.9%	39.4%	99.5%	55.3%	91.3%	73.6%
Female	306	10.8%	88.2%	90.9%	22.2%	99.6%	35.7%	92.4%	72.3%
Inter-observer agreement									
Yes	443	10.4%	93.9%	100.0%	41.3%	100.0%	58.5%	95.2%	79.9%
No	94	34.0%	68.1%	66.7%	12.5%	96.8%	21.1%	74.5%	60.7%

Figure 1 ROC curve and Precision-Recall curve

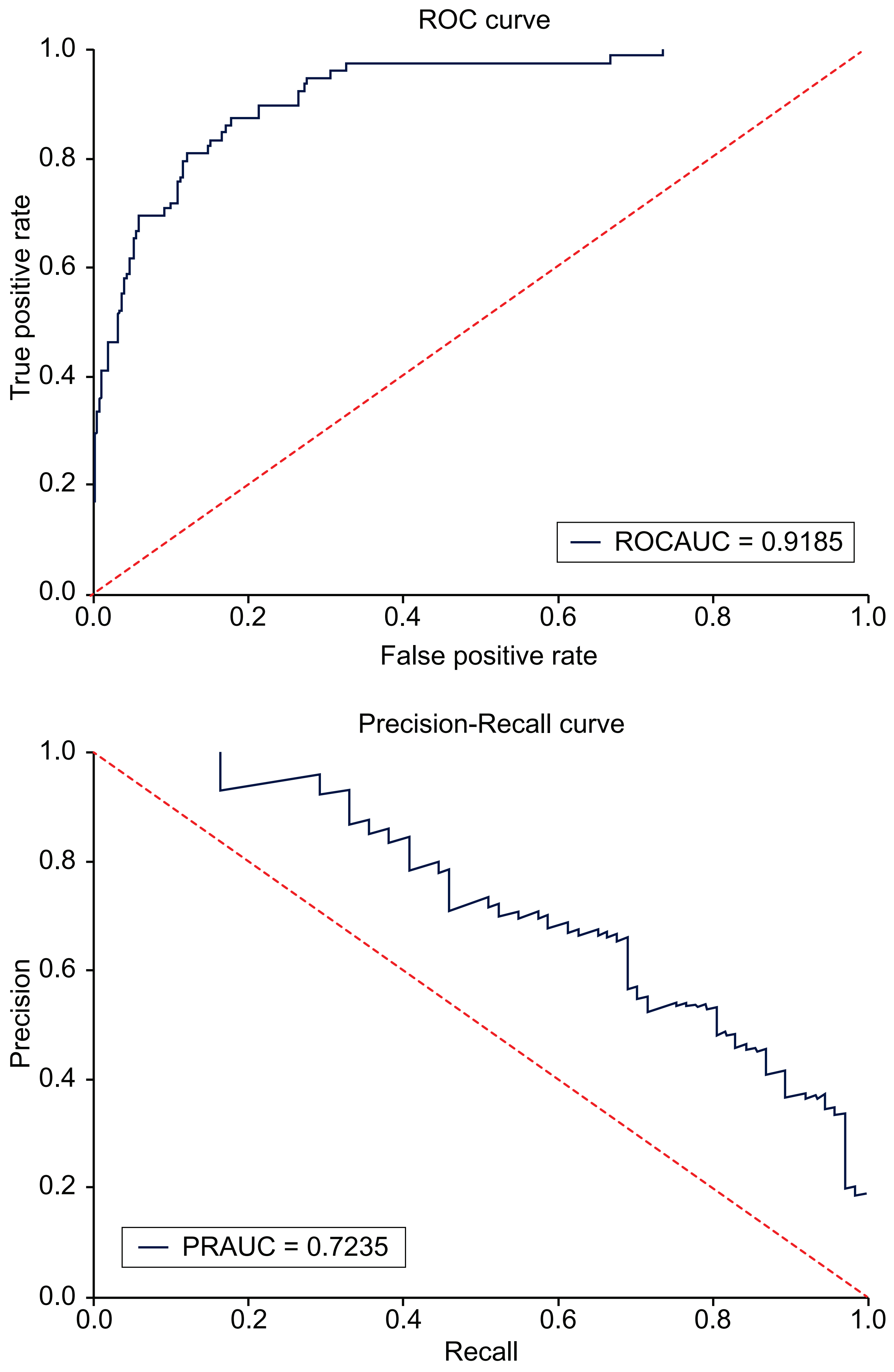

Table 2 Pairwise Cohen's kappa values

	Rater 2	Rater 3	Model
Rater 1	0.540	0.641	0.398
Rater 2	NaN	0.544	0.290
Rater 3	NaN	NaN	0.363

Conclusions

The deep learning-based model identifies pneumonia on pediatric CXRs with 92.0% precision and nearly 100% specificity compared to human interpretation. Lower performance when radiologists disagreed indicates that the model's uncertainty aligns with physician uncertainty. High specificity limits false positives, which can help avoid unnecessary treatments and conserve resources.

Differences in performance across settings reflect variation in healthcare systems: Hong Kong datasets contained later or more severe PEP cases resembling the model's training data, whereas Israeli CXRs more often captured very early, milder PEP, which are less similar to the training examples and harder to detect.

Overall, these results support the model's potential utility for PEP but highlight the importance of local data and further validation to optimize performance across different clinical contexts.

References

1. Chen, Y., Roberts, C. S., Ou, W., Petigara, T., Goldmacher, G. V., Fancourt, N., & Knoll, M. D. (2021). Deep learning for classification of pediatric chest radiographs by WHO's standardized methodology. *PLoS one*, 16(6), e0253239.

2. Wang, D., Ru, B., Lee, E. Y. P., Hwang, A. C. N., Chan, K. C. C., Weaver, J., & Roberts, C. S. (2024). Validation of a deep learning model for classification of pediatric pneumonia in Hong Kong. *Vaccine*, 42(26), 126370.

Disclosures

This Research was supported by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA.

Contact Information

Dong Wang, Ph.D., Merck & Co., Inc., 126 E Lincoln Ave, Rahway, NJ 07065, USA

E-mail: dong.wang10@merck.com

https://bit.ly/4nozZc5

Presented at ISPOR-EU; Glasgow, UK; November 9 - 12, 2025.