Pilot Retrospective Evaluation of an AI Chest X-ray Report Generator by Princess Alexandra Hospital Radiologists

Aaron Nicolson[1], Liz Cooper[1], Hwan-Jin Yoon[1], Claire McCafferty[1], Ramya Krishnan[2],

Michelle Craigie[2], Nivene Saad[2], Jason Dowling[1], Ian Scott[2,3], and Bevan Koopman[1,3]

[1]e-Health Research Centre, CSIRO Health and Biosecurity, Australia
[2]Princess Alexandra Hospital, Metro South Health, Brisbane, Australia
[3]University of Queensland, Brisbane, Australia

Introduction: Chest X-rays (CXRs) are among the most frequently performed imaging investigations worldwide, yet high reporting volumes and workforce constraints contribute to delays and variability in reporting quality. Generative AI systems for automated CXR report generation may support radiologists but require rigorous radiologist-led evaluation before clinical integration.

Aims: The primary objective was to evaluate, in a blinded pairwise comparison, the acceptability of AI-generated radiology reports relative to radiologist-authored reports.

Methods: We conducted a retrospective observational pilot study using 120 studies stratified across eight common findings, evaluated by three consultant radiologists. For each study, raters compared a blinded, randomised pair of reports (AI-generated vs radiologist-authored) and recorded a categorical preference (generated, radiologist, or no preference), with justification (precision, recall, readability).

Results: Across 360 evaluations (120 studies × 3 raters), generated reports were acceptable in 43% of cases (defined as preferred or rated equivalent to radiologist reports). Acceptability varied by finding (e.g., 46% for simple pleural effusion; 33% for pulmonary congestion). Radiologist reports were more often preferred for higher recall and precision, whereas generated reports were more frequently preferred for readability. Complete agreement occurred in 36% of studies; inter-rater reliability was slight ($\kappa = 0.128$, $p = 0.0086$).

Conclusions and Relevance: In this pilot evaluation, AI-generated reports were acceptable in a substantial minority of cases but were commonly judged inferior for recall. As a small, retrospective study, findings should be interpreted cautiously. These results inform sample size calculations and model refinement for a larger evaluation and support further development toward clinically supervised deployment.