Use of NLP techniques to extract and codify radiologist opinions from radiology reports: An "at scale" strategy for delivering post market surveillance of AI tools in clinical practice

Howell Fu[1], Rhydian Windsor[2], Robin Park[2], Peter Aylward[2], Hilal Johnson[2], Mark Beggs[2], Sarim Ather[1,2]

1. Oxford University Hospitals NHS Foundation Trust 2. RAIQC Ltd

Purpose

To automate post-market surveillance of radiology AI tools

Background

- AI tools which interpret images and provide reports are seeing increasing clinical deployment
- Strong post-market surveillance (PMS) is essential, both for assuring efficacy, and for detecting bias (eg poor performance in women or minority ethnic groups)
- AIs are not currently licensed for independent use: *ie* all studies should also be reported by a qualified radiologist (or other professional)
- Therefore radiology reports provide the ground truth for PMS
- Radiologist reports are free-text: therefore traditional methodology requires trained humans to read each report and code whether each finding of interest is present (*eg* consolidation)
- This is intractable at scale: a large teaching hospital may produce 15,000 plain radiograph reports a month
- We have designed and are implementing a post market surveillance scalable PMS tool to assess AI tools concordance within our hospital, using a Natural Language Processing (NLP) model to automate the coding of radiology reports for multiple findings (Figures 1 & 2)
- Optimising the accuracy of the NLP is key to being able to effectively monitor AI tool performance, as the NLP output provides the ground truth for comparison.

Methods

- We developed and evaluated three NLP applications
- Data: anonymised radiologist reports of 300 inpatient chest x-rays and 527 emergency department X-rays for suspected fracture
- Ground truth: the reports were manually coded by radiologists for the presence or absence of 13 relevant findings
  - Chest X-rays: 10 findings including cardiomegaly, consolidation, nodule *etc*
  - Fracture: 3 findings – fracture, joint effusion, dislocation
- The NLP tools evaluated each report, generating a binary "Yes/No" call for each finding. These were compared against manual audits of the radiology reports (Figure 3)
- Hardware: NLP models were run using a single 24 GB Nvidia A10 GPU platform
- Outcome measures:
  - Primary: sensitivity, specificity, and overall accuracy of the NLP tools
  - Secondary: processing times of the NLP tools
- The best-performing NLP tool was selected for integration into the PMS platform and its results are reported here

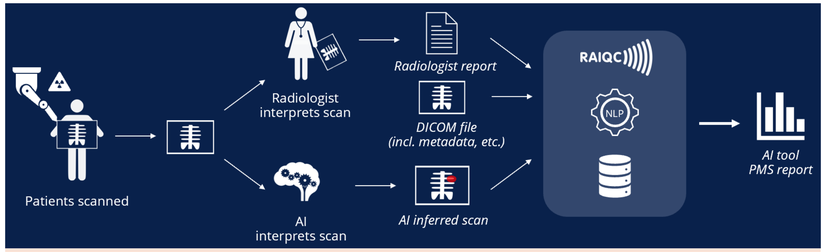


Figure 1: Role of the NLP tool in the image interpretation and PMS workflow

$$\text{AI Concordance (\%)} = \sum \frac{\text{"Condition present" (AI tool)}}{\text{"Condition present" (NLP tool)}}$$

Figure 2: Calculation of AI concordance with radiology reports

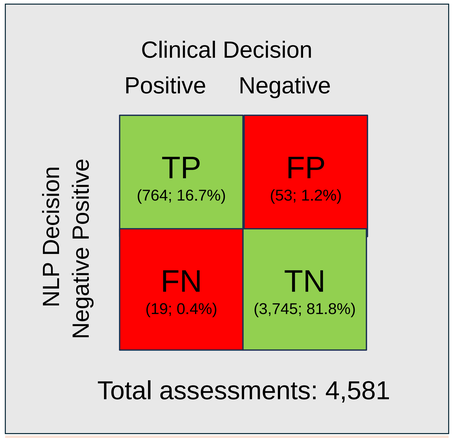


Figure 3: Truth table for the NLP tool, combined over the chest X-rays and fracture X-rays, for a total of 4,581 assessments

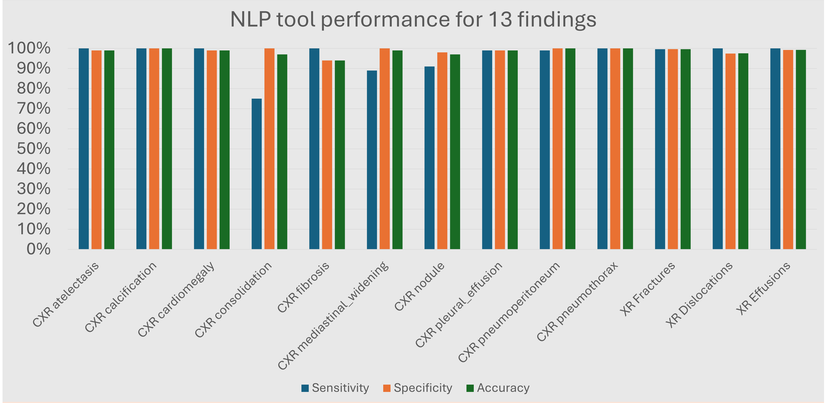


Figure 4: Performance of the best performing NLP tool

Results

- Overall outcomes of the best-performing NLP tool for the 13 findings examined (mean ± SEM):
  - Sensitivity: 96% ± 2.0%
  - Specificity:99% ± 0.4%
  - Accuracy: 98% ± 0.5%
- Nine conditions findings demonstrated accuracy >99% (Figure 4):
  - atelectasis, calcification, cardiomegaly, mediastinal widening, pleural effusion, pneumoperitoneum, pneumothorax, fractures and effusions
- NLP evaluation times:
  - 4-6 seconds per condition finding per report
  - Overall run evaluation times for 10 findings over 500 reports were 2.8-8.3 hours

Conclusion

- We have developed a radiology-specific NLP application which creates coded datasets from free-text radiology reports with high accuracy
- This is embedded in an AI PMS platform being deployed and evaluated in our hospital
- Extrapolation of the NLP performance reported here would allow assessment of 15,000 radiograph reports within a practical 9-27h timeframe with the opportunity to further improve performance
- Automated coding of radiology reports in this way allows "at scale" performance evaluations of deployed imaging AI tools for the first time