The variability of classification labels is an important barrier to the effective comparison of artificial intelligence software between vendors

A Maiter[1,2], E Hesketh[1], P Metherall[1], J Taylor[1], S Alabed[1,2,3], K Dwivedi[1,2,3], W Tindale[1], AJ Swift[1,2,3], CS Johns[1]

[1] Sheffield Teaching Hospitals NHS Foundation Trust, [2] NIHR Sheffield Biomedical Research Centre, [3] The University of Sheffield

Background

- Testing AI-based software options from different vendors and comparing their performance directly are essential for guiding procurement and deployment decisions
- This necessitates a level of consistency in how outputs are presented by software that have performed similar tasks, such as the classification of chest radiographs

Objective

- To assess the consistency of classification labels provided in the outputs of commercial software for the evaluation of chest radiographs from seven vendors (*Annalise.ai*, *Gleamer*, *Infervision*, *Milvue*, *Oxipit*, *Qure.ai* and *Rayscape*) [a]

Methods

Data collection

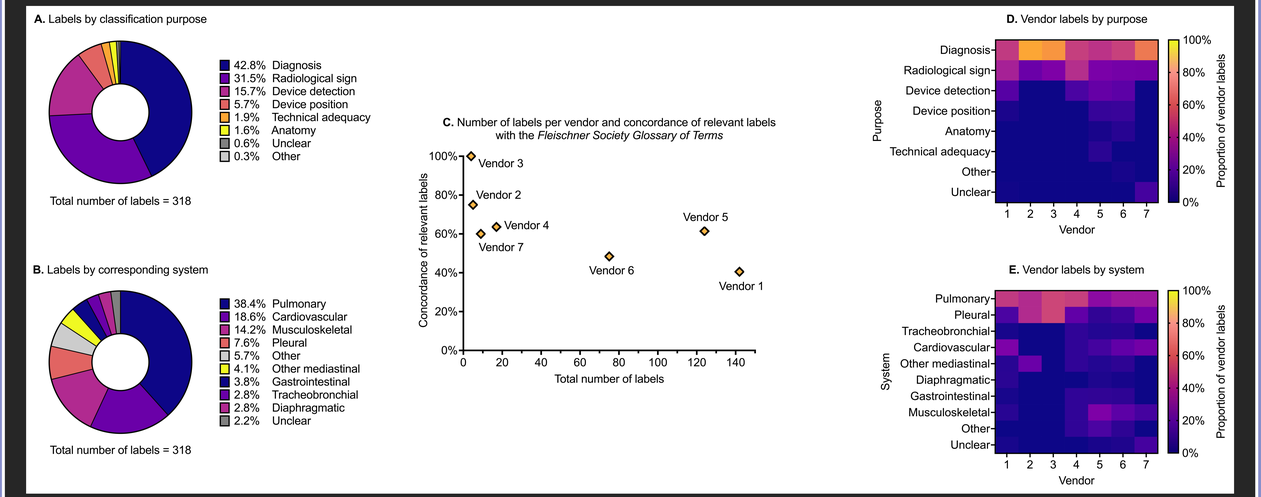
- Software from seven vendors analysed chest radiographs as part of a study assessing diagnostic accuracy
- The classification labels provided in the .csv software outputs from each vendor were retrieved

Label evaluation

- Labels were assessed qualitatively and using descriptive statistics
- Labels were categorised by consensus opinion from three radiologists for:
  1. Their intended purpose (e.g. diagnosis, detection of medical device)
  2. The anatomical system to which they correspond (e.g. pulmonary, cardiovascular)
- If relevant, label terminology was assessed against the *Fleischner Society Glossary of Terms for Thoracic Imaging*
  - Agreement with either the 2024[1] or 2008 versions of the *Glossary* was assessed
  - Relevant labels were defined as corresponding to the lungs, hila or pleura
  - Grammatical variations (e.g. "opacity" and "opacification"), terms with attached qualifiers (e.g. "left-sided" or "multiple") and combined concordant terms (e.g. "pleural mass") were deemed acceptable

Results [b,c]

- 318 unique labels were identified with a marked variation in the number provided by each vendor (range 4 to 142) [C]
- Most labels correspond to the detection of pathology (74%) – but there is inconsistency in whether this is described in terms of diagnosis (e.g. "tuberculosis") or radiological sign (e.g. "airspace opacification") [A & D]
- Fewer labels (21%) correspond to devices, and these are more commonly for detection than a decision about position [A]
- A minority of labels are poorly defined with an unclear purpose (e.g. "cord imaging") [A]
- The commonest systems represented were pulmonary (38%), cardiovascular (19%) and musculoskeletal (14%) [B]
- Some variability was encountered in which systems were represented across the vendors [E]
- Concordance with the *Fleischner Society Glossary of Terms* varied widely across the vendors (range 40% to 100%) [C]



Conclusions

- The number, purpose and terminology of classification labels provided by software can vary considerably between vendors
- Inconsistencies in label purpose (e.g. describing pathology in terms of diagnosis or radiological sign) may represent a significant barrier to comparing performance between software
- Deviation from established terminology (e.g. Fleischner Society Glossary of Terms) could limit the interpretability and clinical utility of outputs
- Better harmonisation of classification labels across vendors and precise label definitions are needed to overcome these issues

No dedicated funding was received for this study.

[a] Listed in alphabetical order.

[b] These represent results from preliminary analysis and are subject to change.

[c] The vendors have been assigned numerical identifiers at random.

[1] Bankier et al. Fleischner Society: Glossary of Terms for Thoracic Imaging. Radiology. 2024.