Augmenting lung cancer diagnosis on chest radiographs: positioning artificial intelligence to improve radiologist performance


AIM: To evaluate the role that artificial intelligence (AI) could play in assisting radiologists as the first reader of chest radiographs (CXRs), to increase the accuracy and efficiency of lung cancer diagnosis by flagging positive cases before passing the remaining examinations to standard reporting.

MATERIALS AND METHODS: A dataset of 400 CXRs including 200 difficult lung cancer cases was curated. Examinations were reviewed by three FRCR radiologists and an AI algorithm to establish performance in tumour identification. AI and radiologist labels were combined retrospectively to simulate the proposed AI triage workflow.

RESULTS: When used as a standalone algorithm, AI classification was equivalent to the average radiologist performance. The best overall performances were achieved when AI was combined with radiologists, with an average reduction of missed cancers of 60%. Combination with AI also standardised the performance of radiologists. The greatest improvements were observed when common sources of errors were present, such as distracting findings.

DISCUSSION: The proposed AI implementation pathway stands to reduce radiologist errors and improve clinician reporting performance. Furthermore, taking a radiologist-centric approach in the development of clinical AI holds promise for catching systematically missed lung cancers. This represents a tremendous opportunity to improve patient outcomes for lung cancer diagnosis.

© 2021 The Royal College of Radiologists. Published by Elsevier Ltd. All rights reserved.
Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide, contributing to over 2 million fatalities annually. In the UK, lung cancers are the second most common malignancy but are the single largest cause of cancer-related deaths, contributing to 21% of female and 23% of male cancer deaths. In Europe and the United States, lung cancer survival is lower than that of any other common malignancy, with 5-year survival rate ranging from 9% in the UK and Ireland to 15% in central Europe. Studies have shown that early diagnosis of lung cancer is critical in improving survival rates, with 5-year survival rates decreasing from 62% for stage 1 diagnosis to 3% for stage 4 diagnosis.

Despite efforts to increase computed tomography (CT) screening for lung cancers, many patients still rely on early diagnosis of their lung cancer via chest radiographs (CXR). This is a complex and subjective task, reflected by the fact that 90% of lung cancer misdiagnosis occurs on CXR. Previous studies have shown that radiologist identification of lung cancers on CXRs can yield low sensitivity and poor interobserver agreement and between 65% and 90% of cancers have been shown to be visible on CXRs in retrospect. The difficulty of tumour identification on CXRs is by far the most common reason for initiating a radiological malpractice suit. Common sources of errors in tumour identification have been shown to include radiologist experience and mindset, tumour characteristics, location, size, and the presence of other distracting pathologies.

Artificial intelligence (AI) has been shown to achieve parity or even surpass expert radiologist performance in certain interpretations of CXRs. Large-scale open-source datasets of thousands of labelled CXRs, such as CheXpert, have served to position the published radiologist performance on these datasets as the primary target for research groups. Indeed, “expert level” is regularly used to describe achieved performance of the top algorithms. Correspondingly, the majority of marketed AI-solutions for the detection of lung cancers focus either on deployment independent of radiologists, as a second reader of examinations, or delivery of supplementary information in the form of algorithmic heat maps.

Contrary to this approach, the combination of well-defined radiologist error characteristics and strong AI-derived performance means that taking a combined approach to the deployment of clinical AI stands to achieve better overall performance in tumour detection. In this sense, AI can be positioned to “plug the gap” of common human errors as well as removing easy cases from worklists, optimising the use of radiological expertise and improving work satisfaction.

The present study evaluated the role of an AI algorithm in an augmented cancer-triaging pathway on a curated dataset of 400 CXRs, 50% of which are positive for lung cancer. This pathway positions AI as the first reader of the CXR, whereby a positive tumour diagnosis is characterised as a high-confidence tumour (HCT) and is immediately triaged for further treatment. All remaining CXRs are then passed to the reporting radiologist to report as normal (Fig 1). This approach would increase overall sensitivity for potential cancers, driving earlier diagnosis and improving patient outcomes.

Materials and methods

This retrospective study was approved by the local radiology board. All patient imaging was fully anonymised for the purposes of the study. Ethical approval was not required as review of anonymised images was retrospective but was endorsed by the local Caldicott guardian.

Study design

Study design and dataset curation was overseen by the research radiologist, who had 9 years of experience as an FRCR-certified consultant radiologist in the NHS. The research radiologist was not one of the reviewing radiologists.

An information request was made to the hospital cancer-tracking team, and the NHS Cancer Registry database was used to yield a list of 7 years’ worth of lung cancers from the hospital site.

Individual patient radiology information system (RIS) packets were hand-searched. Images were reviewed on the picture archiving and communication system (PACS) to find 200 tumours. The presentation CXR was usually harvested; however, prior CXRs were also reviewed to see if the tumour might have been visible in retrospect. “Difficult” cases were sought actively, including those in the apex, behind the heart, overlapping the hilum and so forth, involving subjective value judgment by the research radiologist. Two hundred CXRs with 1, 2 or 3 cm both central and peripheral lung tumours were collected. No tumour was included that was >3.5 cm. A control group of radiographs (where there was no significant finding), aiming to be closely age-matched were also sourced using targeted RIS searches. The initial report was reviewed as well as the image by the research radiologist to collate an appropriate control group of non-cancer images. Tumour mimics, such as non-resolving consolidation, mediastinal mass lesions, and hilar adenopathy were excluded, so that the tumour set included true central or peripheral mass lesions or nodules (i.e., hilar adenopathy was not part of the selected set, but a 3 cm mass overlying the left hilum would be included). A database of these patients was held on the hospital computer network to facilitate any checking-back during the analysis phase. The ground-truth of each examination was established by a combination of the cancer registry database records, the electronic clinical record, and review of both subsequent and prior imaging. The research radiologist determined whether a tumour was visible in retrospect selecting the more challenging studies whenever possible.
CXR review workflow

Images were fully anonymised as follows. They were networked from the hospital PACS to the hospital research iMac into Horus (www.horosproject.org). Images were anonymised in Horus at point of transfer, and all further identifiable metadata was removed using DicomCleaner (http://dclunie.com). Images were exported as DICOM.

CXRs were then reviewed independently by a pool of FRCR-certified consultant radiologists with a minimum of 12 years of experience in the NHS. Reviewing radiologists were aware that the dataset contained a 50% tumour mix. Radiologists were not aware of patient history, clinical presentation, or AI-generated labels when reviewing CXRs. Each radiologist labelled CXRs for all abnormal findings, not only those indicative of cancer, and these labels were then analysed in combination with the AI-generated labels to give an overall assessment of how the performance of multiple radiologists could be augmented by the proposed workflow.

Radiologist findings directly indicative of cancer findings (e.g., nodule, mass) or suggestive of further investigation (e.g., hilar enlargement) were considered positive for cancer findings, on the basis that patients would receive further diagnostic tests and would not be missed at the CXR stage.

Study demographics

Exactly half of the 396 reviewed examinations show a tumour, while the other half formed the control group. The demographic breakdown of the tumour set comprised 92 male and 108 female patients (mean age 72.6 ± 10.4 [range 32–92] years) and the control set comprised 87 male and 113 female patients (mean age 61.8 ± 15.6 [range 34–98] years). Comparison of tumour and normal sets shows no significant difference in patient sex (p > 0.05) and significant differences in patient age (p < 0.05).

Deep neural network

AI-based cancer prediction was performed by a commercially available AI algorithm (Red Dot, Behold.ai, London, UK), an ensemble of deep convolutional neural networks. Included in algorithmic output classes are the following potentially malignant findings: “nodule”, “mass”, “consolidation”, and “hilar enlargement”. The algorithm would therefore be expected to detect these findings in the designation of an HCT finding.

In this study, the AI-generated score for the presence of a tumour (continuous value between 0 and 1, with a score of 1 being a certain tumour) was used to assign the HCT tag to
examinations, with the threshold used to designate HCT-
positivity being selected for optimal performance, which
was 0.21 in this study. All CXRs evaluated in this study were
collected after algorithm development and were not
included in any algorithm training steps. In addition to
tumour diagnosis, the AI algorithm was also used to output
a score for the presence of an abnormality. This score was
used to define overall normality of each CXR using a pre-
defined abnormality threshold.

**Statistical analysis**

Classification performance of the AI algorithm and radi-
ologists was assessed using overall accuracy, sensitivity, specificity, and precision. Further to this, agreement be-
tween radiologists was assessed using observed propor-
tional agreement, Cohen’s kappa score and McNemar’s
test. McNemar’s test was then used to assess if differ-
ences were statistically significant, with a p-value of <0.05
denoting statistically significant differences.

**Results**

Of the initially collected 400 studies, four were discarded
due to incomplete or corrupted data. After collection of 400
CXRs, four were discarded due to incomplete data, leaving
396 examinations (198 positive for tumours, 198 negative).
These were then passed to the reviewing radiologists for
review. This left a total of 396 examinations (56%/44% F/M),
of which 198 were positive for tumour(s) and 198 were
negative.

Standalone radiologist and algorithmic performances for
the detection of tumours on this dataset and combined
performance where the algorithm is positioned as the first
reader of examinations as in Fig 1.

**Radiologist performance**

The performance of three independent FRCR consultant
radiologists in identifying lung cancer on this dataset is
shown in Table 1. The mean accuracy of cancer detection is
87% (84–90%) and the overall mean sensitivity to cancer is
78% (69–86%).

On a patient level, these performances correspond to
between 136 and 171 patients being diagnosed correctly for
tumours and between 62 and 27 patients with missed
cancer pathologies. Notably, all radiologists have a low rate
of false positives, between one and nine examinations in
total (average precision 95.67%).

Analysis of the statistical correlation between the radi-
ologists’ reports shows an average observed proportional
agreement of 86.7% and corresponding average Cohen’s
kappa score of 0.72, denoting good overall agreement.
Agreement between all three radiologists occurs in 80% of
cases. The predictions were analysed for statistically sig-
nificant differences using McNemar’s test. This showed that
predictions made by radiologist 1 and radiologist 3 are
statistically different (p<0.05), whilst all other comparisons
showed only random differences (p>0.05).

**AI performance**

When deployed as a standalone algorithm, the AI algo-
rithm achieved an overall accuracy of 87% on this tumour
dataset, equivalent to the mean performance of the three
reviewing radiologists. The algorithm sensitivity was su-
perior to two of three radiologists at 80% whilst specificity
was marginally lower than radiologists at 93% (Table 1).
There was an increase in false-positive examinations, with
an overall precision of 92%.

**Radiologists+AI**

Combined performances were calculated according to
the proposed “HCT” workflow in Fig 1. In this retrospective
study, this corresponded to using the algorithm-assigned
tumour label if positive (simulating examinations being
removed from radiologists’ worklist), and if it negative,
using the radiologist-assigned label (simulating review of
remaining examinations).

For all radiologists, overall accuracy and sensitivity for
tumour identification were increased by combination with
AI, improving average scores by +3.67% and +13.33%
respectively (Table 2). False-negative cases, where cancer
findings were missed, were reduced by between 15 and 40
cases. Combined performance did show an increase in false-
positive examinations in all cases, with an average precision
change of −5.33% and specificity change of −6%. Overall
changes in accuracy and sensitivity when radiologists are
combined with AI are represented visually in Fig 2.

For all radiologists, improvements in combined scores
were shown to be statistically significant when compared
to their standalone performance (p<0.05). Furthermore,
agreement between radiologists was improved when com-
bined with AI, with all radiologist – AI labels agreeing in 92%
of cases (+12%). Average proportional agreement increased
to 94.33% (+7.63%) and the average Cohen’s Kappa score
was 0.89 (+0.17), suggesting very good agreement.

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standalone tumour classification performance for radiologists and the AI algorithm.</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Rad 1</td>
</tr>
<tr>
<td>Rad 2</td>
</tr>
<tr>
<td>Rad 3</td>
</tr>
<tr>
<td>AI</td>
</tr>
</tbody>
</table>

The top performance for each metric is highlighted in bold.
Rad, radiologist; AI, artificial intelligence.
Correspondingly, all radiologist + AI predictions were shown to be statistically similar using McNemar’s test \((p > 0.05)\).

On average, cases of missed tumours were reduced by 60% by combination of a single radiologist with AI. In comparison, combining the predictions of all three FRCR consultant radiologists and taking any positive prediction of a tumour as the given label, reduced missed tumours by 65.4%.

Algorithm performance on radiologist misses

Of a total of 198 positive tumour cases, 15 tumours were missed by all three radiologists (Table 3). The algorithm detected eight of these cases, which would have been otherwise missed. In total, the algorithm detected 70.2% of all tumours, which were missed by at least one radiologist.

Table 3

<table>
<thead>
<tr>
<th>Radiologist</th>
<th>Exam count</th>
<th>Algorithm positives</th>
<th>Algorithm sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>15</td>
<td>8</td>
<td>0.53</td>
</tr>
<tr>
<td>1</td>
<td>24</td>
<td>12</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>37</td>
<td>30</td>
<td>0.81</td>
</tr>
<tr>
<td>3</td>
<td>122</td>
<td>109</td>
<td>0.89</td>
</tr>
</tbody>
</table>

As part of the study process, the research arbitrator made comments on commonly missed cases. In the 15 examinations where no radiologists spotted the cancer, key reasons given include small tumour size, hidden tumours, poor film quality, and the presence of distractor pathologies.

Distracting findings

Radiologist reporting of examinations in this study required labelling all abnormal pathologies, as would be performed in a clinical setting. Of the 396 examinations, 136 also contained “distracting” findings, abnormalities that are not lung cancers, according to the consensus of radiologist labels, for example, the presence of chronic obstructive pulmonary disease (COPD) or a small pleural effusion. The present results show that radiologist performance for identifying lung cancers was significantly reduced on these examinations, with accuracy and sensitivity decreasing by 18% and 30%, respectively, when compared to examinations without distracting findings (Table 4). The algorithm’s performance on tumours was also decreased, but to a lesser extent, with accuracy and sensitivity decreasing by 12% and 10%, respectively.

Table 4 shows that although a combined radiologist + AI approach improves accuracy and sensitivity in both scenarios, improvements are significantly larger when distracting findings are present. Accuracy increases by 9% with distractors present compared to a 1% increase without. An example of algorithmic identification of tumours when other abnormalities are present is shown in Fig 3; in this case, all radiologists diagnosed the pleural effusion, but only one also caught the nodule in the right upper lobe.

Tumour size

Tumour size was also measured by the research arbitrator to assess the impact of this characteristic on overall performance. Overall sensitivity for tumours increased with tumour size for both radiologists and the algorithm (Table 5); however, combined sensitivity showed performance improvements on cancers of all sizes, with the greatest sensitivity increase (+0.17) coming in tumours 1–2 cm in size.

Discussion

The present study showed that AI-based triage has the potential to improve the accuracy and efficiency of tumour detection on CXRs. The analysis shows that a standalone AI algorithm is equivalent to the average performance of
consultant radiologists in the identification of tumours, with an overall accuracy of 87%; however, the best performances are achieved when an AI algorithm is applied in the proposed triage pathway, with this combination improving performance for every radiologist and achieving an average accuracy, sensitivity, and specificity of 90.67%, 91.33%, and 90%, respectively.

**Improving patient outcomes**

In the present study, implementation of AI-based triage caused a significant reduction in the number of tumours missed by radiologists. The overall reduction in missed cancers of 60% has great promise in improving patient survival rates through the early identification of lung cancers. Currently in the UK, the 5-year net survival is 62% for stage 1 tumours and drops to 3% for stage 4 tumours. Thus, the improved accuracy for cancers on CXRs can help to prioritise these patients onto the National Optimal Lung Cancer Pathway (NOLCP).

The observed variability of radiologist labels on this dataset demonstrates the difficulty and subjectivity of interpreting CXRs, even for experienced and senior readers. This is underlined by the presence of statistically significant differences in the performance of two of the reviewing radiologists (radiologists 1 and 3), representing a potential difference in the standard of patient care. Furthermore, these examinations were reported outside of a clinical setting. It is not unreasonable to suggest that other pressures and less experienced reviewers may lead to greater variability in patient care when CXRs are reviewed in a clinical setting. Therefore, it is particularly encouraging that the implementation of an AI-based triage pathway has a standardising effect on the performance of the reviewing radiologists, with no statistically significant variabilities in their observed performance on combined labels. This demonstrates the potential of AI to improve overall performance and target inconsistencies and inequalities in the standard of patient care.

**Implications for radiology and AI**

Positioning AI as the first reader of examinations stands to improve the overall accuracy and sensitivity to potential cancer cases. In the workflow presented in Fig 1 and examined in this study, direct triaging of positive cases will also reduce the CXR reporting burden; however, the increase in false positives may be passed onto CT and other follow-up procedures.

### Table 4

Performances of radiologists, the AI algorithm, and combined labels on examinations with and without distracting findings.

<table>
<thead>
<tr>
<th></th>
<th>Count % female/male</th>
<th>Tumour count</th>
<th>Radiologist accuracy</th>
<th>Radiologist sensitivity</th>
<th>AI accuracy</th>
<th>AI sensitivity</th>
<th>Combined accuracy</th>
<th>Combined sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-distractors</td>
<td>260 55/45</td>
<td>121</td>
<td>0.93</td>
<td>0.9</td>
<td>0.91</td>
<td>0.84</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td>Distractors</td>
<td>136 57/43</td>
<td>77</td>
<td>0.75</td>
<td>0.6</td>
<td>0.79</td>
<td>0.74</td>
<td>0.84</td>
<td>0.86</td>
</tr>
</tbody>
</table>

AI, artificial intelligence.

![Figure 3](image_url) Algorithmic activation heat map for all abnormalities for a patient presenting with pleural effusion and lung nodule. All radiologists identified the effusion but only radiologist 2 and the algorithm identified the nodule in the right upper lobe (a) with heat map; (b) native.
services. Full clinical implementation of this algorithm may still involve clinical review of HCT-positive examinations, meaning reader disagreement may decrease the false-positive rate. This clinician–AI interaction may give the additional benefit of knowledge transfer from the algorithm to readers, an area of huge potential for clinical algorithms and worthy of further study. Furthermore, in the present study, the threshold for HCT positivity was set for optimal classification performance; however, this can be adjusted at each deployment site to suit clinical requirements.

Radiologist errors in tumour identification are more likely to occur when other “distracting” findings are present. This phenomenon occurred, despite the readers being primed with the knowledge of the dataset containing 50% tumours, with significant other pathologies such as consolidation and other tumour mimics being excluded from the dataset. Correspondingly, it is in this setting that the greatest impact of a combined approach with AI is seen, with the improvements in accuracy being nine-times greater than when no distractor pathology is present. Approaching the challenge of applying AI to CXRs in a more radiologist-centric manner may spur increased focus on common areas of radiological fallibilities and expertise, leading to greater overall improvements in clinical standards from the implementation of AI. The potential of AI to target areas of common radiological errors is demonstrated by the overall reduction in missed tumours with a single radiologist + AI (−60%) being only slightly inferior to the combination of all three consultant radiologists (−65.4%).

Limitations and future work

This study has been designed to validate performance improvements on a curated set of challenging lung cancer cases. Therefore, there are many aspects that do not accurately reflect a real-world clinical setting, primarily the prevalence of lung cancers in the dataset. The 50% distribution of lung cancers in the dataset introduces a risk that both radiologists and AI give the correct classification whilst focussing on the wrong region of the CXR; however, in a clinical setting, this would still represent a patient flagged for follow-up, and therefore, not missed at the CXR stage. Further work correlating radiologist, AI, and ground truth “regions of interest” for possible tumours would address this, but would also stray further from standard reporting conditions for human readers.

Further work on a clinically representative dataset is required to confirm the exact resource saving and performance changes to be expected from the proposed triage pathway. That said, a 200 nodule dataset is a powerful tool that is clinically relevant to any large hospital, reflecting more than most hospitals annual lung cancer cases, addressing the “pain point” in CXR reporting, the missed opportunity to make an early diagnosis of lung cancer. Furthermore, the filtering of positive cancer cancers is not the only potential aid to efficiency and accuracy that AI could offer. A more complete study including other information, such as heat maps, algorithmic scores, and normal triage would be helpful in fully characterising the role of AI.

From the curation of a relatively small tumour dataset from a single hospital site, the present study has demonstrated that useful characterisation of radiologist and algorithm performance is possible, which will prove invaluable in the development of artificial intelligence algorithms of true clinical utility in future. Real progress in developing algorithms that aim to empower radiological resources, not simply to replace them, might be made with an NHS-wide archive of cancers similar to the one used in this study.

Conflict of interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: MT: Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships.

TD, GD: Activities related to the present article: is employed by behold.ai. Activities not related to the present article: is employed by behold.ai. Other relationships: disclosed no relevant relationships.

SR, RR: Activities related to the present article: is employed by behold.ai. Activities not related to the present article: is employed by behold.ai; has stock/stock options in behold.ai. Other relationships: disclosed no relevant relationships.

TNM, MH, JI: Activities related to the present article: is not employed by behold.ai. Activities not related to the present article: has stock/stock options in behold.ai. Other relationships: disclosed no relevant relationships.

Acknowledgements

T.D., G.D. S.R., R.R. T.N.M. are either employed by and/or have stock options in Behold.ai.

References
