The effect of different radiological models on diagnostic accuracy and lung cancer screening performance

ABSTRACT
High false-positive (FP) scan rates associated with low-dose computed tomography (LDCT) lung cancer screening result in unnecessary follow-up tests and exposure to harm. The definition of a ‘positive’ scan can impact FP rates and screening performance. We explored the effect of Lung Imaging Reporting and Data System (Lung-RADS) criteria, PanCan Nodule Malignancy Probability Model and varying nodule size thresholds (≥4 mm, ≥6 mm, ≥8 mm) on diagnostic accuracy and screening performance compared with original trial definitions (National Lung Screening Trial (NLST) criteria) in a secondary analysis of a lung cancer screening cohort. We found Lung-RADS criteria and the PanCan Nodule Malignancy Probability Model could substantially improve screening performance and reduce FP scan rates compared with NLST definitions of positivity but that this needs to be balanced against possible risk of false-negative results.

Trial registration number Australian New Zealand Clinical Trials Registry, ACTRN1261000007033.

BACKGROUND
The landmark National Lung Screening Trial (NLST)1 demonstrated lung cancer mortality reduction by low-dose computed tomography (LDCT) screening, paving the way for US Preventative Services Taskforce and Medicare screening recommendations. However, a major limitation of LDCT screening is the high positive scan rate, averaging 24.2% in NLST lung cancer screening result in unnecessary follow-up tests and exposure to harm. The definition of a ‘positive’ scan can impact FP rates and screening performance. We explored the effect of Lung Imaging Reporting and Data System (Lung-RADS) criteria, PanCan Nodule Malignancy Probability Model and varying nodule size thresholds (≥4 mm, ≥6 mm, ≥8 mm) on diagnostic accuracy and screening performance compared with original trial definitions (National Lung Screening Trial (NLST) criteria) in a secondary analysis of a lung cancer screening cohort. We found Lung-RADS criteria and the PanCan Nodule Malignancy Probability Model could substantially improve screening performance and reduce FP scan rates compared with NLST definitions of positivity but that this needs to be balanced against possible risk of false-negative results.

Trial registration number Australian New Zealand Clinical Trials Registry, ACTRN1261000007033.

METHOD
Participant eligibility and recruitment have been described elsewhere.11 12 Two hundred and sixty-eight healthy current or former smokers (171 men; 85 women) aged 60–74 years were enrolled in a LDCT screening study; median age 64.5 years; median pack-years of smoking 55 and 47% current smokers. Participants received baseline scans (T0) and up to two annual incidence scans (T1 and T2) using a 64-detector helical CT based on NLST protocols. Criteria for scan positivity, based on NLST, were any nodule ≥4 mm diameter (baseline scan) and any new or growing nodule (incidence scan).11 12 Indeterminate nodules underwent 2 years radiological follow-up. Health status follow-up continued for 5 years. Data were censored 1 March 2016.

Scan results were reclassified using Lung-RADS (category 3 or 4 considered positive), the PanCan Model (full model with spiculation, ≥10% risk considered positive) and increasing nodule size thresholds (exploratory cut-points set at maximum diameter of ≥4 mm, ≥6 mm or ≥8 mm and labelled d4, d6 and d8, respectively). Performance metrics were calculated at the participant level using the scan result and NLST Lung Cancer Status definition (cancer present; cancer absent13). Scans were excluded if cancer status could not be determined (eg, participant lost to follow-up). Model discrimination, assessed using the area under the receiver operating characteristic curve (AUC), and calibration assessed by visual plot and Hosmer-Lemeshow test, were calculated at the nodule level.

RESULTS
Baseline scans
Two hundred and fifty-six participants received T0 scans (table 1); 136 participants had 301 nodules (median diameter 5.4 mm, range 4.0–34.0 mm). One hundred and twenty-eight (50.0%) scans were deemed positive (127 scans without historical images demonstrating nodule stability plus one scan with suspicious non-nodule findings). Lung cancer was diagnosed in five (2.0%) participants. One individual with missing lung cancer status was excluded. PanCan, Lung-RADS, d4, d6 and d8 correctly identified all participants with lung cancer. PanCan and d8 produced the greatest reduction in FP scans (89.4% reduction, 95% CI 83.6 to 95.2 and 78.0% reduction, 95% CI 69.4 to 86.6, respectively), improving specificity and positive predictive value without reducing sensitivity or negative predictive value without reducing sensitivity or negative predictive value (table 2).

Incidence scans
One hundred and fifty-six of 472 (33.1%) incidence scans (239 T1; 233 T2) were positive according to our original criteria (new nodule of any size or growth detected) (table 1). Lung cancer was diagnosed in three participants (1.3%) at T1 and four (1.7%) at T2. Six individuals without lung cancer status were excluded. Increasing nodule size threshold for a positive scan result reduced FPR but also reduced sensitivity. Lung-RADS and d4 had similar performance characteristics (table 2).

Downstream effects
At baseline, the PanCan Model would have avoided 110/127 (86.6%) interval CT scans and 3/5 positron emission tomography (PET) scans in participants without cancer; Lung-RADS would have
avoided 75 interval CT scans and no PET scans. At the incidence rounds, The Queensland Lung Cancer Screening Study generated 229 interval CT scans from FP nodules and Lung-RADS would have reduced this number to 44. Across three rounds of screening, Lung-RADS would have avoided 260/356 (73.0%) interval CT scans, 2/7 (28.6%) PET scans and 2/3 (66.7%) bronchoscopies in participants without cancer. However, Lung-RADS would have misclassified one participant with lung cancer present at the T2 screening round (tables 1 and 3).

Model performance

Ten baseline nodules in eight participants were diagnosed as lung cancer during follow-up (median 29.9 months, range 2.0–69.7). PanCan Model discrimination was very good in the 301 baseline nodules (AUC 0.90; 95% CI 0.75 to 1) but not statistically different from Lung-RADS (AUC 0.84; 95% CI 0.69 to 0.98; difference in AUC=0.06, 95% CI –0.04 to 0.16, p value=0.25). Visual plot and Hosmer-Lemeshow goodness-of-fit test (χ²=8.8, df=8, p=0.36) did not indicate significant miscalibration.

**DISCUSSION**

As lung cancer screening gains traction internationally, attention focuses on minimising harm and controlling costs by reducing FP scan rates. In this exploratory comparative study, stricter definitions of positivity decreased FP results, improved performance metrics and reduced downstream tests at the risk of increasing false-negative scans at the incidence round.

**Table 1** Lung cancer screening scan results: comparison of Queensland Lung Cancer Screening Study original definitions, Lung-RADS criteria, PanCan Model and differing nodule size thresholds

<table>
<thead>
<tr>
<th>Baseline round</th>
<th>Positive*</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>d4</td>
<td>128 (50.0)</td>
<td>128 (50.0)</td>
</tr>
<tr>
<td>d6</td>
<td>64 (25.0)</td>
<td>CT scans n=256 (100.0)</td>
</tr>
<tr>
<td>d8</td>
<td>32 (12.5)</td>
<td>224 (87.5)</td>
</tr>
<tr>
<td>Lung-RADS</td>
<td>56 (21.9)</td>
<td>200 (78.1)</td>
</tr>
<tr>
<td>PanCan Model</td>
<td>18 (7.0)</td>
<td>238 (93.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Incidence rounds</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>d-All</td>
<td>156 (33.1)</td>
<td>316 (66.9)</td>
</tr>
<tr>
<td>d4</td>
<td>44 (9.3)</td>
<td>428 (90.7)</td>
</tr>
<tr>
<td>d6</td>
<td>26 (5.5)</td>
<td>446 (94.5)</td>
</tr>
<tr>
<td>d8</td>
<td>17 (3.6)</td>
<td>455 (96.4)</td>
</tr>
<tr>
<td>Lung-RADS</td>
<td>46 (9.7)</td>
<td>426 (90.3)</td>
</tr>
</tbody>
</table>

**Baseline (T0) and incidence (T1 and T2) screening rounds. Data presented as count (%).

Criteria for positive baseline scan: d4 ≥4 mm threshold; d6 ≥6 mm threshold; d8 ≥8 mm threshold; Lung-RADS, Lung Imaging Reporting and Data System categories 3–4 and PanCan Model–Pan-Canadian multivariable logistic-regression risk model threshold ≥10%. Criteria for positive incidence scan: d-All new or growing nodule(s) of any size; d4 ≥4 mm threshold; d6 ≥6 mm threshold; d8 ≥8 mm threshold and Lung-RADS categories 3–4.

* One baseline scan was positive due to non-nodule suspicious findings.

FN, false negative; FP, false positive; NA, not assessed; TN, true negative; TP, true positive.

At baseline, all methods were highly sensitive. Increasing nodule size thresholds and Lung-RADS improved specificity to 76–89%, but the PanCan Model had the highest specificity of 94.8%. At incidence scan rounds, stricter size definitions improved specificity and reduced FP scans at the cost of lower sensitivity. FPR was reduced by a similar degree to that seen in a retrospective application of Lung-RADS to NLST data (52–89% and 74–92% reduction at baseline and incidence rounds, respectively, compared with 52% and 76%, respectively)\(^8\).

Since baseline scans have no historical comparison against which to assess growth, they inherently generate more positive scans and downstream tests than incidence scans.\(^4\) In contrast, incident nodules are more likely to be malignant.\(^16\) In the absence of a validated multivariable risk model for incidence scans, a more conservative, smaller size threshold compared with the baseline scan, such as recommended by Lung-RADS, seems prudent. However, defining an optimal threshold at prevalence and incidence rounds goes beyond the pure metrics of the test and requires a health economic perspective which may differ between countries and healthcare settings. All measurements in this study were based on maximum axial diameter yet volumetric nodule analysis reportedly gives a more accurate assessment of size and interval growth. The NELSON trial compared diagnostic accuracy between diameter-based management and volume-based management and found similar sensitivity but higher specificity for a volumetric approach\(^17\) (diameter sensitivity 92.4% and specificity 90.0%; volumetry sensitivity 92.9% and specificity 94.9%). The diameter measurements were generated from volumetric software results and it is possible this may have underestimated the true difference by avoiding the variability associated with human reader measurements. The true advantage of volumetry, which requires specialised software analysis, against a multivariable risk assessment merits further investigation.

Our study is the first to simultaneously compare these differing methods of scan classification. It has limitations inherent to retrospectivity and relatively small cohort size; however, participants were well characterised and long, near-complete follow-up made cancer status misclassification unlikely.

In conclusion, we demonstrate lung cancer screening performance and FP scan rate can be improved by varying the definition of a positive scan balanced against...
Table 2  
Lung cancer screening scan performance metrics: comparison of Queensland Lung Cancer Screening Study original definitions, Lung-RADS criteria, PanCan Model and differing nodule size thresholds

<table>
<thead>
<tr>
<th>Threshold/method</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>PPV (%)</th>
<th>NPV (%)</th>
<th>FPR reduction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline scan</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d4</td>
<td>100 (47.8 to 100)</td>
<td>50.8 (44.4 to 57.2)</td>
<td>3.9 (1.3 to 8.9)</td>
<td>100 (97.1 to 100)</td>
<td>NA</td>
</tr>
<tr>
<td>d6</td>
<td>100 (47.8 to 100)</td>
<td>76.4 (70.6 to 81.5)</td>
<td>7.8 (2.6 to 17.3)</td>
<td>100 (98.1 to 100)</td>
<td>52.0 (39.4 to 64.6)</td>
</tr>
<tr>
<td>d8</td>
<td>100 (47.8 to 100)</td>
<td>89.2 (84.7 to 92.8)</td>
<td>15.6 (5.3 to 32.8)</td>
<td>100 (98.4 to 100)</td>
<td>78.0 (69.4 to 86.6)</td>
</tr>
<tr>
<td>Lung-RADS</td>
<td>100 (47.8 to 100)</td>
<td>79.6 (74.1 to 84.4)</td>
<td>8.9 (3.0 to 19.6)</td>
<td>100 (98.2 to 100)</td>
<td>58.5 (46.5 to 70.5)</td>
</tr>
<tr>
<td>PanCan Model</td>
<td>100 (47.8 to 100)</td>
<td>94.8 (91.3 to 97.2)</td>
<td>27.8 (9.7 to 53.5)</td>
<td>100 (98.5 to 100)</td>
<td>89.4 (83.6 to 95.2)</td>
</tr>
</tbody>
</table>

Baseline (T0) and incidence scan.  
PanCan Model: Pan-Canadian multivariable logistic-regression risk model threshold ≥10%; d4, d6, d8, size threshold definitions for positive scan—the presence of at least one nodule with maximum axial diameter ≥4 mm, ≥6 mm and ≥8 mm, respectively.

Baseline scan, d4 was the definition used in the Queensland Lung Cancer Screening Study; d-All, Queensland Lung Cancer Screening Study definition of positive scan during incidence round—new or growing nodule of any size.

Possible false-negative results. At baseline, the PanCan Model yielded best results, whereas at incidence scan, d4 followed by Lung-RADS yielded the best specificity and sensitivity.

Henry M Marshall, Henry Zhao, Rayleen V Bowman, Linda H Passmore, Elizabeth M McCaul, Ian A Yang, Kwan M Fong

University of Queensland Thoracic Research Centre and Department of Thoracic Medicine, The Prince Charles Hospital, Queensland, Australia

Correspondence to Dr Henry M Marshall, University of Queensland Thoracic Research Centre, Department of Thoracic Medicine, The Prince Charles Hospital, Rode Rd, Chermside, QLD 4032, Australia; henry.marshall@health.qld.gov.au

Acknowledgements The Queensland Lung Cancer Study team based at The University of Queensland Thoracic Research Centre, The Prince Charles Hospital, Brisbane: Dr John Ayres, BM, FRCR; Dr Jane Crossin, MB BCH, FRCR; Dr Melanie Lau, MBBS, FRANZCR; Adjunct Professor Richard E Slaughter, MBBS, FRANZCR; Stanley Redmond, Dip App Sci; Deborah Courtinney, BN; Dr Steven C Leong, MBBS, FRACP; Dr Morgan Windsor, MBBS, FRACS; Associate Professor Paul V Zimmerman, BSc, MBBS, FRACP, MD, Patients and staff at the Prince Charles Hospital, Brisbane.

Contributors HMM had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.  
Study concept and design: HMM, KMF, IAY, RVB.  
Analysis, acquisition or interpretation of data: all authors.  
Drafting of the manuscript: HMM, HZ.  
Critical revision of the manuscript for important intellectual content: KMF, IAY, RVB.  
Statistical analysis: HMM, HZ.  
Obtained funding: KMF, IAY, RVB.  
Study supervision: KMF, IAY, RVB.

Funding National Health and Medical Research Council (Practitioner Fellowship 1019891 (KMF); Career Development Fellowship 1026215 (IAY); Medical PhD Scholarship 631306 (HMM)); Smart State Project Grant, Queensland Health; National Centre for Asbestos Related Diseases Project Grant and The Prince Charles Hospital Foundation.

Competing interests None declared.

Ethics approval The Prince Charles Hospital human research ethics committee.

Provenance and peer review Not commissioned; externally peer reviewed.


Received 24 October 2016  
Revised 12 February 2017  
Accepted 25 February 2017  
Published Online First 22 March 2017

REFERENCES
3 Herschke CI, Yip R, Yankelwitz DF, et al.  


