ERJ Express. Published on November 8, 2018 as doi: 10.1183/13993003.01359-2018

EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original article

Variable radiologic lung nodule evaluation leads to divergent management recommendations

A. Nair, E.C. Bartlett, S.L.F. Walsh, A.U. Wells, N. Navani, G. Hardavella, S. Bhalla, L. Calandriello, A. Devaraj, J.M. Goo, J.S. Klein, H. MacMahon, C. Schaefer-Prokop, J.B. Seo, N. Sverzellati, S.R. Desai

Please cite this article as: Nair A, Bartlett EC, Walsh SLF, *et al*. Variable radiologic lung nodule evaluation leads to divergent management recommendations. *Eur Respir J* 2018; in press (https://doi.org/10.1183/13993003.01359-2018).

This manuscript has recently been accepted for publication in the *European Respiratory Journal*. It is published here in its accepted form prior to copyediting and typesetting by our production team. After these production processes are complete and the authors have approved the resulting proofs, the article will move to the latest issue of the ERJ online.

Copyright ©ERS 2018

Copyright 2018 by the European Respiratory Society.

Variable radiologic lung nodule evaluation leads to divergent management recommendations

A Nair, MD MRCP FRCR*[1]
EC Bartlett, MD *[2]
SLF Walsh, MD [2]
AU Wells, MD [3]
N Navani, MD PhD[4]
G Hardavella, MD [2]
S Bhalla, MD [5]
L Calandriello, MD [6]
A Devaraj, MD [7]
JM Goo, MD [8]
JS Klein, MD [9]
H MacMahon, MD [10]
C Schaefer-Prokop, MD [11]
JB Seo, MD[12]
N Sverzellati, MD [13]
SR Desai, MD FRCP FRCR [2, 7]
& *The Lung Nodule Evaluation Group (Collaborators)*

[1] Department of Radiology, University College London Hospitals NHS Foundation Trust, 235 Euston Rd, Fitzrovia, London NW1 2BU, UK

[2] Department of Radiology, King's College Hospital NHS Foundation Trust, Denmark Hill, London SE5 9RS, UK

[3] Department of Respiratory Medicine, The Royal Brompton Hospital & Harefield NHS Foundation Trust, Sydney Street, London SW3 6NP, UK

[4] Department of Thoracic Medicine, UCLH & Lungs for Living Centre, UCL Respiratory, University College London

[5]Mallinckrodt Institute of Radiology, 216 S Kings Highway Blvd, St. Louis, MO 63110, USA

[6] Radiologia Diagnostica e Interventistica Generale - Fondazione Policlinico Universitario A. Gemelli IRCCS, Italy

[7] Dept of Radiology, The Royal Brompton Hospital & Harefield NHS Foundation Trust, Sydney Street, London SW3 6NP, UK

[8] Seoul National University Hospital, 101 Daehak-ro, Ihwa-dong, Jongno-gu, Seoul, South Korea

[9] The University of Vermont Medical Center, 111 Colchester Avenue, Burlington, Vermont 05401, USA

[10] Department of Radiology, University of Chicago Medical Center, 5841 S Maryland Ave, Chicago, IL 60637, USA

[11] Department of Radiology, Meander Medical Centre, Amersfoot, the Netherlands

[12] Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43 Gil, Songpa-gu, Seoul 138-736, Korea.

[13] Dept of Clinical Sciences, Division of Radiology, University of Parma, Parma

Author for correspondence:
Dr Arjun Nair
Consultant Radiologist
Department of Radiology
University College London Hospitals NHS Foundation Trust
235 Euston Rd
Fitzrovia
London NW1 2BU, UK

Tel: +447904190377
email: arjun7764@gmail.com

**AN and ECB have contributed equally to this manuscript and request acknowledgement as joint first authors.*

"Take home" message (For social media sharing):

Radiologic variations in pulmonary nodule valuation translate into good overall agreement for guideline-modelled management, but poor agreement for radiologists' actual recommendations; on average, 21% of radiologists can disagree with majority decisions.

ABSTRACT

Radiologic evaluation of incidentally-detected lung nodules on computed tomography (CT) influences management. We assessed international radiologic variation in (1) pulmonary nodule characterization; (2) hypothetical guideline-derived management, and (3) radiologists' management recommendations.

107 radiologists from 25 countries evaluated 69 CT-detected nodules, recording: i) first-choice composition (solid, part-solid [PSN] or ground-glass [GGN], with percentage confidence), ii) morphologic features, iii) dimensions, iv) recommended management and (v) decision-influencing factors. We modelled hypothetical management decisions on: (1) the 2005 and (2) updated 2017 Fleischner Society, and (3) liberal and (4) parsimonious interpretations of the British Thoracic Society 2015 guidelines.

Overall agreement for first-choice nodule composition was good (Fleiss' k=0.65), but poorest for PSNs (weighted Kappa, Kw=0.62; IQR=0.50-0.71). Morphologic variables, including spiculation (k=0.35), showed poor-to-moderate agreement (k=0.23-0.53). Diameter variation was greatest at key thresholds (5mm and 6mm). Agreement for radiologists' recommendations was poor (k=0.30); 21% disagreed with the majority. Although agreement within the four guideline-modelled management strategies was good (k=0.63-0.73), 5-10% of radiologists would disagree with majority decisions if they strictly applied guidelines.

Agreement was lowest for part-solid nodules, while significant measurement variation exists at important size thresholds. These variations resulted in generally good agreement for

guideline-modelled management, but poor agreement for radiologists' actual recommendations.

Funding

This work was partly undertaken at University College London Hospitals (UCLH) / University College London (UCL), by AN and NN who received a proportion of funding from the Department of Health NIHR Biomedical Research Centres Funding Scheme (AN/NN).

INTRODUCTION

Pulmonary nodules detected on thoracic computed tomography (CT), either incidentally [1] or at screening [2, 3], are common and the majority are benign. However, sub-solid nodules (encompassing part-solid and pure ground-glass lesions), are associated with a higher likelihood of malignancy [4, 5]. Published guidelines aim to standardize the imaging follow-up and management [6-9] with initial decisions governed by classification of nodule morphology (i.e. solid versus part-solid versus pure ground-glass) and size (either the maximum diameter for a solid nodules or the solid component of part-solid lesions) [6-10].

Decisions about the management of nodules, especially larger nodules (typically those ≥ 8 mm in maximum diameter) usually follow multidisciplinary team (MDT) discussion; this is especially true in tertiary referral or academic centres where such expertise is available. However, in non-specialist centres, physicians do not rigorously adhere to existing guidelines [11], and the follow-up of nodules are likely, to a greater or lesser degree, to be influenced by the morphological description and recommendations in the radiological report. Previous studies indicate that, while the overall agreement for the characterisation

of pulmonary nodules is fair-to-good, there is not inconsiderable variation [12-14]. However, these have been limited either by the small numbers of radiologists or because of cohort enrichment, in which the original morphological categorisations (as provided by participating radiologists or from pre-annotated lung cancer screening cases), have been used to define a reference standard. Moreover, earlier studies have not formally investigated variation in radiologist's recommendation for further nodule management.

In recognising the importance of international collaboration for enhancing the relevance of research [15], the aim of the present study was to quantify inter-observer agreement and the magnitude of variation among a large international cohort of thoracic radiologists for (1) nodule morphology and dimension; (2) hypothetical guideline-derived management decisions; and (3) the recommended management options chosen by radiologists.

MATERIALS AND METHODS

Case selection

The baseline CT studies of 64 consecutive patients from routine clinical practice (36 female; median age 66.5 years [range 26 to 90 years]), with 70 nodules reviewed by the Pulmonary Nodule Management Service (PNS) at (BLINDED FOR ANONYMITY), were retrospectively identified for the study. All patients with incidentally-detected lung nodules on CT are referred to the PNS for decisions regarding discharge, further follow-up (based on risk factors, size and composition) to be made. Two patients had three nodules each, another two patients had two nodules each, and the remaining 60 patients each had a single nodule under follow-up. For the purposes of this study, each nodule was considered a

separate case; therefore, radiologists were instructed that there was only one nodule per case and that there were 70 nodules in 70 cases to be evaluated. While the upper maximum diameter for a pulmonary nodule is arbitrarily set at 3 cm [16], for the purposes of simplicity, there was no upper diameter threshold in this study. For the retrospective review of clinically indicated data in this fashion, the institutional review board waived the need for ethical approval and informed consent.

Image acquisition, reconstruction and presentation (see also Supplementary Appendix)

For each case, anonymized targeted contiguous CT slices, extracted from the original study and displaying the whole nodule on both lung and soft tissue reconstructions and conforming to current practice, were saved as Digital Imaging and Communication in Medicine (DICOM) datasets. To avoid confusion, each nodule was marked on image sections with a red arrow (**Figure 1**).

In each case, the clinical details comprising age, sex, pack-year smoking status, history of lung cancer (or other malignancy), and the presence/absence of emphysema on CT were provided.

Observer Invitation and Registration

Members of The Fleischner Society, Society of Thoracic Radiology, European Society of Thoracic Imaging, British Society of Thoracic Imaging and the Korean Society of Thoracic Radiology were invited to participate via email. A bespoke secure website was created for registration (as detailed in the **Supplementary Appendix**) and image evaluation. Following registration, radiologists securely downloaded a zipped file containing the DICOM datasets of all 70 cases.

Nodule Characterization – Categorization and Measurement of Dimensions

To mirror 'real life' practice, radiologists were free to review cases on any viewer or Picture Archiving and Communications System (PACS), alter window settings and orientation, and use electronic callipers / magnification tools. For each nodule, radiologists characterized nodule composition, size and morphology (**Table 1).** Radiologists assigned a percentage level of confidence to each category of nodule composition (GGN, PSN and solid). For a given nodule, the category with the highest confidence level was taken as a radiologist's first-choice categorisation.

Nodule Management Recommendations

Management Based on Extant Guidelines

To assess the impact of nodule characterization and measurement on inter-observer agreement for initial and subsequent management of nodules, assuming adherence to existing pulmonary nodule management guidelines, we modelled the management decisions for each nodule and observer based on characterization of that nodule in four scenarios (**Table 2):**

1) the Fleischner Society 2005 [6] and 2013 [7] recommendations for the management of pulmonary nodules (FG1);
2) the updated Flesichner Society 2017 [9] recommendations for the management of pulmonary nodules (FG2);
3) a liberal interpretation of the British Thoracic Society (BTS) 2015 guidelines for the management of pulmonary nodules [8], where any suspicious morphological features warrant further investigation (BTS1); and

4) a parsimonious interpretation of the British Thoracic Society 2015 guidelines where suspicious morphological features were not considered as warranting further investigation (BTS2) (**Table 2**).

We modelled both liberal and parsimonious interpretations of the BTS 2015 guidelines because the designation of morphological features that are suspicious ((lobulation, a concave notch in the nodule margin, bubblelike lucencies, air bronchograms, and pleural indentation or tags) can be subjective, and as such we wished to assess the effect of regarding any such feature as suspicious (in BTS1) or disregarding any of them entirely (in BTS2).

We assigned one of three management strategies to each nodule/observer combination as follows:

i) no further action,

ii) CT surveillance, or

iii) further investigation (including recommendations for PET-CT and invasive work-up).

Further details of our modelling are provided in the **Supplementary Appendix**.

Management Recommendations by Radiologists

Following nodule characterization, radiologists were asked to make recommendations on management (**Table 1, section C**) as they would in usual practice. Specifically, radiologists could select

i) no further action,

ii) CT surveillance,

iii) proceeding directly to invasive management, or

iv) a hybrid strategy comprising CT surveillance followed by invasive management (as indicated by free text responses).

Radiologists were not specifically required to follow any published guideline in their decision-making.

Statistical analysis

We measured overall unweighted inter-observer agreement for nodule composition and binary morphological variables with multirater Fleiss κ statistics, and pairwise weighted inter-observer agreement for nodule composition with Cohen's weighted- kappa coefficient values (κ_w). Weighting was assigned by converting the percentage confidence level for nodule composition into a four point (0-3) scale as follows: 0, not confident; 1, low confidence (5–25%); 2, intermediate confidence (30–65%) and 3, for high confidence (≥70%). κw is expressed as median and interquartile range (IQR) values for all 5671 unique observer pair-combinations per nodule. We categorized Fleiss κ and Cohen's κ_w as poor ($0<\kappa/\kappa_w \leq 0{\cdot}20$), fair ($0{\cdot}20<\kappa/\kappa_w \leq 0{\cdot}40$), moderate ($0{\cdot}40<\kappa/\kappa_w \leq 0{\cdot}60$), good ($0{\cdot}60<\kappa/\kappa_w \leq 0{\cdot}80$), and excellent ($0{\cdot}80<\kappa/\kappa_w \leq 1{\cdot}00$).

For each nodule, we calculated multiple classification and measurement metrics, and assessed relevant relationships between these measurements. Details are provided in the **Supplementary Appendix**.

We analysed inter-observer agreement on management strategies as chosen by the radiologists and within each modelled guideline using multirater Fleiss κ. For the modelled guideline evaluation, we determined the majority decision for each nodule, and calculated

the number and percentage of nodules where there would be disagreement on management even when radiologists agreed on the nodule classification.

Statistical analysis was performed using MedCalc for Windows, version 14.12.0 (MedCalc Software, Ostend, Belgium) and Stata. Multirater Fleiss κ was generated using the Real Statistics Resource Pack (version 5.1,Charles Zaiontz available at www.real-statistics.com). A p value of less than 0.05 was considered statistically significant.

RESULTS

Baseline Data

Of 70 nodules, one was excluded from analysis as some radiologists wished to record a maximum diameter that exceeded the available drop-down options provided on the website. One hundred and seven radiologists (91 attendings/consultants; median experience, 9 years [range, 1-50 years]) from 25 countries (UK/Europe, n=75; North & South America/Canada, n=24; Asia and Africa, n=8) completed the study.

Agreement and diagnostic confidence for nodule characteristics

There was good overall agreement for first-choice categorization (κ=0.652, 95% confidence intervals [CI] 0.650, 0.655) (**Table 3**). However, the weighted agreement for first-choice categorization was best for solid nodules, followed by GGNs, and worst for PSNs. Inter-observer agreement for nodule composition was neither affected by the duration of thoracic observer experience nor by attendance at a lung cancer or a dedicated pulmonary nodule multidisciplinary team meeting (**Supplementary Tables S2 and S3**). The overall median percentage of radiologists disagreeing with the predominant classification was

12.1% (IQR, 2.8%- 30.8%), and was highest amongst nodules predominantly classified as PSNs, followed by GGNs, and lowest for solid nodules.

For all nodules, >50% of the 107 radiologists designated their diagnostic confidence for their first-choice categorization as high (≥70%), with a median of 94/107 (87.9%) of radiologists. However, the number of radiologists with this high level of confidence varied widely, ranging from 55 (51.4%) to 107 (100%) radiologists. The variation in first-choice diagnostic confidence was narrow, with a median coefficient of variation (CoV) of 9.7% (IQR, 5.5%-13.2%).

The multirater Fleiss κ for the presence of the other morphological variables was poor-to-moderate (**Table 3**). In particular, the agreement on spiculation was only fair (κ= 0.365). In 13/69 (18.8%) nodules, all 107 radiologists agreed that spiculation was absent, but there was no nodule in which all 107 radiologists agreed that spiculation was present.

Agreement for nodule size

The median longest nodule diameter measurement was 7.8mm (IQR, 5.9-13.4mm). There were 15 (21.7%), 30 (43.5%), 18 (26.1%) and six (8.7%) nodules with median longest diameters of ≤ 5mm, between 5mm and 10mm, between 10mm and 20mm, and >20mm, respectively. The variation in diameter measurement for each nodule, in these categories, is shown in **Figure 2**.

The median CoV for longest diameter was 15.7% (IQR 12.9%-18.7%). When analysed at clinically relevant median diameter thresholds of 5mm, 6mm, 8mm, 10mm and 20mm,

there were greater magnitudes of variation for smaller (i.e. below the threshold) compared to larger (i.e. above the threshold) nodules at the 5mm and 6mm thresholds, but not at higher thresholds (**Table 4**). Paradoxically, the six nodules with median diameters of >20mm demonstrated (non-statistically significant) greater CoVs than the 63 nodules ≤20mm (18.1% vs 15.6% respectively) (p=0.45).

Across 2091 'nodule-observer' combinations, where radiologists classified a nodule as PSN, the median of the diameter and proportions of the solid component were 14.0mm (IQR, 9.0-19.0mm) and 40.0% (IQR, 20.0-60.0%), respectively, with wide variation (median CoVs of 35.6% [21.7%-44.3%] and 42.5% [27.8%-59.9%) respectively]). In the 18 nodules predominantly classified as PSN, the median of the diameter and proportions of the solid component were 6.5mm (IQR, 4.0-11.0mm) and 32.5% (IQR, 20.0%-50.0%), respectively.

Relationships between nodule size, diagnostic confidence and magnitude of variation are provided in **Table 5.** In brief, there was (1) a moderate negative correlation between the median and CoV of longest diameter (r_s = -0.52, p=0.005) for the 28 majority-classified solid nodules; and (2) a strong negative correlation between the median and CoV of the solid component diameter (r_s = -0.69, p=0.0015) and between the median and CoV of the solid proportion (r_s =- 0.88, p<0.0001) for the 18 majority-classified PSNs. In other words, inter-observer variation increased with decreasing size for both solid nodules and the solid component of PSNs.

Nodule Management Recommendations

Management based on extant guidelines

The overall agreement, majority decisions, and number and proportions that would have hypothetically disagreed with the majority of nodule management decisions as defined by the four modelled guidelines is shown in **Table 6.** All showed good agreement. Applying the latest Fleischner Society recommendations uniformly to this nodule cohort would have resulted in the lowest proportion of nodules requiring further investigation. Almost 1 in 20 radiologists would disagree with the majority management decision according to the more conservative model of the BTS guideline (BTS2), and this rises to almost 1 in 10 radiologists with the latest Fleischner Society recommendations.

Even when radiologists agreed with the majority classification of the nodule, there would have been disagreement on the need for further action in 12/69 (17.4%), 26/69 (37.7%), 31/69 (44.9%) and 31/69 (44.9%) nodules using FG1, FG2, BTS1 and BTS2, respectively. Of those nodules where there was disagreement, 0/12 (0.0%), 8/26 (30.8%), 15/31 (48.4%) and 8/31 (25.8%) would have had variation in the type of management (CT surveillance versus further investigation) using FG1, FG2, BTS1 and BTS2 respectively.

Management Recommendations by Radiologists

The overall agreement for the choice between the four management strategies that radiologists could select was only fair (κ=0.296, 95% CI 0.294, 0.298). In all cases, further investigation was recommended by at least one radiologist. The majority chose no further action in six (8.7%), CT surveillance in 57 (82.6%), and proceeding directly to invasive management in six (8.7%) nodules. A median of 85 (IQR, 72-102) radiologists (79.4%, IQR,

67.1%-95.3%) agreed on the majority decision, and a-median of 22 (IQR, 9-35) radiologists (20.6%; IQR, 8.4%-32.9%) disagreed with the majority decision. Among radiologists disagreeing with the majority decision, the overall median percentages of these observers recommending under- or over-management, relative to the majority decision, was 50% (IQR, 0-100%) for both; therefore, there was no overall trend towards under- or over-management amongst disagreeing radiologists. However, there were 34 (49.2%) nodules in which the number of radiologists agreeing with the majority decision was below the median of 85; in other words, the majority decision in these nodules could be considered less compelling. In this subgroup of nodules, there was a tendency for the disagreeing radiologists to over-manage (median percentage 77.6%, IQR 0-100%) rather than under-manage (22.4%, IQR 0-100%) such nodules, but this difference was not statistically significant (Wilcoxon test for paired samples, $p=0.86$).

For 16/69 (23.2%) nodules, all 107 radiologists agreed on the need for further action. For these nodules, the median (and % of total) observers choosing CT follow-up, proceeding directly to invasive management, or a hybrid strategy comprising CT and invasive management (as indicated by free text responses) were 77 (72.0%), 31 (29.0%), and 3 (2.8%) radiologists, respectively. Overall agreement for management, and for PET-CT, in these 16 nodules, was also only fair ($\kappa=0.33$ and $\kappa=0.32$, respectively).

Nodule composition was considered the dominant factor in determining management by the majority of observers in the majority of nodules (65/69, 94.2%). However, for a given nodule, there was a wide variation in the proportion of radiologists citing nodule composition as the most influential factor, with a median of 64/107 (59.8%)

observers [range 35(32.7%) to 92 (86.0%) radiologists)]. In 4/69 (5.8%) nodules, overall diameter was considered most important.

DISCUSSION

The present study is the largest international evaluation of observer agreement for nodule characterization and, to the best of our knowledge, the only study to have specifically evaluated variation for self-reported management recommendations. Our results show that agreement was lowest for part-solid nodules but that variation is not influenced by observer experience. We have also shown that around one in five radiologists disagreed with the part-solid classification assigned by the majority of radiologists. Moreover, there was an inverse relationship between the magnitude of variation and median diameters as well as proportions of the solid component in part-solid nodules. With respect to radiologists' management recommendations, the overall agreement was only fair and ~20% of radiologists disagreed with the majority decision. These findings are important not only for highlighting difficulties in nodule characterization but also given current issues regarding the management of adenocarcinomas with a small solid component (measuring ≤5 mm) [17, 18], which may mandate more conservative management.

The overall variation for nodule classification in our study mirrors that in the study by Ridge et al [14] for differentiating solid from sub-solid nodules but are higher than for screen-detected nodules reviewed by van Riel and colleagues [13]. The differences between studies may, in part, be attributable to variations in the relative prevalence of nodule types included in the studies. However, by requiring radiologists to assign a level of diagnostic

confidence to their chosen categorization, we were able to weight categorizations to emphasize differences in diagnostic confidence and, thus, arguably obtain a more accurate representation of inter-observer agreement. Furthermore, we analysed variation at the important diameter thresholds which dictate either different nodule management recommendations (i.e. in the case of 5mm, 6mm and 8mm) or tumour stage (in the case of 10mm and 20mm), should a nodule prove to be a primary lung cancer [19]. We found that variation in maximum nodule diameter was greatest for nodules at or below 5mm and 6mm, the key diameter thresholds (based on current guidelines [8, 9]), governing whether discharge or further CT surveillance should be recommended. The problem of inter-observer variation for diameter measurement of small pulmonary nodules using electronic callipers is not new [20, 21]. Variation in measurement at the specified diameter thresholds for follow-up potentially translates into variation in recommendations for further follow-up, even though radiologists may agree on the nodule composition. Indeed, in our study over one-third of nodules would have had a variation in recommended management strategy, irrespective of the guideline used. Such variation in measurement and recommendations may be reduced by applying volumetric analysis,computer-aided detection tools [22, 23] and potentially artificial intelligence (AI) techniques. However, with respect to the latter, the inter-observer variation in our study sounds an equally important note of caution for AI nodule characterisation tools developed using machine-learning techniques; such tools require datasets annotated by radiologists for their development, and may thus mimick the biases and interobserver variation of their contributing radiologists.

To assess how management recommendations might vary if guidelines were adhered to, we modelled hypothetical management decisions based on four guideline

scenarios. Modelling of this type inherently foregoes inter-observer variation in decision-making by assuming that guidelines will be applied consistently based on nodule composition and diameter, and is thus artificial. However, our findings illustrate that even in such artificial situations, agreement is at best good, and up to almost 1 in 10 radiologists would deviate from the majority management decision.

An important ancillary finding of the present study was the poor-to-moderate agreement for other morphological variables. Our results, derived from a much larger observer pool, contrast with the high agreement for assessing nodule margins, air bronchograms, bubble-like lucencies and pleural tags in the study by Lim et al [24]. In part, this may be because different morphologic features are more difficult to identify in smaller lesions: in our study, around two-thirds of nodules were ≤ 10 mm in diameter. This aside, a salutary finding was that observer variation for the presence / absence of spiculation (Fleiss K, 0.37) was below clinically-acceptable limits. Plainly, there are implications for future modifications to existing guidelines not only because the presence of spiculation is regarded as a more indicative sign of cancer but also because of its importance as a variable in risk calculators [5, 25]. It is also noteworthy that pleural indentation (Fleiss K, 0.53) and 'bubble-like' lucencies (Fleiss K, 0.44) — features regarded as 'suspicious' in the BTS guidelines [8] — were associated, at best, with fair observer agreement. Subjectivity in the interpretation of these features will further add to the variation in management strategies. As we were interested in the effect that these other morphological variables could have on the decisions taken in the BTS guideline, we modelled two versions of this guideline, one mandating further investigation when these suspicious features were considered present, and the other with a conservative approach irrespective of their presence. The conservative

approach resulted in only four less nodules requiring further investigation according to the majority of radiologists, suggesting that the effect of these morphological features on recommended nodule management – at least in our cohort – was limited.

Since modelling of management strategies based on guidelines is artificial, it was essential that we also capture the 'real-life' management decisions preferred by radiologists for each nodule. The inter-observer agreement for these management strategies in practice was only fair, and much lower than that for any of the modelled guidelines, despite all radiologists listing nodule composition or nodule size as the most influential factors in their decision-making. This lower level of agreement reinforces the role of radiologists' subjective judgement in deciding nodule management, despite existing recommendations. A recent study demonstrated that such subjective upgrading of management recommendations using the American College of Radiology Lung CT Screening Reporting and Data System (Lung-RADS) assessment 4X category resulted in higher malignancy detection rates in lung cancer screening [26], also underscores this point. It is also worth noting that, although our nodule dataset mirrored 'real-life' clinical practice, the radiologists in our study still performed their evaluations under 'test' conditions, and the majority were of higher grade (consultant/attending). As such, it is conceivable that 'real world' management variation may be considerably higher.

Our results also have potential ramifications for respiratory physicians interpreting radiological reports and recommendations. In particular, physicians practising in non-specialist centres, may wish to obtain a tertiary centre multidisciplinary opinion on the classification and management of nodules i) that have been classified as PSNs (given the

lower agreement for this classification), ii) that are at the potential diameter thresholds for discharge (i.e. 5mm and 6mm); iii) in which management has been influenced by morphological variables (given the poor-to-moderate agreement for some CT features [e.g. spiculation]), and iv) where the radiological recommendation is discordant with that suggested by guidelines.

The current study has limitations: firstly, for the purposes of management strategies, the nodules were assumed to be stable. In practice, growth of nodules often leads to further investigation; however, our prime intention was to model the management of nodules as first discovered. Also, both the prior and current versions of the Fleischner recommendations advocate the use of the average of short- and long-axis diameter for measurement [6, 27], whereas we have modelled management on the longest diameter only. In support of our approach, a recent survey suggests that, at least for sub-solid nodules, only 9% of radiologists manage nodules based on bi-dimensional measurements [28]. While no such data is available for solid nodules, we suggest that our modelling reflects everyday practice. Finally, we did not instruct radiologists as to the type of reconstruction (lung or soft tissue) or viewing window (lung or mediastinal) on which measurements should be made. Although recommendations in this regard have been made [7, 27], variation amongst radiologists in this measurement practice could also have influenced measurement variation.

In conclusion, in a large international cohort of radiologists, we found that the overall agreement for nodule composition was good, but was lowest for part-solid nodules, and significant magnitudes of variation in diameter measurement exist at important diameter

thresholds. As such, radiologists are not consistent in characterising or measuring pulmonary nodules. These variations resulted in generally good agreement for management modelled on current recommendations, but agreement on self-reported management decisions taken by radiologists is poor.

REFERENCES

1 Gould MK, Tang T, Liu IL et al. Recent Trends in the Identification of Incidental Pulmonary Nodules. *Am J Respir Crit Care Med* 2015;192:1208-14.

2 Henschke CI, McCauley DI, Yankelevitz DF et al. Early Lung Cancer Action Project: overall design and findings from baseline screening. *Lancet* 1999;354:99-105.

3 Diederich S, Wormanns D, Semik M et al. Screening for early lung cancer with low-dose spiral CT: prevalence in 817 asymptomatic smokers. *Radiology* 2002;222:773-81.

4 Henschke CI, Yankelevitz DF, Mirtcheva R et al. CT screening for lung cancer: frequency and significance of part-solid and nonsolid nodules. *AJR Am J Roentgenol* 2002;178:1053-7.

5 McWilliams A, Tammemagi MC, Mayo JR et al. Probability of cancer in pulmonary nodules detected on first screening CT. *NEJM* 2013;369:910-9.

6 MacMahon H, Austin JH, Gamsu G et al. Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society. *Radiology* 2005;237:395-400.

7 Naidich DP, Bankier AA, MacMahon H et al. Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner Society. *Radiology* 2013;266:304-17.

8 Callister ME, Baldwin DR, Akram AR et al. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. *Thorax* 2015;70 Suppl 2:ii1-ii54.

9 MacMahon H, Naidich DP, Goo JM et al. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. *Radiology* 2017;284:228-43.

10 American College of Radiology. Lung CT Screening Reporting and Data System (Lung-RADS). *American College of Radiology* 4.

11 Tanner NT, Aggarwal J, Gould MK et al. Management of Pulmonary Nodules by Community Pulmonologists: A Multicenter Observational Study. *Chest* 2015;148:1405-14.

12 Penn A, Ma M, Chou BB et al. Inter-reader variability when applying the 2013 Fleischner guidelines for potential solitary subsolid lung nodules. *Acta Radiol* 2015;56:1180-6.

13 van Riel SJ, Sanchez CI, Bankier AA et al. Observer Variability for Classification of Pulmonary Nodules on Low-Dose CT Images and Its Effect on Nodule Management. *Radiology* 2015;277:863-71.

14 Ridge CA, Yildirim A, Boiselle PM et al. Differentiating between Subsolid and Solid Pulmonary Nodules at CT: Inter- and Intraobserver Agreement between Experienced Thoracic Radiologists. *Radiology* 2015;150714.

15 Rosenkrantz AB, Parikh U, Duszak R, Jr. Citation Impact of Collaboration in Radiology Research. *J Am Coll Radiol* 2018;15:258-61.

16 Hansell DM, Bankier AA, MacMahon H et al. Fleischner Society: glossary of terms for thoracic imaging. *Radiology* 2008;246:697-722.

17 Lee JH, Park CM, Lee SM et al. Persistent pulmonary subsolid nodules with solid portions of 5 mm or smaller: Their natural course and predictors of interval growth. *Eur Radiol* 2016;26:1529-37.

18 Cohen JG, Reymond E, Lederlin M et al. Differentiating pre- and minimally invasive from invasive adenocarcinoma using CT-features in persistent pulmonary part-solid nodules in Caucasian patients. *Eur J Radiol* 2015;84:738-44.

19 Rami-Porta R, Bolejack V, Crowley J et al. The IASLC Lung Cancer Staging Project: Proposals for the Revisions of the T Descriptors in the Forthcoming Eighth Edition of the TNM Classification for Lung Cancer. *J Thorac Oncol* 2015;10:990-1003.

20 Reeves AP, Biancardi AM, Apanasovich TV et al. The Lung Image Database Consortium (LIDC): a comparison of different size metrics for pulmonary nodule measurements. *Acad Radiol* 2007;14:1475-85.

21 Revel MP, Bissery A, Bienvenu M et al. Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? *Radiology* 2004;231:453-8.

22 Devaraj A, van GB, Nair A et al. Use of Volumetry for Lung Nodule Management: Theory and Practice. *Radiology* 2017;284:630-44.

23 Jeon KN, Goo JM, Lee CH et al. Computer-aided nodule detection and volumetry to reduce variability between radiologists in the interpretation of lung nodules at low-dose screening computed tomography. *Invest Radiol* 2012;47:457-61.

24 Lim HJ, Ahn S, Lee KS et al. Persistent pure ground-glass opacity lung nodules >/= 10 mm in diameter at CT scan: histopathologic comparisons and prognostic implications. *Chest* 2013;144:1291-9.

25 Swensen SJ, Silverstein MD, Ilstrup DM et al. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. *Arch Intern Med* 1997;157:849-55.

26 Chung K, Jacobs C, Scholten ET et al. Lung-RADS Category 4X: Does It Improve Prediction of Malignancy in Subsolid Nodules? *Radiology* 2017;284:264-71.

27 Bankier AA, MacMahon H, Goo JM et al. Recommendations for Measuring Pulmonary Nodules at CT: A Statement from the Fleischner Society. *Radiology* 2017;162894.

28 Mets OM, de Jong PA, Chung K et al. Fleischner recommendations for the management of subsolid pulmonary nodules: high awareness but limited conformance - a survey study. *Eur Radiol* 2016;26:3840-9.

29 Lee SM, Park CM, Goo JM et al. Invasive pulmonary adenocarcinomas versus preinvasive lesions appearing as ground-glass nodules: differentiation by using CT features. *Radiology* 2013;268:265-73.

30 Miyata N, Endo M, Nakajima T et al. High-resolution computed tomography findings of early mucinous adenocarcinomas and their pathologic characteristics in 22 surgically resected cases. *Eur J Radiol* 2015;84:993-7.

31 Zhang Y, Qiang JW, Ye JD et al. High resolution CT in differentiating minimally invasive component in early lung adenocarcinoma. *Lung Cancer* 2014;84:236-41.

TABLES

Table 1. Nodule characterization, scenarios for which management based on extant guidelines were modelled, and management recommendations.

A. Nodule characterization*	
Variable	**Requirements for characterization**
Composition	• Categorize as pure ground-glass [GGN], part-solid [PSN], or solid nodule; • assign a percentage level of confidence, in 5% increments, for each categorization (**Figure 1**)
Diameter	• For whole nodule: Record longest axial diameter (nearest whole mm); and • If first choice-categorization was PSN: also record the maximum diameter (whole mm) of the solid component
Solid proportion	• Visually estimate (to the nearest 5%), the proportion of the nodule deemed solid, as a percentage of the total nodule volume (NB for GGNs and solid nodules, radiologists recorded the solid proportion as 0% and 100% respectively)
Morphologic features	• Record the presence/absence of the following morphologic features regarded as more often indicative of a malignant lesion [24, 29-31]: a) spiculation; b) lobulation; c) a concave notch in the nodule margin; d) 'bubble-like' lucencies; e) air bronchograms; f) vessel touching or passing through the nodule; and g) pleural indentation or tags. (NB: Sample images illustrating each feature were also provided on the website.)
B. Modelled extant guideline scenarios	
FG1	The original Fleischner Society recommendations (FG1) for pulmonary nodule management published in 2005 and 2013 [6, 7].
FG2	The new Fleischner Society recommendations (FG2) published in 2017 [9].
BTS1	A liberal interpretation of the British Thoracic Society (BTS1) 2015 pulmonary nodule guidelines [8] where all suspicious morphological features in the case of nodules classified as PSN or GGN would lead to intervention.

C. Questions on management recommendations*

1. *Does this case require follow-up or invasive management?*

    - Options: " No, I would discharge the patient" OR "Yes"
    - If yes: "CT follow-up" OR "Proceed directly to invasive management"

If "CT follow-up":	*If "Proceed directly to invasive management": choose 1 of the following:*
recommend interval for 1st and 2nd low dose follow-up CT scan (NB: assuming persistent but stable nodule on the 1st follow-up CT))	(i) biopsy followed by surgical management if deemed appropriate following MDT discussion; (ii) direct to surgical resection (no biopsy recommended); (iii) thermal ablation; or (iv) other management*

2. *Would you also request a 2-[18^F]-Fluoro-2-deoxy-D-glucose (FDG) positron emission tomography-CT (PET-CT):* No/Yes

3. *List, in order of descending importance (to the nearest 5%), factors influencing your management recommendations as follows:* (1) nodule composition, (2) overall nodule dimension, (3) dimension of the solid component (for PSN), (4) nodule features other than size, (5) smoking status, and (6) other clinical features.

*All responses were restricted to pre-determined drop-down options on the website, except for the "other management" option if selecting "proceed directly to invasive management".

Table 2. Modelling of guideline-based management decisions. FG1= Fleischner Society 2005 and 2013 guidelines. FG2= Fleischner Society 2017 guidelines. BTS1= British Thoracic Society 2015 guidelines with suspicious morphological features warranting further investigation. BTS2= British Thoracic Society 2015 guidelines where suspicious morphological features were not considered as warranting further investigation. GGN= ground-glass nodule. PSN= part-solid nodule.

	Recommendation		
	No further action/not applicable	**CT surveillance**	**Further investigation**
FG1	Not applicable due to age/extrathoracic cancer/history of lung cancer; GGN or PSN ≤5mm	Solid nodules up to 8mm; GGN>5mm PSN with solid component <5mm	Solid nodules >8mm; PSN with solid component ≥5mm
FG2	Solid nodules <6mm, not in upper lobes and with no spiculation; GGN and PSN <6mm	Solid nodules ≥6mm OR Solid nodules <6mm AND either in upper lobes or spiculated; GGN ≥6mm, OR PSN ≥6mm with a solid component <6mm	Solid nodules >8mm; PSN with solid component ≥6mm
BTS1	All nodules <5mm	Solid nodules ≥5mm and <8mm OR solid nodules ≥8mm but with Brock score<10%; GGN or PSN with no suspicious features[a], solid component<6mm AND Brock score <10%	Solid nodules ≥8mm with Brock score ≥10%; GGN or PSN ≥5mm with suspicious morphological features[a] OR solid component ≥6mm OR Brock score≥10%
BTS2	All nodules <5mm	Solid nodules ≥5mm and <8mm OR solid nodules ≥8mm but with Brock score<10%; GGN or PSN with solid component<6mm AND Brock score <10%	Solid nodules ≥8mm with Brock score ≥10%; GGN or PSN ≥5mm with solid component ≥6mm OR Brock score≥10%

[a]Suspicious morphological features were lobulation, a concave notch in the nodule margin, bubble-like lucencies, air bronchograms, and pleural indentation or tags.

Table 3: Characterization of nodule features by radiologists.

A. Categorization

Nodule type	**No (%) of nodules classified by majority of radiologists**	**Weighted agreement for 1st choice categorization, Kw (median, IQR)**	**Percentage of radiologists disagreeing with majority classification (median, IQR)**
Ground-glass	23 (33.3)	0.73 (0.65-0.80)	15.0 (4.9-32.4)
Part-solid	18 (26.1)	0.62 (0.50-0.71)	20.1 (9.3-38.3)
Solid	28 (40.6)	0.80 (0.72-0.85)	3.7 (1.9-17.8)

B. Morphological variables

Feature	**Fleiss κ (95% confidence intervals)**
Spiculation	0.353 (0.350, 0.356)
Lobulated margin	0.233 (0.230, 0.236)
Concave notch along margin	0.243 (0.240, 0.246)
Bubble-like lucencies	0.492 (0.489, 0.496)
Air bronchograms	0.483 (0.480, 0.486)
Vessel touching or passing through	0.333 (0.330, 0.336)
Pleural indentation/ pleural tag	0.532 (0.529, 0.535)

Table 4: Magnitude in variation of longest nodule diameter at various thresholds. Data for coefficients of variation are medians, with interquartile ranges in parentheses. P values are for the Mann-Whitney U test.

	Coefficient of variation (%) for longest diameter measurement		
Diameter threshold (mm)	**At or below threshold**	**Above threshold**	**p value**
5	17.2 (16.5-20.7)	14.6 (12.1-17.9)	0.0035
6	17.1 (15.0-20.6)	14.7 (12.1-17.8)	0.0107
8	16.5 (14.3-18.7)	13.9 (11.8-19.0)	0.1599
10	15.4 (12.7-17.8)	16.7 (13.2-20.0)	0.3479
20	15.6 (12.9-18.3)	18.1 (13.4-20.6)	0.4496

Table 5. Relationships between nodule size, diagnostic confidence and magnitude of variation (measured by coefficient of variation, COV)

	Spearman's rho (95% Confidence intervals)	**p value**
All nodules (n=69):		
Median longest diameter versus COV of longest diameter	-0.146 (-0.370,0.942)	0.23
COV of longest diameter versus COV of first choice diagnostic confidence	0.531 (0.337,0.682)	<0.001
Nodules classified as solid by majority (n=28)		
Median longest diameter versus COV of longest diameter	-0.516 (-0.745, -0.177)	0.005
Nodules classified as part-solid by majority (n=18)		
Median solid diameter versus COV of solid diameter	-0.690 (-0.875, -0.329)	0.002
Median solid proportion versus COV of solid proportion	-0.875 (-0.953, -0.690)	<0.001

Table 6: Variation in management decisions based on hypothetical decision-making according to the Fleischner Society 2005 and 2013 (FG1) recommendations, the Fleischner Society 2017 (FG2) recommendations, the British Thoracic Society 2015 guidelines with suspicious morphological features warranting further investigation (BTS1) and without such features warranting further investigation (BTS2).

	Modelled Guideline			
	FG1	**FG2**	**BTS1**	**BTS2**
Overall agreement on management, Fleiss κ (95% confidence intervals)	0.658 (0.655, 0.660)	0.643 (0.641, 0.645)	0.672 (0.670, 0.675)	0.731 (0.729, 0.734)
Majority decision, no. of nodules (%)				
No further action/ guideline not applicable*	14 (20.3)	19 (27.5)	18 (26.1)	16 (23.2)
CT surveillance	40 (58.0%)	39 (56.5%)	21 (30.4%)	27 (39.1%)
Further investigation	15 (21.7%)	11 (15.9%)	30 (43.5%)	26 (37.7%)
No. of radiologists disagreeing with majority decision (median, IQR)	7 (2-23)	10 (2-28)	8 (2-27)	5 (2-21)
% of radiologists disagreeing with majority decision (median, IQR)	6.5 (1.8-21.5)	9.3 (1.9-26.2)	7.5 (1.9-25.0)	4.7 (1.6-19.4)

*Guidelines were considered not applicable in FG1 and FG2 for solid nodules in patients age <35 years, with a known extra-thoracic malignancy or previous history of lung cancer.

FIGURE LEGENDS

Figure 1. Pulmonary nodule in the right lower lobe demarcated by a red arrow on the lung reconstruction, viewed on lung windows (A), Soft tissue reconstructions on Soft tissue reconstructions on mediastinal windows were also presented (B). The first-choice classification for this nodule was ground-glass by 33 radiologists (all with ≥60% confidence), part-solid (PSN) by 71 radiologists (all with ≥55% confidence), and solid by 3 radiologists (all with ≥75% confidence). As sixty-six radiologists chose PSN as the first-choice categorization with ≥70% confidence, this nodule was assigned a majority classification of PSN.

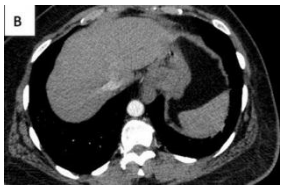
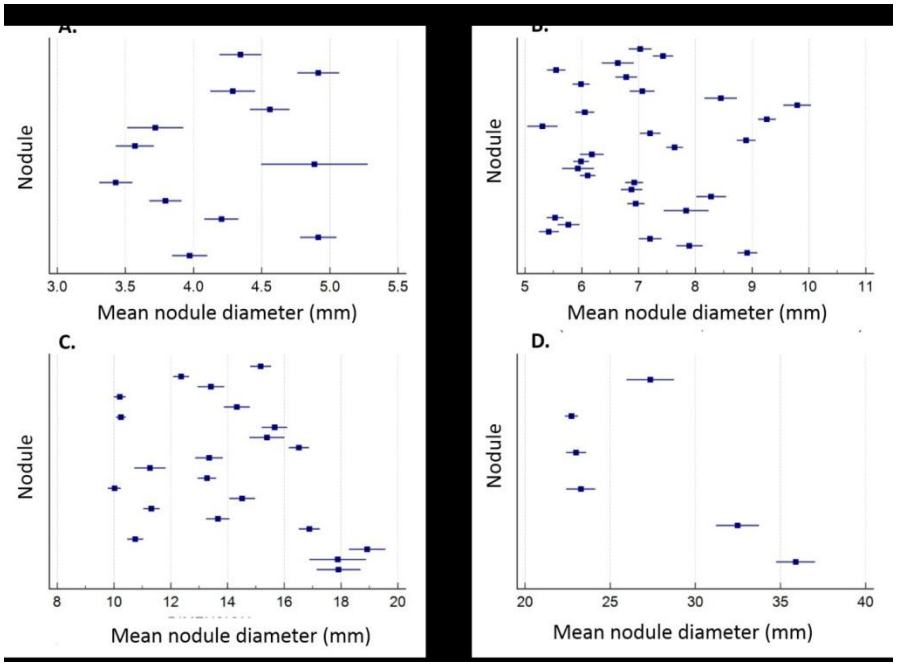


Figure 2. Forest plots depicting the variation in nodule diameters for nodules with mean diameters of ≤5mm (A), >5mm to ≤10mm (B), >10mm to ≤20mm (C) and >20mm (D). Each nodule is represented by a blue horizontal line, with the central square along the line representing the mean diameter, and the two ends of the line representing the lower and upper limits of the 95% confidence interval for nodule mean, respectively.



SUPPLEMENTARY APPENDIX

TABLE OF CONTENTS

Section	Page
CT Acquisition parameters	2
Registration details	3
Supplementary Table S1	4
Statistical analysis	4
Modelling of management recommendations	5
Supplementary Tables S2 and S3	7

CT Acquisition parameters

CT scans of 36/69 and 33/69 nodules were performed with and without contrast enhancement, respectively, during the study period.

All acquisitions were performed on a General Electric multi-slice CT scanner GE Lightspeed VCT 64-slice (General Electric Medical Systems, Milwaukee, Wisconsin, USA) in a craniocaudal direction with the patient in a supine position, and on suspended breathing at full inspiration. The volume and make of contrast material varied during the study period.

Acquisition parameters were as follows: tube current potential 100 kVp; reference tube current-time product 100-700 mAs (for contrast-enhanced CT) and 50-500 mAs (for unenhanced CT), with manufacturer's automated dose modulation; rotation time 500ms.

Contrast-enhanced acquisitions were reconstructed using a 512 x 512 matrix, display field of view 300-400mm on a low frequency (soft tissue) algorithm at 2mm slice thickness, 2mm increment and high frequency (lung) algorithm at 1.25mm slice thickness, 1.25 increment. Unenhanced acquisitions were reconstructed on a low frequency (soft tissue) algorithm at either 1.25mm slice thickness and increment or 0.6mm slicke thickness and increment, and high frequency (lung) algorithm at 1.25mm slice thickness and increment.

Registration details

At registration, participating radiologists provided the following details:

i) Institution and country of employment;

ii) Seniority (attending/consultant versus trainee);

iii) Length of thoracic radiology experience (years);

iv) Membership(s), if any, of a specialist thoracic imaging societies;

v) Regularity of attendance, if applicable, at a lung cancer and/or a dedicated pulmonary nodule multidisciplinary team (MDT) meeting;

vi) the pulmonary nodule management guidelines followed; and

vii) the imaging platform / software used for nodule review.

The full list of participating radiologists is provided in **Supplementary Table 1.**

Supplementary Table S1. List of 107 participating radiologists.

BLINDED

Statistical analysis

For each nodule, we calculated:

i) The predominant first choice classification for nodule composition (GGN, PSN or solid);

ii) The proportion of radiologists disagreeing with the predominant classification, expressed as a percentage (we considered an observer to have agreed with this predominant classification if they had assigned a diagnostic confidence 70% or greater for the first choice, indicating high confidence [32]);

iii) The median and interquartile range (IQR) of the longest diameter, the diameter of the solid component, the solid proportion (where applicable), and of the percentage confidence assigned to first choice classification, assuming non-parametric distributions of these parameters;

iv) The magnitude of variation of the parameters in (iii), expressed as the Coefficient of Variation (CoV) per nodule, defined as: (within-nodule standard deviation ÷ mean) x 100.

Across all nodules, we then calculated the median and IQR of the CoV for each parameter. We also analyzed any differences in CoV of longest diameter at clinically important median diameter thresholds of 5mm, 6mm, 8mm, 10mm and 20mm, using the Mann-Whitney U test [21]. We explored whether the overall diameter of the nodule, and of its solid component where relevant, were associated with the COV in these measurements and in

first choice diagnostic confidence using Spearman's rank correlation (r_s), across all nodules and also subgrouping by predominant nodule classification.

Modelling of management recommendations

The decision to model the Fleischner Society 2005 and 2013 recommendations (REFS) on pulmonary nodule management and the British Thoracic Society 2015 recommendations were based on the fact that the majority of radiologists stated that they most adhered to these guidelines (83/107 and 18/107 radiologists, respectively). During the evaluation period, the updated Fleischner Society recommendations (REFS) had not been published, but to ensure contemporary relevance of these results the updated recommendations were also modelled.

The Fleischner 2005, 2013 and 2017 recommendations were only considered applicable to cases for patients aged 35 years or older, and where no active extrapulmonary cancer or history of lung cancer was known. If the guidelines could be applied, the high-risk options from the Fleischner 2005 and 2017 recommendations were used to make management recommendations at the different diameter thresholds, since all cases in this nodule cohort were high-risk based on their clinical history, and in particular their smoking history.

For the British Thoracic Society (BTS) 2015 pulmonary nodule guidelines, the guideline could be considered applicable to all cases, since the guideline is applicable regardless of route of presentation. For all solid nodules 8mm or greater, and subsolid nodules 5mm or greater, the Brock score (available at http://www.brocku.ca/lung-cancer-risk-calculator) was calculated to aid risk stratification. Family history of lung cancer, a

variable required for Brock score calculation, was not documented in clinical case notes, and was thus assumed to be negative for the purposes of Brock score calculation.

Management decisions based on four guideline models were then assigned as defined in **Table 2** of the main manuscript.

Supplementary Table S2. Weighted Interobserver agreement on nodule composition according to observer experience. Values are median weighted Cohen's kappa coefficients, with interquartile ranges in parentheses.

	≤ 10 years' experience	**> 10 years' experience**
Solid	0.80 (0.72-0.86)	0.80 (0.72-0.85)
Pure ground-glass nodule	0.74 (0.66-0.81)	0.73 (0.65-0.79)
Part-solid nodule	0.62 (0.51-0.70)	0.59 (0.48-0.69)

Supplementary Table S3. Overall interobserver agreement (multirater Fleiss κ, with 95% confidence intervals) for nodule composition according to attendance at lung cancer multidisciplinary team meeting (MDTM) and dedicated pulmonary nodule meeting.

	Yes	**No**
Attends lung MDTM	0.658 (0.655, 0.661)	0.632 (0.623, 0.642)
Attends pulmonary nodule meeting	0.647 (0.639, 0.655)	0.655 (0.652, 0.658)