ORIGINAL ARTICLE

Interobserver agreement for the ATS/ERS/JRS/ALAT criteria for a UIP pattern on CT

Simon L F Walsh,1 Lucio Calandriello,2 Nicola Sverzellati,3 Athol U Wells,4 David M Hansell,5 on behalf of The UIP Observer Consortium

ABSTRACT

Objectives To establish the level of observer variation for the current ATS/ERS/JRS/ALAT criteria for a diagnosis of usual interstitial pneumonia (UIP) on CT among a large group of thoracic radiologists of varying levels of experience.

Materials and methods 112 observers (96 of whom were thoracic radiologists) categorised CTs of 150 consecutive patients with fibrotic lung disease using the ATS/ERS/JRS/ALAT CT criteria for a UIP pattern (3 categories—UIP, possibly UIP and inconsistent with UIP). The presence of honeycombing, traction bronchiectasis and emphysema was also scored using a 3-point scale (definitely present, possibly present, absent). Observer agreement for the UIP categorisation and for the 3 CT patterns in the entire observer group and in subgroups stratified by observer experience, were evaluated.

Results Interobserver agreement across the diagnosis category scores among the 112 observers was moderate, ranging from 0.48 (IQR 0.18) for general radiologists to 0.52 (IQR 0.20) for thoracic radiologists of 10–20 years’ experience. A binary score for UIP versus possible or inconsistent with UIP was examined. Observer agreement for this binary score was only moderate. No significant differences in agreement levels were identified when the CTs were stratified according to multidisciplinary team (MDT) diagnosis or patient age or when observers were categorised according to experience. Observer agreement for each of honeycombing, traction bronchiectasis and emphysema were 0.59±0.12, 0.42±0.15 and 0.43±0.18, respectively.

Conclusions Interobserver agreement for the current ATS/ERS/JRS/ALAT CT criteria for UIP is only moderate among thoracic radiologists, irrespective of their experience, and did not vary with patient age or the MDT diagnosis.

INTRODUCTION

Accurate diagnosis of idiopathic pulmonary fibrosis/usual interstitial pneumonia (IPF/UIP) is essential to ensure prompt initiation of appropriate treatment and enrolment in clinical trials. CT has a key role in making the diagnosis of IPF/UIP. The most recent guidelines published by American Thoracic Society (ATS)/European Respiratory Society (ERS)/Japanese Respiratory Society (JRS)/Latin American Thoracic Association (ALAT) specify the CT appearances of three diagnostic categories: UIP, possible UIP and inconsistent with UIP. Imaging criteria for the diagnosis of UIP include the presence of honeycombing in a basal and subpleural distribution without features considered incompatible with a diagnosis of IPF/UIP. In the correct clinical context, these appearances are considered sufficient to diagnose IPF/UIP without surgical lung biopsy (SLB). If the CT appearances are not those of UIP, the diagnosis of IPF cannot be made on imaging alone. Therefore CT plays a critical role in the evaluation of patients with suspected IPF and once performed, significantly influences subsequent management decisions.

Key messages

What is the key question?

▸ What is the interobserver agreement for the current ATS/ERS/JRS/ALAT CT criteria for UIP among radiologists?

What is the bottom line?

▸ Interobserver agreement among radiologists for the ATS/ERS/JRS/ALAT CT criteria for a UIP pattern on CT is moderate.

Why read on?

▸ CT plays a critical role in the evaluation of patients with suspected idiopathic pulmonary fibrosis and once performed, significantly influences subsequent management decisions.

Received 30 April 2015
Revised 25 September 2015
Accepted 15 October 2015
Published Online First 19 November 2015

Correspondence to
Dr Simon L F Walsh, Department of Radiology, Kings College Hospital Foundation Trust, Denmark Hill, London SE5 9RS, UK; sllfwalsh@gmail.com
METHODS
Case selection and CT protocol
CTs from consecutive patients with a multidisciplinary team (MDT) diagnosis of idiopathic fibrotic lung disease, chronic hypersensitivity pneumonitis (CHP) or fibrotic lung disease associated with a connective tissue disease, diagnosed at the interstitial lung disease unit of the Royal Brompton and Harefield NHS Foundation Trust, UK, between 1 January 2004 and 31 June 2010 were selected. CTs were clinically indicated in all cases and for the purposes of a retrospective examination of these data, informed consent was not required by the institutional review board. Cases with an MDT diagnosis of fibrotic sarcoidosis were excluded from the study as it was considered that these cases might disproportionately increase the number of cases fulfilling the ‘inconsistent with UIP’ CT diagnostic criteria. CTs were performed on a 64-slice multidetector computed tomography (MDCT) (Somatom Sensation 64, Siemens, Erlangen, Germany) or a 4-slice MDCT (Siemens Volume Zoom, Siemens, Erlangen, Germany) in all cases. Images were reconstructed at section thickness of 1.5 mm (4-slice) or 1 mm (64-slice) using a high spatial frequency algorithm. All patients were examined in the supine position from lung apices to lung bases at full-suspended inspiration using standard acquisition parameters: ~90 mA, 120 kVp.

Participating observers
An invitation for observers to participate in the study was approved by the officers of the European Society of Thoracic Imaging (ESTI), Society of Thoracic Radiologists (STR), British Society of Thoracic Imaging (BSTI), the Italian Society of Chest Radiologists (ISCR) and the Korean Society of Thoracic Radiologists (KSTR), and was sent to each society’s membership. Observers were required to provide their specialty (chest radiologist, general radiologist or thoracic radiology fellow) and number of years experience practising in that specialty.

Case distribution and scoring
Scoring of cases was performed in two stages:
1. For the first stage, in order to distribute the cases to a large number of observers from different countries, a web-based image viewing application with 32-bit encrypted password access was used. Fifteen interspaced sections from each anonymised CT were selected. All images were loaded into the viewing application as TIFF (tagged file format) images, uncompressed and with an image resolution of 600 pixels per inch. For each case, observers were asked to assign a diagnosis category score based upon the current ATS/ERS/JRS/ALAT CT criteria for UIP (3 point score—UIP possible UIP; inconsistent with UIP). These guidelines were provided on the web application for reference. The identification of honeycombing is critical in assigning a diagnosis of ‘UIP’ and traction bronchiectasis and emphysema are known to potentially confound this determination. For this reason, observers were also required to score the presence of these three CT patterns (honeycombing, traction bronchiectasis and emphysema) using a 3-point score; definitely present, possibly present or absent (CT pattern score). No definitions for these patterns were given to the observers before participating in the study and no training was given to the observers.
2. For the second stage of the study, two subsets of observers were selected randomly from the participants of first stage of the study—one made up of thoracic radiologists with greater than 20 years experience and one made up of thoracic radiologists with less than 10 years experience. These two groups were given the full volumetric thin-section CT data in digital imaging and communications in medicine (DICOM) format from a new cohort of patients with fibrotic lung disease. As in the first stage of the study, the observers scored the presence of the three CT patterns described above and assigned a diagnosis category based upon the current ATS/ERS/JRS/ALAT CT criteria for UIP.

Statistical analysis
Data are given as means with SDs, medians with IQR, or number of patients and percentage, where appropriate. Statistical analyses were performed using STATAV12 (StataCorp, College Station, Texas). Cohen’s weighted κ coefficient (κw) was used to evaluate interobserver agreement for diagnosis category score and each of the three CT pattern scores. Weighting the κ coefficient allows the degree of disagreement to be quantified by assigning greater emphasis to large differences between scores. Weighted κ coefficients were categorised as follows: poor (0<κ≤0.20), fair (0.20<κw≤0.40), moderate (0.40<κ≤0.60), good (0.60<κw≤0.80) and excellent (0.80<κw≤1.00). As the aim of the study was to evaluate interobserver agreement rather than accuracy, a defined gold standard for the diagnosis and pattern scores was not required.

For binary scores (eg, UIP vs not possible UIP or inconsistent with UIP), interobserver agreement was expressed as an unweighted κ coefficient (κ), expressed as a single figure calculated for multiple observers. For non-binary scores, weighted κ coefficients were calculated for each unique unordered pair of observers and expressed as means with SDs for each high-resolution computed tomography (HRCT) variable.

RESULTS
Patient population and observer groups
For the first stage of the study (scoring of online cases), a total of 472 consecutive patients presenting to the interstitial lung disease unit were identified. From this cohort 92 patients with an MDT diagnosis of fibrotic sarcoidosis were excluded. From the remaining 380 patients, 150 cases were randomly selected. Of these 150 patients, 78 were female. Mean age at the time of CT was 61.5 years (SD=12.2 years), MDT diagnoses of the study group were as follows: Idiopathic fibrotic lung disease (IPF n=34, biopsy proven n=3, fibrotic non-specific interstitial pneumonia (NSIP) n=21, biopsy proven n=3), connective tissue disease related fibrotic lung disease (n=51, biopsy proven n=4) and CHP (n=44, biopsy proven n=12). A total of 112 observers completed the first stage of the study (each scoring all 150 cases). Ninety-six were thoracic radiologists, 16 general radiologists. Thoracic imaging society representation for the 112 radiologists was STR (n=42), ESTI (n=39), Italian society of thoracic radiology (ISTR) (n=15), BSTI (n=12) and KSTR (n=4). Mean experience was 11.9 years (SD=8.5 years) (table 1).

For the second stage of the study (scoring of volumetric thin-section DICOM images), a new cohort of 75 cases of fibrotic lung disease was selected. MDT diagnoses for these cases were as follows: Idiopathic fibrotic lung disease (n=25, biopsy proven n=5), connective tissue disease related fibrotic lung disease (n=25, biopsy proven n=4) and CHP (n=25, biopsy proven n=12). A total of 22 thoracic radiologists completed this stage of the study (<10 years experience n=10, >20 years experience n=12).
Interobserver agreement for scoring of online cases

Diagnosis category scores

For the first stage of the study, interobserver agreement across the diagnosis category scores (UIP, possible UIP, inconsistent with UIP) among the 112 observers was moderate, ranging from 0.48 (IQR 0.18) for general radiologists to 0.52 (IQR 0.20) for thoracic radiologists of 10–20 years experience (table 2, figures 1A–D and 2A–D). At the extremes of the experience spectrum, interobserver agreement among thoracic imaging fellows was 0.50 (IQR 0.10) and for thoracic radiologists of greater than 20 years experience, 0.51 (IQR 0.18). The diagnosis category scores were converted to a binary ‘UIP versus possible UIP or inconsistent with UIP’ score. Mean interobserver agreement for this binary score was moderate ranging from 0.36 for thoracic imaging fellows to 0.42 for thoracic radiologists of less than 10 years experience (table 2). A second analysis was performed to investigate whether patient age, or MDT diagnosis varied with observer agreement for the binary diagnosis score. No significant differences were identified for these subgroups (table 3).

CT pattern scores

Weighted κ values for the presence of honeycomb ing using the 3-point score: definitely present, possibly present or absent, ranged from 0.56 (IQR 0.12) (for thoracic radiologists with more than 20 years experience) to 0.65 (IQR 0.23) (for thoracic radiology fellows) (table 4). Weighted κ values for the 3-point traction bronchiectasis score ranged from 0.32 (IQR 0.25) (for thoracic radiology fellows) to 0.45 (IQR 0.18) (for thoracic radiologists of more than 20 years experience) (table 4).

Interobserver agreement for scoring of volumetric thin-section CT

For the second stage of the study, interobserver agreement across the diagnosis category scores (UIP, possible UIP, inconsistent with UIP) when cases were evaluated on volumetric thin-section CT was for thoracic radiologists with less than 10 years experience, 0.54 (IQR 0.17) and for thoracic radiologists of greater than 20 years experience, 0.40 (IQR 0.12). Interobserver agreement for each of the CT patterns scores is shown in table 5.

DISCUSSION

The key finding in the present study is that interobserver agreement among a large cohort of thoracic radiologists, for the radiological diagnosis of UIP based upon the most recent ATS/ERS/JRS/ALAT guidelines is at best moderate, and is not significantly increased among thoracic radiologists with greater levels of experience.

IPF is a chronic progressive fibrosing interstitial pneumonia, which is characterised by a histopathological and/or a radiological pattern of UIP 1 11 The distinction of IPF/UIP from other chronic fibrosing lung diseases is important because IPF/UIP has a particularly poor prognosis. Diagnosing IPF however, may be challenging because it requires an integrated multidisciplinary approach involving physicians, radiologists and pathologists and indirect evidence suggests that early expert assessment is important. 11 12 Prompt and accurate diagnosis allows commencement of treatment, as well as access to clinical trials and evaluation for lung transplantation. The most recent evidence-based guidelines for the diagnosis and management of IPF represent the collaborative effort of the American Thoracic Society, the European Respiratory Society, the Japanese Respiratory Society and the Latin American Thoracic Society and clearly specifies the CT features which stratify patients into one of three radiologic categories, ‘UIP’, ‘possible UIP’ and ‘inconsistent with UIP’. 13 In the correct clinical context, a radiological diagnosis of ‘UIP’ secures a diagnosis of IPF. 13–16 In patients whose CT diagnosis is ‘possible UIP’ or ‘inconsistent with UIP’, SLB should be considered, although at least one study reports that a diagnosis of ‘possible UIP’ may be sufficient to diagnose IPF/UIP in the proper clinical setting. 10 Therefore CT plays a

Table 1 Observer demographics (n=112) for the first stage of the study (see Methods section)

<table>
<thead>
<tr>
<th>Observer group</th>
<th>n=112</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thoracic radiologist (n=91)</td>
<td></td>
</tr>
<tr>
<td>&gt;20 years experience</td>
<td>22 (19.6%)</td>
</tr>
<tr>
<td>10–20 years experience</td>
<td>27 (24.1%)</td>
</tr>
<tr>
<td>&lt;10 years experience</td>
<td>42 (37.5%)</td>
</tr>
<tr>
<td>General radiologist (n=16)</td>
<td></td>
</tr>
<tr>
<td>&gt;20 years experience</td>
<td>4 (3.5%)</td>
</tr>
<tr>
<td>10–20 years experience</td>
<td>3 (2.7%)</td>
</tr>
<tr>
<td>&lt;10 years experience</td>
<td>9 (8.0%)</td>
</tr>
<tr>
<td>Thoracic imaging fellow (n=5)</td>
<td></td>
</tr>
<tr>
<td>1 year experience</td>
<td>5 (4.4%)</td>
</tr>
</tbody>
</table>

Interobserver agreement for scoring of online cases: Interobserver agreement for the diagnosis categories, ‘UIP’, ‘possible UIP’ and ‘inconsistent with UIP’ expressed as Cohen’s weighted κ coefficient stratified according to observer experience and specialty

<table>
<thead>
<tr>
<th>Interobserver agreement</th>
<th>Means±SD</th>
<th>Median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UIP diagnosis categories</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(UIP, possible UIP, inconsistent with UIP)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thoracic radiology fellows (n=5)</td>
<td>0.47±0.05</td>
<td>0.50 (0.10)</td>
</tr>
<tr>
<td>Thoracic radiologists (experience &lt;10 years, n=42)</td>
<td>0.50±0.12</td>
<td>0.51 (0.16)</td>
</tr>
<tr>
<td>Thoracic radiologists (experience 10–20 years, n=27)</td>
<td>0.51±0.11</td>
<td>0.52 (0.20)</td>
</tr>
<tr>
<td>Thoracic radiologists (experience &gt;20 years, n=22)</td>
<td>0.48±0.14</td>
<td>0.51 (0.18)</td>
</tr>
<tr>
<td>General radiologists (n=16)</td>
<td>0.45±0.13</td>
<td>0.48 (0.18)</td>
</tr>
<tr>
<td>Binary diagnosis score (Typical UIP or Possible UIP/inconsistent with UIP)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thoracic radiology fellows (n=5)</td>
<td>0.36*</td>
<td></td>
</tr>
<tr>
<td>Thoracic radiologists (experience &lt;10 years, n=42)</td>
<td>0.42*</td>
<td></td>
</tr>
<tr>
<td>Thoracic radiologists (experience 10–20 years, n=27)</td>
<td>0.39*</td>
<td></td>
</tr>
<tr>
<td>Thoracic radiologists (experience &gt;20 years, n=22)</td>
<td>0.40*</td>
<td></td>
</tr>
<tr>
<td>General radiologists (n=16)</td>
<td>0.41*</td>
<td></td>
</tr>
</tbody>
</table>

The ‘possible UIP’ and ‘inconsistent with UIP’ categories were combined to generate a binary ‘typical UIP or possible UIP/inconsistent with UIP’ score. Interobserver agreement expressed as Cohen’s weighted κ coefficient for this binary categorisation, stratified according to observer experience and specialty.
* Unweighted κ.
UIP, usual interstitial pneumonia.
critical role in the evaluation of patients with suspected IPF and once performed, significantly influences subsequent management decisions.

Reasonable levels of observer agreement are a requisite for diagnostic criteria to be clinically useful. As up to two-thirds of patients with IPF/UIP are diagnosed based upon CT appearances alone, the issue of interobserver agreement between radiologists for this diagnosis is important.\(^2\)\(^{16}\) Despite this, the interobserver agreement for the most recent ATS/ERS/JRS/ALAT CT criteria for IPF/UIP is not known. In a study by Aziz et al, observer agreement between 11 thoracic radiologists was evaluated for diagnosis of 131 cases of diffuse lung diseases. In this study, agreement for a diagnosis of IPF/UIP was reported as good ($\kappa_w=0.63$).\(^7\) In contrast, a study by Lynch et al,\(^5\) which involved 315 cases of IPF/UIP, reported low levels of observer agreement between two expert thoracic radiologists for a CT pattern considered ‘consistent with IPF’ ($\kappa_w=0.33$). Thomeer et al\(^6\) reported similar, low levels of agreement between three expert thoracic radiologists for a radiological diagnosis of typical UIP ($\kappa_w=0.40$). Most recently, in a paper by Assayag et al,\(^17\) CT appearances of 69 cases of biopsy proven rheumatoid-related interstitial lung were evaluated by two experienced thoracic radiologists applying a binary score of ‘definite UIP’ or ‘not’ (based upon current ATS/ERS/JRS/ALAT CT criteria), and reported a $\kappa$ coefficient of 0.67 for this score. A limitation of these studies is that all, with the exception of one,\(^17\) predate the most recent diagnostic guidelines and therefore may not easily be interpreted in the context of current recommendations. In addition, all employed small numbers of academic radiologists from tertiary referral centres with specific expertise in the evaluation of patients with IPF.\(^3\)\(^6\) In contrast, the results of our study demonstrate that, in a large diverse group of thoracic radiologists, the interobserver agreement for the current CT criteria for a diagnosis of IPF/UIP is only moderate. Furthermore, no significant differences in agreement were demonstrated between observer subgroups of different levels of experience. These results are reinforced by the finding that interobserver agreement was not improved among thoracic radiologists.

**Figure 1** (A) Biopsy proven usual interstitial pneumonia (UIP) in a patient with a multidisciplinary diagnosis of rheumatoid arthritis related fibrotic lung disease. Assigned CT diagnoses expressed as a percentage of 116 observers were: definite UIP 20%, possible UIP 36.5%, inconsistent with UIP 42.6%. In this case 62.6% of observers assigned a grade of definitely present for honeycombing. (B) Biopsy proven UIP in a patient with a multidisciplinary diagnosis of rheumatoid arthritis related fibrotic lung disease. Assigned CT diagnoses expressed as a percentage of 116 observers were: definite UIP 20%, possible UIP 36.5%, inconsistent with UIP 42.6%. In this case 62.6% of observers assigned a grade of definitely present for honeycombing. (C) Biopsy proven UIP in a patient with a multidisciplinary diagnosis of rheumatoid arthritis related fibrotic lung disease. Assigned CT diagnoses expressed as a percentage of 116 observers were: definite UIP 20%, possible UIP 36.5%, inconsistent with UIP 42.6%. In this case 62.6% of observers assigned a grade of definitely present for honeycombing. (D) Biopsy proven UIP in a patient with a multidisciplinary diagnosis of rheumatoid arthritis related fibrotic lung disease. Assigned CT diagnoses expressed as a percentage of 116 observers were: definite UIP 20%, possible UIP 36.5%, inconsistent with UIP 42.6%. In this case 62.6% of observers assigned a grade of definitely present for honeycombing.


Interstitial lung disease
radiologists of greater than 20 years experience, when the study was repeated using thin-section volumetric CT. The rationale for converting the 3-point diagnosis score (UIP, possible UIP, inconsistent with UIP) to a binary score (UIP or possible UIP/inconsistent with UIP) was that based upon current guidelines, the key radiological determination is identifying patients whose CT allows a confident diagnosis of IPF/UIP to be made, therefore avoiding the need for SLB. In addition, this distinction has prognostic implications in the setting of IPF—a typical UIP pattern on CT may confer a worse prognosis than those cases of IPF who present with atypical CT appearances of UIP. The overall level of interobserver agreement for the entire cohort of observers for this binary diagnosis score was moderate, regardless of observer experience, whether the observer was a thoracic radiologist or general radiologist. Recently, Gruden et al reported that in the absence of honeycombing, a heterogeneous pattern of fibrosis on CT may be sufficient to secure a diagnosis of UIP. In our study, a second binary score of ‘UIP and possible UIP’ versus ‘inconsistent with UIP’ was generated and interobserver agreement for this categorisation was also

Table 3 Scoring of online cases: The ‘possible UIP’ and ‘inconsistent with UIP’ categories were combined to generate a binary ‘typical UIP or possible UIP/inconsistent with UIP’ score

<table>
<thead>
<tr>
<th>Disease subgroups</th>
<th>Interobserver agreement*</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDT diagnosis—CHP (n=44)</td>
<td>0.39</td>
</tr>
<tr>
<td>MDT diagnosis—CTD-ILD (n=51)</td>
<td>0.35</td>
</tr>
<tr>
<td>MDT diagnosis—Idiopathic FLD (n=55)</td>
<td>0.38</td>
</tr>
<tr>
<td>Age subgroups</td>
<td></td>
</tr>
<tr>
<td>Patient age &lt;50 years (n=25)</td>
<td>0.31</td>
</tr>
<tr>
<td>Patient age 50–60 years (n=40)</td>
<td>0.39</td>
</tr>
<tr>
<td>Patient age 60–70 years (n=48)</td>
<td>0.42</td>
</tr>
<tr>
<td>Patient age &gt;70 years (n=37)</td>
<td>0.40</td>
</tr>
</tbody>
</table>

*Interobserver agreement expressed as Cohen’s κ coefficient for this binary categorisation stratified according to multidisciplinary diagnosis and patient age. 

CHP, chronic hypersensitivity pneumonitis; CTD-ILD, connective tissue disease related interstitial lung disease; FLD, fibrotic lung disease; MDT, multidisciplinary team; UIP, usual interstitial pneumonia.
moderate. Given the results of the current study, it might be that the best use of HRCT in this setting may be as the starting point for a logical follow-up study. Third, in common with other studies of interobserver agreement, we did not use an independent ‘gold standard’ against which observers’ scores were evaluated.

The primary goal of the study was to quantify levels of observer agreement for a CT diagnosis of UIP based upon current ATS/ERS/JRS/ALAT CT criteria rather than accuracy of CT diagnosis, which is a separate, albeit related, issue. Fourth, our patient population was a selection of consecutive cases of fibrotic lung disease referred to our interstitial lung disease unit. We excluded cases of fibrotic sarcoidosis on the basis that this diagnosis is in most cases straightforward and because inclusion of these patients might disproportionately increase the number of cases with an ‘inconsistent with UIP’ CT diagnosis. The usual diagnostic difficulty encountered on CT, in the context of fibrotic lung disease, is the separation of patients with IPF/UIP from those with non-classical CT appearances alone in approximately 50% of cases. Furthermore, when IPF/UIP presents with non-classical CT appearances, the usual alternative diagnoses are fibrotic NSIP or CHP. Fifth, we did not confine cases to those with a biopsy proven diagnosis, as this would effectively eliminate patients with typical UIP features.
We had no control over the observers who participated in the study. An open invitation was made to one national (BSTI) and four international (ESTI, STR, KSTR, ISCR) thoracic imaging societies without any exclusion criteria. However, in order to evaluate the performance of general thoracic radiologists who routinely provide opinions on CT studies, our approach for enrolling observers was necessarily broad. This is in contrast to most previous studies where observers are preselected because of their expertise, which conceivably may reduce their clinical applicability to the general thoracic radiologist population. As it has been suggested by some that the current ATS/ERS/JRS/ALAT guidelines require expertise that may not always be available, evaluating the performance of thoracic radiologists of varying levels of expertise is clinically important.

In conclusion, we have demonstrated in a large number of thoracic radiologists, of varying levels of experience, that interobserver agreement for the most recent ATS/ERS/JRS/thoracic radiologists, of varying levels of expertise is clinically important. This is in contrast to most previous studies where observers are preselected because of their expertise, which conceivably may reduce their clinical applicability to the general thoracic radiologist population. As it has been suggested by some that the current ATS/ERS/JRS/ALAT guidelines require expertise that may not always be available, evaluating the performance of thoracic radiologists of varying levels of expertise is clinically important. Based upon the results of this study, modification of these criteria may be necessary to improve observer agreement.


**Funding** This study was supported by the NIHR Respiratory Disease Biomedical Research Unit at the Royal Brompton and Harefield NHS Foundation Trust and Imperial College London. DMH is the recipient of an NIHR Senior Investigator award.

**Competing interests** None declared.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**REFERENCES**