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 Consort

[1]Department of Radiology, Kings College Hospital Foundation Trust, London, UK
[2]Department of Bioimaging and Radiological Sciences, Institute of Radiology, Catholic University, "A. Gemelli" Hospital, Rome, Italy
[3]Department of Clinical Sciences, Section of Radiology, University of Parma, Parma, Italy
[4]Interstitial Lung Diseases Unit, Royal Brompton Hospital, London, UK
[5]Department of Radiology, Royal Brompton Hospital, London, UK

Correspondence to
Dr Simon L F Walsh, Department of Radiology, Kings College Hospital Foundation Trust, Denmark Hill, London SE5 9RS, UK; slfwalsh@gmail.com

Received 30 April 2015
Revised 25 September 2015
Accepted 15 October 2015

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.

Key messages

What is the key question?
- What is the interobserver agreement for the current ATS/ERS/JRS/ALAT CT criteria for usual interstitial pneumonia (UIP) among radiologists?

What is the bottom line?
- Interobserver agreement among radiologists for the ATS/ERS/JRS/ALAT 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.

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.[1 2] 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 patterns.[2] 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).[2–4] 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.

Reasonable levels of observer agreement are a requisite for diagnostic criteria to be clinically useful. Several studies have reported on the observer agreement for a CT diagnosis of IPF/UIP based upon earlier guidelines, with conflicting results.[5–7] However, all of these studies involved expert thoracic radiologists, selectively chosen for their experience in the interpretation of diffuse lung diseases on CT. In contrast, thoracic radiologists, working in non-specialist centres may also be required to provide opinions on CTs of patients with suspected IPF/UIP. Therefore, the aim of this study was to evaluate interobserver agreement for the ATS/ERS/JRS/ALAT CT criteria for UIP among an international group of thoracic radiologists of varying levels of experience. As it has been reported that traction bronchiectasis and emphysema may confound the identification of honeycombing,[8] observer agreement for the presence of honeycombing, traction bronchiectasis and emphysema were also evaluated in this study.

To cite: Walsh SLF, Calandriello L, Sverzellati N, *et al*. *Thorax* Published Online First: [*please include* Day Month Year] doi:10.1136/thoraxjnl-2015-207252

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.[2] 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 STATA V.12 (StataCorp, College Station, Texas). Cohen's weighted κ coefficient (κ_w) 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.[9] Weighted κ coefficients were categorised as follows: poor ($0<\kappa_w\leq0.20$), fair ($0.20<\kappa_w\leq0.40$), moderate ($0.40<\kappa_w\leq0.60$), good ($0.60<\kappa_w\leq0.80$) and excellent ($0.80<\kappa_w\leq1.00$). [9] 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=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).

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

Observer group	n=112
Thoracic radiologist (n=91)	
>20 years experience	22 (19.6%)
10–20 years experience	27 (24.1%)
<10 years experience	42 (37.5%)
General radiologist (n=16)	
>20 years experience	4 (3.5%)
10–20 years experience	3 (2.7%)
<10 years experience	9 (8.0%)
Thoracic imaging fellow (n=5)	
1 year experience	5 (4.4%)

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). Emerging reports suggest that SLB might not be required in patients with a 'possibly UIP' pattern on CT.[10] Therefore, a second binary score ('UIP or possible UIP' vs inconsistent with UIP) was also evaluated. Interobserver agreement for this distinction was also moderate ranging from 0.39 for thoracic imaging fellows to 0.45 for thoracic radiologists of greater than 20 years experience.

Table 2 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

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

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 κ coefficient for this binary categorisation, stratified according to observer experience and specialty.
*Unweighted κ.
UIP, usual interstitial pneumonia.

CT pattern scores

Weighted κ values for the presence of honeycombing 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). Weighted κ values for the 3-point emphysema score ranged from 0.40 (IQR 0.13) (for thoracic radiology fellows) to 0.55 (IQR 0.22) (for thoracic radiologists of less than 10 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.[1 2 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'.[2] 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

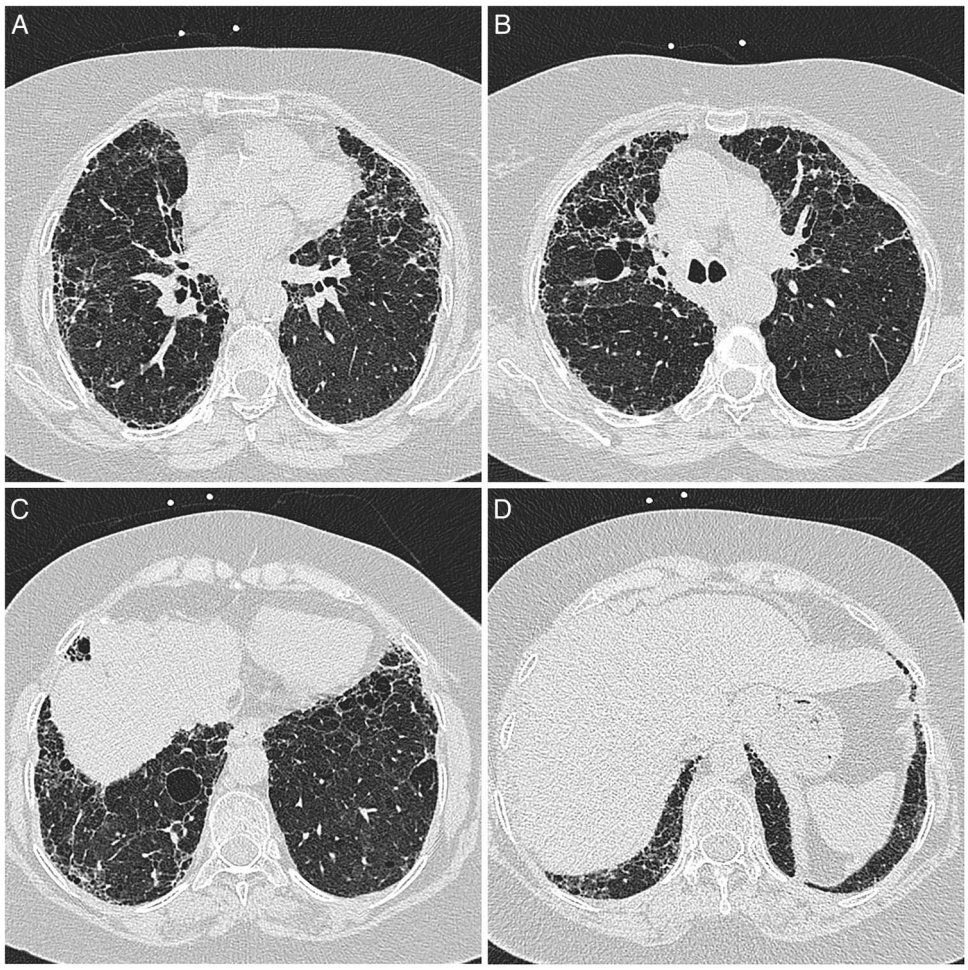

Figure 1 (A) Biopsy proven usual interstitial pneumonia (UIP) in a patient with a multidisciplinary diagnosis of rheumatoid arthritis related fibrotic lung disease. Assigned CT diagnoses expressed as a percentage of 116 observers were: definite UIP 20%, possible UIP 36.5%, inconsistent with UIP 42.6%. 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.

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 3 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 different diffuse lung diseases. In this study, agreement for a diagnosis of IPF/UIP was reported as good (κ_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' (κ_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 (κ_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 κ 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.[5 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

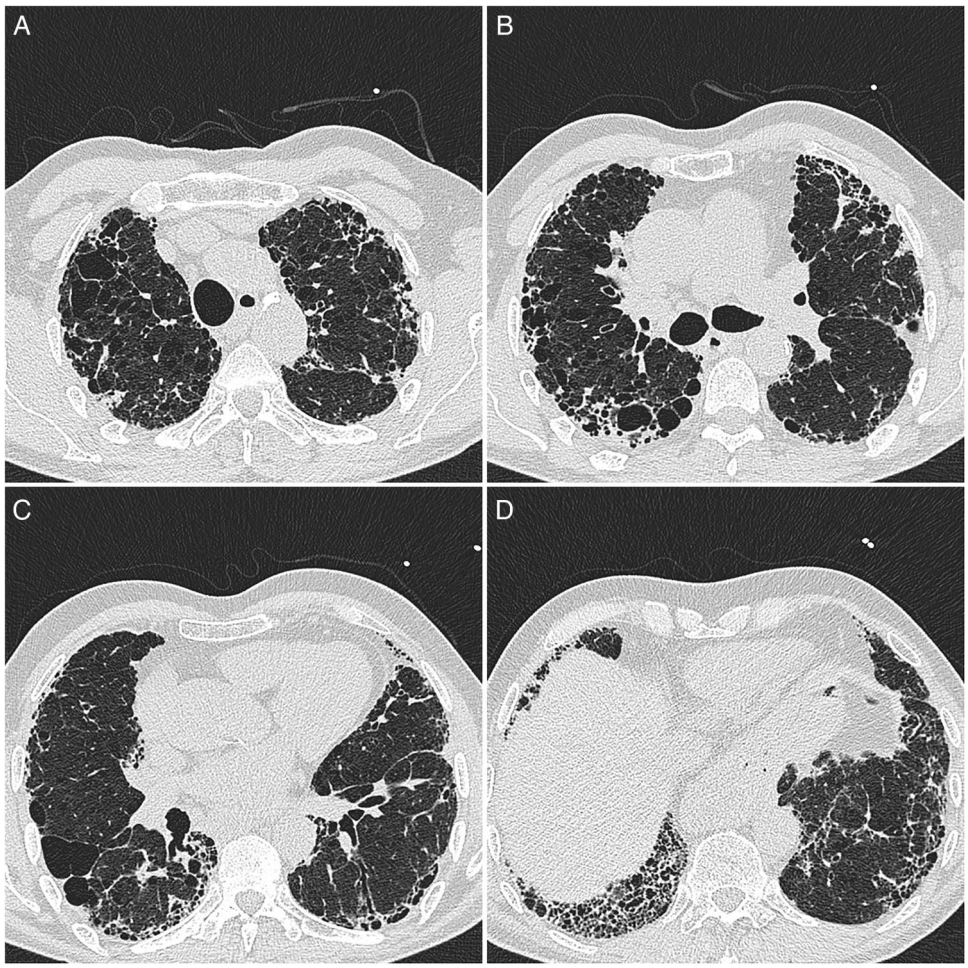

Figure 2 (A) Biopsy proven usual interstitial pneumonia (UIP) in a patient with a multidisciplinary diagnosis of usual interstitial pneumonia (IPF). Assigned CT diagnoses expressed as a percentage of 116 observers were: definite UIP 73.9%, possible UIP 21.7%, inconsistent with UIP (4.3%). In this case 91.3% of observers assigned a grade of definitely present for honeycombing. (B) Biopsy proven UIP in a patient with a multidisciplinary diagnosis of IPF. Assigned CT diagnoses expressed as a percentage of 116 observers were: definite UIP 73.9%, possible UIP 21.7%, inconsistent with UIP (4.3%). In this case 91.3% of observers assigned a grade of definitely present for honeycombing. (C) Biopsy proven UIP in a patient with a multidisciplinary diagnosis of IPF. Assigned CT diagnoses expressed as a percentage of 116 observers were: definite UIP 73.9%, possible UIP 21.7%, inconsistent with UIP (4.3%). In this case 91.3% of observers assigned a grade of definitely present for honeycombing. (D) Biopsy proven UIP in a patient with a multidisciplinary diagnosis of IPF. Assigned CT diagnoses expressed as a percentage of 116 observers were: definite UIP 73.9%, possible UIP 21.7%, inconsistent with UIP (4.3%). In this case 91.3% of observers assigned a grade of definitely present for honeycombing.

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

Binary diagnosis score (Typical UIP or Possible UIP/inconsistent with UIP)	Interobserver agreement*
Disease subgroups	
MDT diagnosis—CHP (n=44)	0.39
MDT diagnosis—CTD-ILD (n=51)	0.35
MDT diagnosis—Idiopathic FLD (n=55)	0.38
Age subgroups	
Patient age <50 years (n=25)	0.31
Patient age 50–60 years (n=40)	0.39
Patient age 60–70 years (n=48)	0.42
Patient age >70 years (n=37)	0.40

*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.

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.[18] 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*[10] reported that in the absence of honeycombing, a heterogeneous pattern of fibrosis on CT maybe 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 4 Scoring of online cases: Interobserver agreement for the 3-point HRCT pattern scores for honeycombing, traction bronchiectasis and emphysema expressed as Cohen's weighted κ coefficient stratified according to observer experience and specialty

	Interobserver agreement	
	Mean±SD	Median (IQR)
Honeycombing score (Definite, possible, absent		
Thoracic radiology fellows (n=5)	0.60±0.14	0.65 (0.23)
Thoracic radiologists (experience <10 years, n=42)	0.61±0.11	0.63 (0.14)
Thoracic radiologists (experience 10–20 years, n=27)	0.60±0.12	0.63 (0.16)
Thoracic radiologists (experience >20 years, n=22)	0.59±0.10	0.60 (0.13)
General radiologists (n=16)	0.57±0.11	0.58 (0.14)
Traction bronchiectasis score (Definite, possible, absent)		
Thoracic radiology fellows (n=5)	0.28±0.13	0.32 (0.25)
Thoracic radiologists (experience <10 years, n=42)	0.40±0.15	0.42 (0.22)
Thoracic radiologists (experience 10–20 years, n=27)	0.45±0.15	0.48 (0.23)
Thoracic radiologists (experience >20 years, n=22)	0.45±0.13	0.46 (0.18)
General radiologists (n=16)	0.41±0.14	0.44 (0.17)
Emphysema score (Definite, possible, absent)		
Thoracic radiology fellows (n=5)	0.41±0.08	0.40 (0.13)
Thoracic radiologists (experience <10 years, n=42)	0.54±0.15	0.55 (0.22)
Thoracic radiologists (experience 10–20 years, n=27)	0.53±0.13	0.54 (0.19)
Thoracic radiologists (experience >20 years, n=22)	0.42±0.17	0.43 (0.28)
General radiologists (n=16)	0.43±0.15	0.40 (0.22)

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 of a diagnostic process which is dynamic, includes disease behaviour and prognostic biomarkers, and which focuses less stringently on specific HRCT patterns.

The level of agreement for honeycombing might be an explanation for the source of disagreement between the UIP diagnosis categories. Interobserver agreement for honeycombing was moderately better than for the UIP diagnosis categories. This suggests that although some disagreement for the diagnosis of UIP may have been caused by discrepancies with regards to the presence of honeycombing (the difference between a 'UIP' and 'possible UIP'), a smaller proportion of disagreement may have related to the distribution of the honeycombing change (the difference between 'UIP' and an 'inconsistent with UIP' CT pattern). For example, in cases where emphysema and fibrosis coexist, observers might agree that honeycombing was present, but might disagree on nature of upper lobe cystic change, with some observers calling this honeycombing (and therefore regarding the distribution of honeycombing as atypical for UIP) but others calling these changes emphysema and deciding that the true honeycomb change predominated in the lower lobes (figure 1A–D). Recently, Watadani *et al* reported levels of observer agreement for honeycombing similar to those demonstrated in the current study.[8] The current study extends the findings of Watadani *et al* by relating interobserver agreement for honeycombing to the evaluation of CT appearances in patients with suspected IPF/UIP. Although speculative, agreement on the individual features considered 'inconsistent with a UIP pattern' might also have impacted overall agreement on the diagnosis categories. For example, the presence of subtle mosaic attenuation or ground glass opacification (both of which may be seen in cardiac failure—a complication of IPF/UIP) on a HRCT which otherwise has typical UIP features might result in a diagnosis category of 'inconsistent with UIP'. A limitation of the current study is that in cases considered inconsistent with UIP, we did not require observers to specify why.

Table 5 Interobserver agreement for the diagnosis categories and CT pattern scores assessed on volumetric thin-section CT, expressed as Cohen's weighted κ coefficient stratified according to observer experience

	Interobserver agreement	
	Mean±SD	Median (IQR)
UIP diagnosis categories (UIP, possible UIP, inconsistent with UIP)		
Under 10 years experience	0.50±0.17	0.54 (0.17)
Over 20 years experience	0.39±0.15	0.40 (0.12)
Honeycombing score (Definite, possible, absent)		
Under 10 years experience	0.51±0.18	0.53 (0.11)
Over 20 years experience	0.47±0.13	0.48 (0.09)
Traction bronchiectasis score (Definite, possible, absent)		
Under 10 years experience	0.62±0.14	0.65 (0.09)
Over 20 years experience	0.47±0.16	0.45 (0.14)
Emphysema score (Definite, possible, absent)		
Under 10 years experience	0.58±0.11	0.57 (0.11)
Over 20 years experience	0.36±0.20	0.34 (0.18)

UIP, usual interstitial pneumonia.

A number of issues with regards to our methodology warrant discussion. First, the study was performed in two stages: the first stage involved the scoring of cases online and the second involved the scoring of full volumetric thin-section CT data. The decision to perform the second stage of the study was made *after* the results of the first stage of the study were available and was designed to test if the level of interobserver agreement was improved when observers had access to full volumetric thin-section CT. Second, we intentionally did not supply observers with clinical data. The primary aim of the study was to evaluate interobserver agreement for the ATS/ERS/JRS/ALAT CT criteria for IPF/UIP, which do not specify clinical criteria.[2] Interobserver agreement for a MDT diagnosis (which is a related but separate issue and would include all available clinical information) in the setting of fibrotic lung disease might be 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.[7] 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, fibrotic NSIP and CHP which can only be achieved based on CT appearances alone in approximately 50% of cases.[19] Furthermore, when IPF/UIP presents with non-classical CT appearances, the usual alternative diagnoses are fibrotic NSIP or CHP.[20] 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/ALAT CT criteria for a diagnosis of IPF/UIP is at best only moderate. As CT diagnosis plays an important role in influencing management decisions during the initial evaluation of patients with suspected IPF, accurate and consistent application of these guidelines among thoracic radiologists is clinically important. Based upon the results of this study, modification of these criteria may be necessary to improve observer agreement.

Collaborators The UIP Observer Consort are Joseph Jacob, Anand Devaraj, Giampaola Gavelli, Hrudaya Nath, Joseph Tashjian, Denis Tack, Zsuzsanna Monostori, Takeshi Johkoh, Anne Davies, Maurizio Zompatori, Roberta Polverosi, Ennio Vincenzo Sassani, Gilbert R Ferretti, Can Zafer Karaman, Gianluigi Sergiacomi, Tomas Franquet, Kavita Garg, Richard Mannion, Paula Campos, Robert Karl, Paul Burrowes, Franco Quagliarini, Louise Norlen, John T Murchison, Antoine Khalil, Sundeep M Nayak, Alain Nchimi, Nevzat Karabulut, Anna Rita Larici, Suzanne Matthews, Eva Castañer, Juan Arenas-Jimenez, Ralph Drosten, Ryoko Egashira, Hyun-Ju Lee, Santiago Rossi, Mosleh Al-Raddadi, Anastasia Oikonomou, Anagha P Parkar, Pierre Yves Brillet, Wagner Diniz de Paula, Laurent Medart, Florian Poschenrieder, Igor Pozek, Taiwo Senbanjo, Katharina Marten-Engelke, Juntima Euathrongchit, Cal Delaplain, Giancarlo Cortese, Gian Alberto Soardi, Aleksandar Grgic, Jeffrey Kanne, Michelle Muller, John Bruzzi, Adrien Jankowski, Gracijela Bozovic, Andrea Goncalves, Justus Roos, Daniel Vargas, Mark Elias, Djenaba Bradford-Kennedy, Danielle Seaman, Brett Memauri, Humera Chaudhary, Foong Wong, Julia Alegria, H Henry Guo, Helmut Prosch, Luce Cantin, Goffredo Serra, Elisa Baratella, Rosalba Silecchia, Aziza Icksan, Adam Wallis, Sidhdharth Damani, Rahul Renapurkar, Samanjit Hare, Annemilia del Ciello, Jennifer Jung, Mario Silva, Sushilkumar Sonavane, Sue Thomas, Anna Beattie, Benedikt Kislinger, Leo Ca, Girish Shroff, Hester Gietema, Yuranga Weerakkody, Stephen Hobbs, Shinyu Izumi, Tomoo Kishaba, Akira Shiraki, Yuko Waseda, Maria Chiara Castoldi, Vincent Herpels, Antonio Pais, Roger Matus, Enrico Lubin, Ben Wilson, Mathias Mueller, Shadha Ahmed Alzubaidi, Alain Ortiz, Iara Sequeiros, Nicola Boscolo Bariga, Mark Wills, Sze Mun Mak, Giuseppe Aquaro, Esin Cakmakci Midia, Nicola Schembri, Joseph A M Sheehan, Alexia Farrugia, Jennifer Tomich, J M Fernandez Garcia-Hierro.

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

1 Travis WD, Costabel U, Hansell DM, *et al*. An official American Thoracic Society/European Respiratory Society statement: update of the international multidisciplinary classification of the idiopathic interstitial pneumonias. *Am J Respir Crit Care Med* 2013;188:733–48.
2 Raghu G, Collard HR, Egan JJ, *et al*. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med* 2011;183:788–824.
3 Hunninghake GW, Zimmerman MB, Schwartz DA, *et al*. Utility of a lung biopsy for the diagnosis of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2001;164:193–6.
4 Raghu G, Mageto YN, Lockhart D, *et al*. The accuracy of the clinical diagnosis of new-onset idiopathic pulmonary fibrosis and other interstitial lung disease: a prospective study. *Chest* 1999;116:1168–74.
5 Lynch DA, Godwin JD, Safrin S, *et al*. High-resolution computed tomography in idiopathic pulmonary fibrosis: diagnosis and prognosis. *Am J Respir Crit Care Med* 2005;172:488–93.
6 Thomeer M, Demedts M, Behr J, *et al*. Multidisciplinary interobserver agreement in the diagnosis of idiopathic pulmonary fibrosis. *Eur Respir J* 2008;31:585–91.
7 Aziz ZA, Wells AU, Hansell DM, *et al*. HRCT diagnosis of diffuse parenchymal lung disease: inter-observer variation. *Thorax* 2004;59:506–11.
8 Watadani T, Sakai F, Johkoh T, *et al*. Interobserver variability in the CT assessment of honeycombing in the lungs. *Radiology* 2013;266:936–44.
9 Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992;304:1491–4.
10 Gruden JF, Panse PM, Leslie KO, *et al*. UIP diagnosed at surgical lung biopsy, 2000–2009: HRCT patterns and proposed classification system. *AJR Am J Roentgenol* 2013;200:W458–67.
11 American Thoracic Society/European Respiratory Society International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias. This joint statement of the American Thoracic Society (ATS), and the European Respiratory Society (ERS) was adopted by the ATS board of directors, June 2001 and by the ERS Executive Committee, June 2001. *Am J Respir Crit Care Med* 2002;165:277–304.
12 Lamas DJ, Kawut SM, Bagiella E, *et al*. Delayed access and survival in idiopathic pulmonary fibrosis: a cohort study. *Am J Respir Crit Care Med* 2011;184:842–7.
13 Mathieson JR, Mayo JR, Staples CA, *et al*. Chronic diffuse infiltrative lung disease: comparison of diagnostic accuracy of CT and chest radiography. *Radiology* 1989;171:111–16.
14 Grenier P, Valeyre D, Cluzel P, *et al*. Chronic diffuse interstitial lung disease: diagnostic value of chest radiography and high-resolution CT. *Radiology* 1991;179:123–32.
15 Lee KS, Primack SL, Staples CA, *et al*. Chronic infiltrative lung disease: comparison of diagnostic accuracies of radiography and low- and conventional-dose thin-section CT. *Radiology* 1994;191:669–73.
16 Swensen SJ, Aughenbaugh GL, Myers JL. Diffuse lung disease: diagnostic accuracy of CT in patients undergoing surgical biopsy of the lung. *Radiology* 1997;205:229–34.
17 Assayag D, Elicker BM, Urbania TH, *et al*. Rheumatoid arthritis-associated interstitial lung disease: radiologic identification of usual interstitial pneumonia pattern. *Radiology* 2014;270:583–8.
18 Flaherty KR, Thwaite EL, Kazerooni EA, *et al*. Radiological versus histological diagnosis in UIP and NSIP: survival implications. *Thorax* 2003;58:143–8.
19 Silva CI, Müller NL, Lynch DA, *et al*. Chronic hypersensitivity pneumonitis: differentiation from idiopathic pulmonary fibrosis and nonspecific interstitial pneumonia by using thin-section CT. *Radiology* 2008;246:288–97.
20 Sverzellati N, Wells AU, Tomassetti S, *et al*. Biopsy-proved idiopathic pulmonary fibrosis: spectrum of nondiagnostic thin-section CT diagnoses. *Radiology* 2010;254:957–64.