Inter-reader reproducibility of a radiographic grading system for usual interstitial pneumonitis validates its use as a surrogate endpoint in clinical trials

KATHLEEN M. CAPACCIONE* ☞, HONG MA, LYNDON LUK and MARY M. SALVATORE

Department of Radiology, Columbia University Irving Medical Center, New York, NY, USA

Received: October 28, 2023 ● Revised manuscript received: January 11, 2024 ● Accepted: January 21, 2024

ABSTRACT

Background: The primary purpose of this study was to assess the interreader reliability of a grading system for UIP based on the quantification of normal lung. This grading system considers each of the following lung regions: right upper and middle lobes, right lower lobe, left upper lobe, and left lower lobe. Each is assigned a grade based on the following: 0: 0% normal lung; 1: 1–49% normal lung; 2: 50–74% normal lung; 3: 75–89% normal lung; 4: 90–99% normal lung; 5: 100% normal lung. The secondary purpose was to compare the grades rendered by non-cardiothoracic subspecialty trained radiologists to grades established by cardiothoracic radiologists, which were considered the gold standard.

Methods: Chest CT images of patients were obtained by searching the radiology record system for the terms “usual interstitial pneumonia” and “UIP”. Each case was confirmed by radiologist review; pathology was not assessed given the small fraction of cases that underwent biopsy due to the high risk of complications in patients with fibrotic lung disease. Two cardiothoracic radiologists evaluated each CT and reached a consensus grade. Two different radiologists who were not subspecialty trained in cardiothoracic radiology independently graded each case. Spearman correlation analysis was performed to compare the two reader’s grades as well as each reader’s grade independently to the gold standard score.

Results: Our analysis demonstrated a strongly positive statistically significant interreader correlation coefficient (RS) = 0.7062, P < 0.001. Our analysis of each reader compared to the gold standard demonstrated an RS = 0.77559, P < 0.001 and an RS = 0.69958, P < 0.001 for readers 1 and 2, respectively, both representing statistically significant strongly positive correlations.

Conclusions: These results demonstrate strong interreader reproducibility and show that radiologists without subspecialty training in cardiothoracic radiology render grades that correlate strongly with those given by cardiothoracic radiologists. These findings support the use of this grading system for UIP both to monitor clinical progression and as a surrogate endpoint for antifibrotic drug trials.

KEYWORDS

UIP, pulmonary fibrosis, grading system, clinical trials

Introduction

Usual interstitial pneumonia (UIP)/idiopathic pulmonary fibrosis (IPF) is a devastating pulmonary disease with relentless progression and mortality outcomes comparable to lung cancer [1]. It is one of over one hundred diseases that fall under the umbrella of interstitial lung disease, diseases that have varying known and unknown etiologies but progress down a final common pathway of pulmonary parenchymal scarring which ultimately results
in decreased oxygen exchange and respiratory insufficiency [2–4]. Other common types of interstitial lung disease include nonspecific interstitial pneumonia (NSIP) and chronic hypersensitivity pneumonitis (CHP) [5]. Assessing progression of UIP/IPF is an important part of disease management because change over time has prognostic value and may trigger changes in treatment strategy.

Given the poor outcomes for patients with UIP/IPF, there have been extensive efforts to develop and test antifibrotic agents to slow or ideally arrest the progression of the disease. Two antifibrotic agents have been approved clinically to slow the progression of UIP, pirfenidone and nintedanib [6–9]. Numerous agents are in various stages of preclinical development, including phosphodiesterase inhibitor BI1015550 [10] and pamrevlumab, a first-in-class anti-connective tissue growth factor (CTGF) antibody [11]. These and other agents have the potential to transform the treatment landscape for patients with UIP/IPF. However, the establishment of clinically meaningful surrogate endpoints for pulmonary fibrosis drug trials has been a significant challenge to the field.

The diagnosis of UIP is made radiographically by the presence of specific radiographic features. According to an official ATS/ERS/JRS/ALAT clinical practice guideline, the diagnosis of UIP is made with the findings of basilar predominant subpleural reticulation, traction bronchiectasis and/or bronchiolectasis, and honeycombing. In the absence of honeycombing, the pattern is called “probable UIP” [12, 13]. A subsequent statement defined that progressive pulmonary fibrosis can be diagnosed when at least two of the following three criteria are met: worsening symptoms, radiological progression, and physiological progression [14]. Therefore, the ability to accurately diagnose radiological progression has significant implications for patient care.

We have previously developed a grading system for UIP based on the quantity of remaining normal lung and validated that it correlates with PFTs [15] (Fig. 1). This grading system considers two regions: “normal lung”, and “pathologically affected lung”, which can include regions of fibrosis, ground glass, mosaic attenuation, consolidation, emphysematous change, cysts, and other lung parenchymal abnormalities. The reasoning for this is that pathologically affected lung from any causes will result in diminution of pulmonary function, therefore the overall lung score should reflect the amount of healthy lung. We demonstrated that this grading system correlates with patient mortality [16]. Our system has several advantages: 1) CT imaging is highly reproducible and is not patient or operator dependent; 2) it is based on quantitating normal lung, which is similar to the way that PFTs are reported; 3) it can account for multiple lung diseases that occur concurrently; 4) it does not require special software or machines as do some of the other proposed UIP grading systems [17–19].

Fig. 1. Demonstrates representative examples of each grade in our system: (A) Grade 0, (B) Grade 1, (C) Grade 2, (D) Grade 3, (E) Grade 4, (F) Grade 5
Patients and methods

Grading system

We employed the grading system for normal lung as previously published [15]. Briefly, the system considers 4 regions (right upper and middle, right lower, left upper, left lower) and gives an individual grade based on percentage of normal lung. Percentage scores are as follows: 0: 0% normal lung; 1: 1–49% normal lung; 2: 50–74% normal lung; 3: 75–89% normal lung; 4: 90–99% normal lung; 5: 100% normal lung. Scores were summed to generate a single score representing total amount of normal lung.

CT scans used for grading

This study was conducted under Columbia University IRB-approved HIPAA-compliant protocol (AAAT6351, approved 03/31/2021). The requirement for consent was waived given the retrospective nature of this study. CT scans used for grading in this study were obtained by searching the radiology record system MModal for the search terms “UIP” and “usual interstitial pneumonia”. This cohort was used previously to assess mortality, and inclusion criteria were patients that were deceased at the time of review and had a date of death could be identified. The first 100 patients by serial medical record number were reviewed and ultimately 94 met inclusion criteria. CT scans analyzed in this study were performed on a variety of different machines both within our department and on scanners at outside hospitals brought to our institution for review. The parameters were variable mimicking real life conditions. A minimum of 5 mm slices were used for image review, and in the majority of cases image slice thickness of 1.25 mm was available and used for review.

CT grading

All studies were initially reviewed by two cardiothoracic radiologists (M.M.S and K.M.C) with 28- and 6-years experience, respectively. CT grade for each lobe was obtained by consensus. Readers were blinded to all other clinical and mortality data during the time of CT review.

All cases were subsequently reviewed and graded independently by two radiologists who were not subspecialty trained in cardiothoracic radiology (H.M. and L.L.), with 11- and 9- years’ experience, respectively. Both were subspecialty trained in abdominal imaging and had general radiology experience reading cross sectional imaging including cardiothoracic studies, however received no additional training in this grading system beyond a general introduction, as explained in the introduction of this manuscript. Readers scored each of the four regions as described above to generate a composite score.

Statistical analysis

Statistical analysis was performed using Spearman correlation using an online calculator [20]. First, correlation was performed between the two readers to establish the interreader reproducibility of the grading system score. Subsequently, each reader’s score was correlated to the scores established by the cardiothoracic trained radiologists who developed the system, which was taken as the gold standard grade. For all analyses, a two-tailed p value less than or equal to 0.05 was considered significant.

Results

Radiologists display strong interreader reliability using this grading system

We first sought to evaluate the interreader reproducibility of our grading system. Results of the two readers’ grades were analyzed using Spearman correlation to assess for the degree of interreader reliability as measured by correlation coefficient. The interreader correlation coefficient ($R_s$) = 0.7062, $P < 0.001$ representing a statistically significant strong positive correlation (Fig. 2). A small number of outliers were present, for example a single case rated by one reader as 20 and the second reader as 6. Similarly, when evaluating the reader’s grades compared to the established gold standard, a small number of outliers were present in each case. These were included in the comparison because there was no systematic disagreement among readers, and occasional disagreement in final reports closely mirrors the real practice setting. Despite these few outliers, correlations were strong between readers and between each reader and the gold standard.

Non-subspecialty trained radiologists demonstrate strong correlation with gold standard grading performed by cardiothoracic radiologists

We next sought to determine whether the grades rendered by radiologists not subspecialty-trained in cardiothoracic radiology were comparable to those given by cardiothoracic radiologists, which were considered the gold standard for this study. We performed Spearman correlation analysis...
between the gold standard cardiothoracic radiologist consensus score and the scores from each of the readers. Our analysis demonstrated an $R_s = 0.77559$, $P < 0.001$ and $R_s = 0.69958$, $P < 0.001$ for readers 1 and 2, respectively, both representing strong positive correlations (Fig. 3).

Discussion

The development of a reproducible grading system to monitor radiologic progression of UIP/IPF represents a major advance both for clinical decision making and use in clinical trials. Here, we demonstrate that our grading system is highly reproducible by radiologists who are not subspecialty trained in cardiothoracic radiology. The results support the use of this system to monitor clinical progression as well as in clinical trials given that it provides a quantitative measure that is neither patient nor operator dependent, and is highly reproducible among readers.

While other groups have tried to develop grading systems for interstitial lung disease, many of them involve artificial intelligence and the use of complex computer systems [17–19]. These software programs may be unavailable to many clinicians limiting their utility particularly in the context of general radiology settings. An advantage of this highly reproducible system is that it does not require any special programs or systems and could be implemented across medical systems to standardize the assessment of progression of pulmonary fibrosis. Many patients who are ultimately diagnosed with UIP/IPF are initially treated in the community setting and are eventually referred to a center of excellence for management of their fibrotic lung disease. Use of this quantitative grade could help them transition from one practice setting to the other with quantitative radiologic data so that the monitoring of their disease would be continuous.

Establishing surrogate endpoints in clinical trials for pulmonary fibrosis is important to drive forward the development of new anti-fibrotic agents [21, 22]. Given the high reproducibility of this system demonstrated herein, in conjunction with its correlation with gold standard mortality [16] and PFTs [15], this grading system is an excellent candidate. We propose its incorporation into clinical trials monitoring followed by comparison with current standard of care endpoints to prove its utility in this context.

Limitations of this study included that all of the cases were drawn from a single institution, which was a tertiary care center with extensive expertise in interstitial lung disease. Although the readers were not subspecialty trained in cardiothoracic radiology, they had exposure to a greater number of cases of interstitial lung disease than a radiologist in general practice. The “gold standard” grade was established by consensus of two cardiothoracic radiologists. The use of a computerized method to generate a score could potentially have resulted in a more standardized measure. Finally, the additional inclusion of a non-radiologist with...
clinical expertise in interstitial lung disease as a reader may have demonstrated even more expanded reproducibility of this system if results were concordant with the gold standard reads, however such a reader was not included at this time. Despite these limitations, these data establish this grading system based on radiologic quantification of normal lung as highly reproducible among radiologists without cardiothoracic radiology subspecialty training.

Conclusion

Here, we validate the interreader reproducibly of this grading system for normal lung, as well as demonstrate that there is strong correlation between individual readers and the gold standard grades. This provides support for the use of this system as a surrogate endpoint in clinical trials and as a clinical metric to follow progression of UIP/IPF over time.

Authors’ contributions: Conceptualization: KMC, MMS; Data collection: KMC, HYM, LL, MMS; Data analysis: KMC; Manuscript-writing: KMC; Manuscript-editing: KMC, HYM, LL, MMS. All authors reviewed the final version of the manuscript and agreed to submit it to IMAGING for publication.

Funding sources: No financial support was received for this study.

Conflict of interest: Dr. Mary Salvatore has received grant support from Boehringer Ingelheim and Genentech and has served as a consultant for AbbVie, Bioclinica, and LungLife AI. Dr. Capaccione has served as an advisor for Cardinal Health. The remaining authors have no conflicts of interest to disclose.

Ethical statement: This work was conducted under an IRB-approved, HIPAA-compliant protocol (Columbia University, AAAS18289) to study patients with pulmonary fibrosis.

REFERENCES


Open Access statement. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (https://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted use, distribution, and reproduction in any medium for non-commercial purposes, provided the original author and source are credited, a link to the CC License is provided, and changes – if any – are indicated.