Eur Radiol (2015) 25:2558–2566
DOI 10.1007/s00330-015-3685-4

CHEST

Chest Computed Tomography-Based Scoring of Thoracic Sarcoidosis: Inter-rater Reliability of CT Abnormalities

D. A. Van den Heuvel · P. A. de Jong · P. Zanen · H. W. van Es · J. P. van Heesewijk · M. Spee · J. C. Grutters

Received: 31 December 2014 / Revised: 16 February 2015 / Accepted: 18 February 2015 / Published online: 9 April 2015

Abstract
Purpose To determine inter-rater reliability of sarcoidosis-related computed tomography (CT) findings that can be used for scoring of thoracic sarcoidosis.
Materials and methods CT images of 51 patients with sarcoidosis were scored by five chest radiologists for various abnormal CT findings (22 in total) encountered in thoracic sarcoidosis. Using intra-class correlation coefficient (ICC) analysis, inter-rater reliability was analysed and reported according to the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) criteria. A pre-specified sub-analysis was performed to investigate the effect of training. Scoring was trained in a distinct set of 15 scans in which all abnormal CT findings were represented.
Results Median age of the 51 patients (36 men, 70 %) was 43 years (range 26 – 64 years). All radiographic stages were present in this group. ICC ranged from 0.91 for honeycombing to 0.11 for nodular margin (sharp versus ill-defined). The ICC was above 0.60 in 13 of the 22 abnormal findings. Sub-analysis for the best-trained observers demonstrated an ICC improvement for all abnormal findings and values above 0.60 for 16 of the 22 abnormalities.
Conclusions In our cohort, reliability between raters was acceptable for 16 thoracic sarcoidosis-related abnormal CT findings.
Key Points
- *Thoracic sarcoidosis is common; knowledge on reliability of CT scoring is limited.*
- *Scoring CT abnormalities in pulmonary sarcoidosis can achieve good inter-rater agreement.*
- *CT scoring validation in thoracic sarcoidosis is important for diagnostic and prognostic studies.*

Keywords Sarcoidosis · Pulmonary · Computed tomography · Reliability · Scoring

Abbreviations and acronyms

CT	Computed tomography
ICC	Intra-class correlation coefficient
GRRAS	Guidelines for Reporting reliability and Agreement Studies
PACS	Picture Archiving and Communication System
WASOG	World Association of Sarcoidosis and Other Granulomatous disorders
ANOVA	Analysis of variance

Introduction

Thin-section computed tomography (CT) of the chest is increasingly used in the diagnosis of sarcoidosis. A multitude of abnormalities can be detected [1–4], but little is known about inter-rater agreement. There appears to be good reported agreement between raters in scoring a limited number of abnormal CT findings in sarcoidosis with kappa values ranging from 0.81 to 0.89. These results are based on studies focussing on correlations between

D. A. Van den Heuvel (✉) · H. W. van Es · J. P. van Heesewijk · M. Spee
Department of Radiology, St. Antonius Hospital Nieuwegein, Koekoekslaan 1, 3435 CM Nieuwegein, The Netherlands
e-mail: d.van.den.heuvel@antoniusziekenhuis.nl

P. A. de Jong
Department of Radiology, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands

P. Zanen · J. C. Grutters
Division Heart & Lungs, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands

P. Zanen · J. C. Grutters
Center of Interstitial Lung Diseases, Department of Pulmonology, St. Antonius Hospital Nieuwegein, Koekoekslaan 1, 3435 CM Nieuwegein, The Netherlands

abnormal CT findings and pulmonary function impairment; however, the number of raters as well as the number of abnormal CT findings studied has been limited [5–9].

Little is known about prognostic and clinical implications of abnormal CT findings in sarcoidosis. The best evidence available comes from studies investigating changes in disease extent or abnormal CT findings in serial imaging [10–14]. In order to investigate the potential role of CT in clinical decision-making in sarcoidosis, the inter-rater agreement of CT scoring has to be determined, as reasonable agreement is a first requisite for clinical usefulness. It is also essential for applicability of study results in other centres and thus enables comparisons of clinical outcomes between sarcoidosis centres.

The aim of our study was to measure the inter-rater reliability of a large number of thoracic sarcoidosis-related abnormal CT findings. With these results, we will have a firm basis for future investigations.

Methods

Patient cohort

Our local medical ethical committee approved this retrospective study.

A random sample of 51 patients was drawn from a cohort of Dutch sarcoidosis patients from a single sarcoidosis referral centre in the Netherlands. All patients had available CT images in the Picture Archiving and Communication System (PACS). The cohort consisted of a mix of all radiographic stages [15] (including stage IV with fibrotic end-stage disease). All patients were diagnosed with sarcoidosis based on histology or met the criteria of the World Association of Sarcoidosis and Other Granulomatous disorders (WASOG). This entails appropriate clinical presentation, granulomas identified in one or more organs, multiple organ involvement, and no other cause for granulomatous reaction [16].

Thin section CT acquisition protocol

All examinations were performed on a 16-detector-row computed tomography (CT) system (Brilliance-16; Philips, Cleveland, OH, USA). Inspiratory scans were volumetric acquisitions with a slice thickness 1.0, collimation 16×0.75 (120 kVp, 130 mA.s, 512 FOV) and reconstructed with a high spatial frequency algorithm. Expiration scans were static, post-expiratory acquisitions at three predefined levels (aortic arch, carina, and lung bases) at 2-cm intervals. In some patients, expiration scans were low-dose volumetric acquisitions. Imaging direction was from the apices to the diaphragm. All patients had an expiration scan. Of these, six were of insufficient quality to assess air trapping (12 %).

Abnormal CT findings

When deciding which abnormalities to score, it was considered important to a) prevent selection bias and b) form a wide basis on which future investigations could be conducted. Therefore, all possible sarcoidosis related abnormal CT findings previously described in interstitial lung and airway disease defined by the Fleischner Society glossary [17] were included. Also, more sarcoidosis-specific abnormalities like the galaxy and cluster sign were included. The result of this approach was a list of 22 items that had to be scored in each CT (Table 1). The abnormalities were scored per lobe in a semi-quantitative way when they involved parenchymal disease, or listed as present/absent in the case of airway and pleural disease. The lingula was regarded as a separate lobe. Quantification of parenchymal disease was grouped into five categories: 0, absent; 1, 1 – 25 %; 2, 26 – 50 %; 3, 51 – 75 %;

Table 1 Overview of CT findings subject of scoring per lobe

		RUL	RML	RLL	LUL	LIN	LLL
Nodules							
Bronchovascular	(0/1)						
Parenchymal	(0/1)						
Subpleural	(0/1)						
Confluent	(0/1)						
Predominant sharp	(0/1)						
Predominant size	(0/1)						
% of lobe	(0-4)						
Other abnormalities							
Peripheral consolidation	(0-4)						
Central consolidation	(0-4)						
Ground Glass	(0-4)						
Mosaic pattern	(0-4)						
Air trapping	(0-4)						
Architectural distortion	(0-4)						
Septal lines	(0-4)						
Honeycombing	(0-4)						
Bronchial deformation	(0/1)						
Bronchiectasis	(0/1)						
Pleural thickening	(0/1)						
Parenchymal calcifications	(0/1)						
Aspergiloma	(0/1)						
Measurements	mm						
Nodes hilum right							
Nodes hilum left							
Nodes mediastinal							

(0/1) abscent or present

(0-4) percentage lobar involvement

RUL right upper lobe, *RML* right middle lobe, *RLL* right lower lobe, *LUL* left upper lobe, *LIN* lingua, *LLL* left lower lobe

and 4, 76 – 100 % of lobar involvement. The quantification of parenchymal disease was based upon expert opinion consulting both experienced raters and a statistician (PZ).

Nodules

Nodules are rounded opacities up to 3 cm in diameter and were classified on their peribronchovascular, parenchymal, or subpleural location. The frequently encountered thickening of the bronchovascular interstitium was scored as peribronchovascular nodules, although in some cases there will also be some fibrosis involved. Parenchymal nodules are nodules that are neither peribronchovascular nor subpleural in location. In addition, the aspect of the nodules was characterized as coalescent or as having a sharp or ill-defined margin. Ground glass caused by diffuse micronodules (granular groundglass) was scored both as parenchymal nodules and as groundglass (Fig. 1). Furthermore, the galaxy and cluster sign were assessed and scored as nodules. Because these classical items represent patterns of, respectively, confluent and clustered nodular disease they were captured under the confluent and parenchymal nodular items.

Consolidations

Consolidations are areas with increased opacity and obscuration of underlying structures. They were classified as central and peripheral regarding one-third of the cross-sectional distance from the hilum to the peripheral pleural surface as the central zone. Round opacities lager than 3 cm with or without air bronchogram were scored as consolidations and not as masses (Fig. 2). In the literature, fibrotic masses are also regarded as important features for pulmonary sarcoidosis. We captured these under central or peripheral consolidations.

Bronchial abnormalities

Bronchial abnormalities were classified as bronchiectasis and bronchial deformation. Bronchiectasis was scored when the findings were compliant with the Fleischner definition. Bronchial deformation is not defined by the Fleischner glossary, but is considered an important item in thoracic sarcoidosis. It was scored when bronchi or bronchioles were deformed with or without the presence of bronchial dilatation. Although angulated or crossing bronchi are typical features of deformation in the presence of fibrosis, deformation could also be scored if the bronchial lumen was narrowed as a result of external compression for instance due to enlarged lymph nodes (Fig. 3).

Pleural abnormalities

Pleural thickening had to be differentiated from pseudoplaques. Pseudoplaques are coalescent small nodules in contact with the visceral pleura. In this study these plaques were scored as subpleural nodules (Fig. 4).

Septal lines

Septal lines were scored if there was thickening of the interlobular septa. The thickening could be smooth as well as irregular. Lines not related to the interlobular septa were not scored (Fig. 5).

Air trapping

In order to determine the presence of air trapping, expiration scans were used. If these were not available or inadequate according to the observers' discretion, not available (NA) was scored. The presence of air trapping was

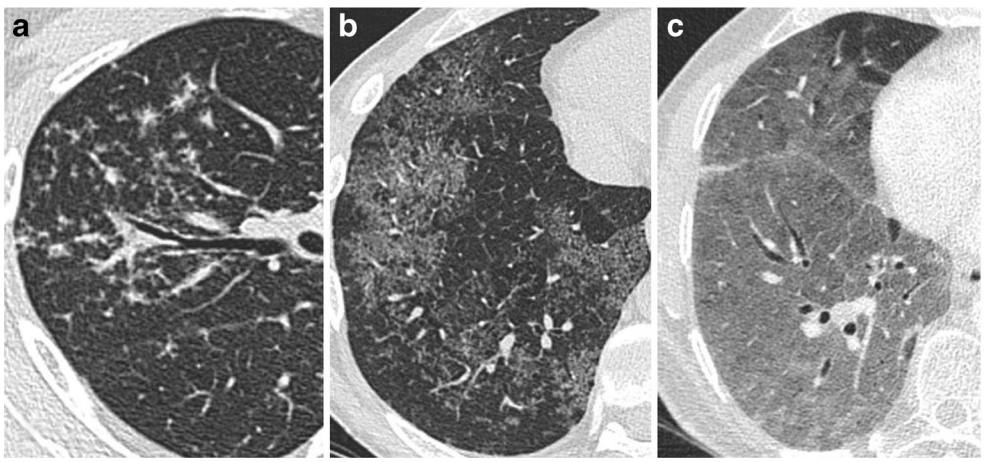

Fig. 1 Different presentations of a nodular pattern in pulmonary sarcoidosis. (*A*) Typical peribronchovascular, subpleural, parenchymal, and coalescent nodules. There is also thickening of the peribronchial interstitium, which in this study, is also scored as peribronchovascular nodules. (*B*) Micronodules resulting in a granular groundglass pattern. (*C*) Groundglass pattern caused by profuse micronodules. Findings in this patient were scored as nodules and as groundglass abnormalities. Note the thickening of interlobular septa and of the oblique fissure

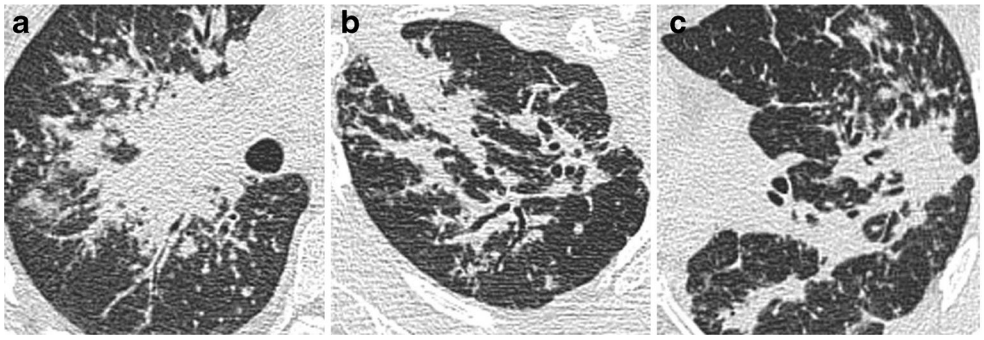

Fig. 2 Different presentations of consolidations in pulmonary sarcoidosis. (*A*) Large central consolidation with surrounding nodules, bronchovascular bundles thickening, and groundglass. (*B*) Peripheral consolidation representing active inflammation. As in *A*, there is also nodular disease and thickening of the bronchovascular bundles. (*C*) Combined central and peripheral consolidations. In this patient there are also signs of fibrosis reflected as architectural distortion and bronchiectasis

based on visual assessment; no measurements were performed. Air trapping was considered present when lung parenchyma failed to increase in attenuation and failed to decrease in volume.

Lymph node measurements

The largest lymph node was measured at three levels: mediastinal, right, and left hilar. Although no evidence exists, 10 mm short axis for mediastinal and 5 mm short axis for hilar lymph nodes was considered enlarged.

Training and thin-section CT scoring

In this study there were five raters. Raters 1 and 2 had several meetings discussing the qualitative and semi-quantitative way of scoring. Fifty unrelated CTs from sarcoidosis patients were used during these discussion and training sessions, and as a result, raters 1 and 2 received the most training. From this set of 50 CT scans, 15 were selected by raters 1 and 2 to serve as training scans. This selection was performed in such a way that all possible abnormal CT findings encountered and described in sarcoidosis were present. Rater 3 received a training session according to the continuous learning method in which all scans of the test cohort were scored by rater 3 observed and corrected by rater 1 [18]. Raters 4 and 5 only received a brief instruction on how to score explaining the qualitative and semi-quantitative approach. After scoring, ambiguities within the test cohort were discussed. The effect of training on the agreement of scoring was not assessed; there were no pre- and post-training scoring sessions.

In the actual scoring process all 51 scans were independently scored by the five observers who were employed at two different institutions. Four observers were working at a sarcoidosis referral centre and the fifth observer at a tertiary care centre involved in lung transplantation. Observers 1 and 2 (DH and PJ) are chest radiologists involved in the design of the study. Observer 3 (MS) is a chest radiology clinical fellow, and observers 4 and 5 (HE and JH) are experienced (>10 years) chest radiologists not involved in designing the study protocol. All raters were aware of the fact that the scans concerned patients diagnosed with sarcoidosis.

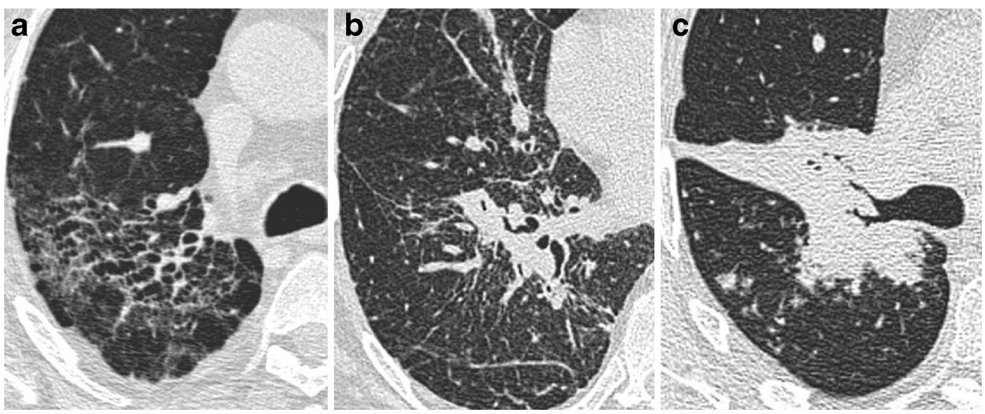

Fig. 3 Examples of bronchial involvement in sarcoidosis. (*A*) Irregular bronchial dilatation with surrounding architectural distortion, groundglass, and volume loss representing fibrosis. (*B*) Central bronchial deformation with angulation and stenosis representing fibrotic disease. Note the more diffuse architectural distortion of the lung parenchyma. (*C*) Central bronchial stenosis and obstruction caused by massive consolidation and lymphadenopathy. There is atelectasis caused by the central obstruction of bronchi. Because of the absence of angulated crossing bronchi or other fibrotic abnormalities this is probably an active inflammation representing reversible disease

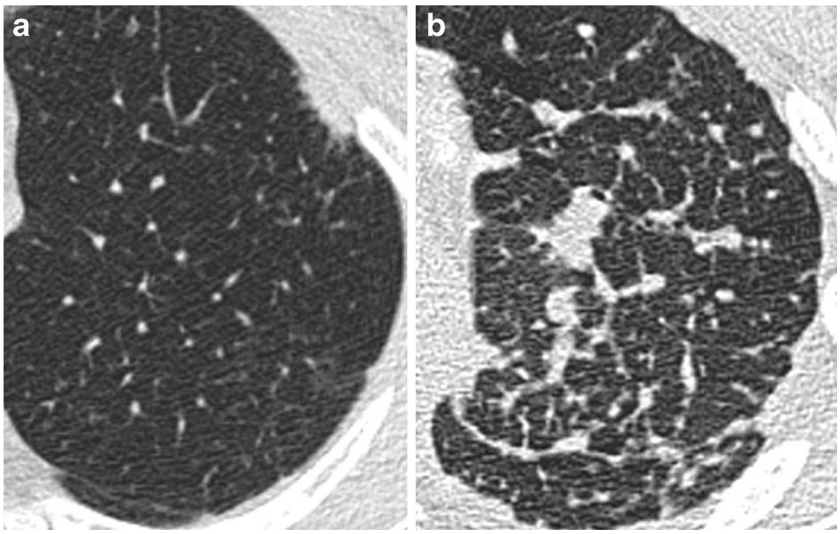

Fig. 4 Pleural involvement in sarcoidosis. (*A*) Pseudoplaque formed by confluent subpleural nodules creating a plaque. Note the acute angles with the pleura differentiating it from a real pleural thickening. (*B*) Diffuse pleural thickening in combination with subpleural nodules and interlobular thickening

Statistical analysis

In compliance with the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) criteria, variability of scoring sarcoidosis related abnormal CT findings was analysed by determining the inter-rater agreement [19]. As the spatial distribution of the abnormal CT findings over the lung was not the subject of investigation, the sum of the scores per lobe was used to create a continuous data set. This made analysis with an intra-class correlation coefficient (ICC) based on an analysis of variance (ANOVA) model possible. In addition, the two-way mixed approach was chosen; the scoring raters form a random sample from all possible raters so that the results apply to a general population of chest radiologists. This analysis was performed for all raters and also in a sub- analysis leaving out the two raters who only recieved a short intruction on how to score. A generally accepted scale considers an ICC value of <0.40 as poor reproducibility; ICC values in the range 0.40 to 0.75 indicate fair to good reproducibility, and an ICC value of greater than 0.75 shows excellent reproducibility [20]. For clinical use, an ICC value of at least 0.60 is recommended [21]. Data were analyzed with SPSS 22.

Fig. 5 Thickening of the interlobular septa, pleura and oblique fissure. The abnormalities present in this image were scored as: septal lines, pleural thickening, and subpleural nodules. There is also a small area with groundglass opacity

Results

Patient demographics and prevalence of CT findings

Patient demographics are presented in Table 2. Median age of the 51 patients (36 men, 70 %) was 43 years (range 26 – 64 years). All radiographic stages were present. Stage 2 was most frequently encountered followed by stage 3 (37 and 27 %, respectively). In Table 3 the prevalence of the scored items of observer 1 (DH) is shown. The most common abnormalities scored in this cohort are subpleural, peribronchovascular, and parenchymal nodules (86, 80, and 76 %, respectively), followed by architectural distortion and bronchial deformation (69 and 67 %, respectively). Lymphadenopathy was also frequently encountered with mediastinal, right hilar, and left hilar location in, respectively, 92, 76, and 73 % of the cases.

Inter-rater reliability

ICCs of all scored abnormal CT findings are presented in Table 4. ICCs varied substantially between abnormal CT findings ranging from 0.804 for honeycombing to 0.112 for

Table 2 Patient Demographics

Total number of patients	n =51	
Age, years		
Median	43	
Range	26-64	
Sex, n (%)		
Male	36	(70)
Female	15	(30)
Radiographic stage, n (%)		
0	2	(4)
1	11	(22)
2	19	(37)
3	14	(27)
4	5	(10)

nodular margin. Only honeycombing showed an excellent agreement, 19 other abnormalities had a fair to good reproducibility, and only two had a poor agreement. The importance of training is suggested by the improvement of the ICC values after excluding the the raters who only recieved a short instruction on how to score (ICC for trained observers). Excellent agreement for the three most trained raters is now achieved in nine, fair to good reproducibility in 12, and poor reproducibility in only one abnormal CT finding. Combining the closely related abnormal CT findings, bronchiectasis and bronchial deformation improves the ICC values from 0.540 and 0.592 to 0.674 (bronchial distortion). The merging of central and peripheral consolidations changes the ICC values from 0.875 and 0.630 to 0.834.

Table 3 Prevalence of abnormal CT findings

	Prevalence n (%)
Nodules	
Bronchovascular	41 (80.4)
Parenchymal	39 (76.4)
Subpleural	44 (86.3)
Coalescent	20 (39.2)
Sharp margin	36 (70.6)
Other abnormalities	
Peripheral consolidation	30 (58.8)
Central consolidation	20 (39.2)
Ground Glass	24 (47.1)
Mosaic pattern	10 (19.6)
Air trapping	26 (51)
Architectural distortion	35 (68.6)
Septal lines	30 (58.8)
Honeycombing	2 (3.9)
Bronchial deformation	34 (66.7)
Bronchiectasis	17 (33.3)
Pleural thickening	21 (41.2)
Parenchymal calcifications	8 (15.7)
Aspergiloma	0 (0)
Lymphnodes	
Right hilum	39 (76.5)
Left hilum	37 (72.5)
Mediastinal	47 (92.2)

Discussion

In this study, we investigated the reliability of scoring chest CT findings in patients with thoracic sarcoidosis. The results show that although there is a considerable variation in ICCs between the various CT findings, scoring of most of these abnormalities is possible with good inter-rater reliability. Most items of the

Table 4 Inter-rater reliability of scoring abnormal CT findings in a cohort of sarcoidosis patients

CT abnormality	ICC	ICC for trained observers	*p*-value for trained observers
Honeycombing	0.804	0.905	0.783
Central consolidation	0.728	0.875	0.102
Nodules subpleural	0.719	0.802	0.177
Size in mm	0.710	0.756	<0.001
Architectural distortion	0.706	0.844	<0.001
Confluent nodules	0.700	0.676	<0.001
Mediastinal lymphnode diameter	0.692	0.799	<0.001
Right hilar lymphnode diameter	0.676	0.754	<0.001
Nodules parenchymal	0.670	0.878	0.286
Percentage	0.662	0.743	0.006
Nodules bronchovascular	0.638	0.648	<0.001
Left hilar lymphnode diameter	0.634	0.710	<0.001
Septal thickening	0.609	0.647	0.034
Calcifications in lung parenchyma	0.537	0.757	0.288
Air trapping	0.526	0.629	0.045
Ground glass	0.508	0.552	0.014
Peripheral consolidation	0.488	0.630	0.003
Bronchiectasis	0.470	0.540	<0.001
Pleural thickening	0.426	0.511	<0.001
Bronchial deformation	0.407	0.592	<0.001
Mozaic pattern	0.368	0.557	0.076
Margin	0.112	0.098	<0.001
Bronchial distortion	0.634	0.674	0.077
bronchiectasis and bronchial deformation			
Consolidation	0.665	0.834	0.029
central and peripheral consolidation			

ICC intra class correlation

scoring system can therefore be used in diagnostic, prognostic and outcome studies in sarcoidosis.

Current CT scoring systems are coarse and, in our opinion, do not use all available data from CT analysis. In this study it was our aim to develop a reproducible unbiased scoring system. It is important to stress that the aim of this study was not to correlate the scoring of CT findings with physiological parameters or outcome measures. It is possible that some of the CT items scored do not have a clinical significance or, in contrast, are clinically important, but lack a good agreement. At a later stage we will address this by correlating CT items with physiological parameters and other outcome measures. This way we hope to answer relevant clinical questions including comparing the prognostic value of HRCT versus PET/CT. The scoring method described in this manuscript is time consuming (10 – 20 min per scan) and thus is not clinically applicable, but it does provide valuable data on unbiased scoring of all abnormal CT findings encountered in pulmonary sarcoidosis. We believe this article shows promising results enabling further investigations towards a clinically applicable scoring system.

To our knowledge, this is the first study focussing on the scoring reliability of all possible CT findings in patients with proven sarcoidosis. Because we were interested in whether any chest radiologist could score these abnormalities, we used five raters with different levels of training and experience in the field of chest radiology, an approach that has not been taken before. The strength of this study is that it primarily focussed on reliability and was conducted and presented according to the GRRAS guidelines [19]. This enables a straightforward interpretation and comparison with other reliability studies. Furthermore, all CT findings were defined according to generally accepted criteria [17] and we aimed to score morphology without interpretation. We accepted the fact that different pathological processes could be represented by the same abnormal CT finding. For instance, hilar radiating masses, although appearing to be fibrotic, often consist of fibrosis as well as active inflammation. Whether or not masses are fibrotic cannot always be differentiated on HRCT. Only serial imaging or PET/CT can demonstrate if they are fibrotic or also represent active inflammatory disease [22]. These hilar masses were regarded as consolidations and scored as being central or peripheral in location. The same applies for bronchial deformation. We accepted the fact that bronchial deformation in our scoring system can represent fibrosis, but also external compression by enlarged lymph nodes. In line with this, we also decided not to include descriptive items like the galaxy and cluster sign in our scoring system. Although these signs are considered to be virtually pathognomic and thus helpful in diagnosis sarcoidosis they can also be encountered in other diseases and are in fact interpretations of nodular disease and captured by the items confluent and parenchymal nodules [23]. We believe that for a scoring system to be successful and eventually clinically applicable, abnormalities need to be well-described, leaving little room for interpretation. For the above-mentioned reasons, we decided not to use existing semi-quantative scales used for HRCT assessment because these are biased by using a selection of CT items or focus on the disease outcome [6, 7, 24]. Because of the superiority of HRCT over chest radiography in detecting lymphadenopathy and parenchymal abnormalities, we also did not use existing chest radiography systems.

CT scoring systems

Although scoring methods are different there are a few studies describing the scoring of CT findings to compare our results with [5–9, 14]. In 2003, Drent et al described inter-rater agreement using the scoring system proposed by Oberstein [6, 25]. In this study, agreement was investigated between two observers for bronchovascular bundle thickening or irregularity (BVB), intra-parenchymal nodules (ND) and consolidations including groundglass opacification (PC), septal and non septal lines (LS), pleural thickening (PL), and lymphnode enlargement defined as having a short axis diameter of 1 cm or more (LN). ICCs in this study, which included 80 patients with proven sarcoidosis, ranged from 0.36 for LN and BVB to 0.43, 0.57, 0.62, and 0.78 for PL, LS, PC, and ND, respectively. Earlier, Abehsera et al investigated CT patterns of fibrosis in 80 patients with fibrotic pulmonary sarcoidosis [7]. They identified a bronchial distortion pattern including bronchial deformation and traction bronchiectasis with or without surrounding masses; a honeycombing pattern and a linear pattern including hilar peripheral lines, distorted septal reticulation and translobular lines. Agreement between two observers scoring these patterns was good with a kappa of 0.87 for recognizing the main CT pattern. However, no subanalysis was performed for the separate CT findings that defined the main patterns. In our study, four fibrosis-related abnormalities were scored: honeycombing, architectural distortion, bronchial deformation, and bronchiectasis. The ICCs of the latter two were lower compared to Abehsera. A possible explanation could be that it is difficult to distinguish bronchial deformation from bronchiectasis when they are located centrally with a surrounding mass (a common finding in pulmonary sarcoidosis). Abehsera et al. addressed this by defining both bronchiectasis and bronchial deformation as bronchial distortion. When we used that approach too, the ICC for bronchial distortion improved to 0.64. The question remains whether these abnormalities should be combined or scored separately. Although the combination yields a higher reliability, important information could be lost because deformation can be due to compression by lymph nodes and does not necessarily indicate fibrosis, while bronchiectasis usually does.

Chest radiographic scoring systems

For decades, the classification system presented by Scadding has been used because of its prognostic value [15]. It has also been used as a scoring tool in studies determining therapy response, but is limited by its wide range in inter-rater agreement (kw 0.43 – 0.80) [26, 27]. Other scoring systems used in studies investigating therapy response are the International Labor Organization (ILO) classification for radiographs for pneumoconiosis and the Muers scoring system (modified ILO). These perform slightly better than the Scadding staging (kw 0.33 – 0.87) and, in addition have demonstrated a correlation between the extent and profusion of components scored and functional parameters [28]. Our study has shown that the reliability of CT scoring is more or less comparable to that of the radiographic scoring. Considering the fact that CT has a higher sensitivity for detecting abnormalities it may, therefore, be more suited for scoring sarcoidosis-related abnormalities and developing a scoring system for future investigations.

Study limitations

One of the limitations of this study is that some CT findings are relatively uncommon in our cohort, which makes it difficult to reliably estimate inter-rater agreement. Honeycombing for instance is an abnormality present in only 4 % of the cases. Although the agreement of scoring this item in our study is high (ICC 0.91), a recent publication on this important feature of fibrotic lung disease reports a considerable variation [29]. An explanation for this could be that in pulmonary sarcoidosis honeycombing is easier to detect compared to fibrotic lung disease in which Usual Interstitial Pneumonia (UIP) is suspected. Second, scoring in this study was performed in a semi quantitative way. This generally results in higher agreement compared to a real continuous or nearest to 5 % scoring system, but comes at the expense of discriminatory power [30]. The question is, however, whether a nearest to 5 % method is desirable, because this could hamper the applicability in daily clinical practice. Last, the effect of training of raters was not adequately investigated due to the lack of pre- and post-training comparison. Although analysis of the data shows that agreement improves when the results of two experienced but untrained radiologists are left out, no firm conclusions can be made on the effect of training on agreement of scoring CT items.

In conclusion, scoring of thoracic sarcoidosis-related abnormal CT findings in a visual semi quantitative manner resulted in fair to excellent inter-rater reliability. Although there is some variation, most of the abnormalities have an ICC greater than 0.60 and can, therefore, be used in future investigations using this scoring system.

Acknowledgements The scientific guarantor of this publication is prof. J.C. Grutters. The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article. The authors state that this work has not received any funding. One of the authors has significant statistical expertise. Institutional review board approval was obtained. Written informed consent was obtained from all subjects (patients) in this study. None of the study subjects or cohorts have been previously reported. Methodology: retrospective, cross sectional study, multicenter study.

References

1. Chiles C (2002) Imaging features of thoracic sarcoidosis. Semin Roentgenol 37:82–93
2. Hennebicque AS, Nunes H, Brillet PY, Moulahi H, Valeyre D, Brauner MW (2005) CT findings in severe thoracic sarcoidosis. Eur Radiol 15(1):23–30
3. Criado E, Sanchez M, Ramirez J et al (2010) Pulmonary sarcoidosis: Typical and atypical manifestations at high-resolution CT with pathologic correlation. Radiographics 30:1567–1586
4. Spagnolo P, Sverzellati N, Wells AU, Hansell D (2014) Imaging aspects of the diagnosis of sarcoidosis. Eur Radiol 24(4):807–816
5. Naccache JM, Lavole A, Nunes H et al (2008) High-resolution computed tomographic imaging of airways in sarcoidosis patients with airflow obstruction. J Comput Assist Tomogr 32:905–912
6. Drent M, De Vries J, Lenters M et al (2003) Sarcoidosis: Assessment of disease severity using HRCT. Eur Radiol 13: 2462–2471
7. Abehsera M, Valeyre D, Grenier P, Jaillet H, Battesti JP, Brauner MW (2000) Sarcoidosis with pulmonary fibrosis: CT patterns and correlation with pulmonary function. AJR Am J Roentgenol 174:1751–1757
8. Bergin CJ, Bell DY, Coblentz CL et al (1989) Sarcoidosis: Correlation of pulmonary parenchymal pattern at CT with results of pulmonary function tests. Radiology 171:619–624
9. Hansell DM, Milne DG, Wilsher ML, Wells AU (1998) Pulmonary sarcoidosis: Morphologic associations of airflow obstruction at thin-section CT. Radiology 209:697–704
10. Murdoch J, Muller NL (1992) Pulmonary sarcoidosis: Changes on follow-up CT examination. AJR Am J Roentgenol 159:473–477
11. Brauner MW, Lenoir S, Grenier P, Cluzel P, Battesti JP, Valeyre D (1992) Pulmonary sarcoidosis: CT assessment of lesion reversibility. Radiology 182:349–354
12. Fazzi P, Sbragia P, Solfanelli S, Troilo S, Giuntini C (2001) Functional significance of the decreased attenuation sign on expiratory CT in pulmonary sarcoidosis. Chest 119:1270–1274
13. Akira M, Kozuka T, Inoue Y, Sakatani M (2005) Long-term follow-up CT scan evaluation in patients with pulmonary sarcoidosis. Chest 127:185–191
14. Zappala CJ, Desai SR, Copley SJ et al (2014) Accuracy of individual variables in the monitoring of long-term change in pulmonary sarcoidosis as judged by serial high- resolution CT scan data. Chest 145: 101–107
15. Scadding JG (1961) Prognosis of intrathoracic sarcoidosis in england. A review of 136 cases after five years' observation. Br Med J 2:1165–1172

16. Costabel U, Hunninghake GW (1999) ATS/ERS/WASOG statement on sarcoidosis. sarcoidosis statement committee. american thoracic society. european respiratory society. world association for sarcoidosis and other granulomatous disorders. Eur Respir J 14:735–737
17. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Muller NL, Remy J (2008) Fleischner society: Glossary of terms for thoracic imaging. Radiology 246:697–722
18. Sverzellati N, Devaraj A, Desai SR, Quigley M, Wells AU, Hansell DM (2011) Method for minimizing observer variation for the quantitation of high-resolution computed tomographic signs of lung disease. J Comput Assist Tomogr 35:596–601
19. Kottner J, Audige L, Brorson S et al (2011) Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol 64:96–106
20. Rosner B (2010) Fundamentals in biostatisctics 7th, edn. Brooks/Cole, Pacific Grove California United States
21. Rankin G, Stokes M (1998) Reliability of assessment tools in rehabilitation: An illustration of appropriate statistical analyses. Clin Rehabil 12:187–199
22. Mostard RL, Verschakelen JA, van Kroonenburgh MJ (2013) Severity of pulmonary involvement and (18)F-FDG PET activity in sarcoidosis. Respir Med 107(3):439–447
23. Marchiori E, Zanetti G, Barreto MM, de Andrade FT, Rodrigues RS (2011) Atypical distribution of small nodules on high resolution CT studies: patterns and differentials. Respir Med 105(9):1263–1267
24. Walsh SL, Wells AU, Sverzellati N et al (2014) An integrated clinicoradiological staging system for pulmonary sarcoidosis: a case-cohort study. Lancet Respir Med 2(2):123–130
25. Oberstein A, von Zitzewitz H, Schweden F, Muller-Quernheim J (1997) Non invasive evaluation of the inflammatory activity in sarcoidosis with high-resolution computed tomography. Sarcoidosis Vasc Diffuse Lung Dis 14:65–72
26. Baughman RP, Shipley R, Desai S et al (2009) Changes in chest roentgenogram of sarcoidosis patients during a clinical trial of infliximab therapy: comparison of different methods of evaluation. Chest 136(2):526–535
27. Zappala CJ, Desai SR, Copley SJ et al (2011) Optimal scoring of serial change on chest radiography in sarcoidosis. Sarcoidosis Vasc Diffuse Lung Dis 28(2):130–138
28. Muers MF, Middleton WG, Gibson GJ et al (1997) A simple radiographic scoring method for monitoring pulmonary sarcoidosis: relations between radiographic scores, dyspnoea grade and respiratory function in the British Thoracic Society Study of Long-Term Corticosteroid Treatment. Sarcoidosis Vasc Diffuse Lung Dis 14(1):46–56
29. Johkoh T, Sakai F, Noma S et al (2014) Honeycombing on CT; its definition, pathologic correlation, and future direction of its diagnosis. Eur J Radiol 83:27–31
30. Ng CS, Desai SR, Rubens MB, Padley SP, Wells AU, Hansell DM (1999) Visual quantitation and observer variation of signs of small airways disease at inspiratory and expiratory CT. J Thorac Imaging 14:279–285