Reliability of a Scoring System for Qualitative Evaluation of Lymphoscintigraphy of the Lower Extremities

Mojgan Ebrahim[1], Irina Savitcheva[1,2], and Rimma Axelsson[1,2]

[1]Division of Medical Imaging and Technology, Department of Clinical Science, Intervention and Technology (CLINTEC), Karolinska Institutet, Solna, Sweden; and [2]Division of Function and Imaging, Department of Medical Physics and Nuclear Medicine, Karolinska University Hospital, Huddinge, Sweden

Lymphoscintigraphy is an imaging technique to diagnose and characterize the severity of edema in the upper and lower extremities. In lymphoscintigraphy, a scoring system can increase the ability to differentiate between diagnoses, but the use of any scoring system requires sufficient reliability. Our aim was to determine the inter- and intraobserver reliability of a proposed scoring system for visual interpretation of lymphoscintigrams of the lower extremities. **Methods:** The lymphoscintigrams of 81 persons were randomly selected from our database for retrospective evaluation. Two nuclear medicine physicians scored these scans according to the 8 criteria of a proposed scoring system for visual interpretation of lymphoscintigrams of the lower extremities. Each scan was scored twice 3 mo apart. The total score was the sum of the scores for all criteria, with a potential range of 0 (normal lymphatic drainage) to 58 (severe lymphatic impairment). The intra- and interobserver reliability of the scoring system was determined using the Wilcoxon signed-rank test, percentage of agreement, weighted κ, and intraclass correlation coefficient with 95% confidence interval. In addition, for 7 categories, differences in total scores between and within observers were determined. **Results:** We found some insignificant differences between observers. Percentage agreement was high or very high, at 82.7%–99.4% between observers and 84.6%–99.4% within observers. For each criterion of the scoring system, the κ-correlations showed moderate to very good inter- or intraobserver reliability. The total scores for all criteria had good inter- and intraobserver reliability. Regarding the interobserver comparison, 66% and 64% of the difference in total scores were within ±1 scale point (−1, +1), and regarding the intraobserver comparison, 68% and 72% of the difference in total scores were within ±1 scale point. **Conclusion:** The proposed scoring system is a reliable tool for visual qualitative evaluation of lymph transport problems in patients with lymphedema of the lower extremities.

Key Words: lymphoscintigraphy; edema; scoring system; inter- and intra-observer reliability

J Nucl Med Technol 2017; 45:219–224
DOI: 10.2967/jnmt.116.185710

Received Nov. 9, 2016; revision accepted Apr. 11, 2017.
For correspondence or reprints contact: Rimma Axelsson, Karolinska Institutet, Radiology Department, Huddinge Hospital, Stockholm, 141 86, Sweden.
E-mail: rimma.axelsson@ki.se
Published online May 4, 2017.

Lymphoscintigraphy, the imaging test most commonly used to evaluate the lymphatic system, offers objective evidence to diagnose and characterize the severity of edema in the upper and lower extremities and to distinguish lymphatic from nonlymphatic edema (*1*). Quantitative and qualitative lymphoscintigraphy may complement each other, but in the clinical setting a qualitative assessment of morphologic features is used more frequently (*2*).

In qualitative analysis, lymphoscintigraphy can be used to obtain a detailed description of many characteristics. The most important criteria for identifying dysfunction of the lower extremities are delayed, asymmetric, or absence of visualization of regional lymph nodes, or divergence of lymph either through skin lymph vessels (i.e., dermal backflow) or into the deep lymphatic system (e.g., visualization of popliteal lymph nodes) (*3*). These abnormalities may have additional findings such as asymmetric visualization of lymphatic channels, collateral lymphatic channels, and interrupted vascular structures (*1*).

Scoring systems increase diagnostic differentiation when the results of qualitative lymphoscintigraphy are borderline (*1*). However, any scoring system or diagnostic test should be proven reliable and reproducible, as measured by inter- and intraobserver correlation, before general use in the population of interest.

We have developed a scoring system that includes 8 criteria for visual interpretation of lymphoscintigrams of the lower extremities. The aim of this study was to determine the inter- and intraobserver reliability of this proposed scoring system.

MATERIALS AND METHODS

This retrospective study at the Nuclear Medicine Department of Huddinge Hospital was performed in accord with the Declaration of Helsinki and was approved by the Ethics Committee (approval 2014/1964-31/1) on December 10, 2014.

Study Design

All patients who had undergone lymphoscintigraphy of the lower extremities between January and October 2013 were included in this study. Lymphoscintigraphy was performed after subcutaneous injection of 20 MBq of ^{99m}Tc-nanocolloid (Nanocoll; GE Healthcare, Amersham Health) between the first and second toes of each foot.

TABLE 1
The Scoring System

Criterion	Category	Options	Score
C_1	Display of lymphatic vessels	Visible entirely (extending to pelvis)	0
		Visible partially	3
		Not visible	10
C_2	Pattern of lymphatic vessels	Straight with ordinary course	0
		Straight with abnormal course	3
		Tortuous/prominent with some points of ordinary course	3
		Tortuous/prominent with some points of abnormal course	5
C_3	Uptake in inguinal nodes	Uptake before stress	0
		Uptake after stress	1
		No uptake at third hour	10
C_4	Uptake in pelvic nodes	Uptake	0
		No uptake	5
C_5	Uptake in lumbar nodes	Uptake	0
		No uptake	5
C_6	Uptake in leg nodes outside vessel: foot, knee, lower leg, thigh	No uptake	0
		Uptake	3
C_7	Focal accumulation	No focal accumulation	0
		Focal accumulation, increasing with time	10
C_8	Dermal backflow	No dermal backflow	0
		Dermal backflow	10

Both the swollen leg and the healthy leg were imaged, so that the two sides could be compared. Whole-body γ-camera imaging was performed (e.cam; Siemens). Images were recorded with a 20% window centered on the 140-keV photopeak of ^{99m}Tc, using a scanning speed of 10 cm/min. The lymphoscintigraphic assessment included images of the lower extremities at 4 times during the resting state (5, 20, 35, and 50 min after injection) and twice during exercise (60 and 180 min after injection) to show passive and active lymphatic flow, respectively.

These images were reviewed according to the criteria of the proposed scoring system by two nuclear medicine physicians, one with experience in reading lymphoscintigrams and the other without, at 2 scoring times 3 mo apart. To reduce the risk of bias, the images were anonymized and no clinical information about the patients was provided to the observers.

The proposed scoring system included 8 criteria (C_1–C_8): display of lymphatic vessels (C_1); pattern of lymphatic vessels (C_2); uptake in inguinal lymph nodes (C_3); uptake in pelvic lymph nodes (C_4); uptake in lumbar lymph nodes (C_5); uptake in leg lymph nodes outside the vessels, including foot, knee, lower leg, and thigh (C_6); focal accumulation (C_7); and dermal backflow (C_8) (Table 1). The use of a discontinuous scale for each criterion was based on existing knowledge and 30 y of experience in our hospital. In the literature, these criteria in the evaluation of lower-extremity lymphedema have been reported as a range of findings on lymphoscintigrams (*1,3*). Although the clinical meaning of, for example, C_6 is uncertain, it could not be scored in the same way as C_7 or C_8. Mere existence of focal tracer accumulation (C_7) or dermal backflow (C_8) is enough to diagnose lymphedema, and therefore weighted values were applied, with the highest value being applied for signs pathognomonic of lymphedema.

TABLE 2
Scores for Each Observer

	Observer 1		Observer 2	
Criterion	Scoring time 1	Scoring time 2	Scoring time 1	Scoring time 2
C_1	0.72 ± 2.36	0.88 ± 2.58	0.95 ± 2.51	0.85 ± 2.65
C_2	0.71 ± 1.36	0.79 ± 1.42*	0.57 ± 1.23	0.59 ± 1.29*
C_3	0.53 ± 1.58	0.56 ± 1.58	0.67 ± 1.89	0.65 ± 1.74
C_4	0.34 ± 1.26	0.31 ± 1.21	0.31 ± 1.21	0.22 ± 1.02
C_5	0.78 ± 1.82	0.68 ± 1.72	0.83 ± 1.87	0.77 ± 1.81
C_6	0.37 ± 0.99	0.42 ± 1.03	0.40 ± 1.01	0.39 ± 1.01
C_7	1.11 ± 3.15†	0.74 ± 2.63	0.43 ± 2.04†	0.43 ± 2.04
C_8	0.80 ± 2.72	1.04 ± 3.07	0.86 ± 2.82	0.99 ± 2.99
Total score	5.37 ± 8.49	5.43 ± 8.45	5.02 ± 7.89	4.88 ± 8.11

*Significant difference between observers at scoring time 2.
†Significant difference between observers at scoring time 1.
Data are mean ± SD.

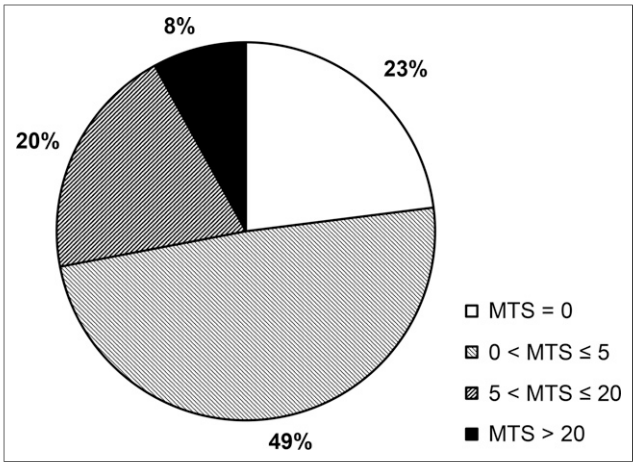

FIGURE 1. Percentage of lymphoscintigrams in 4 different groups of MTS.

C_1–C_3 were judged on images up to 60 min after injection, and C_4–C_8 were judged on 180-min images. C_1–C_3 were 3-point response scales, and C_4–C_8 were 2-point response scales. The total score was the sum of C_1–C_8. Potential total scores were in the range 0–58, spanning the spectrum from normal lymphatic drainage (0) to the most severe lymphatic impairment (58).

Statistics

Each criterion is considered an ordinal variable, but the total score is considered interval data. Ordinal data should not be analyzed with parametric measures, but the summations of all criteria in this scoring system can be analyzed parametrically. An overall mean and SD was computed for each criterion at both scoring times and for both observers. Using the Wilcoxon signed-rank matched-pairs test, interobserver scores were evaluated for each scoring time. The effect size for the Wilcoxon test was calculated by $r = |Z|/\sqrt{n}$, where r is effect size, $|Z|$ is the absolute value of normal approximation of the Wilcoxon test statistic, and n is the number of subjects in the study. An effect size of less than 0.30 was considered small (*4*). The Student t test was used to compare total scores between and within observers.

To evaluate the intraobserver and interobserver reliability of all criteria, percentage agreement and weighted κ were used. For the κ-values, 95% confidence intervals (CIs) were calculated. A κ of less than 0.4 was interpreted as poor agreement; a κ of 0.4–0.6, as moderate agreement; a κ of 0.6–0.8, as good agreement; and a κ of more than 0.8, as very good agreement (*5*). To investigate the intraobserver and interobserver reliability of the total scores, intraclass correlation coefficients (ICCs) with 95% CIs were determined (2-way mixed model together with single-measure opinion and type of absolute agreement). An ICC of less than 0.4 was interpreted as weak; an ICC of 0.4–0.74, as moderate; an ICC of 0.75–0.9, as strong, and an ICC of more than 0.9, as very strong (*6*). For intra- and interobserver comparisons, the difference in total score (DTS) for each of 7 criteria was categorized (DTS = 0, |1|, |2|, |3|, |4|, |5|, and >|5|). SPSS, version 22.0 (IBM), was used for all analyses. A P value of less than 0.05 was chosen as the significance level.

RESULTS

For this retrospective evaluation, lymphoscintigrams were available for 81 patients (66 women and 15 men; mean age ± SD, 57.5 ± 13.1 y) in our database. Fifty-four of these patients had no scintigraphic findings indicating lymphedema or a blockage in the lymphatic system in either leg. Twenty-two patients had some scintigraphic findings corresponding to lymphedema in the right or left leg, with the opposite leg having an ordinary status. Only 5 patients had some scintigraphic findings of the disease in both legs.

The scores for each observer at the 2 scoring times are shown in Table 2. The data were not normally distributed. The Wilcoxon test showed no significant differences between scoring times in either observer for any criterion. Significant differences between observers were found only for C_2 and C_7: at scoring time 2 for C_2 and at scoring time 1 for C_7, but the effect sizes of the Wilcoxon test regarding these criteria were small or very small (in both cases, $r \leq 0.25$), indicating no substantial differences between observers with respect to these criteria. Further analysis showed that the medians of the differences between scores were zero for all criteria. Overall, no significant differences in total score were found between scoring times for either observer using either the Wilcoxon test or the Student t test (in both tests, $P > 0.05$). There were also no significant differences in total score between observers at either scoring time ($P > 0.05$) (Table 2).

TABLE 3
Interobserver Reliability

Criterion	Scoring time 1		Scoring time 2	
	Agreement	κ	Agreement	κ
C_1	90.7%	0.635 (0.559–0.711)	90.7%	0.596 (0.518–0.647)
C_2	82.7%	0.490 (0.418–0.562)	83.3%	0.526 (0.448–0.604)
C_3	90.1%	0.782 (0.718–0.849)	88.9%	0.767 (0.703–0.831)
C_4	99.4%	0.949 (0.899–1.00)	98.1%	0.814 (0.710–0.918)
C_5	93.8%	0.790 (0.723–0.857)	96.9%	0.876 (0.822–0.930)
C_6	92.6%	0.673 (0.610–0.736)	92.0%	0.668 (0.608–0.728)
C_7	92.0%	0.448 (0.383–0.513)	93.2%	0.411 (0.347–0.475)
C_8	95.7%	0.719 (0.658–0.780)	91.9%	0.560 (0.504–0.616)

Data in parentheses are 95% CI.

TABLE 4
Intraobserver Reliability

Criterion	Observer 1		Observer 2	
	Agreement	κ	Agreement	κ
C_1	95.1%	0.781 (0.712–0.850)	85.8%	0.458 (0.413–0.503)
C_2	84.6%	0.606 (0.534–0.678)	85.7%	0.543 (0.505–0.581)
C_3	91.4%	0.809 (0.744–0.874)	88.3%	0.519 (0.464–0.574)
C_4	99.4%	0.949 (0.899–1.00)	98.1%	0.814 (0.710–0.918)
C_5	97.5%	0.925 (0.882–0.968)	96.3%	0.863 (0.808–0.918)
C_6	91.4%	0.637 (0.583–0.691)	95.1%	0,823 (0.756–0.890)
C_7	93.8%	0.633 (0.558–0.708)	95.1%	0.410 (0.331–0.489)
C_8	93.8%	0.659 (0.584–0.734)	97.5%	0.839 (0.758–0.920)

Data in parentheses are 95% CI.

The mean of the total scores (MTS) was categorized as representing normal findings (MTS = 0), very mildly altered findings ($0 < MTS \leq 5$), mildly or moderately altered findings ($5 < MTS \leq 20$), or greatly altered findings ($MTS > 20$), and the percentages of observations in these 4 groups were calculated (Fig. 1). As seen in Figure 1, on average, the total scores were less than or equal to 5 in 72% of the lymphoscintigrams, reflecting absence of disease or disease in very early stages in most patients.

Interobserver reliability (Table 3) was high to very high (82.7%–99.4% agreement). According to the interpretation of Altman (*5*), the scoring system showed good or very good interobserver correlations for 6 criteria (C_1, C_3, C_4, C_5, C_6, and C_8) and moderate correlations for 2 criteria (C_2 and C_7) at scoring time 1. At scoring time 2, the scoring system showed good or very good interobserver correlations for 4 criteria (C_3, C_4, C_5, and C_6) and moderate correlations for the other criteria (C_1, C_2, C_7, and C_8). According to the interpretation of Fleiss (*6*), the total scores from all criteria showed moderate or strong interobserver reliability. The ICC was 0.884 (95% CI, 0.845–0.913) for scoring time 1 and 0.709 (95% CI, 0.604–0.786) for scoring time 2.

Intraobserver reliability analysis (Table 4) revealed high or very high intraobserver agreement (84.6%–99.4%). Using the interpretation of Altman and Fleiss (*5,6*), the scoring system had 3 very good κ-correlations (C_3, C_4, and C_5) and 5 good κ-correlations (C_1, C_2, C_5, C_6, and C_7) for scoring time 1 and 4 very good κ-correlations (C_4, C_5, C_6, and C_8) and 4 moderate κ-correlations (C_1, C_2, C_3, and C_7) for scoring time 2. By the criteria of Fleiss (*6*), the total scores for all criteria also showed strong intraobserver reliability. The ICC was 0.805 (95% CI, 0.734–0.857) for observer 1 and 0.906 (95% CI, 0.874–0.930) for observer 2.

The DTS between and within observers is shown in Figures 2–5.

In the interobserver comparison, 66% and 64% of DTS were within ±1 scale point (i.e., DTS = 0 or |1|). This means that the total score for both observers at scoring times 1 and 2 was the same or nearly the same in 66% and 64% of all lower extremities, respectively (Figs. 2 and 3). The proportion of lower extremities with a DTS of more than |5| was 12% at scoring time 1 and 14% at scoring time 2. No significant percentage differences in the categorized DTS were seen between observers.

In the intraobserver comparison, 68% and 72% of DTS were within ±1 scale point. This means that the total score for each observer at scoring times 1 and 2 was the same or nearly the same in 68% and 72% of all lower extremities,

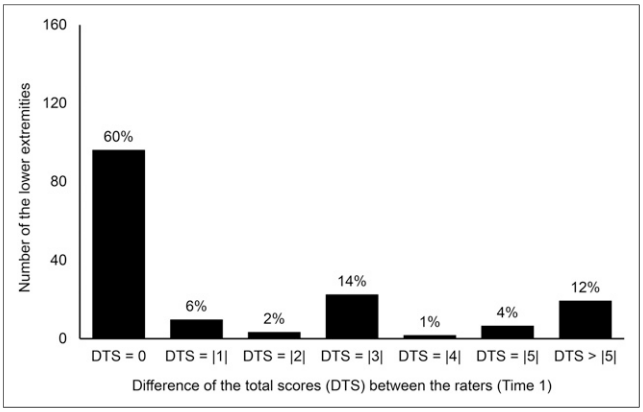

FIGURE 2. Histogram of DTS between observers at scoring time 1. Data are absolute values of DTS.

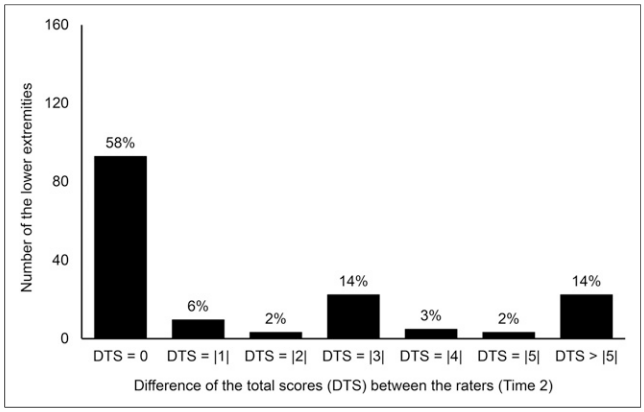

FIGURE 3. Histogram of DTS between observers at scoring time 2. Data are absolute values of DTS.

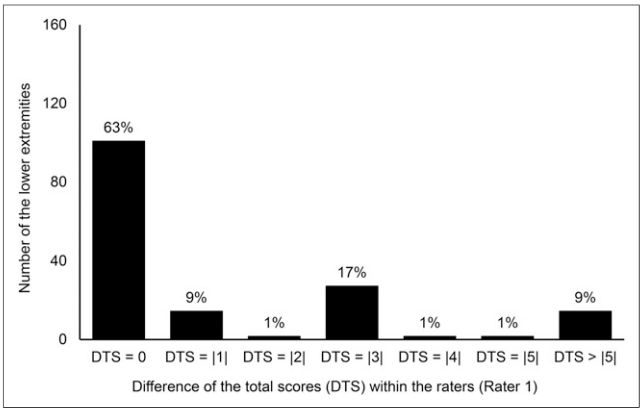

FIGURE 4. Histogram of DTS between scoring times for observer 1. Data are absolute values of DTS.

respectively (Figs. 4 and 5). The proportion of lower extremities with a DTS of more than |5| was 9% for observer 1 and 10% for observer 2. No significant percentage differences in the categorized DTS were seen within observers.

DISCUSSION

Despite the recent emphasis on the advantages of lymphoscintigraphy for detection of lymphedema, a standardized and reliable method of evaluating and reporting imaging results is still needed. We previously showed that there is a need for a simple tool to use in everyday practice (*7*). We have compiled several important criteria for lymphedema into a new scoring system for visual interpretation of lymphoscintigrams, but before this scoring system can be applied in clinical practice, its reliability and reproducibility require testing. Such testing was the purpose of the current study.

All assessments of criteria will be affected by the presence of random errors. Thus, when assessments are repeated, some or perhaps even most of the scores for individual subjects will change. It is also likely that the mean score and SD will differ between different time points or different raters (Table 2), but a measurement tool with sufficient reliability should result in fewer differences in scores between repeated measurements.

Our analysis showed no statistically significant intraobserver differences in any criteria of the scoring system, and interobserver differences were found for only 2 criteria (C_2 and C_7).

The mean difference in each criterion between observers at each scoring time, or between scoring times for each observer, was smaller than its respective SD, reflecting a skew in our data (*8*). The results of the Wilcoxon test were not suitable because the medians of the differences between scores were zero for all criteria (*9*). On the other hand, the traditional significance tests cannot assess the size or importance of effects. For example, in a large sample, even a small effect could be statistically significant. Therefore, in this situation it is important to report measures of effect size (*10*). We could find significant differences in some criteria between observers (i.e., C_2 and C_7), but the effect sizes of these differences were small or very small ($r \leq 0.25$). These measures of effect size reflected no substantial differences between these scores. Overall, no significant DTS was found between or within observers.

For some criteria, we found moderate κ-correlations for reliability, as can be explained by the skew in the score distribution, especially in the context of the high percentage of agreement (*11,12*). In addition, we found very high percentages of agreement between scoring times for observer 2 regarding C_7 and C_8 (i.e., 95.1% for C_7 and 97.5% for C_8; Table 4), but the κ-coefficient was fairly low for C_7 ($\kappa = 0.410$) and, in contrast, was very good or excellent for C_8 ($\kappa = 0.839$). Both these criteria were dichotomized (i.e., with a 2-point response scale: 0 or 10; Table 1), and therefore the difference between the κ-coefficients could not be explained by a difference in possibility. The number of legs in our sample that received 10 points from observer 2 was very small regarding C_7 (7 received 10 points and 155 received 0 points), compared with the number of legs that received 10 points from the same observer regarding C_8 (16 received 10 points and 148 received 0 points). This great discrepancy between the percentages of agreement and κ-coefficients revealed a disadvantage to using κ as a measure of reliability (*13*). The prevalence of a finding in an observed sample influences κ-coefficients in a manner similar to the way the prevalence of a disease under clinical consideration influences predictive values (*13,14*). The κ-statistic alone may have less interpretive value in the data analysis because of the low prevalence of a certain score in our sample and the disproportionate number of zero values in our data.

Combination of the high or very high percentages of agreement, the moderate or strong ICCs, and the moderate to very good κ-values of inter- and intraobserver reliability found for our scoring system suggests that the system is reproducible. We found that 64% or 66% of the DTS between observers was within ±1 scale point (−1, +1) and that 68% or 72% of the DTS within observers was within ±1 scale point. The percentages of all 7 categorized DTS were almost the same between observers at each scoring

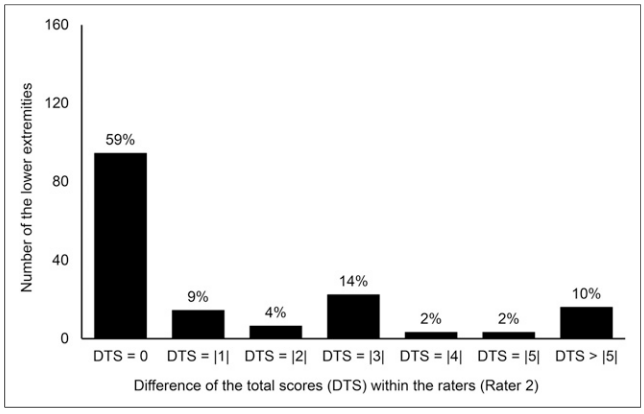

FIGURE 5. Histogram of DTS between scoring times for observer 2. Data are absolute values of DTS.

time and between scoring times for each observer. A DTS of more than |5| was slightly more common between observers than within observers. Overall, this scoring system demonstrated slightly better intraobserver reliability than interobserver reliability.

Lymphoscintigrams make up a very small proportion of all scans that a nuclear medicine physician reviews. Therefore, nuclear medicine physicians without sufficient experience in reading this type of scan will show a lower intraobserver correlation, as can explain the variance in inter- and intraobserver reliability observed in our study.

A wide variation in the reliability of lymphoscintigraphy of the upper extremities was reported for a study from 2014 (*15*). In that study, quantitative elements of lymphoscintigraphy had weak to moderate reproducibility but qualitative elements had excellent reproducibility. In another study, moderate inter- and intraobserver reliability was reported for evaluation of dermal backflow in qualitative lymphoscintigraphy of the upper extremities (*2*). On the other hand, some studies (*16–19*) have shown a high variation in reliability for interpretation of different types of scans. In one of these studies (*18*), it was pointed out that difficult cases can create a larger proportion of disagreement. The severity and extent of disease in patients may also influence the degree of agreement (*20*). It is easier for nuclear medicine physicians to diagnose abnormalities seen during the late phase of lymphedema. In our study, because the scans were unselected and many of the patients were not in the late phase of the disease, the prevalence of pathologic findings in the scans was low—potentially negatively affecting interobserver agreement. Poor imaging technique, lack of knowledge or experience, and clinical misjudgment have been found to be the 3 factors most affecting the reliability of image interpretation (*2*).

CONCLUSION

Our data show that the proposed scoring system for scintigraphic evaluation of patients with lymphedema of the lower extremities is easily applied, has good to excellent reproducibility in experienced hands, and can be recommended for further validation.

DISCLOSURE

No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Szuba A, Shin WS, Strauss HW, Rockson S. The third circulation: radionuclide lymphoscintigraphy in the evaluation of lymphedema. *J Nucl Med.* 2003;44:43–57.
2. Dylke ES, McEntee MF, Schembri GP, et al. Reliability of a radiological grading system for dermal backflow in lymphoscintigraphy imaging. *Acad Radiol.* 2013;20:758–763.
3. Keramida G, Humphrys M, Ryan N, Peters AM. "Stocking effect" in lymphoscintigraphy. *Lymphat Res Biol.* 2014;12:194–196.
4. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates; 1988:149–150.
5. Altman D. *Practical Statistics for Medical Research.* London, U.K.: Chapman & Hall; 1991:400–405.
6. Fleiss JL. *The Design and Analysis of Clinical Experiments.* New York, NY: Wiley; 1986:17–20.
7. Ebrahim M, Axelsson R. National inventory of lymphoscintigraphy procedure. *Med Res Arch.* 2015;2:1–9.
8. Altman DG, Bland JM. Detecting skewness from summary information. *BMJ.* 1996;313:1200.
9. Twomey PJ, Viljoen A. Limitations of the Wilcoxon matched pairs signed ranks test for comparison studies. *J Clin Pathol.* 2004;57:783.
10. Grissom RJ, Kim JJ. *Effect Sizes for Research: A Broad Practical Approach.* Mahwah, NJ: Lawrence Erlbaum Associates; 2005:280–281.
11. Post MW, de Witte LP. Good inter-rater reliability of the Frenchay Activities Index in stroke patients. *Clin Rehabil.* 2003;17:548–552.
12. Hirsch O, Keller H, Albohn-Kühne C, Krones T, Donner-Banzhoff N. Pitfalls in the statistical examination and interpretation of the correspondence between physician and patient satisfaction ratings and their relevance for shared decision making research. *BMC Med Res Methodol.* 2011;11:71.
13. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med.* 2005;37:360–363.
14. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol.* 1990;43:543–549.
15. Devoogdt N, Van den Wyngaert T, Bourgeois P, et al. Reproducibility of lymphoscintigraphic evaluation of the upper limb. *Lymphat Res Biol.* 2014;12:175–184.
16. Bellamy N, Klestov A, Muirden K, et al. Perceptual variation in categorizing individuals according to American College of Rheumatology classification criteria for hand, knee, and hip osteoarthritis (OA): observations based on an Australian Twin Registry study of OA. *J Rheumatol.* 1999;26:2654–2658.
17. Nelitz M, Guenther KP, Gunkel S, et al. Reliability of radiological measurements in the assessment of hip dysplasia in adults. *Br J Radiol.* 1999;72:331–334.
18. Beam CA, Conant EF, Sickles EA. Factors affecting radiologist inconsistency in screening mammography. *Acad Radiol.* 2002;9:531–540.
19. Moran M, Ryan J, Higgins M, et al. Poor agreement between operators on grading of the placenta. *J Obstet Gynaecol.* 2011;31:24–28.
20. Caglar M, Kiratli PO, Karabulut E. Inter- and intraobserver variability of ^{99m}Tc-DMSA renal scintigraphy: impact of oblique views. *J Nucl Med Technol.* 2007;35:96–99.