A Multicenter Trial on Interobserver Reproducibility in Reporting on ^{99m}Tc-DMSA Planar Scintigraphy: A Belgian Survey

Carlos De Sadeleer, Marianne Tondeur, Koen Melis, Marie-Benedict Van Espen, Jean Verelst, Hamphrey Ham, and Amy Piepsz

St. Pierre Hospital, Université Libre de Bruxelles, Brussels; and Akademisch Ziekenhuis, Vrije Universiteit Brussel, Brussels, Belgium

Conflicting opinions have been expressed regarding reproducibility in ^{99m}Tc-dimercaptosuccinic acid (DMSA) planar renal image interpretation. The purpose of this investigation was to determine the level of interobserver variability among a large group of Belgian nuclear medicine physicians who evaluated a randomly selected series of DMSA planar scintigraphic examinations performed on children and adults. **Methods:** All Belgian nuclear medicine centers (n = 82) were invited to participate in a reproducibility study on ^{99m}Tc-DMSA scintigraphy. ^{99m}Tc-DMSA scans obtained on 10 adults and 40 children were randomly selected from the databases of 2 hospitals. Those participating in this investigation (65 centers = 79%) received a series of computer disks containing 50 ^{99m}Tc-DMSA studies. To avoid potential problems related to unfamiliar display, the disks were formatted to be interpretable using the participants' own computer systems. Each participant was then free to use his or her usual display (hard copies, contrast enhancement, color scale, gray scale, and so forth). For each kidney, the observers had to choose between the following answers: normal, abnormal, equivocal, and lack of quality. **Results:** Forty-two responses were obtained from a wide variety of institutions and from observers with different levels of experience in interpreting ^{99m}Tc-DMSA scintigraphy. Altogether, the following data were obtained: 60.8% normal, 25.2% abnormal, 7.0% equivocal, and 3.2% lack of quality. The median percentage of agreement (overall reproducibility) for the 42 observers was 92%. When the results of all 42 observers were compared, the median agreements on normality and abnormality were 93.5% and 90.5%, respectively. In a small number (n = 4) of kidneys, reproducibility was poor and ranged from 51% to 70%. Except for 2 outliers, all observers had almost the same level of performance. **Conclusion:** A large number of Belgian nuclear medicine physicians participated in evaluating a large randomly selected sample of ^{99m}Tc-DMSA studies, and excellent interobserver agreement was found.

Key Words: ^{99m}Tc-DMSA; interobserver reproducibility; planar scintigraphy

J Nucl Med 2000; 41:23–26

Conflicting opinions have been expressed regarding reproducibility in ^{99m}Tc-dimercaptosuccinic acid (DMSA) planar renal image interpretation (*1–7*). The purpose of this investigation was to determine the level of interobserver variability among a large group of Belgian nuclear medicine physicians who evaluated a randomly selected sample of ^{99m}Tc-DMSA planar studies performed on children and adults.

Received Jan. 4, 1999; revision accepted May 18, 1999.

For correspondence or reprints contact: Carlos De Sadeleer, MD, Borrekent 51, 9450 Haaltert, Belgium.

MATERIALS AND METHODS

Study Design

On the basis of a listing of the Belgian Society of Nuclear Medicine, all Belgian nuclear medicine centers (n = 82) were invited to participate and to report on a series of ^{99m}Tc-DMSA studies. Those who were willing to participate in the study (n = 65, 79%) received a series of computer disks containing 50 ^{99m}Tc-DMSA studies.

The original images were converted in Interfile and then were reformatted to be interpretable on each participating center's computer system. These data translations were unsuccessful for 2 commercial systems and, for this reason, 16 of the 65 centers could not participate to the study. Moreover, data of 1 of 50 patients could not be included in the study because of a translation error.

Forty-two responses were obtained from a wide variety of institutions and from observers with different levels of experience in reporting ^{99m}Tc-DMSA scintigraphic findings. Most responses were given by individual observers; in some departments, the responses were given by consensus within the department.

Patients

^{99m}Tc-DMSA scans of 50 patients were randomly selected from the routine databases (optical disks) of 2 nuclear medicine departments: 25 studies from a general hospital (St. Pierre Hospital) and 25 studies from an academic hospital (Akademisch Ziekenhuis).

The only selection criterion was to include 10 adults and 40 children (age, 3 mo to 15 y). No attempt was made to select studies on the basis of technical quality.

Urinary tract infection was the reason for performing the scintigraphic examinations on almost all patients. In about half of the patients, the examination was performed during the acute phase of infection; in the other half, it was performed during follow-up.

Imaging

All patients were imaged on a digital γ camera computer system equipped with a low-energy, high-resolution collimator (Elscint SP4 [Elscint, Haifa, Israel] or Sopha DSX Rectangular [Sopha Medical Vision, Buc, France]). Scintigraphy was performed 2–4 h after intravenous injection of ^{99m}Tc-DMSA. The adult dose was 110 MBq (3 mCi); the doses administered to children were reduced

according to guidelines of the European Pediatric Task Group (*8*). All patients were examined in supine position, collimator side up. Young children were positioned directly on the collimator. The total acquisition time per image was variable. At least 3 images (1 posterior view and 2 posterior oblique projections) were obtained in a 256 × 256 matrix for a minimum of 300,000 counts each. Zoomed or pinhole images (or both) were occasionally obtained in young children.

Reporting

All participants received a series of disks containing all images formatted to be interpretable using their own computer systems. The participants were asked to use their usual processing program and display (hard copies, contrast enhancement, color scale, gray scale, and so forth).

Each observer had to choose, for each kidney, between 4 answers: normal, abnormal, equivocal, and lack of quality. The abnormal answer included all types of abnormalities (including malformation). The equivocal answer had to be chosen when the observer could not decide between normal and abnormal; the lack of quality answer had to be chosen when the observer considered that the quality of the image was so poor that no valuable interpretation could be given. The remaining cases were considered normal.

No patient history was supplied. No guidelines concerning interpretation of the images were given to the observers.

Data were collected centrally and were withheld from all observers.

RESULTS

In total, 4116 answers relating to 98 kidneys were given by 42 observers. There were 2501 normal answers (60.8%), 1037 definitely abnormal answers (25.2%), 290 equivocal reports (7.0%), and 131 quality insufficient (3.2%). Figure 1 illustrates the observers' responses. One hundred fifty-seven answers were missing, primarily because of image deformation or deterioration related to translation errors (3.8%).

Analysis of Overall Reproducibility

For each kidney, interobserver variability was calculated by simply determining the percentage of normal and abnormal answers (the equivocal and quality insufficient responses were excluded). The higher of these 2 percentages was considered the index of interobserver agreement for that kidney. A median percentage agreement on the whole population of 98 kidneys was then calculated.

A high level of interobserver agreement was found. The median percentage of agreement (overall reproducibility) was 92%. The median agreements on normality and abnormality were 93.5% and 90.5%, respectively. In a small number of kidneys (n = 4), poor reproducibility (range, 50%–70%) was observed (Fig. 2).

Influence of Observers

To test whether some observers reported differently than others, a score was given to each observer. For each kidney, each observer received several points corresponding to his or her response compared with the response of the majority. If, for a given kidney, the observer gave an abnormal answer, whereas 70% of the observers gave a normal answer, the observer received 30 points; in the opposite case, a score of 70 points was given. The same procedure was followed for each kidney, and the total number of points obtained by the observer reflected his or her conformity to most of the observers. Only the normal and abnormal responses were considered, and a correction of the total score was introduced for other types of responses, including the lack of response.

The median score for the observers was 87 points. Except for 2 outliers, all observers had almost the same level of performance. One of these outliers considered almost all kidneys as being abnormal (Fig. 3).

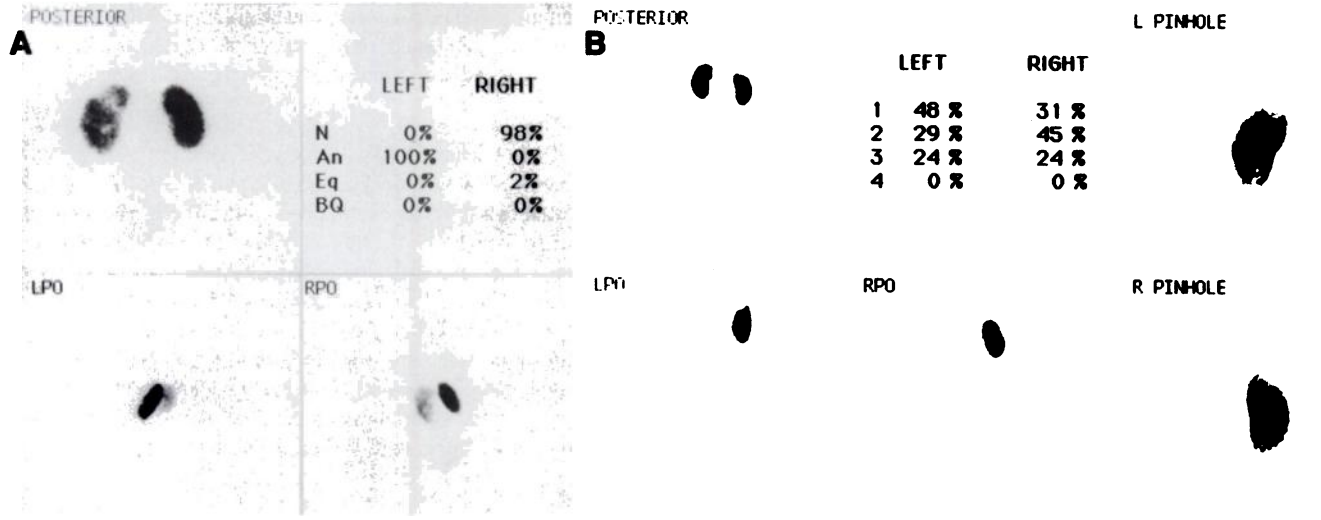

FIGURE 1. (A) ^{99m}Tc-DMSA planar images (posterior, right posterior oblique [RPO], and left posterior oblique [LPO]) and corresponding responses of observers (N = normal, An = abnormal, Eq = equivocal, and BQ = bad quality) are shown. Left kidney was considered abnormal by all observers; right kidney was considered normal by almost all observers. (B) ^{99m}Tc-DMSA planar (posterior, right posterior oblique, and left posterior oblique) and pinhole images of both kidneys are shown. Equivocal answer was chosen by 24% of observers for both kidneys because they could not decide between normal and abnormal responses. Other observers were divided between normal and abnormal responses.

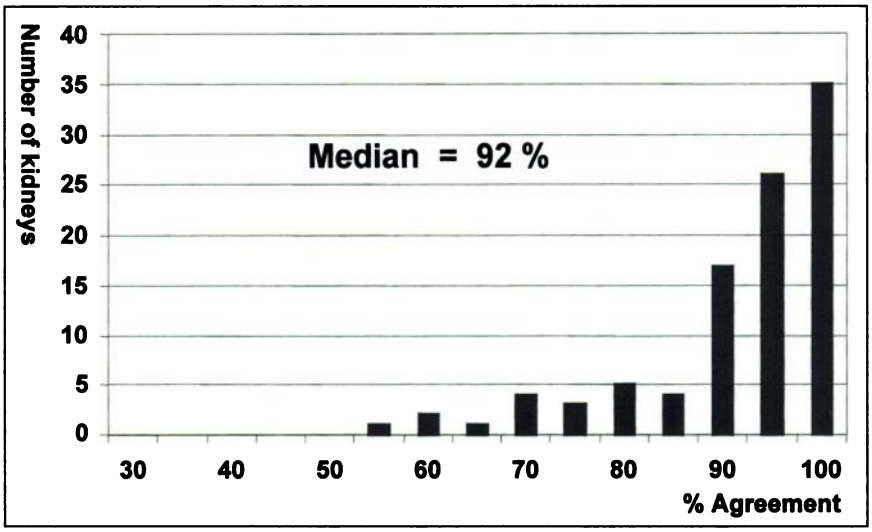

FIGURE 2. Overall interobserver reproducibility. Median percentage of agreement was 92%. Except for 4 kidneys (reproducibility range, 51%–70%), high level of reproducibility was found.

DISCUSSION

Planar renal cortical scintigraphy with ^{99m}Tc-DMSA is widely used, primarily for the evaluation of renal sequelae after acute renal infection and also for the diagnosis of acute disease (*9–13*). The technique is readily available in most nuclear medicine departments. However, few data on interobserver reproducibility are available. Some investigators have reported good interobserver reproducibility (*1–3*), whereas others considered that the reproducibility was poor (*6,7*). These major divergences are not surprising because of a large number of factors that may affect reproducibility:

Display

An unfamiliar display may represent a source of difficulty for nuclear medicine physicians who have to report on clinical studies. They are accustomed to their own displays, hard copies, background subtraction, and contrast enhancement parameters. For instance, image contrast may play a noticeable role in the detection of defects (*14*). Furthermore, analog or digital images, zoomed displays, and contours all contribute to different types of data presentation (*4*).

Contrary to the findings of previous studies (*1–7*), the observers in this study were not asked to interpret the scintigrams from hard-copy radiographs; the raw dataset, offered to the observers for interpretation, was entered into each observer's computer. Each observer was then free to use his or her usual processing program and display (hard copies, contrast enhancement, color scale, gray scale, and so forth); potential problems relating to unfamiliar display were therefore avoided.

It is worthwhile to note that transferring data to different types of processing systems is not straightforward. Indeed, different manufacturers use floppy disks of different dimensions and different operating systems, and most have different Interfile "dialects." In this study, 16 centers were unable to participate because of these technical problems, and 1 study was lost because of a translation error.

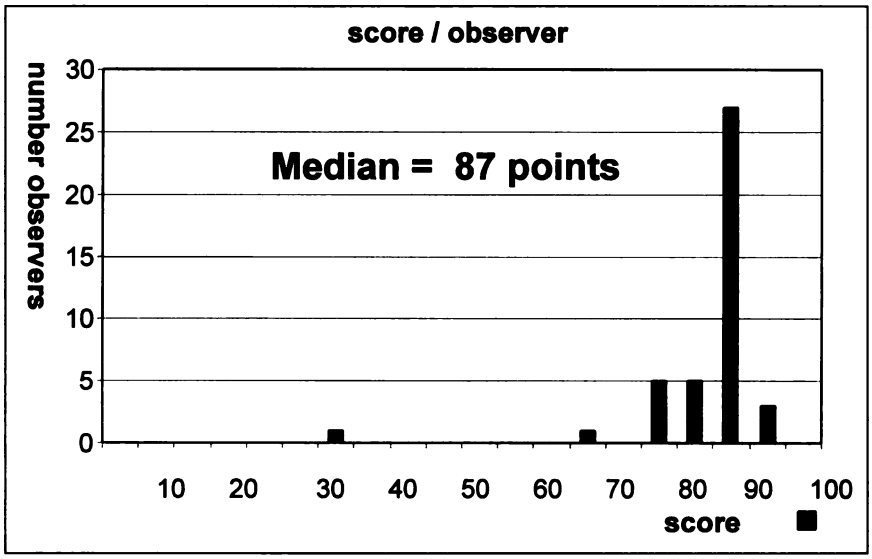

FIGURE 3. Influence of observers on overall reproducibility. Median score for all observers was 87 points. Except for 2 outliers (scores, 31 and 67 points), all observers had almost same level of performance.

Type of Scintigraphic Abnormalities and Patient's Age

The degree of reproducibility may depend on the intensity and the extension of the lesions: Abnormalities seen during an acute phase of infection are often more striking than are the residual lesions observed 6 mo later. The age of the patient may play a role, and reproducibility may be different in infants, older children, and adults.

In this study, we tried to simulate, as much as possible, the routine clinical workload by selecting at random a number of clinical studies from the routine databases (optical disks) of 2 different centers. No selection was made on the basis of the quality of the images or the type of lesions observed.

Selection of Observers

It is not unreasonable to postulate that reproducibility might differ depending on the experience of the observers or on the fact that they are (or are not) working in the same institution. Previous studies generally underline the fact that the observers are experts (*1,4,5*) or are working in the same institution (*2,3,5*).

Because ^{99m}Tc-DMSA is performed in almost all nuclear medicine centers, the real value of the test depends on the reproducibility of interpretation among all active nuclear medicine physicians, whatever the level of experience. Therefore, a large number of Belgian nuclear medicine centers were included in this study.

Scintigraphic Criteria

In a recent consensus on ^{99m}Tc-DMSA scintigraphy (*15*), attention was paid to the multiple normal variants and also to the various types of renal abnormalities that can be observed in clinical situations. These different patterns are not always easy to describe precisely, and opinions may differ about whether a given abnormality should be classified in 1 or another category. For instance, heterogeneity may be normal as well abnormal; scarring is poorly defined on the basis of pure scintigraphic criteria. Therefore, it is not surprising that reproducibility studies based on such classifications may produce some poor results. In this study, the observer was simply asked to choose between normal and abnormal; these 2 choices constitute the simplest, yet clinically meaningful, categories.

Two other categories, poor quality and equivocal, were also proposed to evaluate the quality of the data and the degree of certainty of the observers' responses. Because the number of poor-quality and equivocal answers was rather low, these categories were ignored in the analysis.

Measuring Reproducibility

Good reproducibility was reported by Piepsz et al. (*1*), Patel et al. (*2*), and Everaert et al. (*3*) when the observers (n = 2–4) had to choose between normal, abnormal, or equivocal. On the contrary, poor correlation was found by Gacinovic et al. (*6*) and Jakšić et al. (*7*) when the observers (n = 6 or 7) had to quantify the number of scars or to analyze 7 different parameters. The divergences between these results are associated primarily with differences in the number of observers and the number of scintigraphic criteria.

In all of these studies, the reproducibility was estimated on the basis of a complete agreement between the observers. The rate of complete agreement will necessarily decrease when the number of possible answers or the number of observers increases.

In this study, the methodology was different: Interobserver variability was calculated by counting the number of observers who gave the same answer. This rather simple statistic has the merit of being easy to understand. An agreement of 90% means that 90 of 100 observers gave the same answer. According to the large number of observers, particularly low reproducibility would have been obtained if complete agreement had been used for that purpose.

CONCLUSION

In this study we have attempted to reproduce, as much a possible, the conditions encountered in clinical situations: large numbers of observers without high expertise in this area, large numbers of clinical cases, various ages, clinical indications, and image quality. We have avoided the drawback associated with unusual displays. Under these circumstances, remarkable interobserver agreement was found.

ACKNOWLEDGMENTS

We thank all Belgian nuclear medicine physicians who agreed to participate in this study.

REFERENCES

1. Piepsz A, Clarke SEM, Mackenzie R, Gordon I. A study on the interobserver variability in reporting on ^{99m}Tc-DMSA scintigraphy [abstract]. *Eur J Nucl Med.* 1993;20:194.
2. Patel K, Charron A, Hoberman A, Brown ML, Rogers KD. Intra- and interobserver variability in interpretation of DMSA scan using a set of standardized criteria. *Pediatr Radiol.* 1993;23:506–509.
3. Everaert H, Flamen P, Franken PR, Peeters P, Bossuyt A, Piepsz A. ^{99m}Tc-DMSA renal scintigraphy for acute pyelonephritis in adults: planar and/or SPET imaging? *Nucl Med Commun.* 1996;17:884–889.
4. Howman-Giles R, Craig J, Uren R, et al. Variability in the interpretation of DMSA scintigraphy and application of oblique SPECT reconstruction [abstract]. *J Nucl Med.* 1998;5:27P.
5. Shanon A, Feldman W, McDonald P, et al. Evaluation of renal scars by technetium-labeled dimercaptosuccinic acid scan, intravenous urography, and ultrasonography: a comparative study. *J Pediatr.* 1992;120:399–403.
6. Gacinovic S, Buscombe J, Costa DC, Hilson A, Bomanji J, Ell PJ. Interobserver agreement in the reporting of Tc-99m DMSA renal studies. *Nucl Med Commun.* 1996;17:596–602.
7. Jakšić E, Beatović S, Žagar I, et al. Interobserver variability in the interpretation of ^{99m}Tc-DMSA renal scintigraphy: multicentric study [abstract]. *Nucl Med Commun.* 1997;18:325.
8. Paediatric Task Group: European Association of Nuclear Medicine. A radiopharmaceuticals schedule for imaging in paediatrics. *Eur J Nucl Med.* 1990;17:127–129.
9. Majd M, Rushton HG. Renal cortical scintigraphy in the diagnosis of acute pyelonephritis. *Semin Nucl Med.* 1992;22:98–111.
10. Conway JJ. The role of scintigraphy in urinary tract infection. *Semin Nucl Med.* 1988;18:308–319.
11. Handmaker, H. Nuclear renal imaging in acute pyelonephritis. *Semin Nucl Med.* 1982;12:246–253.
12. Sty JR, Wells RG, Starshak RJ, Schroeder BA. Imaging in acute renal infection in children. *AJR.* 1987;148:471–477.
13. Bjorgvinsson E, Majd M, Eggli KD. Diagnosis of acute pyelonephritis in children: comparison of sonography and (99m)Tc-DMSA scintigraphy. *AJR.* 1991;157:539–543.
14. Rodriguez JL, Perera A, Fraxedas R, Reyes L, Hernandez A, Solaro ME. Renal 99Tc(m)-DMSA SPET and planar imaging: Are they really the same? *Nucl Med Commun.* 1997;18:556–561.
15. Piepsz A, Blaufox MD, Gordon I, et al. Consensus on renal cortical scintigraphy in children with urinary tract infection. *Semin Nucl Med.* 1999;29:160–174.