

Consensus in Reading DMSA Scintigrams

Suchitra Thongmak MD, Prayong Vachvanichsanong MD, Chirawat Utamakul MD, Ugrit Markmanee Bed and Tada Yipintsoi MB.

Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla, 90110

Address correspondence to S. Thongmak

ABSTRACT

Background: To evaluate agreement in reading dimercaptosuccinic acid (DMSA) scintigrams and the potential reasons for disagreement.

Material: DMSA scintigrams were read by 4 observers. Consensus in reading was evaluated and also compared to potential for the presence of kidney scars as projected by the clinical conditions. These conditions excluded recent pyelonephritis. Two sets of DMSA scintigrams were evaluated.

Result: In the first set of 59 kidneys, intrapersonal correlation ranged from 72 to 97%. Agreement (by all 4 or 3 out of 4) readers as to defect grades and defect areas was in 48/59 kidneys = 81%. In the second set (i.e. after a session of discussion as to criteria in reading defects) of 76 kidneys, agreement as to defect score, size and now including site was 43/76 kidneys = 57%.

Conclusion: These suggest that the discussion was inadequate in depth that increasing parameters in reading (i.e. the addition of site of defect on to defect scores and size of defect) and low volume reading probably account for the lower agreement in the second set of reading.

Key words: renal scar, scan defect, correlation, vesicoureteral reflux, Tc-99m, defect score.

INTRODUCTION

DMSA (dimercaptosuccinic acid) is now an accepted tracer for detecting varied abnormalities of the kidneys, particularly renal parenchymal involvement in acute pyelonephritis (UTI) and subsequent renal scars, such as related to vesico-ureteral reflux (VUR)⁽¹⁻³⁾. The ability to tag it with Tc-99m facilitates its imaging which further popularises its use. As expected, with subjective interpretation of images, this is then followed by questioning the accuracy

and consistency in 'seeing' and 'interpreting' photon defects at varied anatomical sites. Aside from a single study in pigs⁽⁴⁾ using histological scars in the kidney as the gold standard, other reports on accuracy rests on inferior imaging devices for scars such as ultrasound⁽⁵⁾, contrast pyelography⁽⁶⁾ or using the nephrographic phase of contrast arteriography to delineate normality⁽⁷⁾. As often, reliance was placed on comparing correlations among

readers⁽⁸⁻¹⁰⁾. These gave conflicting results. In two, there were a very high degree of agreement^(9, 10) while the other reported wide differences despite a session at a consensus. The reasons for observer differences may be intuitively guessed at: number of scintigraphic criteria [as discussed by De Sadeleer et al.⁽¹⁰⁾], methods of obtaining the image [e.g. pinhole vs full or partial SPECT^(4, 11)], and lack of proper guidelines particularly among the low volume readers. This latter rests on high acquaintancy with non-homogeneous pattern seen on normal kidneys⁽⁷⁾ and or being able to compare the result to a gold standard, which in human is non-existent.

At our unit, we read DMSA scan for evaluating kidney scar. Potentials for incorrect and inconsistent reading of DMSA scan come from what were written above namely: no consensus, low volume, no repeated study to allow evaluating appearance or disappearance of defects and no reference towards a gold standard.

The aim of this study was to determine variability (Intra and interpersonal variation) in the interpretation of DMSA scintigraphy among 4 observers with different level of experiences, before and after "exchange of criteria" where the exchange consisted of discussing criterias for accepting defects as being pathological in DMSA scan.

MATERIAL AND METHODS

Study Design: Four observers at our institution, 3 are nuclear medicine physicians (1 of whom read scintigram only occasionally) and 1 pediatric nephrologist were asked to read and report two sets of Tc-99m-DMSA scintigraphic study selected independently. The original request for and the report on each scintigram were only

used to select the cases for review and were not available to the observer. All patients had been imaged in the same institution and the data were displayed in the uniform way.

Patient selection

The first set consisted of 30 patients. No attempt was made to select studies on the basis of technical quality. Most patients were sent for investigation by the pediatric nephrologist as part of a follow-up of vesicoureteral reflux (VUR). Among these patients, urinary tract infection (UTI) had to be absent for at least 6 months. Few had recurrent UTI. Others were investigated for anatomical abnormalities of the kidney and a few were kidney donors. The second set of scintigrams consisted of 38 cases, also randomly selected. Four were repeats from the first set. All cases were masked to the observers.

Imaging and interpretation

Patients were imaged on a digital gamma camera computer system equipped with a low-energy, high-resolution collimator (Toshiba GCA-901 A/HG). The images were obtained 2 hours after intravenous injection of Tc-99m-DMSA. The adult dose was 111-185 MBq (3-5 mCi); the doses administered to children were reduced to 0.05-0.1 mCi/kg. Planar anterior, posterior and laterally right and left posterior oblique images of both kidneys (800k-1,000k counts in a 512 x 512 matrix) were obtained. The images were stored in the computer such that they can be read off the screen with all the attendant image enhancement and magnification.

To be certain that the reading was of similar format, we agreed on these definitions. A DMSA scintigraphic defect was defined

as relative decrease in radioactivity at sites different from normal. The defects had to be scored as to site and severity along the line of a format reported by Laguna et al.⁽¹²⁾ The gradings were: grade 0, normal; grade 1, equivocal to slight; grade 2, moderate defect; grade 3 severe reduction in activity, and grade 4, practically absent uptake.

The images from the first set of patients [30 cases and 59 kidneys (one of the patients had a single kidney)] were evaluated by the 4 observers independently. Each observer had to record the grade of the defect and the area involved twice, several days apart. From these, intraobserver correlation and interobserver variation were calculated. After that, the 4 observers all read most of images together and exchanged their criteria for scoring.

Within two months of the first reading, the second set of DMSA scan was separately recorded by the same 4 observers. At this time, additional data on the isotopic density of each kidney was provided to circumvent potential error from eyeballing or averaging the gray or colored images. Also, aside from the defect score and its approximate size, they had to record the defect location relying on the posterior view of the kidney, e.g. upper, middle or lower and each further divided into lateral, central or medial portions.

Data analysis

The gradings (defect score) and extent of involvement (site and size in terms of per cent whole kidney) for each kidney were summarized, averaged and tabulated for each reader without a prior knowledge of the patient.

Agreement on the interpretation of each kidney scintigram were divided thus: 4:0

implied that all 4 agreed; 3:1 implied one differed; 2:2 implied 2 pairs, each pair had defect scores differing by at least 1 grade; and "all" disagreed implied that 3 or 4 disagreed from one another (e.g. defect scores of 0, 0, 1.5, 3.0 for each of the four readers). Defect scores that differed by less than 1 was considered equivalent. A defect with an area of 5% or less of the whole kidney was considered inconsequential. Correlations were calculated in percentages.

The scoring for concordance in the second set was made more strict in that the defect score, the size of the defect and the site had to match. Absolute difference in defect area cannot exceed 20%. The site had to be in the same upper, middle or lower part of the kidney although it can differ slightly whether it is medial, central or lateral. For example, defects of 1.0-1.8 among 3 readers were accepted as similar if the sizes of defects were 10%, 25% and 30%, and all were in the upper half of that kidney.

The decision with regards to agreement was made by one reader (TY). The result was partitioned into the varied concordances according to the defect scores and accompanying diseases (not known until after the table was constructed).

The probability of having scars on the kidneys was judged from the clinical diagnosis for each kidney. The possibility was predicted by the clinical nephrologist (PV). Hence we have 4 subdivisions with regards to kidney scar: no scars (N); possibly no scar (PN) was judged from those with previous UTI or with contralateral VUR; probable scar (PS) was assumed among those with grade 1-2 VUR, or with neurogenic bladder or recurrent UTI;

lastly, the definite scar (S) included those with dilated VUR (grade 3-5), duplex kidney, post partial nephrectomy or hydro-nephrosis and the one patient with ileal conduit.

RESULTS

The median age at DMSA scintigraphy for all patients in the first set was 4.0 years (range 3 months-13 years) and in the second set this was 9.1 years (range 2 months-41 years). The 5 kidney donors were 22 to 41 years old.

In the first group of 59 kidneys in 30 subjects, there were 4 subjects with UTI, 8 kidneys with non-dilated VUR (grade 1-2), 29 kidneys with dilated VUR (grade 3-5), 12 normal kidneys with contralateral VUR, one subject had a duplex kidney and partial nephrectomy.

In the second group of 76 kidneys in 38 subjects, there were 5 kidney donors, 6 subjects with previous cystitis or UTI, 13 kidneys with non-dilated VUR (4 subjects had involvement of both kidney), 19 kidneys with dilated VUR (5 subjects had involvement of both kidneys), 12 normal kidneys with contralateral VUR. The remainder included one with neurogenic bladder, one with an ileal conduit and 3 subjects with duplex kidneys. Among those with duplex kidneys, one had a 50% right heminephrectomy. Contrast pyelogram showed a smaller right kidney. Most read that the DMSA scintigram of the right kidney showed scar at the upper pole involving 12% of the kidney area. The scintigram of the left kidney was read as showing low defect score averaging 6% of the kidney. Another of the duplex kidney had bilateral heminephrectomy also 50% each. The pyelogram from this showed slightly reduce kidney size, left not different from

right. All readers disagreed on the scintigraphic defects of both kidney varying from grade 0 to 3 but occupying areas of 10% or less.

In the first set, the intrapersonal correlation was 90, 95, 72 and 97%. The lowest correlation was from the pediatric nephrologist, this was her first exposure to reading scintigrams. Table 1 shows the result from the first reading. The total agreement in 26 kidneys among the 4 readers (4:0) could be further subdivided into an agreement that there were no scar (defect score = 0) and an agreement that there were marked scarring occupying an average of over half the kidney (average area 56%). In the next rows, the majority agreed (3:1) that no defect existed in 17 kidneys, the dissenter was the only one seeing significant scar (1.6 defect score) and averaging nearly a fifth of the kidney shadow (19%). In the other five of the 3:1 agreement, the difference was in the degree of severity, the average absolute difference was 1.3 ± 0.3 (i.e. the difference in defect score between the 3 readers versus the one reader independent of the arithmetic sign) and the dissenter gave higher scores than the majority in 4 out of the 5 readings.

From the table, one calculated the inter-personal agreement as 44%, 37%, 7% and 12% for the 4:0, 3:1, 2:2 and no agreement respectively. Note as well that there appeared to be very little relationship between potential scars on the kidney versus the defect score given by the majority of the readers.

The result from the interpretations of DMSA scan from the second set are summarized in table 2. Similar to the first set, total agreement in 24 kidneys were either because they were read as no scar (14 kid-

Table 1 Results of 4-observer reporting of DMSA scintigraphy (first set)

Agreement	N Kidney (subject)	Defect (Score)	Area (%)	Potential scar			
				N	PN	PS	S
4 : 0	21(13)	0	0	-	9	4	8
	5(5)	2.7 ± 0.9	56 ± 29	1	-	-	4
3 : 1	17(12)	0 vs 1.6 ± 0.4	19 ± 8	-	3	3	11
	5(5)	2.3 ± 0.4 vs 3.2 ± 0.7	32 ± 8 vs 38 ± 16	-	1	1	3
2 : 2	3(3)	0 vs. 2.0 ± 0.3	33 ± 15	-	2	-	1
	1(1)	2.5 vs 3.7	38 vs 40	-	-	-	1
All	7(6)	-	-	1	1	1	4

Legend to table 1. N refers to number of kidneys (number subjects contributing towards these kidneys). [NB. The numbers of subjects in the brackets together will be greater than the real number studied since these are not mutually exclusive]. Defect score represents the average defect + SD. Area represents the percentage of the kidney size that was judged to be occupied by the defect.

The abbreviation for potential scars according to the clinical conditions are : N = no scar, PN = possible no scar, PS = probable scar, S = high probability of having kidney scar.

There were 2 subdivisions to the 4 : 0, and 3:1 agreement depending on whether the average reading of the defect suggests no defect (score = 0) or definite defect (score > 1). (eg. for the 3:1, 0 vs. 1.6 implies that 3 read as no defect but the one who read different saw significant defects in all the 17 kidneys).

Table 2 Results of 4-observer reporting of DMSA scintigraphy (second set)

Agreement	N Kidney (subject)	Defect (Score)	Area (%)	Potential scar			
				N	PN	PS	S
4 : 0	14(11)	0	0	1	3	6	4
	10(8)	2.7 ± 0.7	29 ± 15	-	-	1	9
3 : 1	15(14)	0 vs 1.4 ± 0.6	0 vs 20 ± 23	6	4	2	3
	4(3)	1.8 ± 0.7 vs 0.3 ± 0.5	14 ± 8 vs 3 ± 6	1	-	-	3
2 : 2	18(14)	1.5 ± 0.5 vs 0	15 ± 8 vs 0	4	5	6	3
All	15(12)	-	-	2	5	3	5

Legend similar to table 1.

neys with zero defect score) or heavily scarred (2.7 defect score occupying 29% of the kidney area). In this latter the clinical predictions tallied with the defect scores in that 9 was expected to show definite scar because of the dilated VUR.

The next 2 rows summarize the category where there was a discordant reader. In the first of these (15 kidneys in 14 subjects), the majority read as showing no defect while the dissenter read defects averaging 1.4 ± 0.6 with an area of $20 \pm 23\%$ (the size of the area suggested that the difference was unlikely to be a misread by chance). A clinical assessment suggested that in 6 kidneys out of the 15, there was a low likelihood of having scars. It should be pointed out that a single reader (TY) accounted for 11 of the 15 differences. In contrast to the preceding, the next row shows that in 3 out of the 4 kidneys, the dissenting reader observed no defect while the majority averaged the defect at 1.8 ± 0.7 with an average area for scar of $14 \pm 8\%$ (14% area comprised a seventh of the area of the kidney). Clinically, most of these kidneys should be scarred.

The next portion of the table (2:2) shows differences between 2 pairs of readers where one pair saw no defect while the other pair gave an average score for the defect as 1.5. Note as well that the dissenter in the previous 3:1 agreement was also in the pair that consistently read scars. The clinical estimation of scars did not help to decide the probability as to which pair of readers was more likely to be correct. Lastly, in 15 kidneys, all or 3 out of 4 read differently.

From this table, it is seen that interpersonal agreement is now 32%, 25%, 24%, and 20% respectively, ranging from full agreement (4:0) to all disagreed (all).

DISCUSSION

We attempted to answer whether low volume operators (our nuclear medicine staff) and non-experienced readers would be able to agree on the reading of scars (defects) in DMSA scintigrams, and whether following a short exchange of criterias, the agreement can be improved. The data used were retrospective records from patients where the majority were investigated because of VUR and where 33 out of 59 (set 1) and 27 out of the 76 (set 2) kidneys had high probability of scars. Farnsworth et al.⁽⁶⁾ showed results suggesting correlation between presence of DMSA scar and severity of VUR. The results as seen in table 1 and 2 were disappointing but not unexpected in that neither the agreement among the 4 readers improved following a dialogue (despite an intrapersonal agreement of greater 90% in 3 out of 4 readers) nor did the reading of scars tally with the disease process. Surprisingly, given the composition of our reader, this result parallels that of Gacinovic et al.⁽⁸⁾ using 7 observers and exact same 32 scintigrams for the first and second readings 6 months apart and also interspersed by a consensus. Furthermore, they were able to show that one of the 2 who 'over-read' and 1 of the 2 who 'under-read' persisted in the pattern of their scorings in the second attempt, very similar to one of us who often over-detect scars on DMSA scintigrams. These results are in marked contrast to the report of Craig et al.⁽⁹⁾ In that report, 2 very experienced readers showed very high degree of matching in 882 pediatric kidneys where there were about 50 moderately severe defects (by DMSA). Disagreement (they scored their defects as we did) was seen in only 0.7%. De Sadeleer et al.⁽¹⁰⁾ utilised 42 readers who were allowed only 4 options: nor-

mal, abnormal, equivocal and lack of quality, but no account was made as to site nor size. They reported that 90% or more of the readers agreed in the given options in 78 out of a possible 98 kidneys. These were patients investigated for acute pyelonephritis. They commented that agreement depended also on the number of criteria used for the interpretation eg. grades, site, size etc. of defects.

From the present data and review of literature, we wish to make the following comments and suggestions as to ways to improve accuracy and consistency in reading and interpreting scars in kidneys by DMSA scintigraphy. This may also be relevant in post-grad training.

1. Carefully arrived at a consensus of normal DMSA scintigraphy^(2,3).
Note that the appearances of cut off or hypoactive poles were reported as uncommon.
2. It is worth noting that Gordon et al.⁽¹³⁾ showed that injected DMSA in kidneys reach a plateau at 6 hours while most nuclear physicians agreed on a 3 hours image after intravenous introduction of the agent⁽²⁾. We did ours in 2 hours. Would timing after injection made any difference in accentuating or attenuating photon deficient areas?. Mandell et al.⁽³⁾ suggested collecting delayed-images in certain instances.

3. Would the technic of imaging increase its resolution and hence allow better recognition of photon defects?. As stated, Rossleigh et al.⁽⁴⁾ found that pinhole collimator was highly accurate when the gold standard was macroscopic experimentally-induced kidney scar in pigs. (out of 24 possible scars, the imaging by pin-hole collimator detected 23, while planar parallel holes and SPECT picked up only 19 and 18 respectively). Other consensus^(2,3) also made similar suggestion but balked at the acquisition time as being too long. Peng et al.⁽¹¹⁾ suggested that SPECT may provide better delineation and if so, then 180° posterior would be better than the 180° anterior (because of abdominal attenuation) or 360° rotation. Such manouver may reduce the apperances of 'hypoactive' upper poles.
4. We think that the advice on not using color enhancement^(2,3) is worth investigating. We also think that reading the scan by using a format for scoring the defect⁽²⁾ and denoting size and site would ensure better commitment and subsequent re-evaluation.

Acknowledgement

We thank Ms. Patcha Nuychim who typed this manuscript.

REFERENCE

1. Gordon I. Indications for 99mtechnetium dimercaptosuccinic acid scan in children. *J Urol* 1987;137:464-7.
2. Piepsz A, Blafox MD, Gordon I, Granerus G, Majd M, et al. Consensus on renal cortical scintigraphy in children with urinary tract infection. *Semin Nucl Med* 1999;29:160-74.
3. Mandell GA, Eggli DF, Gilday DL, Heyman S, Leonard J C, et al. Procedure guideline for renal cortical scintigraphy in children. *J Nucl Med* 1997;38:1644-6.
4. Rossleigh MA, Farnsworth RH, Leighton DM, Yong JLC, Rose M and Christian CL. Technetium-99m dimercaptosuccinic acid scintigraphy studies of renal cortical scarring and renal length. *J Nucl Med* 1998;39:1280-5.
5. Bjorgvinsson E, Majd M, and Eggli KD. Diagnosis of acute pyelonephritis in children: comparison of sonography and Tc-99m-DMSA scintigraphy. *AJR* 1991;157:539-43.
6. Farnsworth RH, Rossleigh MA, Leighton DM, Bass SJ and Rosenberg AR. The detection of reflux nephropathy in infants by 99m technetium dimercaptosuccinic acid studies. *J Urol* 1991;145:542-6.
7. Pusuwan P, Reyes L and Gordon I. Normal appearances of technetium-99m dimercaptosuccinic acid in children on planar imaging. *Eur J Nucl Med* 1999;26:483-8.
8. Gacinovic S, Buscombe J, Costa DC, Hilson A, Bomanji J and Ell PJ. Inter-observer agreement in the reporting of 99Tcm-DMSA renal studies. *Nucl Med Commun* 1996;17:596-602.
9. Craig JC, Irwig LM, Howman-Giles RB, Uren RF, Bernard EJ, et al. Variability in the interpretation of dimercaptosuccinic acid scintigraphy after urinary tract infection in children. *J Nucl Med* 1998;39:1428-32.
10. De Sadeleer C, Tondeur M, Melis K, Van Espen M, Verelst J, et al. A multicenter trial on interobserver reproducibility in reporting on Tc-99m-DMSA planar scintigraphy: A Belgian survey. *J Nucl Med* 2000;41:23-6.
11. Peng NJ, Kwok CG, Chiou YH, Jao GH, Tsay DG, et al. Posterior 180o Tc-99m-dimercaptosuccinic acid renal SPECT. *J Nucl Med* 1999;40:60-3.
12. Laguna R, Silva F, Orduna E, Conway JJ, Weiss S and Calderon C. Technetium-99m-MAG 3 in early identification of pyelonephritis in children. *J Nucl Med* 1998;39:1254-7.
13. Gordon I, Evans K, Peters AM, Kelly J, Morales BN, et al. The quantitation of 99Tcm-DMSA in paediatrics. *Nucl Med Commun* 1987;8:661-70.