Evaluation of Semi-Quantitative Scoring System for Metaiodobenzylguanidine (mIBG) Scans in Patients With Relapsed Neuroblastoma

Julia A. Messina, BS,[1] Su-Chun Cheng, ScD,[2] Benjamin L. Franc, MD,[3] Martin Charron, MD,[4] Barry Shulkin, MD,[6] Bao To, MD,[3] John M. Maris, MD,[5] Gregory Yanik, MD,[7] Randall A. Hawkins, MD, PhD,[3] and Katherine K. Matthay, MD[1]*

Background. The purpose of this study was to determine the accuracy of two semi-quantitative scoring systems to assess response to ^{131}I-metaiodobenzylguanidine (mIBG) therapy in recurrent neuroblastoma. ***Procedures.*** Diagnostic mIBG scan pairs (n = 57) were collected for patients who underwent ^{131}I-mIBG therapy for relapsed neuroblastoma. Two scoring systems were designated: Method 1, which divided the body into nine segments to view osteomedullary lesions with an additional tenth segment to assess soft tissue involvement; and Method 2, which divided the body into seven segments without a corresponding compartment for soft tissue involvement. Four nuclear medicine physicians independently assigned extension and intensity scores utilizing both methods, and separately recorded their impression of whether the post-therapy scan had improved, not changed, or worsened. Inter- and intra-observer concordance and correlation with overall response and progression-free survival (PFS) were performed. ***Results.*** Method 1 produced the highest inter-observer concordance and was used to calculate the relative extension scores (post-therapy score divided by pre-therapy score), which correlated significantly with overall response. Patients who achieved complete response (CR) or partial response (PR) (n = 21) had lower relative extension scores, compared to those without response ($P < 0.001$). The readers' overall impression associated highly ($P < 0.001$) with the relative extension scores though results were less quantitative. Concordance was higher if initial scores were >5. Relative extension score did not predict PFS. ***Conclusion.*** Semi-quantitative scoring of mIBG scans provides a more reliable method of assessing response in patients with relapsed neuroblastoma than qualitative impression. The reproducibility and high inter-observer concordance makes mIBG score an important component of overall response criteria in patients with recurrent neuroblastoma.Pediatr Blood Cancer

Key words: ^{131}I-metaiodobenzylguanidine; international neuroblastoma response criteria; mIBG score; neuroblastoma

INTRODUCTION

Neuroblastoma, the most common extra-cranial solid tumor in children, originates as a primary tumor of the sympathetic nervous system but metastasizes often to bone and bone marrow, resulting in a poor prognosis. Approximately 15% of patients who present with metastatic disease at diagnosis are refractory to induction chemotherapy while 40% will eventually relapse after having a complete response (CR) or partial response (PR) [1].

Intravenous administration of radiolabeled metaiodobenzylguanidine (mIBG), a norepinephrine analog that specifically targets malignant cells of the sympathetic nervous system, is an effective therapy for patients with refractory disease, with response rates of 30–40% [2–8]. Many of the patients who undergo therapy with ^{131}I-mIBG or other treatments for relapsed neuroblastoma have sites of disease only apparent on mIBG scans or in bone marrow biopsies and cannot be evaluated by standard response criteria for solid tumors, such as the RECIST criteria [9]. Therefore, a standardized scoring system to predict the clinical response and progression-free survival (PFS) with ^{131}I-mIBG treatment is needed to help quantitate therapeutic efficacy of agents used in treatment of refractory neuroblastoma.

In recent analyses of high-risk, metastatic neuroblastoma, semi-quantitative scoring systems that divide the body into anatomical sections were developed to assign numeric scores to patients' diagnostic ^{123}I and ^{131}I-mIBG scans [10–14]. Assessing extent of disease before, during, and after induction chemotherapy, three studies have shown a correlation between semi-quantitative scores either at diagnosis or during induction with response at the end of induction chemotherapy [10,12,13] while another showed good concordance among scan readers but poor correlation with response [14].

Although some of these studies showed a significant correlation between the change in semi-quantitative score

[1]Department of Pediatrics, University of California, San Francisco, California; [2]Department of Epidemiology and Biostatistics, University of California, San Francisco, California; [3]Department of Nuclear Medicine, University of California, San Francisco, California; [4]Department of Nuclear Medicine, Hospital for Sick Children, University of Toronto, Toronto, Canada; [5]Department of Pediatrics, Children's Hospital of Philadelphia and the University of Pennsylvania, Philadelphia, Pennsylvania; [6]Department of Nuclear Medicine, St. Jude Children's Research Hospital, University of Michigan, Ann Arbor, Michigan; [7]Department of Pediatrics, University of Michigan, Ann Arbor, Michigan

Grant sponsor: National Institute of Health; Grant numbers: PO1 CA81403, 2MO1 RR0127; Grant sponsor: Campini Foundation; Grant sponsor: Conner Research Fund; Grant sponsor: Katie Dougherty Foundation; Grant sponsor: Kasle and Tkalcevik Neuroblastoma Research Fund; Grant sponsor: Alex's Lemonade Stand and the Thrasher Foundation.

*Correspondence to: Katherine K. Matthay, Department of Pediatrics, Box 0106, University of California, San Francisco, CA 94143.
E-mail: matthayk@peds.ucsf.edu

Received 5 November 2005; Accepted 27 December 2005

DOI 10.1002/pbc.20777

and patient response to induction chemotherapy, each used slightly different methods to assign scores. In addition, these scoring systems have only been tested in newly diagnosed neuroblastoma patients. No investigation has been performed to correlate results from these scoring systems to response among neuroblastoma patients with refractory or relapsed disease later in the treatment course. In such patients, changes in score might be expected to be smaller, and the number of lesions at initiation of treatment would frequently be less than at the time of original diagnosis for stage 4 neuroblastoma.

Based on an analysis of patients who have been treated with ^{131}I-mIBG in a Phase II study for refractory disease, the present study sought to determine whether or not methods from two semi-quantitative scoring systems correlated with response to therapy based on International Neuroblastoma Response Criteria (INRC) [15] and PFS in a relapse population. This study also evaluated inter- and intra-observer concordance among the four independent readers to determine the reliability of the two semi-quantitative scoring systems.

METHODS

Study Subjects

This study included 49 patients with relapsed or refractory neuroblastoma who were treated at the University of California, San Francisco (n = 36), the Children's Hospital of Philadelphia (n = 7), and the University of Michigan (n = 6) between February 20, 1998 and December 11, 2003 on the clinical trial, ^{131}I-Metaiodobenzyguanidine Therapy for Neuroblastoma: a Phase II study [7]. Patient characteristics are shown in Table I. Patients were treated with 444–666 MBq/kg of ^{131}I-mIBG for one (n = 42), two (n = 6), or three (n = 1) courses. Although some patients had only a few lesions, they all had relapsed metastatic disease, with high likelihood of other sub-clinical lesions, and therefore were candidates for the targeted radionuclide rather than local radiaotherapy. Patients were evaluated as described below for response approximately 2 weeks prior to each therapy and 8 weeks after each therapy. Appropriate informed consent was obtained for all patients with approval at each center by the institutional human research review board and radiation safety committee.

mIBG Scoring Method

mIBG scans were performed using a standard protocol, as described in [16]. Four nuclear medicine physicians (B.T., B.F., M.C., B.S.) assigned scores to 57 pre- and post-therapy mIBG diagnostic scan pairs based upon two semi-quantitative scoring systems (Table II) [10,12]. Extension scores were assigned to each segment to quantify the extent of mIBG-positive lesions within a given segment. Intensity scores were assigned to each segment to quantify the degree of mIBG uptake within the lesions of a given segment. The present study differentiated between the two scoring systems as Method 1 and Method 2. For Method 1, the patient's skeleton was divided into nine segments to view osteomedullary lesions with an additional tenth segment to assess soft tissue involvement (Fig. 1A) [10,13]. For Method 2, the skeleton was divided into seven segments to view osteomedullary lesions without a corresponding compartment to assess soft tissue involvement (Fig. 1B) [12,14].

Absolute pre-therapy extension, post-therapy extension, pre-therapy intensity, and post-therapy intensity scores were calculated for Methods 1 and 2 by summing the segmental scores assigned by the readers. In each region, the lesions

TABLE I. Patient Characteristics

Characteristic	Number of patients
Gender	
Male	29
Female	20
MYCN gene amplification at diagnosis	
Amplified	12/34 tested (35.29%)
Time from diagnosis to metaiodobenzylguanidine (mIBG) treatment (years)	
Median [range]	2.6 [0.6–11.8]
Age at mIBG treatment (years)	
Median [range]	8.2 [1.9–30.2]
Patients receiving multiple treatments with mIBG	7
Sites of disease at time of mIBG treatment	
Bone/bone marrow, soft tissue	26
Bone/bone marrow only	13
Soft tissue only	10

TABLE II. mIBG Scan Characteristics

Characteristic	Number of mIBG scans
Number of scans reviewed[a]	
Paired pre- and post-therapy scans	57
Total number of individual scans	110
Type of scan	
^{123}I	104
^{131}I	6
Time interval between pre- and post-therapy scans (weeks)	
Median	10.07
Range	[3–22]
Time interval between MIBG therapy and post-therapy scan (weeks)	
Median	7
Range	[3–12]
Scans with SPECT Views[b]	8
Scans with lateral views of skull	55

[a]In four cases, the post-therapy scan from the first MIBG treatment served as the pre-therapy scan for the second treatment.
[b]Scores from the SPECT views were not used in the analysis.

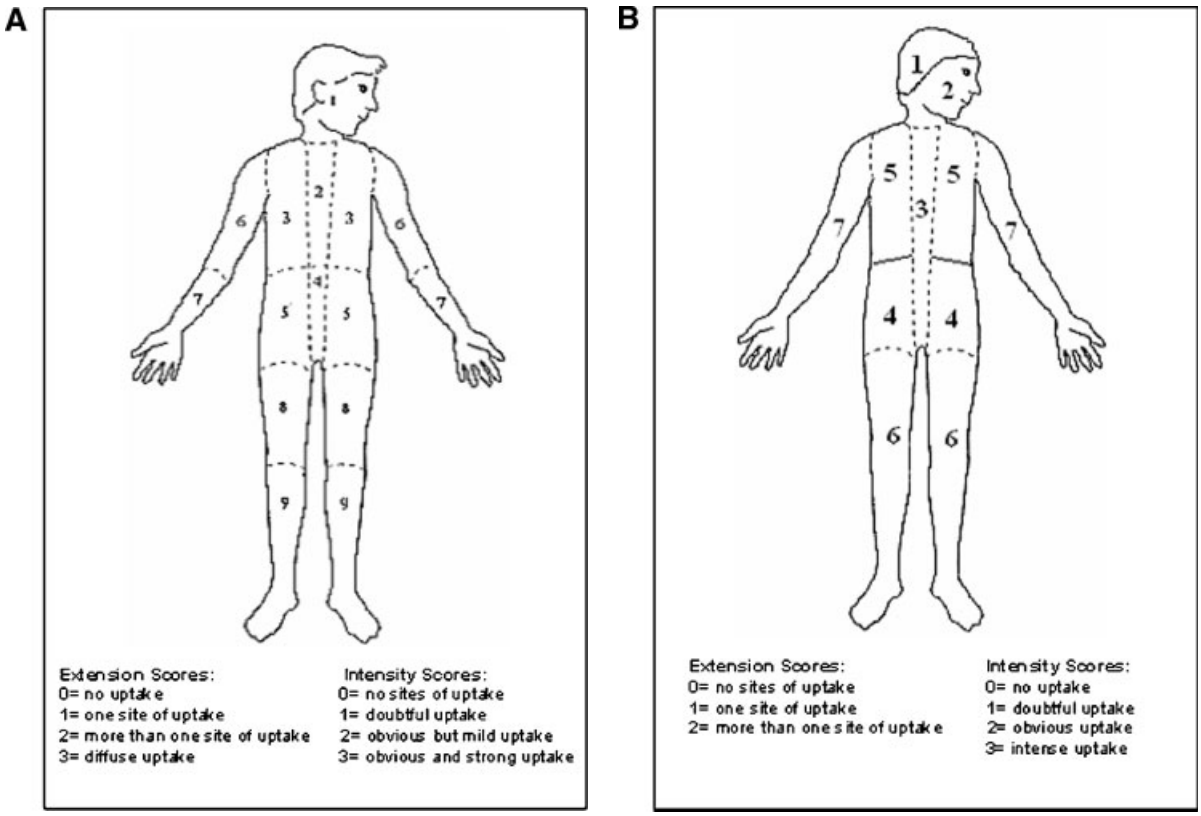


Fig. 1. **A**: Method 1, The patient's skeleton is divided into nine segments to view osteomedullary lesions [11] with an additional tenth segment to assess soft tissue osteomedullary lesions [11] with an additional tenth segment to assess soft tissue. **B**: Method 2, The patient's skeleton is divided into seven segments to view osteomedullary lesions [10].

were scored as follows for extension of metastases. The extension score for Method 1 was graded as: 0, no sites per segment; 1, one site per segment; 2, more than one site per segment; and 3, diffuse involvement (>50% of the segment). An example of the scoring by the four readers is shown in Figure 2A,B. The extension score for Method 2 was graded as: 0, no sites per segment; 1, one site per segment; 2, more than one site per segment. The intensity score for both methods was graded as: 0, for no uptake; 1, for doubtful uptake; 2, for obvious uptake; and 3, for strong uptake. Thus, the maximum extension and intensity score for Method 1 would be 30 and 30, and the maximum scores for Method 2 would be 14 and 21, respectively. The relative extension and intensity scores were calculated by dividing the absolute post-therapy score by the absolute pre-therapy score.

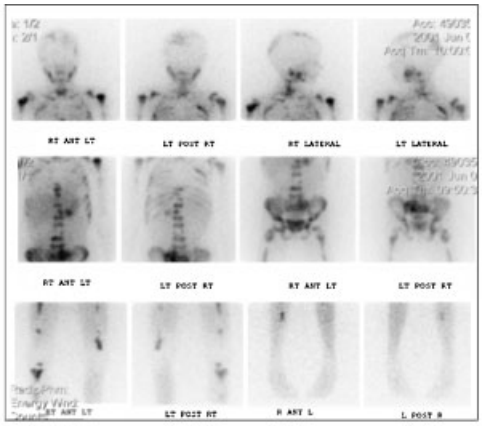
The four readers scored the scan pairs independently from identical hard copies of the scans as digital copies were not available in many of the patients. Readers were blinded to both the clinical history of the patient's disease and to the patient's overall response to ^{131}I-mIBG therapy. To maintain consistency in the scoring process, general guidelines were established before the readers began scoring scans: (1) information as to whether or not a patient had a prior adrenalectomy or nephrectomy due to tumor involvement was provided, so that the readers could differentiate between physiologic uptake in an adrenal gland or kidney and active disease on the diagnostic mIBG scan; (2) the pre-treatment scan was scored before the post-treatment scan using Method 1 followed by Method 2; (3) physiologic uptake of mIBG in the liver was used as a point of reference to determine whether or not areas of abnormal uptake were of high or low intensity; and (4) when multiple sites of disease with varying intensities were present within a given segment, the intensity score for that segment was based upon the lesion with the greatest intensity. Finally, the readers also rated a patient's overall qualitative response to ^{131}I-mIBG based on their analysis of the diagnostic mIBG scan pair. Disregarding numeric score, the readers recorded their impression of whether the post-therapy scan had improved, not changed, or worsened in comparison to the pre-therapy scan.

To measure intra-observer consistency, three patients' scan pairs were chosen to be scored on a second occasion in a blinded fashion by the four readers. These three patients' scan pairs were selected based upon extent of disease to determine whether or not variations arose within a reader's results in scoring widespread versus localized disease.

Response Evaluation

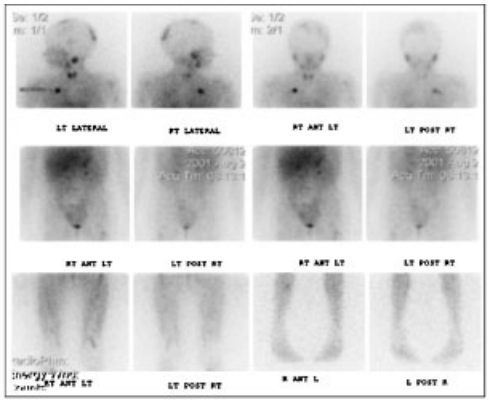
Response to ^{131}I-mIBG therapy based upon INRC was rated as CR, very good partial response (VGPR), PR, mixed

A



Reader	1	2	3	4	5	6	7	8	9	10	SUM
A	3	3	3	3	3	3	0	3	2	1	24
B	3	2	3	3	3	3	0	2	2	0	21
C	3	3	3	3	3	3	0	3	2	1	24
D	3	3	3	3	3	3	3	3	2	0	26

B



Reader	1	2	3	4	5	6	7	8	9	10	SUM
A	3	0	0	0	0	0	0	2	1	2	8
B	2	0	2	0	0	2	0	2	1	0	9
C	3	0	0	0	0	2	0	2	1	2	10
D	2	0	0	0	0	0	0	0	1	0	3

Fig. 2. **A**: Pre-therapy ^{123}I-metaiodobenzylguanidine (MIBG) scan and extension score by Method 1 by four readers. **B**: Post-therapy ^{123}I-MIBG scan and extension score by Method 1 by four readers.

response (MR), no response (NR), or progressive disease (PD). Response was determined (prior to the current scoring study) by radiological review of post-therapy mIBG, CT and MRI scans, morphologic analyses of bone marrow, and vanillylmandelic acid (VMA) and homovanillic acid (HVA) levels in urine [15]. Pre-therapy evaluation was done after a patient finished any prior therapeutic regimen and no earlier than 6 weeks before ^{131}I-mIBG treatment. Post-therapy response evaluation was required at 6–8 weeks post-mIBG treatment and before a patient moved onto a new therapeutic regiment.

One patient's scores were excluded from the INRC response and overall impression analyses after determining that the pre-therapy mIBG scans were inadequate for response assessment. This patient still had scans included for the inter-observer concordance and was included in the overall PFS analysis.

Statistical Analysis

To quantify the concordance in scoring among readers, the analysis used the generalized concordance correlation coefficient (CCC) [17,18]. The 95% confidence interval for CCC and the *P*-value for comparing concordance were calculated based on bootstrap bias-corrected confidence limits [19]. To assess the reader agreement on overall response impression, the analysis used the κ-type statistics of O'Connell and Dobson [20]. We assigned impression scales 1, 0, and −1 to categories improved, not changed, and worsened, respectively, and used two choices of disagreement function (ω_1 and ω_0 in O'Connell and Dobson: with and without partial agreement).

Given the definitions of the relative scores in the mIBG Scoring Method section, a value of zero for the pre-therapy absolute score would result in an ill-defined relative score. When a reader assigned a pre-therapy absolute extension and intensity scores of 0 because he did not see any evidence of disease on the pre-therapy mIBG scan, this reader's score was excluded from the analysis of relative scores. This was particularly a problem for Method 2, which excluded score assignment for soft tissue lesions.

Correlation of scores with response was examined using logistic regression and generalized additive models [21]. The Wilcoxon rank sum test was used to compare continuous scores. Proportions of response were compared by the Fisher exact test and trend in proportions was assessed by the Cochran–Armitage test. The accuracy of using the relative score or impression as a prognostic test for response was summarized in terms of false positive fractions, positive predictive values as well as positive and negative diagnostic likelihood ratios (DLR), and the accuracy of two tests was compared by their relative DLR [22].

PFS was defined as the amount of time in months between ^{131}I-mIBG therapy and disease progression or death. For patients who received multiple therapies, the date of their first therapy was used to determine the time between therapy and disease progression. Survival and PFS curves were estimated by the Kaplan–Meier method and compared by the log-rank test [23]. In the PFS analysis, the time to disease progression was censored at the time a patient went on new therapy or at the time of the last follow-up visit. The proportion of patients changed to new therapy was estimated by its cumulative incidence rate with disease progression being a competing risk [24].

RESULTS

Inter- and Intra-Observer Score Concordance

Table III summarizes the pre- and post-therapy scores for all patients and gives the inter-observer concordance. In all cases, the inter-observer concordance for absolute scores was greater than or equal to 0.88. Method 1 absolute extension scores produced the highest inter-observer concordance among the four raters. The absolute extension score concordance was marginally higher than the absolute intensity score concordance for both methods. The relative score concordance was lower for both methods in comparison to the absolute score concordance.

To determine whether relative extension score concordance differed based upon extent of disease pre-therapy, the patient population was stratified into two groups based upon pre-therapy Method 1 absolute extension scores. Figure 3 shows the distribution of median (among four readers) pre-therapy extension scores. There were 25 scans with the median score ≤ 3, 4 scans with the score >3 but ≤ 5, and 28 scans with the score >5. Thus, the stratification of five sites of disease pre-therapy was chosen as the cut-point for the analysis. The relative extension and intensity concordance for Method 1 scoring was significantly higher for the groups with more extensive disease, either scores >3 ($P < 0.02$) or >5 ($P < 0.01$) (Table IV).

To evaluate the association of relative scores with the outcomes (INRC response and PFS), the median of the four readers' Method 1 extension scores was used to compute the relative extension score for each scan. These scores were utilized because they produced very high inter-observer concordance, accounted for soft tissue lesions, and the median statistic is more robust than the mean. We also compared the median relative extension scores with the median relative intensity scores (correlation coefficient = 0.88) and found no significant difference. In terms of high (>0.5) versus low (≤ 0.5) relative scores, the relative intensity and extension scores gave the same classifications except for one case (extension = 0.42, intensity = 0.71).

TABLE III. Scores and Inter-Observer Concordance on 57 Paired mIBG Scans

mIBG Score	Median	Range	Concordance (95%CI)
Absolute extension scores			
Pre-therapy Method 1	4.50	[0–30]	0.92 (0.86, 0.96)
Post-therapy Method 1	3.00	[0–30]	0.92 (0.82, 0.96)
Pre-therapy Method 2	2.50	[0–14]	0.91 (0.84, 0.94)
Post-therapy Method 2	1.50	[0–14]	0.89 (0.80, 0.94)
Absolute intensity scores			
Pre-therapy Method 1	6.50	[0–30]	0.91 (0.86, 0.94)
Post-therapy Method 1	4.00	[0–30]	0.88 (0.80, 0.93)
Pre-therapy Method 2	4.50	[0–21]	0.90 (0.85, 0.93)
Post-therapy Method 2	2.50	[0–21]	0.89 (0.82, 0.92)
Relative extension scores[a]			
Method 1	0.90	[0.00–3.50]	0.47 (0.35, 0.61)
Method 2	0.85	[0.00–3.00]	0.48 (0.31, 0.65)
Relative intensity scores[a]			
Method 1	0.79	[0.00–3.33]	0.58 (0.43, 0.73)
Method 2	0.69	[0.00–5.00]	0.51 (0.39, 0.69)

[a]For Method 1, six pre-therapy scans received absolute scores of 0: two scans by two readers and four scans by one reader. For Method 2, which did not include scores for soft tissue lesions, 18 pre-therapy scans received absolute scores of 0: 2 scans by all four readers, 2 scans by three readers, 7 scans by two readers, and 7 scans by one reader.

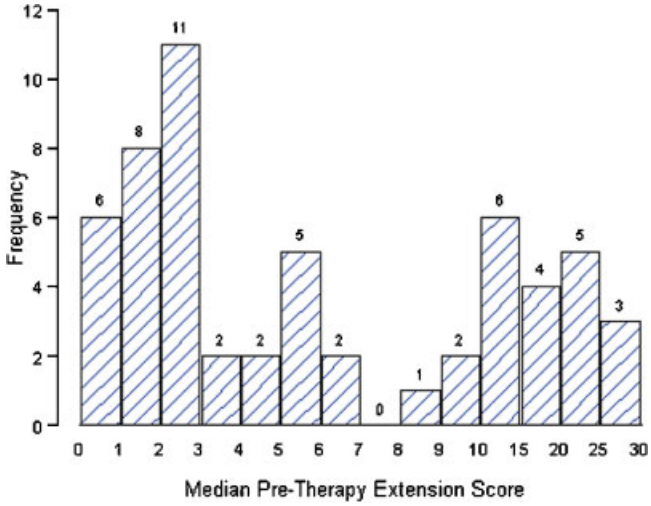


Fig. 3. Distribution of median pre-therapy extension scores. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

For the intra-observer analysis of the four readers, three pairs of pre- and post-therapy scans were read on two separate blinded occasions. The scans had median pre-therapy scores of 2, 14, and 19 in order to have a range of low (≤ 5) and high (>5) scores. Consistent with the results of the inter-observer concordance analysis, there was better agreement in scoring scans with high scores in comparison to low scores. Only 2 of these 12 pairs of relative scores resulted in a change in category between low (≤ 0.5) and high (>0.5) relative scores. Similarly, for the response by qualitative impression, 10/12 pairs were concordant and only 2 pairs of impression scores were discrepant for response.

When assessing soft tissue lesions, certain inconsistencies arose in scoring for both methods. For Method 1, 13 patients received pre-therapy scores from multiple readers for the 10th compartment without any documented soft tissue involvement by other imaging modalities. In addition, 10 patients, who only had evidence of soft tissue disease pre-therapy, should have received soft tissue scores for the 10th compartment from all readers, but in 7 of the 10 cases, at least one reader gave a soft tissue score of 0. For Method 2, this same group of 10 patients should have received pre-therapy scores of 0 for all compartments due to lack of bone and bone marrow disease but 8 had scores mistakenly assigned as skeletal disease by at least one of the four readers.

TABLE IV. Concordance for Method 1 mIBG Score After Stratification by Extent of Disease (mIBG Score of ≥3 or ≥5) Pre-Therapy

	Extension concordance				Intensity concordance			
	Pre	Post	Relative	*P*	Pre	Post	Relative	*P*
Low (≤3)	0.23	0.35	0.33	<0.02	0.38	0.55	0.46	<0.02
High (>3)	0.88	0.91	0.66		0.86	0.88	0.78	
Low (≤5)	0.27	0.34	0.28	<0.01	0.42	0.54	0.43	<0.01
High (>5)	0.87	0.91	0.80		0.84	0.89	0.86	

Correlation of Score With Response

There were 56 scan pairs associated with valid INRC response data, including 4 with CR, 17 with PR, 27 with NR (including 2 with MR), and 8 with PD. Figure 4A shows the relative extension score by Method 1 grouped by response. As shown by the outliers, seven therapies, that were scored as having a response with a relative score of ≤0.5, did not achieve CR or PR by other modalities, such as CT scan or bone marrow analysis (Table V). Conversely, six therapies failed to achieve response by relative score but still achieved response by INRC. Among these six therapies, five patients had low absolute pre-therapy extension scores (≤5) and also intermediate relative scores, and one therapy, with an absolute pre-therapy extension score of 7 and a median relative score of 1.04, resulted in NR by mIBG scan but PR by CT. However, for this study, we used the median score, where the individual response by INRC was previously derived from an impression by a single nuclear medicine reader.

The Method 1 relative extension score correlated with response to ^{131}I-mIBG therapy based on INRC, such that patients who achieved CR (n = 4) or PR (n = 17) had significantly lower relative extension scores ($P < 0.001$), with median score of 0.31 (CI 0.11, 0.51), compared to NR/PD (n = 35), with median of 0.98 (CI 0.90, 1.06) (Fig. 4A). Based on a linear logistic model, patients with low relative extension scores were significantly more likely to have a CR or PR to ^{131}I-mIBG therapy ($P < 0.001$) such that the odds of a CR or PR became about sixfold when relative extension scores drop by 0.5. Proportions of response (CR or PR) by INRC showed a significant decreasing trend ($P < 0.001$) among patients with low (≤0.5), intermediate (>0.5 but ≤1), and high (>1) relative extension scores. Of those who had a low relative score, 68% (15/22) had a response by INRC while only 20% (5/25) of those with an intermediate relative score and 11% (1/9) of those with a high relative score still achieved response. Compared to those with an intermediate score, patients with a low score were estimated to be 3.4 times (CI 1.5, 8.9) likely to achieve a response. The likelihood of response for patients with a high score was 0.56-fold (CI

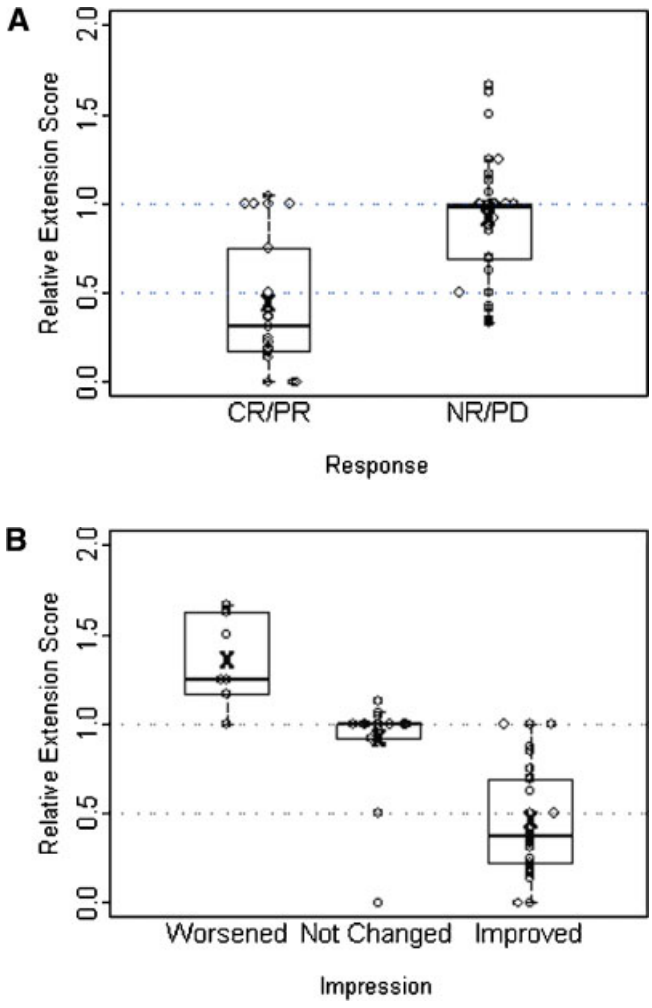


Fig. 4. **A**: Relationship between median relative score and tumor response to mIBG therapy by International Neuroblastoma Response Criteria (INRC). X marks the mean; the bar marks the median; the lower and upper border mark the first (Q1) and the third (Q3) quartiles, respectively, complete response (CR)/partial response (PR) (n = 21), no response (NR)/progressive disease (PD) (n = 35); $P < 0.001$. The five outliers with high relative score (iÝ1) in the CR/very good partial response (VGPR)/PR group had very low initial scores and poor agreement. **B**: Impression and relative extension score. Improved (n = 29), not changed (n = 20), worsened (n = 7).

TABLE V. Patients With Relative Scores ≤0.5 but No Response (NR) by International Neuroblastoma Response Criteria (INRC)

Patient #	Median relative extension score	Response by INRC	Comments
49	0.50	Progressive disease (PD)	Previously negative bone marrow became positive for disease post-therapy
48 (Third therapy)	0.50	PD	New nodule in breast positive for neuroblastoma by fine needle aspirate
102	0.33	NR	Bone marrow remained positive post-therapy
159	0.35	NR	Bone marrow remained positive post-therapy
84	0.41	PD	Increase in tumor size on CT
158	0.34	PD	Previously negative bone marrow became positive for disease post-therapy
58	0.43	PD	Previously negative bone marrow became positive for disease post-therapy

0.02, 4.02) compared to those with an intermediate score, although the difference was not significant.

We also examined the correlation of response with relative score according to pre-therapy disease score using the stratification of ≤5 (low) and >5 (high). Proportions of response did not differ significantly ($P > 0.9$) between the two pre-therapy strata for patients with a low (≤0.5) relative score, and likewise ($P > 0.4$) for patients with a high (>1) relative score. However, for those with an intermediate relative score, 33% (5/15) achieved response (1 CR and 4 PR) in the low stratum compared to 0% (0/10) in the high stratum ($P = 0.06$).

Inter-Observer Overall Response Impression Agreement

Impression of each scan was summarized as improved, not changed, or worsened for the INRC response and PFS analyses. Of the 57 scans, 23 (40%) scan pairs had a 'perfect' agreement (all four raters agreed); 20 (35%) had a 'majority' agreement (the same impression by three readers with the impression by the other reader differing by one category); 8 scans had a 'closely-split' agreement (two readers agreed with each other, and the two readers differed by one category); the other 6 scan pairs had mixed impressions. Allowing for partial agreement, agreement is perfect (=1.0) for 40%, excellent for 49%, and good (between 0.6 and 0.8) for 11% of the scans.

Using the impression scales (described above), of the scans in perfect or majority agreement, 24, 14, and 4 scans (excluding 1 inevaluable for response) were classified as improved, not changed, and worsened, respectively. The readers' overall impression of response associated well ($P < 0.001$ with or without excluding the scans in mixed agreement) with the relative extension scores although these results were less quantitative (Fig. 4B).

However, the use of the impression score was less accurate than the semi-quantitative scoring system for prediction of a good (positive) response. The relative accuracy of the relative score compared to the impression score was evaluated by comparing their DLRs to predict response. For the relative score, the positive DLR was 3.57 compared to 2.36 for the impression score, and the negative DLR was 0.36 compared to 0.29. That is, the odds of response increased by 3.57-fold with knowledge of a low (≤0.5) relative score compared to only 2.36-fold by an improved impression. The relative positive DLR was 1.52, implying that a positive result on relative score is more indicative of response than impression. However, the relative negative DLR was 1.23, implying that a negative result on relative score is less convincing for non-response than a negative impression result. The lower accuracy of a negative result on relative score can be improved by taking the pre-therapy score into account for patients with a relative score between 0.5 and 1.0 as discussed in the previous section.

Score and PFS

Survival and PFS for all patients are shown in Figure 5A. The PFS curve is truncated as patients were censored when they went on to new therapy. PFS according to pre-therapy stratum of score ≤5 or >5 showed a trend for improved PFS with a lower pre-therapy score ($P = 0.20$) (Fig. 5B). The relative extension score of ≤0.5 or >0.5 did not appear to affect PFS (Fig. 5C) and became even less significant when considering both disease progression and switching to new therapy as endpoint ($P = 0.73$). The reversal of the expected order of the curves in Figure 4C is explained by the fact that patients with higher relative scores, and therefore poorer response to the mIBG therapy, were about twice as likely to change to new therapy earlier, and therefore, be censored in the PFS analysis when compared to those with a lower relative score. In fact, by 4 months, 57% of patients with a high relative score >0.5 had changed to new therapy, and by 6 months post-mIBG therapy, 70% had changed, whereas 28% and 39% of patients with a low relative score had changed to new therapy by 4 and 6 months, respectively. Ultimately, overall response by INRC to mIBG therapy had no effect on PFS in this group of 49 refractory patients (data not shown). This finding was consistent with the lack of effect of relative score on PFS.

DISCUSSION

Several studies have shown that incorporation of mIBG scan results into analysis of response is critical both for

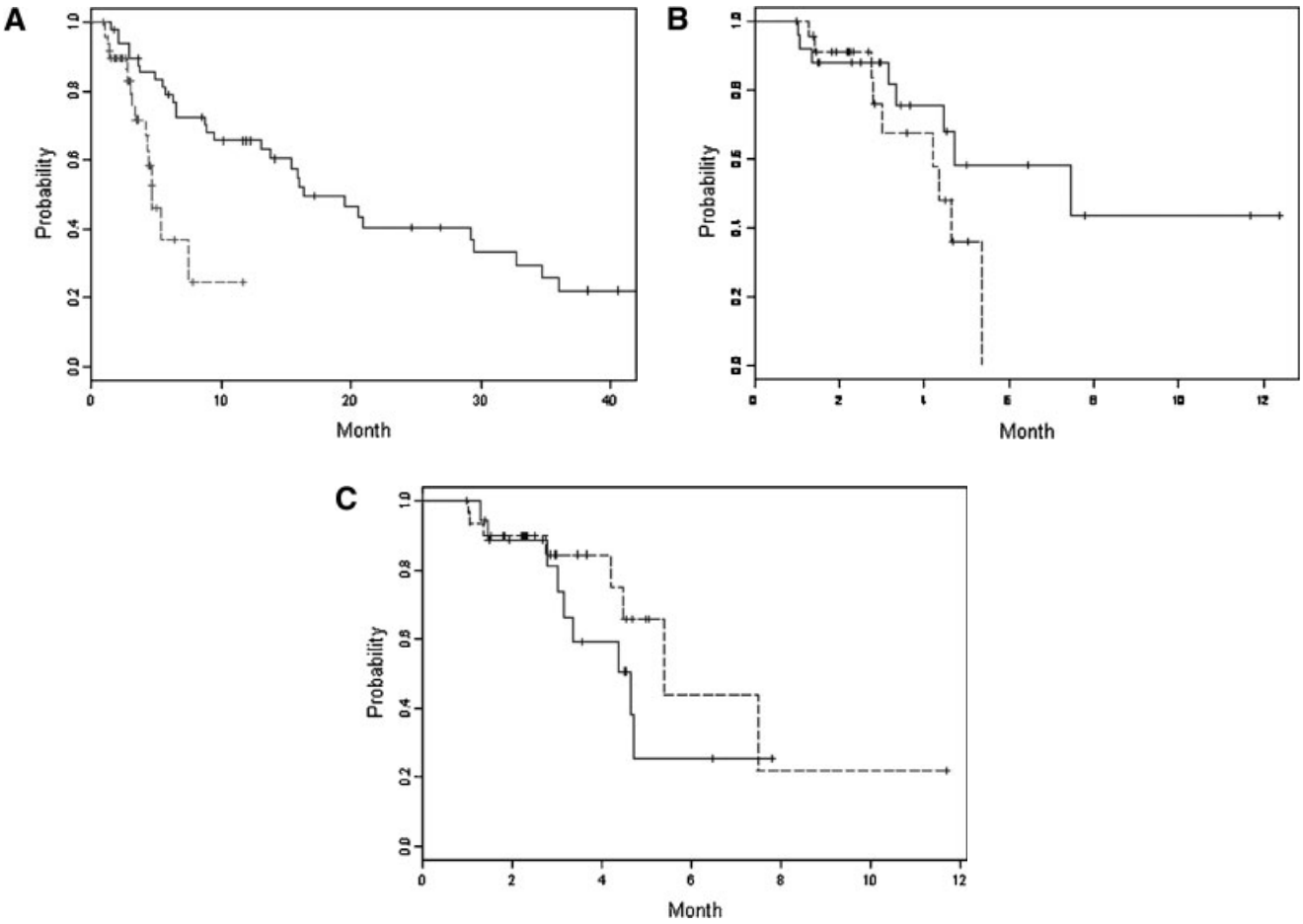


Fig. 5. **A**: Overall survival (OS) (solid line) and progression-free survival (PFS) (dashed line) for all patients (n = 49) from time of first mIBG therapy. **B**: PFS by pre-therapy mIBG extension score for low ($\leq$5, solid line) or high score (>5, dashed line); $P = 0.20$. **C**: PFS by relative extension score post-therapy of $\leq$0.5 (solid line) or >0.5 (dashed line); $P = 0.34$.

determining response and survival [11,25–28]. An objective and consistent method for interpretation of these scans is therefore critical for comparison of clinical outcomes. A prior study [10], using Method 1 to score patients at diagnosis, had a median pre-therapy extension score of 18, compared to the current study, with a median score of 4.5. The results of the current study show that semi-quantitative scoring of mIBG scans is reliable in relapsed neuroblastoma despite the lower median score, with excellent inter- and intra-observer concordance, similar to results for newly diagnosed patients [10,13,14]. We have shown further that the calculation used in the relative score for response continues to be reliable though more error is introduced in patients with fewer than three to five lesions prior to therapy. Extension and intensity score concordances were very similar. Overall, the data showed no advantage in using both scores to assess response, and there was only one instance where the relative intensity score fell into a different category for response than the relative extension score.

Method 1 relative extension scores can be used accurately for semi-quantitative analysis of response by mIBG scan in refractory patients and correlate significantly with whether or not a patient had an overall response (CR, VGPR, or PR) or no response (NR, MR, or PD) by INRC to therapy. Based upon the results displayed in Figure 4A, the present study shows that a relative extension score of $\leq$0.5 indicates that a patient likely had a response to mIBG therapy by INRC, and a relative extension score of >0.5 indicates that a patient likely had NR.

Relative extension scores had the highest concordance among readers for patients with more than three to five sites of disease. For patients with fewer sites of disease, because the post-therapy score is divided by the pre-therapy score to calculate the relative score, a single digit difference assigned by the four raters in either of these scores will have a significant effect on this ratio. The data suggest that any decrease in absolute score is indicative of response in patients with scores below three to five at the commencement of therapy, but that the semi-quantitative scoring system is most useful in those with pre-treatment score $\geq$3.

Method 2 (without assigning soft tissue scores) produced many pre-therapy absolute scores of 0, resulting in a

substantially decreased sample size since these scores were excluded from the analysis. Ideally, only 10 patients should have received pre-therapy scores of 0 by Method 2 because they were documented as having no evidence of soft tissue disease pre-therapy. However, only two patients received pre-therapy scores of 0 from all four readers. Even though these patients' scans were dropped from the concordance analysis, failing to assess soft tissue lesions could result in an inaccurate response evaluation by semi-quantitative score. Discrepancies in assignment of score for soft tissue lesions might be eliminated by simply allowing readers to score any lesion within its appropriate anatomical segment instead of trying to differentiate between soft tissues or skeletal lesions.

Although the impression of the reader as to whether a scan was improved, worsened, or stabilized showed reasonable concordance and correlated with the relative score and the response by INRC, the accuracy was lower than that of the semi-quantitative scoring system. For example, only 7/22 (these 7 patients are listed in Table V) patients with a low relative score of $\leq$0.5 had NR or PD, whereas 12/29 (6 of these 12 are in Table V) patients rated as ''improved'' on MIBG scan, fell into the NR or PD category by INRC. This corresponded to a positive predictive value of 68% for the relative score versus 59% for the impression. It also resulted in a false positive fraction of only 20% for the relative score compared to 34% for the impression.

Despite a strong correlation with response by INRC, the relative extension scores carried no significance in predicting PFS. This may be a result of two factors. First, there was no difference in this group of 49 patients in PFS between those with response to therapy and those without response. Secondly, those that were left with residual disease were also often entered on a new therapy soon after the MIBG treatment.

The semi-quantitative scoring system must be used in combination with other response criteria in refractory neuroblastoma as it only assesses disease visible on mIBG scan and fails to include analysis of bone marrow and mIBG-negative disease only evaluable on CT or MR. Current cooperative high-risk neuroblastoma protocols both in Europe and in North America are incorporating semi-quantitative scoring by mIBG scan into their central review of response, and a recent International Neuroblastoma Risk Group meeting of cooperative study groups from Asia, Europe, and North America in September, 2005 reached a consensus on incorporation of a common scoring system into their evaluation of response for a more objective measure and to facilitate comparison between studies. This new system will assign soft tissue lesions within their appropriate skeletal segment, and divide the body into seven segments: head, spine, ribs and chest, abdomen and pelvis, arms, legs.

Semi-quantitative scoring of mIBG scans provides a more reliable method of assessing response in mIBG positive lesions in patients with relapsed neuroblastoma than a standard impression reading, especially in patients with initial scores $>$3. The scoring system is reproducible with good intra- and inter-observer concordance and will provide an important component of overall response criteria for this disease in patients with recurrent as well as newly diagnosed neuroblastoma.

ACKNOWLEDGMENT

We thank Sylvia Corpuz and John Huberty in the Department of Nuclear Medicine at UCSF for valuable technical assistance and Janet Veatch for coordination of the Phase II therapy study. We also acknowledge the financial support (in part) of the National Institute of Health grant PO1 CA81403, 2MO1 RR0127, as well as donations from the Campini Foundation, the Conner Research Fund, the Dougherty Foundation, Kasle and Tkalcevik Neuroblastoma Research Fund, Alex's Lemonade Stand, and the Thrasher Foundation.

REFERENCES

1. Matthay KK, Villablanca JG, Seeger RC, et al. Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-cis- retinoic acid. Children's Cancer Group. N Engl J Med 1999;341:1165–1173.
2. Garaventa A, Bellagamba O, Lo Piccolo MS, et al. 131I-metaiodobenzylguanidine (^{131}I-MIBG) therapy for residual neuroblastoma: A mono-institutional experience with 43 patients. Br J Cancer 1999;81:1378–1384.
3. Kang TI, Brophy P, Hickeson M, et al. Targeted radiotherapy with submyeloablative doses of ^{131}I-MIBG is effective for disease palliation in highly refractory neuroblastoma. J Pediatr Hematol Oncol 2003;25:769–773.
4. Troncone L, Riccardi R, Montemaggi P, et al. Treatment of neuroblastoma with ^{131}I-metaiodobenzylguanidine. Med Pediatr Oncol 1987;15:220–223.
5. Hoefnagel CA, Voute PA, De Kraker J, et al. [131I]metaiodobenzylguanidine therapy after conventional therapy for neuroblastoma. J Nucl Biol Med 1991;35:202–206.
6. Klingebiel T, Berthold F, Treuner J, et al. Metaiodobenzylguanidine (mIBG) in treatment of 47 patients with neuroblastoma: Results of the German Neuroblastoma Trial. Med e palliation in highly refractory neuroblastoma. J Pediatr Hematol Oncol 2003; 25:769–773.
7. Matthay KK, Panina C, Huberty J, et al. Correlation of tumor and whole-body dosimetry with tumor response and toxicity in refractory neuroblastoma treated with (131)I-MIBG. J Nucl Med 2001;42:1713–1721.
8. Matthay KK, DeSantes K, Hasegawa B, et al. Phase I dose escalation of ^{131}I-metaiodobenzylguanidine with autologous bone marrow support in refractory neuroblastoma. J Clin Oncol 1998; 16:229–236.
9. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. J Natl Cancer Inst 2000;92:205–216.
10. Matthay KK, Edeline V, Lumbroso J, et al. Correlation of early metastatic response by ^{123}I-metaiodobenzylguanidine scintigraphy with overall response and event-free survival in stage IV neuroblastoma. J Clin Oncol 2003;21:2486–2491.
11. Katzenstein HM, Cohn SL, Shore RM, et al. Scintigraphic response by ^{123}I-metaiodobenzylguanidine scan correlates with event-free

survival in high-risk neuroblastoma. J Clin Oncol 2004;22:3909–3915.
12. Suc A, Lumbroso J, Rubie H, et al. Metastatic neuroblastoma in children older than one year: Prognostic significance of the initial metaiodobenzylguanidine scan and proposal for a scoring system. Cancer 1996;77:805–811.
13. Ady N, Zucker JM, Asselain B, et al. A new ^{123}I-MIBG whole body scan scoring method—Application to the prediction of the response of metastases to induction chemotherapy in stage IV neuroblastoma. Eur J Cancer 1995;31A:256–261.
14. Frappaz D, Bonneu A, Chauvot P, et al. Metaiodobenzylguanidine assessment of metastatic neuroblastoma: Observer dependency and chemosensitivity evaluation. The SFOP Group. Med Pediatr Oncol 2000;34:237–241.
15. Brodeur GM, Pritchard J, Berthold F, et al. Revisions of the international criteria for neuroblastoma diagnosis, staging, and response to treatment. J Clin Oncol 1993;11:1466–1477 (see comments).
16. Shulkin BL, Shapiro B. Current concepts on the diagnostic use of MIBG in children. J Nucl Med 1998;39:679–688.
17. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255–268.
18. Barnhart HX, Haber M, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. Biometrics 2002;58:1020–1027.
19. Efron B, Tibshirani RJ. Better bootstrap confidence intervals. An introduction to the Bootstrap. New York: Chapman & Hall; 1993. pp 184–188.
20. O'Connell DL, Dobson AJ. General observer-agrreement measures on individual subjects and groups of subjects. Biometrics 1984;40: 973–983.
21. Hastie T, Tibshirani R. Generalized additive models: Logistic regression. Generalized additive models. New York: Chapman and Hall; 1990. pp 95–101.
22. Pepe. Measures of accuracy for binary tests. New York: Oxford; 2003. pp 14–34.
23. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–481.
24. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. Ann Stat 1988;16:1141–1164.
25. Kushner BH, Yeh SD, Kramer K, et al. Impact of metaiodobenzylguanidine scintigraphy on assessing response of high-risk neuroblastoma to dose-intensive induction chemotherapy. J Clin Oncol 2003;21:1082–1086.
26. Ladenstein R, Philip T, Lasset C, et al. Multivariate analysis of risk factors in stage 4 neuroblastoma patients over the age of one year treated with megatherapy and stem-cell transplantation: A report from the European Bone Marrow Transplantation Solid Tumor Registry. J Clin Oncol 1998;16:953–965.
27. Shah Syed GM, Naseer H, Usmani GN, et al. Role of iodine-131 MIBG scanning in the management of paediatric patients with neuroblastoma. Med Princ Pract 2004;13:196–200.
28. Frappaz D, Combaret V, Desuzinges C, et al. Can MIBG scan replace the need for bone marrow assessment at diagnosis and reassessment in stage 4 neuroblastomas? Bull Cancer 2004;91: E253–E260.