Computer-assisted Curie Scoring for Metaiodobenzylguanidine (MIBG) Scans in Patients with Neuroblastoma

Elizabeth A. Sokol[a], Roger Engelmann[b], Wenjun Kang[c], Navin Pinto[d], Adam Starkey[b], Hollie Lai[e], Helen Nadel[f], Barry L. Shulkin[g], Yonglin Pu[b], Daniel Appelbaum[b], Gregory A. Yanik[h], Susan L. Cohn[a], Samuel Armato[#b], Samuel Volchenboum[#a,c]

[a]Department of Pediatrics, University of Chicago, Chicago IL
[b]Department of Radiology, University of Chicago, Chicago, IL
[c]Center for Research Informatics, University of Chicago, Chicago IL
[d]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA
[e]Department of Radiology, Children's Hospital of Orange County, Orange, CA
[f]Department of Radiology, University of British Columbia, Vancouver, BC
[g]Department of Diagnostic Imaging, St. Jude Children's Research Hospital, Memphis, TN
[h]Department of Pediatrics, University of Michigan School of Medicine, Ann Arbor, MI

[#]Contributed equally.

Corresponding Author:
Samuel Volchenboum
900 E 57th Street
KCBD 5130
Chicago, IL 60637
Phone: (773) 769-7433
Fax: (773) 834-1329
Email: slv@uchicago.edu

Word count
- Abstract: 249
- Main text: 3485

Tables, Figures, and Supplemental Files: 3 Tables, 3 Figures, 3 Supplemental

Tables, and 2 Supplemental Figures

Short running title: Computer-assisted Curie Scoring of MIBG Scans

Keywords: MIBG, neuroblastoma, Curie score

Abbreviations:

MIBG	Metaiodobenzylguanidine
COG	Children's Oncology Group
INRG	International Neuroblastoma Risk Group
EFS	Event-free Survival
SIOPEN	International Society of Pediatric Oncology European Network

INRC	International Neuroblastoma Response Criteria
QARC	Quality Assurance Review Center
SPECT	Single-photon Emission Computed Tomography

This work was presented previously at Advances of Neuroblastoma Research in 2016, American Society of Clinical Oncology in 2016, and American Society of Clinical Pharmacology and Therapeutics in 2018.

Abstract

Background: Radiolabeled metaiodobenzylguanidine (MIBG) is sensitive and specific for detecting neuroblastoma. The extent of MIBG-avid disease is assessed using Curie scores. Although Curie scoring is prognostic in patients with high-risk neuroblastoma, there is no standardized method to assess the response of specific sites of disease over time. The goal of this study was to develop approaches for Curie scoring to facilitate the calculation of scores and comparison of specific sites on serial scans.

Procedure: We designed three semi-automated methods for determining Curie scores, each with increasing degrees of computer assistance. Method A was based on visual assessment and tallying of MIBG-avid lesions. For Method B, scores were tabulated from a schematic that associated anatomic regions to MIBG-positive lesions. For Method C, an anatomic mesh was used to mark MIBG-positive lesions with automatic assignment and tallying of scores. Five imaging physicians experienced in MIBG interpretation scored 38 scans using each method, and the feasibility and utility of the methods were assessed using surveys.

Results: There was good reliability between methods and observers. The user-interface methods required 57-110 seconds longer than the visual method. Imaging physicians indicated that it was useful that Methods B and C enabled tracking of lesions. Imaging physicians preferred method B to method C because of its efficiency.

Conclusions: We demonstrate the feasibility of semi-automated approaches for Curie score calculation. Although more time was needed for strategies B and C, the ability to track and document individual MIBG-positive lesions over time is a strength of these methods.

Introduction

Neuroblastoma is an embryonal tumor of the sympathetic nervous system responsible for 15 percent of pediatric cancer deaths in the United States[1]. It displays genetic and clinical heterogeneity. Based on clinical and biologic variables, patients are assigned to risk groups and treatment regimens[2]. Despite excellent outcomes for some, survival remains poor for high-risk patients despite intensive, multi-modal therapies[3-5]. Survival has been shown to be superior for those who respond to induction therapy[6-9]. ^{123}I-MIBG whole-body scintigraphy is a powerful imaging technique for detecting neuroblastoma and evaluating treatment response.

The current standard in the Children's Oncology Group (COG) for comparing successive ^{123}I- MIBG scans to assess treatment response is the Curie method[10], as detailed in a recent consensus report from the International Neuroblastoma Risk Group (INRG) Task Force[11]. The scoring algorithm divides the body into nine skeletal sections with a tenth soft-tissue section. The ten sections are graded for extent of MIBG avidity on a 0-3 scale: 0=no involvement, 1=one site, 2=more than one site, 3=diffuse involvement (>50% of the segment). This method is considered "semi-subjective". The Curie score is the sum of all ten segments. Serial patient Curie scores are compared to assess treatment response.

Matthay *et al.* reported significantly worse outcomes for patients with total Curie scores >2 following induction chemotherapy compared to those with scores of <=2[6,12]. Subsequent analyses determined Curie scores >2 after induction but not at diagnosis in patients enrolled on a high-risk COG study were associated with significantly worse event-free survival (EFS)[7]. There was no correlation between Curie score at diagnosis and survival. More recently, in an International Society of Pediatric Oncology European Network (SIOPEN) high-risk study, researchers were able to validate that patients with Curie score of ≤2 post induction have significantly better EFS[13]. Although the prognostic value of Curie scores have not yet been validated in high-risk patients receiving current COG standard treatment including tandem stem cell transplants and immunotherapy following induction, MIBG relative scores on bone sectors have been integrated into the recent revision of the International Neuroblastoma Response Criteria (INRC)[14]. The relative bone score is the ratio of the Curie scores at response assessment to diagnosis (without the soft tissue component). Resolution of MIBG activity defines a complete response. A partial response is defined as a reduction of 50% or greater in MIBG bone score. Reduction by less than 50% is stable disease. Any new lesion represents progressive disease.

Although the prognostic significance of Curie scores has been established, the manual methods currently used to calculate total scores do not provide mechanisms to longitudinally track specific lesions or easily compare regions of diffuse involvement over time. Often, only the total Curie score is provided in the MIBG report by the imaging physician, without documentation of the specific lesions and sites of disease. We hypothesized that by integrating a computerized user interface and automation, Curie scores would be more accurately quantified and documented for longitudinal review. To test this hypothesis, we designed three semi-automated methods to calculate Curie scores with increasing degrees of computer assistance. A survey was administered to evaluate the imaging physicians' opinions regarding the feasibility and utility of each method in clinical practice. The aim of the study was to compare the three methods to determine: 1) the efficiency of each method for determining Curie scores; 2) the variance of Curie scores between readers; and 3) the feasibility of tracking specific MIBG-positive lesions over time.

Methods

Patient Cohort

Patients with neuroblastoma and available MIBG scans were identified through University of Chicago and COG. Institutional review board approval was obtained to collect imaging and clinical information from patients at University of Chicago. Consents were obtained from patients available to consent, and a waiver was granted for patients unavailable to consent. De-identified scans were also obtained from the Quality Assurance Review Center (QARC) through COG under a data use agreement with University of Chicago. The use of these scans was additionally given a waiver of consent. Participating imaging physicians also signed research consents to collect data surrounding their use of the methods and their survey information.

MIBG Semiquantitative Curie Scoring

Three semi-automated mechanisms to calculate Curie scores were designed to allow imaging physicians to view and score planar anterior and posterior 24-hour MIBG scans. In each method, a computer interface (developed in University of Chicago's Abras system[15]) allowed the imaging physician to view the images, perform windowing and zooming of images, and determine Curie scores. To evaluate the three methods, MIBG scans were reviewed without any accompanying clinical information by five nuclear imaging physicians experienced in interpreting MIBG scans, from four academic institutions: University of Chicago, St. Jude Children's Research Hospital, University of British Columbia, and Children's Hospital of Los Angeles. Only planar images were reviewed. Accompanying single photon emission computed tomography (SPECT) or SPECT/CT images, if performed, were not made available.

Method A: Manual Curie Score Within a Computer Interface

Similar to traditional Curie scoring, this method simply facilitated a sum of scores from ten different anatomic sites, including skeletal (cranio-facial, cervical and thoracic spine, chest (ribs/sternum/clavicles/scapula), lumbar and sacral spine, pelvis, humeri, lower arms, femurs, and lower legs) and soft-tissue. As in traditional Curie scoring, skeletal sites were individually scored from 0 to 3 as described above. A score of 3 was assigned for the soft-tissue region if disease occupied **>**50% of the chest or abdomen. A patient’s Curie score at each time point was calculated as the sum of scores over all individual sites, with a maximum score of 30.

For this method, imaging physicians reviewed each set of images and clicked on buttons to directly indicate a 0 - 3 score for each of the nine anatomic segments (“regions”) and the soft tissue segment (Fig 1, Supplemental Figure S2). This method most closely approximates the current scoring method of subjectively evaluating each segment and adding the scores. The interface eases this process by allowing the imaging physician to quickly click the score for each segment, while the overall tally is maintained. Furthermore, in contrast to traditional methods, the ten individual region scores are preserved, allowing the clinician to return later to review the contribution from each segment.

Method B: Computer-Assisted Curie Scores

In Method B, the imaging physicians marked the lesion locations on the images and then indicated the corresponding Curie segment by clicking a schematic figure. The score of each segment (0, 1, 2, or 3) was automatically computed (Fig 2, Supplemental Figure S2). The imaging physician could indicate a lesion in one of three ways: simple point, line for a linear bone lesion, or a loop to show the area of a lesion. When a line or loop was drawn, the interface collected user input on whether the lesions belonged to a segment with > 50% tumor involvement. If so, the segment was automatically scored a 3 by the system. Otherwise, one lesion marked in a segment resulted in a 1 score for the segment, and two or more marks resulted in a 2 score. To indicate a soft-tissue lesion, the imaging physician held down a modifier key on the keyboard while drawing the lesion instead of clicking a skeletal segment on the schematic.

After drawing one or more lesions, the imaging physician indicates the corresponding anatomic segment. The scores were then automatically updated. The schematic was color-coded as a visual reference for the imaging physician to reflect the current score for each anatomic segment (red:3, orange:2, yellow:1). The system tracked the individual segment scores as well as the current total Curie score. The lesions and contributing segments are then preserved for later review.

Method C: Computer-Assisted Curie Scores

In Method C, the imaging physician defined a Curie anatomic segment region map or "mesh" on the patient images by specifying key anatomic points that corresponded to points shown on a skeleton schematic (Figs 3A and 3B, Supplemental Figure S2), making adjustments to the region map as necessary by dragging the intersecting handles between the segments (Fig 3C). An anterior image mesh defined seven of the nine skeletal regions, and a smaller posterior image mesh defined the other two skeletal regions corresponding to the spine. The ability to manipulate the mesh is especially useful in children, as there may not be uniformity in anatomic landmarks identified by the mesh generation software. The imaging physician could then mark lesions in either image (as points, lines, or loops), except that spine lesions were required to be marked in the posterior image, while medial rib, sternum, and pelvic lesions were required to be marked in the anterior image, in order to be assigned to the proper anatomic segment. Based on location, each lesion mark was automatically assigned to the corresponding Curie anatomic region, and scores were adjusted by the system without requiring additional effort from the imaging physician. A key difference between the Methods B and C is that in the latter, the imaging physician need not click on the schematic to assign an anatomic segment to a lesion. Rather, the lesion's position within the region map automatically determined its anatomic segment. The lesions could be specified at any time during a case, before or after the region map itself was specified, and the region map could likewise be modified as needed.

Comparison of the Curie Scoring Methods A, B, and C

Imaging physicians participated in three sessions, each separated by at least two weeks to reduce recall bias. The first session involved learning Method A and then scoring all 38 scans. The second session involved learning Methods B and C and then scoring 19 scans using Method B and 19 scans using Method C, reversing this for the third session with randomization of scan order. The 19 scans were randomized for each imaging physician, so they all read each scan with each method, but in different sessions. Data collected included scores for each region and the time it took to complete scoring of each scan.

A University of Chicago analyst was physically or virtually present (through Skype or WebEx) with the imaging physicians at each session and provided technical guidance to assist the imaging physicians when needed, being careful not to influence any of the clinical decisions. This guidance included noting and helping users correct technical mistakes in using the interface, occasionally noting anatomic segment omissions, and indicating features of the interface (such as using the color-coded schematic to see which anatomic segments have no lesions assigned) that could help the imaging physicians use the interfaces to their full extents. The guidance was most often needed near the beginning of a session, and observers generally became very proficient and confident in using the interfaces as they progressed in experience. The software is designed to be used unaided.

Survey

An 18-question survey was administered to characterize ease-of-use, clinical utility, and potential adoption of each of the methods (Supplemental Figure S1). The secure REDCap survey was administered *via* e-mail to all participating imaging physicians and was completed by all five imaging physicians.

Statistical Analysis

ANOVA testing was completed by a senior statistician in University of Chicago's Center for Research Informatics to analyze variance between imaging physicians and between methods. Cohen's kappa statistics and Weighted Fleiss' kappa statistics were then calculated to evaluate intra-observer and inter-observer reliability. The survey data were analyzed descriptively.

Results

Patient Cohort

Thirty-four patients from University of Chicago with available MIBG scans were enrolled on this study. Additionally, MIBG scans from 4 patients were obtained from QARC for analysis. All patients had a diagnosis of neuroblastoma except for one with metastatic paraganglioma (Supplemental Table S1). Twenty-seven University of Chicago patients had metastatic disease. The patients ranged in age from six weeks to 22 years (median 3 years). Eleven scans were obtained at initial diagnosis, ten during or after induction chemotherapy, three during or prior to immunotherapy, one at end of therapy, and nine during therapy for relapsed disease. The four patients with scans obtained through QARC have unknown clinical information.

Curie Scoring

Reliability between methods

We first wanted to study whether the same observer obtained similar results using Methods A, B, and C. All MIBG scans were scored by each imaging physician using all three methods (Supplemental Table S2). Reliability between methods was assessed with Cohen’s kappa coefficient, a statistic that measures inter-rater agreement for categorical items. A Cohen’s kappa greater than 0.6 denotes good agreement and greater than 0.8 suggests very good agreement between methods. Scores of 0, 1, 2, and 3 in each region were considered categorical variables as a score of 3 represents the degree to which a region is involved with disease, not the number of lesions. First analysis was performed assuming that all Curie regions are rated with similar reliability. The kappa statistic between Method A and Method B was 0.869, indicating very good agreement (Table 1). Similarly, the kappa statistic between Method A and Method C was 0.847 and between Method B and Method C was 0.861. To compare all three methods, Fleiss' kappa was used, which is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings. Fleiss' kappa for Methods A, B, and C was 0.702, suggesting substantial agreement among the methods. Similar analyses were performed by Curie region, and kappa statistics were calculated between Methods A and B (kappa range 0.678-0.951), Methods A and C (kappa range 0.618-0.917) and Methods B and C (kappa range 0.689-0.935). These data suggest excellent reliability across methods. The poorest reliability existed in regions seven (the lower arms) and region ten (soft tissue) but even these showed substantial reliability across methods.

Inter-observer reliability

We then calculated the reliability among observers for each method. Inter-observer reliability was calculated using weighted Fleiss' kappa statistics to show consistency between the imaging physicians (Table 2). The kappa statistic was first calculated assuming that all Curie regions were of similar reliability and was 0.840 for Method A, 0.811 for Method B, and 0.804 for Method C, demonstrating excellent reliability between observers. We then evaluated by Curie region and found kappa statistics ranging from 0.743 to 0.933 for Method A, from 0.699 to 0.918 for Method B, and from 0.728 to 0.901 for Method C. Overall inter-observer reliability is excellent, similar to intra-observer reliability.

Time analysis

On average, reading scans with Method B took 72% longer than with Method A. Method C took 141% longer than Method A and 68% longer than Method B. For Method A, the time for scoring each scan ranged from 15 seconds to 437 seconds, for Method B, the time to analyze each scan ranged from 12 seconds to 737 seconds, and for Method C, the time to analyze each scan ranged from 41 seconds to 693 seconds (Supplemental Table S3). On average, Method B took 57 seconds longer per scan than Method A, and Method C took 53 seconds longer than Method B and 110 seconds longer than Method A.

Physician Assessment of the Curie Scoring Methods

All five participating imaging physicians completed the survey (Supplemental Table S4). The imaging physicians noted how likely they would be to utilize each method (Method A: 3 very likely, 1 somewhat likely, 1 unlikely; Method B: 4 very likely, 1 somewhat unlikely; Method C: 2 very likely, 2 somewhat likely, 1 unlikely). The imaging physicians indicated that their preferred method would be somewhat or very useful for routine MIBG scan reading. The majority specified that the data provided by the semi-automated methods would be very useful for central reviewers evaluating MIBG scans as part of a clinical study. All of the imaging physicians agreed that it would be valuable for the treating oncologists to have information about the response of individual MIBG lesions. Comments provided as part of the survey revealed concern that Method C took longer than the other methods, and the added time needed to determine the Curie score limited its utility as a clinical tool. However, one imaging physician commented that a semi-automated method of any kind would be better than current practice. Another reported that these strategies could be incorporated into practice. Four thought it was somewhat or very important to keep a record of the lesions that contributed to the score in each region. Three indicated that the information provided by Methods B and C regarding each of the component scores would be useful to oncologists for response assessment and treatment decisions.

Discussion

In this study, we designed and tested three semi-automated methods for evaluating Curie scores with varying degrees of computer assistance. Method A most closely mimics current practice with assignment of Curie scores to each region by visual inspection, with the system providing a running tally. Method B involves marking MIBG-avid lesions on the scan and assigning them to Curie regions on a schematic figure. Method C involves creation of a mesh to define Curie regions on the image and then the marking of MIBG-avid lesions. We showed it is feasible to utilize a user interface and semi-automated method to apply Curie scores. Furthermore, there was consistency across providers and methods. The inter-method and inter-observer reliability is very good when evaluating both total scores and each individual region. The imaging physicians easily learned to use each method, indicating these methods could be broadly employed to aid in assigning Curie scores. Currently, many imaging physicians evaluating MIBG scans provide only a total score. Several studies have demonstrated the scores assigned after induction therapy are prognostic of outcome of high-risk patients treated with prior treatment regimens. Curie scores are included in the recently published INRC response criteria and provide important information for treatment decisions[14]. Tracking specific MIBG-positive lesions over time is likely to enhance accurate response assessment, provide additional prognostic information, and may ultimately lead to more informed treatment decision-making. These semi-automated methods have the potential to standardize Curie scoring in clinical practice. Methods B and C took 57 seconds (72%) and 110 seconds (141%)

respectively longer to use than Method A. However, our survey results indicated that imaging physicians preferred Methods B and C to Method A, because these methods enabled longitudinal tracking of lesions. The imaging physicians also noted that a disadvantage of Method C was the increased time required when compared to Methods A and B.

Analysis of patients enrolled on a previous COG clinical trial conducted from 2001 to 2006, demonstrated that Curie scores following induction therapy were prognostic of outcome in patients with Stage 4 high-risk neuroblastoma[7]. Extremely poor outcomes were observed for patients with *MYCN* non-amplified tumors with Curie scores >2 and for patients with *MYCN*-amplified disease with Curie scores >0. Decarolis and colleagues confirmed the prognostic value of Curie scores >2 following induction and showed that a SIOPEN MIBG score >4 following induction was also associated with inferior outcome[8]. These studies highlight the prognostic importance of MIBG scoring. The current manual method of determining Curie scores limits the ability to longitudinally monitor specific lesions or regional disease in a standardized manner. We hypothesized that by integrating computational techniques, Curie scores would be more reliably quantified and specific sites of disease could be accurately assessed for response.

These semi-automated methods represent the first step toward making Curie scoring more consistent. Further automation may be achieved via the process previously reported at University of Chicago with Tc-99m bone scans[16,17]. A computer-aided diagnostic approach was designed to identify differences in scans from multiple time points using a non-linear image warping technique. Shiraishi *et al.*, created a computational algorithm involving image density normalization and downstream processing to successfully identify new and resolved lesions over time. This method was subsequently found to be beneficial 84.6% of the time and has the potential to significantly aid in the evaluation of Tc-99 bone scans. Although bone scans are no longer used in patients with neuroblastoma, a similar approach could be applied to MIBG scans to help identify very subtle changes in metastatic disease patterns, thus making this modality more quantitative and leading to a more precise prognostication method for children with neuroblastoma. Future versions of the computer-assisted methods will need to adjust for improvements in technology, including SPECT imaging.

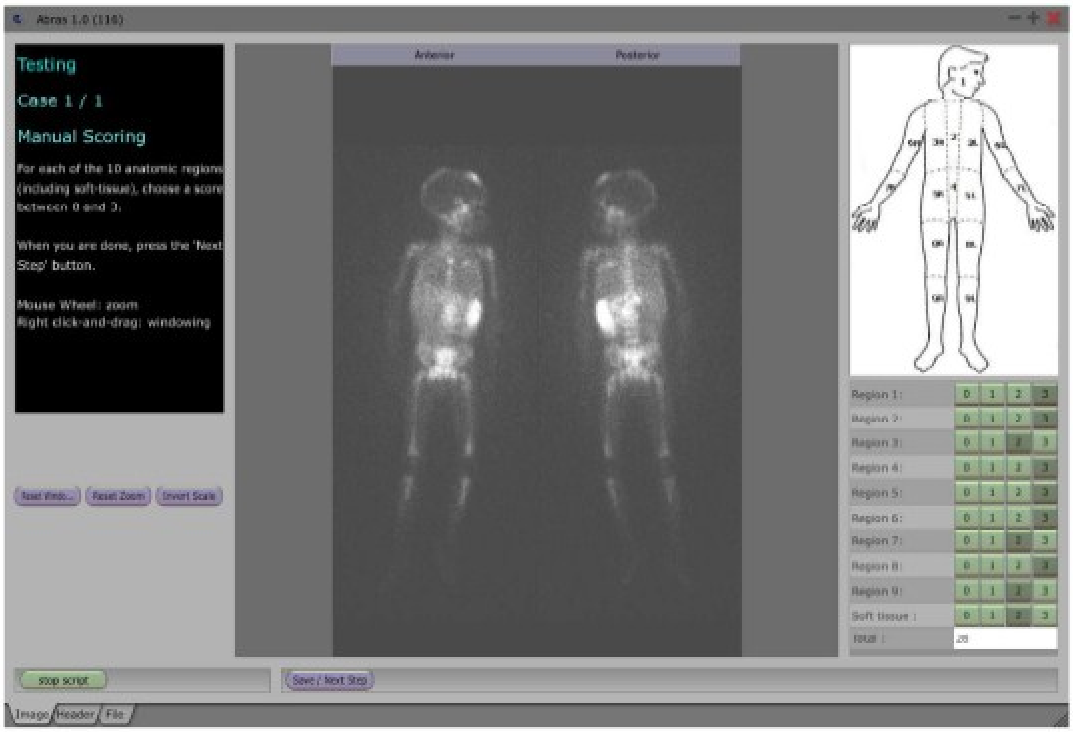
To bring these methods to clinical practice, the method must be validated in a larger study by comparing scores obtained using semi-automated scoring to scores given by consensus review of expert readers. A prospective study can be used to determine the feasibility of using a semi-automated method in regular clinical practice. Ultimately, broad utilization of these methods could help to standardize the application of Curie scores and aid in monitoring the response of MIBG-avid neuroblastoma over time.

Conflict of Interest Statement: Samuel Armato and Roger Engelmann receive royalties and licensing fees for CAD technology through The University of Chicago and Samuel Armato is a consultant for Aduro Biotech, Inc and Boehringer-Ingelheim.

Acknowledgements: This work was supported by [NIH CTSA UL1 TR000430] and [NICHD-NIGMS T32GM007019]. Supported, in part, by [NSF REU Award No.1062909]. The authors would like to thank Ling Teng, Jason Tam, and Sabina Nilakhe for her assistance in early development of the user interface.

References

1. Park JR, Bagatell R, London WB, et al. Children's Oncology Group's 2013 blueprint for research: neuroblastoma. *Pediatr Blood Cancer.* 2013;60(6):985-993.
2. Matthay KK, Maris JM, Schleiermacher G, et al. Neuroblastoma. *Nat Rev Dis Primers.* 2016;2:16078.
3. Grupp SA, Stern JW, Bunin N, et al. Tandem high-dose therapy in rapid sequence for children with high-risk neuroblastoma. *J Clin Oncol.* 2000;18(13):2567-2575.
4. Matthay KK, Reynolds CP, Seeger RC, et al. Long-term results for children with high-risk neuroblastoma treated on a randomized trial of myeloablative therapy followed by 13-cis-retinoic acid: a children's oncology group study. *J Clin Oncol.* 2009;27(7):1007-1013.
5. Pinto NR, Applebaum MA, Volchenboum SL, et al. Advances in Risk Classification and Treatment Strategies for Neuroblastoma. *J Clin Oncol.* 2015;33(27):3008-3017.
6. Matthay KK, Edeline V, Lumbroso J, et al. Correlation of early metastatic response by 123I-metaiodobenzylguanidine scintigraphy with overall response and event-free survival in stage IV neuroblastoma. *J Clin Oncol.* 2003;21(13):2486-2491.
7. Yanik GA, Parisi MT, Shulkin BL, et al. Semiquantitative mIBG scoring as a prognostic indicator in patients with stage 4 neuroblastoma: a report from the Children's oncology group. *J Nucl Med.* 2013;54(4):541-548.
8. Decarolis B, Schneider C, Hero B, et al. Iodine-123 metaiodobenzylguanidine scintigraphy scoring allows prediction of outcome in patients with stage 4 neuroblastoma: results of the Cologne interscore comparison study. *J Clin Oncol.* 2013;31(7):944-951.
9. Yanik GA, Parisi MT, Naranjo A, et al. Validation of Postinduction Curie Scores in High-Risk Neuroblastoma: A Children's Oncology Group and SIOPEN Group Report on SIOPEN/HR-NBL1. *J Nucl Med.* 2018;59(3):502-508.
10. Ady N, Zucker JM, Asselain B, et al. A new 123I-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(2):256-261.
11. Matthay KK, Shulkin B, Ladenstein R, et al. Criteria for evaluation of disease extent by (123)I-metaiodobenzylguanidine scans in neuroblastoma: a report for the International Neuroblastoma Risk Group (INRG) Task Force. *Br J Cancer.* 2010;102(9):1319-1326.
12. Katzenstein HM, Cohn SL, Shore RM, et al. Scintigraphic response by 123I-metaiodobenzylguanidine scan correlates with event-free survival in high-risk neuroblastoma. *J Clin Oncol.* 2004;22(19):3909-3915.
13. Yanik GA, Parisi MT, Naranjo A, et al. Validation of post-induction Curie scores in high risk neuroblastoma. A Children's Oncology Group (COG) and SIOPEN group report on SIOPEN/HR-NBL1. *J Nucl Med.* 2017.
14. Park JR, Bagatell R, Cohn SL, et al. Revisions to the International Neuroblastoma Response Criteria: A Consensus Statement From the National Cancer Institute Clinical Trials Planning Meeting. *J Clin Oncol.* 2017;35(22):2580-2587.
15. Starkey A SW, Armato SG III. Abras: A portable application for observer studies and visualization. *International Journal of Computer Assisted Radiology and Surgery.* 2011;6(suppl. 1):S193-S195.
16. Shiraishi J, Li Q, Appelbaum D, Pu Y, Doi K. Development of a computer-aided diagnostic scheme for detection of interval changes in successive whole-body bone scans. *Med Phys.* 2007;34(1):25-36.
17. Shiraishi J, Appelbaum D, Pu Y, Engelmann R, Li Q, Doi K. Clinical utility of temporal subtraction images in successive whole-body bone scans: evaluation in a prospective clinical study. *J Digit Imaging.* 2011;24(4):680-687.



Testing

Case 1 / 1

Manual Scoring

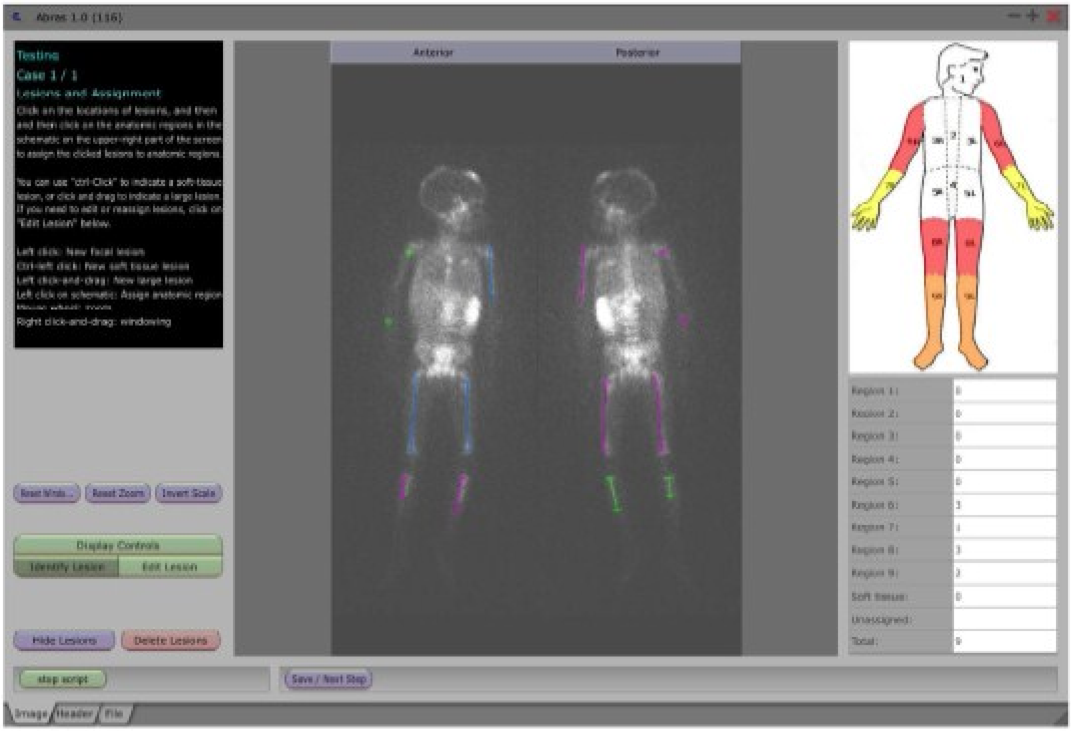
For each of the 10 anatomic regions (including soft-tissue), choose a score between 0 and 3.

When you are done, press the 'Next Step' button.

Mouse Wheel: zoom
Right click-and-drag: windowing

Region 1: 0 1 2 [✓ 3]
Region 2: 0 1 2 [✓ 3]
Region 3: 0 1 [✓ 2] 3
Region 4: 0 1 2 [✓ 3]
Region 5: 0 1 2 [✓ 3]
Region 6: 0 1 2 [✓ 3]
Region 7: 0 1 [✓ 2] 3
Region 8: 0 1 2 [✓ 3]
Region 9: 0 1 [✓ 2] 3
Soft tissue : 0 1 [✓ 2] 3
Total : 26

FIGURE 1. In the interface for Method A, the radiologist indicated a score of 0-3 for each of the 10 Curie regions by using the buttons in the lower right part of the screen. See Supplemental Figure S2 for text in figure.



Testing

Case 1 / 1

Lesions and Assignment

Click on the locations of lesions, and then and then click on the anatomic regions in the schematic on the upper-right part of the screen to assign the clicked lesions to anatomic regions.

You can use "ctrl-Click" to indicate a soft-tissue lesion, or click and drag to indicate a large lesion. If you need to edit or reassign lesions, click on "Edit Lesion" below.

Left click: New focal lesion
Ctrl-left click: New soft tissue lesion
Left click-and-drag: New large lesion
Left click on schematic: Assign anatomic region
Mouse wheel: zoom
Right click-and-drag: windowing

Region 1:	0
Region 2:	0
Region 3:	0
Region 4:	0
Region 5:	0
Region 6:	3
Region 7:	1
Region 8:	3
Region 9:	2
Soft tissue:	0
Unassigned:	
Total:	9

FIGURE 2. In Method B, the radiologist clicks on lesions in the anterior or posterior images and then indicates their corresponding anatomic segment by clicking on the schematic in the upper right portion of the screen. See Supplemental Figure S2 for text in figure.

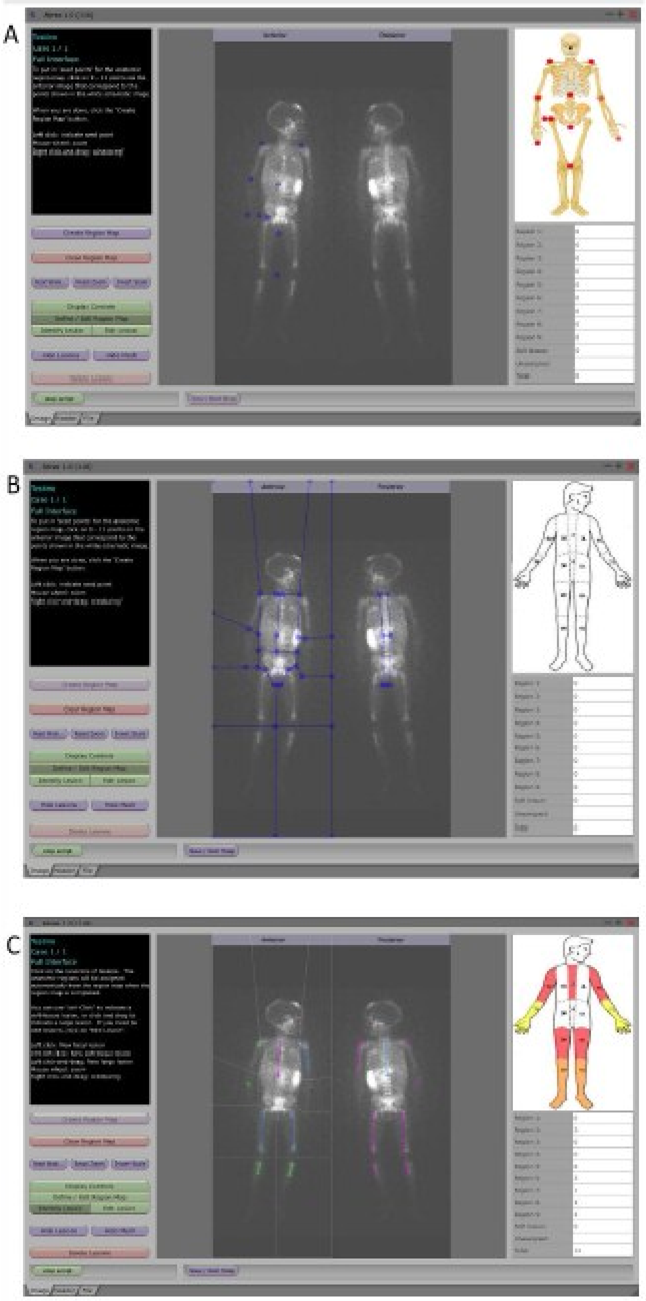


FIGURE 3. A) In Method C, the radiologist first indicates key points that will define the region map ("mesh"). The radiologist clicks on 9 - 11 points on the anterior image (in blue) that correspond to the points shown on the skeleton image in the upper right corner of the screen, with the left elbow and left fingertip points being optional. B) After the radiologist specifies the key points, the region map ("mesh") is created, which outlines the Curie anatomic segments for this case, and the radiologist can then adjust the region map if necessary by moving the circular handles between

each anatomic segment. C) After the mesh has been created, the radiologist can draw lesions (as points, lines, or loops), and the system automatically assigns anatomic segments and scores based on drawn lesion locations. See Supplemental Figure S2 for text in figure.

TABLE 1: Comparison of Scores Between Methods

	Method A vs Method B	Method A vs Method C	Method B vs Method C
Total	0.869	0.847	0.861
Region 1 Cranio-facial	0.887	0.877	0.909
Region 2 Cervical and Thoracic Spine	0.867	0.842	0.866
Region 3 Ribs/sternum/clavicles/scapula	0.776	0.774	0.775
Region 4 Lumbar and Sacral Spine	0.867	0.865	0.832
Region 5 Pelvis	0.887	0.873	0.896
Region 6 Upper Arms	0.865	0.858	0.803
Region 7 Lower Arms and Hands	0.678	0.618	0.737
Region 8 Femurs	0.951	0.917	0.935
Region 9 Lower Legs and Feet	0.908	0.888	0.932
Region 10 Soft Tissue	0.721	0.623	0.689

Cohen's kappa statistics are shown comparing total Curie scores and Curie scores for each region comparing each pair of Methods. All kappa scores are higher than 0.6 indicating very good across methods.

TABLE 2: Comparison of Scores Between Observers

	Method A	Method B	Method C
Total	0.840	0.811	0.804
Region 1 Cranio-facial	0.792	0.811	0.773
Region 2 Cervical and Thoracic Spine	0.851	0.834	0.838
Region 3 Ribs/sternum/clavicles/scapula	0.827	0.762	0.756

Region 4 Lumbar and Sacral Spine	0.773	0.789	0.787
Region 5 Pelvis	0.743	0.724	0.728
Region 6 Upper Arms	0.893	0.811	0.768
Region 7 Lower Arms and Hands	0.851	0.918	0.867
Region 8 Femurs	0.905	0.865	0.834
Region 9 Lower Legs and Feet	0.933	0.901	0.901
Region 10 Soft Tissue	0.834	0.699	0.783

Weighted Fleiss' kappa statistics are shown comparing total Curie scores and Curie scores for each region for each method. All kappa scores are higher than 0.6 indicating very good inter-observer reliability.

TABLE 1: Cohen's kappa statistics are shown comparing total Curie scores and Curie scores for each region comparing each pair of Methods. All kappa scores are higher than 0.6 indicating very good across methods.

TABLE 2: Weighted Fleiss' kappa statistics are shown comparing total Curie scores and Curie scores for each region for each method. All kappa scores are higher than 0.6 indicating very good inter-observer reliability.

SUPPLEMENTAL FIGURE S1:

Survey utilized to evaluate ease-of use, clinical utility, and potential adoption of the semi-automated methods.

SUPPLEMENTAL TABLE S1:

The ages at the time of scan and of disease diagnosis, disease stages, risk groups, *MYCN* status, and therapy received during the time of imaging are included for the 34 patients from the University of Chicago.

SUPPLEMENTAL TABLE S2:

Scores for each Curie region applied by each radiologist with each method.

SUPPLEMENTAL TABLE S3:

The mean time to score the scans for each scan using each method is shown. In all cases, Method A or B were the fastest method. Method C took longest in all but 5 cases.

SUPPLEMENTAL TABLE S4:

Survey responses to all survey questions as answered by all participating radiologists.

SUPPLEMENTAL FIGURE S2:

The text present in each figure is provided here so it can be read more clearly. These are the instructions provided to the radiologists as they learned each interface.