original report

Comparative Effectiveness of Tumor Response Assessment Methods: Standard of Care Versus Computer-Assisted Response Evaluation

Brian C. Allen
Edward Florez
Reza Sirous
Seth T. Lirette
Michael Griswold
Erick M. Remer
Zhen J. Wang
Jacob E. Bieszczad
Kelly L. Cox
Ajit H. Goenka
Candace M. Howard-Claudio
Hyunseon C. Kang
Sadhna B. Nandwana
Rupan Sanyal
Atul B. Shinagare
J. Clark Henegan
Judd Storrs
Matthew S. Davenport
Balaji Ganeshan
Amit Vasanji
Brian Rini
Andrew D. Smith

Author affiliations appear at the end of this article.

Corresponding author: Andrew D. Smith, MD, PhD, University of Alabama at Birmingham, Department of Radiology, 619 19th St South, Birmingham, AL 35249; e-mail: andrewdennissmith@uabmc.edu.

abstract

Purpose To compare the effectiveness of metastatic tumor response evaluation with computed tomography using computer-assisted versus manual methods.

Materials and Methods In this institutional review board–approved, Health Insurance Portability and Accountability Act–compliant retrospective study, 11 readers from 10 different institutions independently categorized tumor response according to three different therapeutic response criteria by using paired baseline and initial post-therapy computed tomography studies from 20 randomly selected patients with metastatic renal cell carcinoma who were treated with sunitinib as part of a completed phase III multi-institutional study. Images were evaluated with a manual tumor response evaluation method (standard of care) and with computer-assisted response evaluation (CARE) that included stepwise guidance, interactive error identification and correction methods, automated tumor metric extraction, calculations, response categorization, and data and image archiving. A crossover design, patient randomization, and 2-week washout period were used to reduce recall bias. Comparative effectiveness metrics included error rate and mean patient evaluation time.

Results The standard-of-care method, on average, was associated with one or more errors in 30.5% (6.1 of 20) of patients, whereas CARE had a 0.0% (0.0 of 20) error rate ($P < .001$). The most common errors were related to data transfer and arithmetic calculation. In patients with errors, the median number of error types was 1 (range, 1 to 3). Mean patient evaluation time with CARE was twice as fast as the standard-of-care method (6.4 minutes *v* 13.1 minutes; $P < .001$).

Conclusion CARE reduced errors and time of evaluation, which indicated better overall effectiveness than manual tumor response evaluation methods that are the current standard of care.

Clin Cancer Inform.

INTRODUCTION

In clinical trials and clinical practice, objective tumor response as evaluated on computed tomography (CT) images defines critical end points in patients with metastatic disease who are treated with systemic agents. Response Evaluation Criteria in Solid Tumors (RECIST; version 1.1) is based on tumor length measurements and is the most commonly used criteria with which to longitudinally assess metastatic tumor response for a wide variety of solid malignancies.[1,2] In the era of targeted therapy, metastatic tumor shrinkage has become less common or is delayed, which limits the utility of RECIST version 1.1.[3-7]

Many targeted agents are classified as antiangiogenic and cause tumor devascularization, which leads to changes in tumor size, attenuation, and morphology on CT images.[8,9] Choi criteria, which uses the percent change in tumor length and attenuation to derive objective response, was initially successful in predicting tumor response and survival in metastatic GI stromal tumors; however, less consistent results were observed when Choi criteria were applied to metastatic renal cell carcinoma (RCC) that was treated with antiangiogenic targeted agents.[8,9] MASS (Morphology, Attenuation, Size, and Structure) criteria, which accounts for objective changes in tumor length and attenuation and subjective development of tumor necrosis, was subsequently developed and found to be predictive of progression-free survival in patients with metastatic RCC who were treated with antiangiogenic targeted agents.[9]

Computer-assisted detection, picture archiving and communications systems, voice recognition systems, and electronic medical records have been designed to provide guidance, automate tasks, reduce errors, and improve efficiency and documentation in diagnostic imaging.[10,11] Despite advances in tumor segmentation methods and data management, the current standard of care for assessing metastatic tumor response is to manually apply tumor response criteria to derive objective response and manually document the findings.[12-15] These manual methods are prone to human errors in target lesion selection, target lesion measurement, data transfer, data calculations, response categorization, and archiving of data and annotated images.[16-19] Furthermore, these manual methods are potentially inefficient and become particularly complex when evaluating images by multiple imaging criteria, which is becoming more common in oncologic clinical trials.[20,21]

A computer-assisted response evaluation (CARE) system has been developed to guide readers through and to automate many of the steps in tumor response assessment with the goal of reducing errors and improving efficiency and documentation.[22] In this study, we compared the effectiveness of metastatic tumor response evaluation in patients with RCC using standard of care versus a CARE method.

MATERIALS AND METHODS

Study Design

Informed consent was waived in this institutional review board–approved, Health Insurance Portability and Accountability Act–compliant, retrospective, multi-institutional comparative-effectiveness observational study.

Participants

An existing imaging data set from a multinational, multi-institutional, prospective phase III trial of adult patients with metastatic clear-cell RCC who were treated with sunitinib or interferon alfa was used.[23] Three hundred seventy-five participants were included in the sunitinib arm. CT images from this study were prospectively archived for central independent review. Patients with unavailable baseline imaging or initial post-therapy imaging (n = 61) and those with non-digitized images (n = 39) were excluded, which left 275 participants available for analysis. For our study, images from 20 of 275 participants were randomly selected for additional evaluation. A power analysis for 11 readers with a cluster correlation of 50% and expected patient-level error percentage of 20% for standard of care and 2% for CARE yielded greater than 90% power to detect differences between the two methods.

Image Acquisition

Baseline CT (CT0) and initial follow-up CT (CT1) imaging of the chest, abdomen, and pelvis were performed per routine institutional acquisition parameters with slice thickness at $\leqslant$ 5 mm and administration of intravenous iodinated contrast unless contraindicated. Other CT acquisition and contrast material injection parameters were not specified, although CT manufacturer, number of detector rows, and tube voltage were obtained from images (Table 1). Images were deidentified with a unique coded identifier that was applied to each CT examination.

Readers

Eleven readers from 10 institutions participated in this study. Readers had a median of 5 years of post-training clinical experience (range, 2 to 23 years) and were practicing American Board of Radiology–certified academic radiologists with fellowship training in body imaging. Readers were provided with a written overview of RECIST version 1.1 and Choi and MASS criteria as well as relevant literature concerning the various tumor response criteria they were asked to review.[1,9,24]

Standard-of-Care Method

Readers were provided remote access to a standard image viewer (Philips iSite Enterprise, version 3.6.150.0 test environment; Philips, Andover, MA). Readers used coded identification numbers to locate and view CT0 and CT1 images. A Web-based Qualtrics survey platform (Provo, UT) was used to record data on target lesions and was designed to mimic the use of electronic data capture devices that are used in modern oncologic clinical trials. Target lesion data entered by each reader included the type of target lesion (primary tumor, metastasis, or lymph node), target lesion length (centimeters), and mean attenuation (Hounsfield units), calculated percent change in length and mean attenuation, presence or absence of marked decreased attenuation or marked central necrosis per MASS criteria, nontarget lesion response, and objective response per RECIST version 1.1 and Choi and MASS criteria. All readers reviewed a 30-minute instructional video that detailed the approach to the standard-of-care method, use of the Qualtrics survey to record results, and a requirement to archive all images that were annotated during review. Qualtrics was used to time the tumor response assessment for

Table 1. Patient, Tumor, and Imaging Characteristics

Variable	Whole Cohort (N = 275)	Interobserver Cohort (n = 20)
Patient characteristic		
Demographics		
Median age (Q1, Q2), years	62 (53, 68)	64 (57.5, 71.5)
Male sex, No. (%)	205 (75)	16 (80)
MSKCC risk groups, No. (%)		
Favorable risk	185 (67)	11 (55)
Intermediate risk	85 (31)	9 (45)
Poor risk	6 (2)	0 (0)
Survival, median (95% CI)		
PFS, years	0.87 (0.35 to 1.45)	0.93 (0.63 to 1.58)
OS, years	2.23 (1.03 to 3.21)	2.59 (1.05 to 3.20)
Tumor characteristic		
RCC pathologic type, No. (%)		
Clear cell only	241 (88)	20 (100)
Mix with clear cell component	34 (12)	0 (0)
Target lesion locations, No. (%)		
Lung	132 (48)	7 (35)
Mediastinal/hilar lymph nodes	107 (39)	6 (30)
Adrenal glands	64 (23)	1 (5)
Liver	49 (18)	3 (15)
Bones	42 (15)	4 (20)
Retroperitoneal lymph nodes	26 (9)	2 (10)
Other organ site	23 (8)	5 (25)
Other lymph nodes	4 (1)	0 (0)
Percent change in tumor metrics, median (Q1, Q3)		
Length, cm (N = 275)	−10.44 (−19.17, −1.15)	−6.33 (−14.55, −2.50)
Mean attenuation, HU (n = 229 contrast enhanced)	−32.06 (−60.36, 0.06)	−35.91 (−59.68, −16.93)
Imaging characteristic		
CT manufacturer, No. (%)		
GE (Milwaukee, WI)	141 (51)	14 (70)
Siemens (Forchheim, Germany)	110 (40)	5 (25)
Philips (Amsterdam, the Netherlands)	22 (8)	1 (5)
Toshiba (Tokyo, Japan)	2 (1)	0 (0)
CT scanner detector rows, No. (%)		
4	186 (68)	16 (80)
8	17 (6)	1 (5)
16	69 (25)	3 (15)
64	3 (1)	0 (0)
Peak kilovoltage, No. (%)		
120	238 (87)	18 (90)
130	30 (11)	2 (10)
140	7 (2)	0 (0)

(continued on following page)

Table 1. Patient, Tumor, and Imaging Characteristics (continued)

Variable	Whole Cohort (N = 275)	Interobserver Cohort (n = 20)
Time of CT scans relative to start of therapy, median (Q1, Q3)		
Baseline (CT0), days	−11 (−15, −7)	−8.5 (−12.5, −6.5)
Initial post-therapy (CT1), days	28 (26, 29)	27.5 (27, 29.5)
Contrast enhancement, No. (%)		
CT0 and CT1 contrast enhanced	229 (83)	18 (90)
CT0 and CT1 nonenhanced	21 (8)	2 (10)
CT0 nonenhanced, CT1 contrast enhanced	10 (4)	0 (0)
CT0 contrast enhanced, CT1 nonenhanced	15 (5)	0 (0)
Phase of IV contrast (n = 229 contrast enhanced), No. (%)		
Same contrast phase for CT0 and CT1	220 (96)	20 (100)
CT0 contrast phase earlier than CT1	5 (2)	0 (0)
CT0 contrast phase later than CT1	4 (2)	0 (0)

Abbreviations: CT, computed tomography; CT0, baseline CT; CT1, initial post-therapy CT; HU, Hounsfield units; IV, intravenous; MSKCC, Memorial Sloan Kettering Cancer Center; OS, overall survival; PFS, progression-free survival; Q, quartile; RCC, renal cell carcinoma.

each patient interpretation, from the loading of the survey to export of the data.

CARE Method

A custom image-viewing and semiautomated advanced postprocessing software platform (eMASS) was developed by A.D.S. and iteratively improved upon in collaboration with software engineers from ImageIQ.[22] In brief, eMASS software was designed to facilitate CARE of target lesions on baseline and initial post-therapy imaging studies by providing stepwise guidance of the required measurements and observations, interactive error identification and correction methods, and automated tumor metric extraction, calculations, response categorization, and data/image archiving. The software identifies common errors in tumor response assessment and implements corrective measures as detailed in Table 2.[16-19] Efficiency is improved via automation of multiple steps, including the simultaneous extraction of bidimensional tumor length and mean attenuation from a manual tumor segmentation process in which free-form regions of interest are drawn around the periphery of target lesions by the reviewer. The software automatically archives all data, annotates images, and instantly generates a summary output display for the reader to review (Fig 1).

All readers reviewed a 30-minute instructional video that detailed the use of the CARE method. A password-restricted encrypted data sharing platform (Dropbox, San Francisco, CA) was used to share eMASS software (version 1.0.24; eMASS LLC, Birmingham, AL) and deidentified images. Readers used their local personal computer running Windows 7 or higher (Microsoft, Redmond, WA) to download the stand-alone software package. The software automatically timed the tumor response assessment for each patient interpretation from the time of image loading until data export. Exported data are automatically stored in a comma-separated value format database. Exported images are automatically stored in Digital Imaging and Communications in Medicine and Portable Network Graphics formats in the main directory of the software.

Reading Sessions

All patients were evaluated once by using the standard-of-care method and twice by using the CARE method, with differences in how target lesions were identified. A crossover design, patient random assignment, and 2-week washout period were used to reduce recall bias (Appendix, online only). In brief, a total of 11 readers participated. Two reader pools (1 and 2) and three reading sessions (A, B, and C) were used. For reading sessions A and B, readers were provided with CT0 and CT1 images and independently picked their own target lesions as they evaluated images. Reading session C used the CARE method only, was designed to eliminate interobserver variability as a result of target lesion selection, and used a set

Table 2. Computer-Assisted Solutions to Common Sources of Error in Tumor Response Assessment

Common Errors in Tumor Response Assessment	Computer-Assisted Response Evaluation Solution
Error in target lesion selection per RECIST version 1.1	
Selection of more than five total target lesions	Assign all target lesions a number and prohibit the selection of more than five target lesions.
Selection of more than two target lesions per organ system	Require assignment of all target lesions to an organ system by using a dropdown list, provide user with a warning when selection of more than two target lesions per organ system is made, and prevent additional analysis until error is corrected.
Selection of a metastasis measuring < 1.0 cm in long axis	Require labeling of all target lesions as primary mass, metastasis, or lymph node. Notify reader of an error if primary mass or metastasis measures < 1.0 cm in long axis or if lymph node measures < 1.5 cm in short axis, and prevent additional analysis until error is corrected.
Selection of a lymph node measuring < 1.5 cm in short axis	
Error in target lesion measurement	
Metastasis measured in short axis	Eliminate manual data transfer. The user places a free-form ROI around the peripheral margin of the tumor on the axial image where the target lesion is the largest (per RECIST version 1.1 guidelines). The long axis length is automatically derived and archived in a database for all target lesions labeled as a primary mass or metastasis. The short axis length is automatically derived and archived in a database for all target lesions labeled as lymph nodes.
Lymph node measured in long axis	
Error in data transfer	
> 0.5 cm in length between annotated image and database	Eliminate manual data transfer by automatically extracting and archiving all tumor metric data.
> 5 HU difference between annotated image and database	
Error in application of MDA* per MASS criteria	
Failure to identify MDA despite decreased attenuation ⩾ 40 HU	Eliminate manual detection of MDA by automatically calculating the absolute change in attenuation of all target lesions and automatically recording MDA as present if a target lesion decreases by ⩾ 40 HU. Note that the user is required to enter information on the presence or absence of intravenous contrast before selecting any target lesions, and MDA is not assessed if either study is not contrast enhanced.
MDA applied to a lung target lesion	Prevent assessment for MDA for target lesions with organ system labeled as lung.
Error in calculation on the basis of tumor measurements	
Incorrect calculation of percent change in size	Eliminate manual calculations by automatically calculating percent change in size, attenuation, and other tumor metrics.
Incorrect calculation of percent change in tumor attenuation	

(continued on following page)

Table 2. Computer-Assisted Solutions to Common Sources of Error in Tumor Response Assessment (continued)

Common Errors in Tumor Response Assessment	Computer-Assisted Response Evaluation Solution
Error in objective response categorization on the basis of calculations	
RECIST version 1.1 categorization error	Eliminate manual objective response categorization. During the computer-assisted tumor response assessment process, the user is required to enter information on nontarget lesion response, presence or absence of one or more new metastases, and presence or absence of marked central necrosis (for MASS criteria). The software automatically calculates the percent change in all tumor metrics and combines this information with the above to automatically derive objective response per RECIST version 1.1, Choi criteria, MASS criteria, and a number of other tumor response criteria.
Choi criteria categorization error	
MASS criteria categorization error	
Error in image archiving	
Failure to store annotated images	Eliminate the need for manual image storage by automatically storing annotated images.

Abbreviations: HU, Hounsfield units; MASS, Morphology, Attenuation, Size, and Structure; MDA, marked decreased attenuation; ROI, region of interest.

*MDA is defined as > 40-HU decrease in the mean attenuation of a target lesions compared with baseline. According to MASS criteria, MDA should not be applied to lung target lesions.

arrangement of target lesion images that were provided to all readers.

For reading session A, reader pool 1 evaluated participants 1 to 10 by using the standard-of-care method, followed by participants 11 to 20 by using the CARE method. Reader pool 2 evaluated participants 11 to 20 by using the standard-of-care method, followed by participants 1 to 10 by using the CARE method.

To reduce recall bias, readers waited at least 14 days before beginning session B. For reading session B, reader pool 1 evaluated participants 1 to 10 in a randomized order by using the CARE method, followed by participants 11 to 20 in a randomized order by using the standard-of-care method. Reader pool 2 evaluated participants 11 to 20 in a randomized order by using the CARE method, followed by participants 1 to 10 in a randomized order by using the standard-of-care method (Appendix Fig A1, online only).

Reading session C was designed to eliminate interobserver variability as a result of target lesion selection. Target lesions were identified by A.D.S., and single Digital Imaging and Communications in Medicine format images with each target lesion from CT0 and CT1 were provided to readers. All readers used an imaging atlas—digital slides that contained pictures of each target lesion with an arrow pointing to the target—and evaluated participants 1 to 20 by using the CARE method. For the purposes of this article, CARE1 indicates that the readers picked their own target lesions, and CARE2 indicates that the target lesions were preselected for the readers.

Tumor Response Error Rate and Type Analysis

Two imaging research fellows—E.F. and R.S.—independently used the output data from the Qualtrics platform and the output file from the CARE method, along with the archived annotated images, to manually identify and record the presence or absence of all common tumor response error types identified in Table 2. Only a single error event per patient was recorded for each error type, even if multiple errors of the same type occurred in the same patient. A third reviewer—B.C.A., a reader from the study—evaluated all potential errors that were identified by E.F. and R.S. In the case of any discrepancy between the three reviewers, a panel that consisted of five readers from the study evaluated the potential errors by consensus.

Post-study Survey

A post-study Qualtrics survey that was unrelated to the data capture mechanism and designed to assess reader opinion about the CARE method was completed by all readers (N = 11) before an analysis of the study data and results.

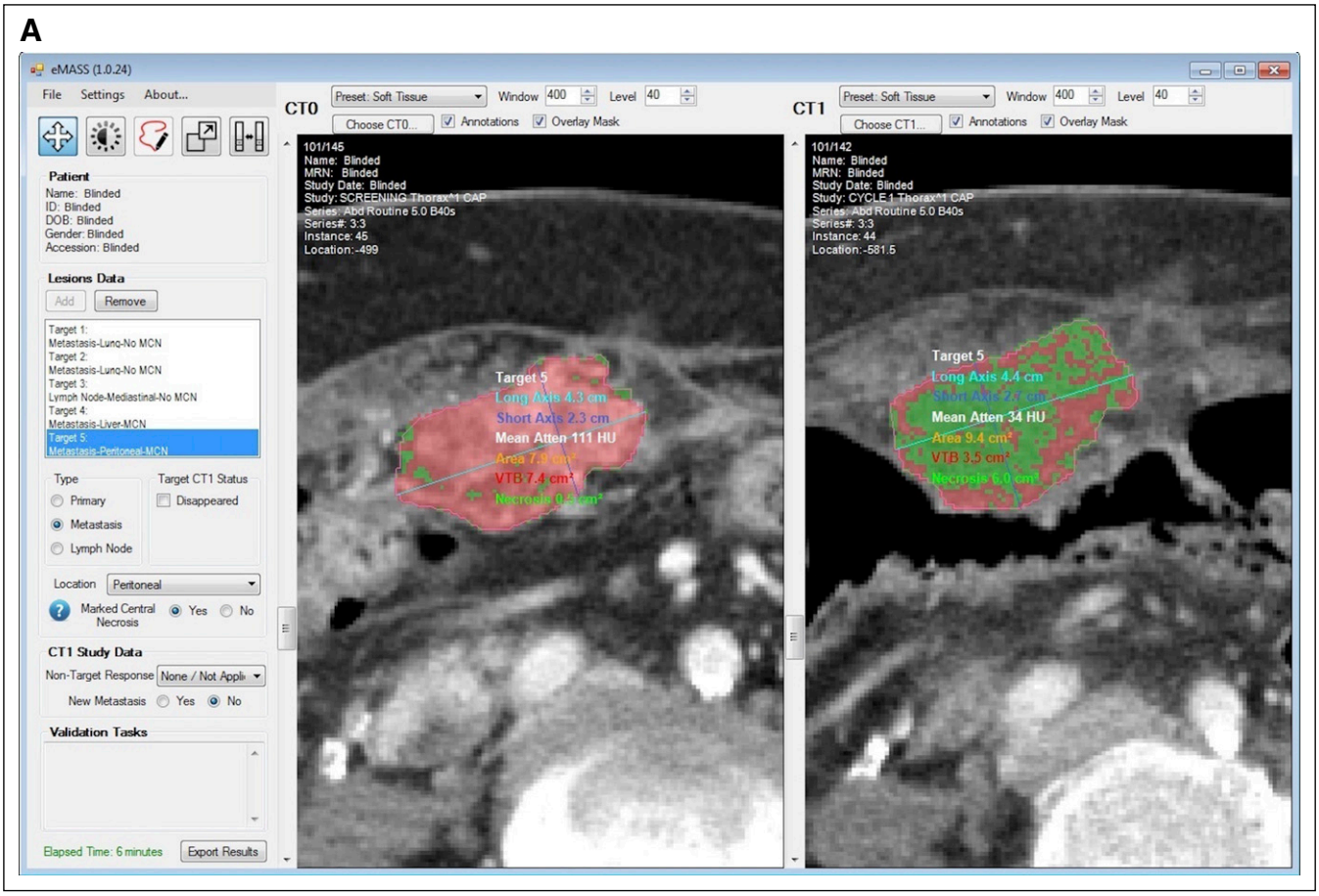

Fig 1. (A) Screenshot of eMASS software viewer and (B) summary output display. (A) The software viewer allows the reader to simultaneously view the baseline computed tomography (CT) images (CT0; on left) and initial post-therapy CT images (CT1; right) and make measurements. The target lesion in the peritoneum was selected by using a free-form region of interest, and the tumor metrics are automatically derived and displayed. The red color highlights the vascular tumor burden (VTB; enhancing tumor), and the green color indicates tumor necrosis (nonenhancing tumor). Left panel facilitates computer-assisted response evaluation by guiding the user through the tumor assessment process. As the software identifies errors in tumor response

Statistical Analysis

Patient, imaging, and measurement characteristics were summarized as median and interquartile range for continuous variables and count (percentage) for categorical variables for both the 275 patients who satisfied the inclusion criteria and the subset of 20 patients used in primary results.

Side-by-side boxplots were constructed to compare interpretation time between standard-of-care and CARE1 methods. Scatter plots were constructed to visualize the distributions of the percent change in tumor length and mean attenuation in each patient by using the standard-of-care, CARE1, and CARE2 methods.

Average percent of errors per patient was calculated by using the marginal effect from a multilevel mixed model, clustered by reader. Multiplying this by 20—the number of patients read by each reader—yields the expected number of errors per reader. Comparisons between the standard-of-care and CARE1 methods were drawn from the fixed effect of this model. Agreement across readers was assessed via intraclass correlation coefficients (ICCs) computed from two-way random effects models. ICC values were characterized as poor (0.00 to < 0.25), moderate (0.25 to < 0.50), good (0.50 to < 0.75), and very good (0.75 to 1.00). All analyses were performed with STATA (version 14.1; STATA, College Station, TX; Computing Resource Center, Santa Monica, CA).

RESULTS

Patient, Tumor, and Imaging Characteristics

The study cohort (n = 20) was randomly selected from the full cohort (N = 275) and had similar baseline characteristics (Table 1).

Tumor Response Error Rate and Type

The standard-of-care method was associated with errors types in all categories, including target lesion selection, target lesion measurement, data transfer, application of MASS criteria, tumor metric calculations, objective response categorization, and annotated image archiving (Table 3). The most common errors were related to data transfer and calculation.

The standard-of-care method, on average, was associated with one or more errors in 30.5% (6.1 of

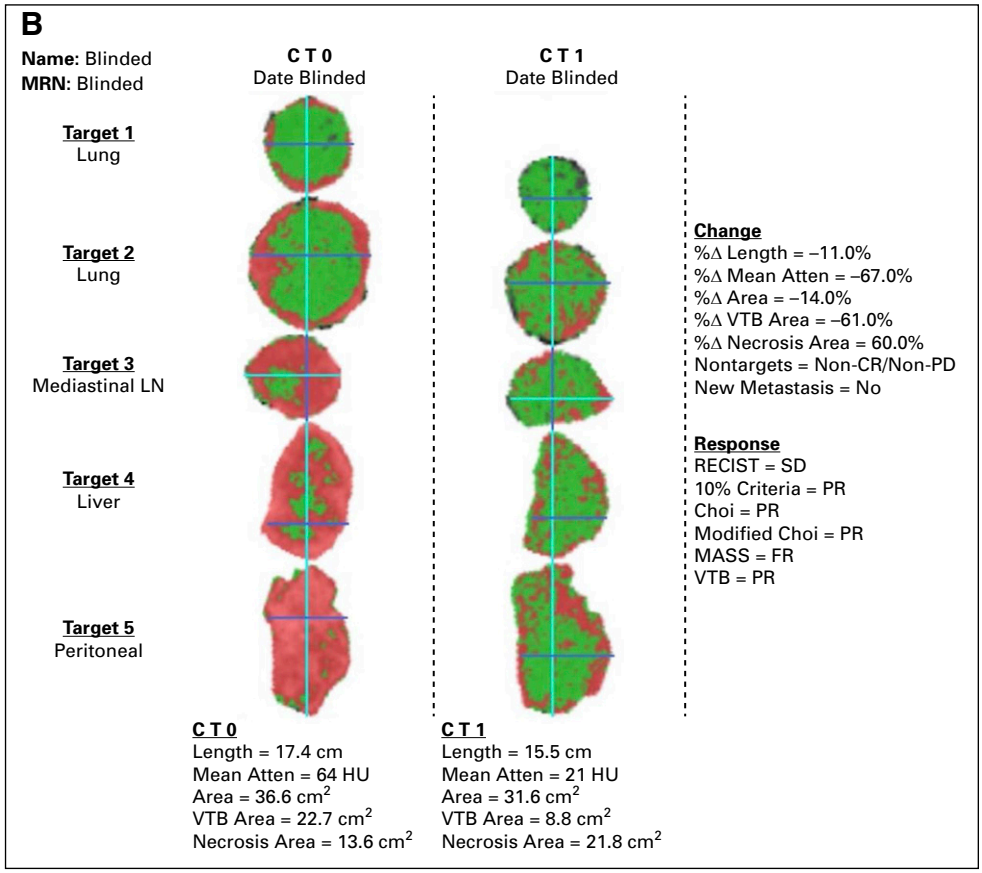

Fig 1. (Continued). assessment, a dialogue box in the side panel indicates the required corrective action and the user must correct the error before moving to the next task. The software also automates a number of features to reduce errors and improve efficiency. After exporting the results, a customizable summary output (B) is automatically generated and displayed in < 1 second. Target lesions are displayed in scaled vertical columns to visually display the tumor burden. Tumor metrics and objective response per multiple criteria are automatically derived and provided in the output display. Atten, attenuation; CR, complete response; FR, favorable response; HU, Hounsfield units; LN, lymph node; MASS, Morphology, Attenuation, Size, and Structure; MRN, medical record number; PD, progressive disease; PR, partial response; SD, stable disease.

20) of patients, whereas the CARE method had a 0.0% (0.0 of 20) error rate ($P < .001$). The error rate for the standard-of-care method was similar between reading sessions 1 and 2 (Appendix Table A1). The median error rate for 11 readers who interpreted 20 cases using the standard-of-care method was 25% (range, 15% to 55%). In patients with errors, the median number of error types was 1 (range, 1 to 3). When considering total errors per response criteria using the standard-of-care method, errors were more common when applying Choi criteria (24.5% [4.9 of 20]) and MASS criteria (23.0% [4.6 of 20]) compared with RECIST version 1.1 (11.0% [2.2 of 20]; $P < .001$ for both).

Interpretation Time

Mean patient interpretation time was twice as fast when using the CARE compared with the standard-of-care method: 6.4 minutes (95% CI, 6.11 to 6.66) *v* 13.1 minutes (95% CI, 12.5 to 13.7), respectively ($P < .001$; Appendix Fig A2, online only).

Interobserver Agreement

Interobserver agreement results are summarized in Table 4. There was very good interobserver agreement for measuring length and mean attenuation when readers used the CARE method to evaluate preselected target lesions (CARE2 range of ICC, 0.95 to 0.98), which resulted in a higher level of agreement than when readers chose their own target lesions (standard-of-care range of ICC, 0.59 to 0.89; CARE1 range of ICC, 0.69 to 0.88). Similarly, improved interobserver agreement for measuring the percent change in length was observed when readers used the CARE method to evaluate preselected target lesions (CARE2 ICC, 0.89), which resulted in a higher level of agreement than when readers chose their own target lesions (standard-of-care ICC, 0.79; CARE1 ICC, 0.59); however, when using either reader-selected or preselected target lesions, there was poor agreement for the percent change in mean attenuation (range of ICC, 0.00 to 0.06), which is used in Choi criteria. Distributions of the percent change in tumor length and mean attenuation for each patient according to tumor response evaluation method are depicted graphically in Fig 2.

Post-study Survey

The post-study survey indicated that 100% of readers (11 of 11) strongly preferred the CARE method and found CARE to be much easier to use than the

Table 3. Errors in Tumor Response Assessment According to Evaluation Method

Errors in Tumor Response Assessment	Standard of Care	CARE	*P*
Error in target lesion selection per RECIST version 1.1			
Selection of more than five total target lesions	0.0% (0.0/20)	0.0% (0.0/20)	—
Selection of more than two target lesions per organ system	1.5% (0.3/20)	0.0% (0.0/20)	.074
Selection of a metastasis measuring < 1.0 cm in long axis	0.0% (0.0/20)	0.0% (0.0/20)	—
Selection of a lymph node measuring < 1.5 cm in short axis	2.0% (0.4/20)	0.0% (0.0/20)	.043
Error in orientation of target lesion measurement			
Metastasis measured in short axis	0.0% (0.0/20)	0.0% (0.0/20)	—
Lymph node measured in long axis	2.0% (0.4/20)	0.0% (0.0/20)	.035
Error in data transfer			
> 0.5 cm in length between annotated image and database	1.5% (0.3/20)	0.0% (0.0/20)	.079
> 5 HU difference between annotated image and database	6.5% (1.3/20)	0.0% (0.0/20)	< .001
Error in application of MDA* per MASS criteria			
Failure to identify MDA despite decreased attenuation ⩾ 40 HU	3.0% (0.6/20)	0.0% (0.0/20)	.007
MDA applied to a lung target lesion	2.5% (0.5/20)	0.0% (0.0/20)	.013
Error in calculation on the basis of tumor measurements			
Incorrect calculation of percent change in size	2.0% (0.4/20)	0.0% (0.0/20)	.043
Incorrect calculation of percent change in tumor attenuation	6.0% (1.2/20)	0.0% (0.0/20)	< .001
Error in objective response categorization on the basis of calculations			
RECIST version 1.1 categorization error	1.0% (0.2/20)	0.0% (0.0/20)	.155
Choi criteria categorization error	3.0% (0.6/20)	0.0% (0.0/20)	.007
MASS criteria categorization error	5.0% (1.0/20)	0.0% (0.0/20)	.001
Error in image archiving			
Failure to store annotated images	2.5% (0.5/20)	0.0% (0.0/20)	.012
Total errors per response criteria			
RECIST version 1.1 errors	11.0% (2.2/20)	0.0% (0.0/20)	< .001
Choi criteria errors	24.5% (4.9/20)	0.0% (0.0/20)	< .001
MASS criteria errors	23.0% (4.6/20)	0.0% (0.0/20)	< .001
Total errors for complete assessment by three response criteria	30.5% (6.1/20)	0.0% (0.0/20)	< .001

NOTE. Summary statistics are presented as the average percentage of errors per patient (average number of errors per reader/average number of patient assessments per reader), across 11 independent readers. Only a single error event per patient was recorded for each error type, even if multiple errors of the same type occurred in the same patient.

Abbreviations: CARE, computer-assisted response evaluation; HU, Hounsfield units; MASS, Morphology, Attenuation, Size, and Structure; MDA, marked decreased attenuation.

*MDA is defined as > 40-HU decrease in the mean attenuation of a target lesions compared with baseline. According to MASS criteria, MDA should not be applied to lung target lesions.

Table 4. Interobserver Agreement According to Response Evaluation Method

Descriptor	Standard of Care	CARE 1	CARE 2
Length			
Sum at CT0, cm	0.89 (0.81 to 0.94)	0.88 (0.80 to 0.94)	0.98 (0.97 to 0.99)
Sum at CT1, cm	0.89 (0.81 to 0.94)	0.88 (0.80 to 0.94)	0.98 (0.97 to 0.99)
Change, cm	0.75 (0.62 to 0.87)	0.60 (0.44 to 0.77)	0.90 (0.82 to 0.96)
Percent change	0.79 (0.67 to 0.89)	0.59 (0.43 to 0.76)	0.89 (0.81 to 0.95)
Attenuation			
Mean at CT0, HU	0.79 (0.70 to 0.90)	0.69 (0.53 to 0.84)	0.95 (0.90 to 0.98)
Mean at CT1, HU	0.59 (0.45 to 0.78)	0.80 (0.69 to 0.90)	0.97 (0.95 to 0.99)
Change, HU	0.47 (0.33 to 0.69)	0.38 (0.23 to 0.61)	0.85 (0.74 to 0.94)
Percent change	0.00 (−0.06 to 0.03)	0.06 (0.00 to 0.22)	0.00 (−0.04 to 0.14)

NOTE. Data are presented as intraclass correlation coefficient (95% CI). For the standard-of-care and CARE1 methods, readers independently picked all target lesions. For the CARE 2 method, target lesions were provided to all readers before evaluation. A comparison of CARE 1 versus CARE 2 methods allows for an assessment of the role of target lesion selection in tumor response assessment.
Abbreviations: CARE, computer-assisted response evaluation; CT0, baseline computed tomography exam; CT1, initial follow-up computed tomography exam; HU, Hounsfield units.

standard-of-care method. Most readers (73%; eight of 11) indicated that CARE would improve overall productivity in the clinical setting, with 18% (two of 11) undecided and 9% (one of 11) indicating that productivity would not be improved. Additional survey results are provided in the Appendix.

DISCUSSION

The current standard-of-care for assessing metastatic tumor response is to manually apply tumor response criteria to derive the objective response and manually document the results. In our study, the standard-of-care method was associated with a high error rate, despite the use of expert readers and training before reading sessions. Similar findings have been reported by using RECIST version 1.1 in clinical practice, but have not been confirmed in a multi-institutional study.[17] Past efforts to reduce reader errors and interobserver variability in tumor response assessment have included reader educational training and the use of central independent image review in oncologic clinical trials, which have mildly reduced errors and improved interobserver variability.[17,25-27]

In our study, the use of the CARE method with step-wise guidance and interactive error identification and correction methods eliminated reader errors and achieved good to very good interobserver agreement. This is a clinically important finding, as errors in tumor response evaluation may adversely impact treatment decisions. CARE was used to automate tumor metric extraction, calculations, response categorization, and data and image archiving, which resulted in markedly improved efficiency as the mean interpretation time was cut in half compared with the standard-of-care method.

We observed high inter-reader agreement for measuring multiple tumor metrics by using the CARE method, despite the use of 11 readers from 10 different institutions with varying experience and practice patterns with less than 1 hour of training with the CARE method. Two observations are noteworthy. First, interobserver agreement using either the standard-of-care or CARE methods for measuring mean attenuation was good, but was dismal for percent change in mean tumor attenuation, even when all readers measured the same preselected target lesions. This implies that the mathematical conversion to percent change in mean attenuation is responsible for the poor interobserver agreement, which is likely related to the fact that zero mean attenuation is an arbitrary—not absolute—definition and is not comparable to a length measurement of zero, which is an absolute value that indicates that no tumor is present. These findings suggest that the percent change in mean attenuation may not be reproducible in Choi criteria and modified Choi criteria. Second, interobserver agreement for measuring length with the CARE method was higher when target lesions were preselected for readers than when readers selected their own target lesions, which indicated that target lesion selection is a major contributor to interobserver variability.

This study had several limitations. First, the study was retrospective in design, although CT images were a representative sample from a landmark multinational, multi-institutional, phase III prospective study and the standard-of-care tumor response

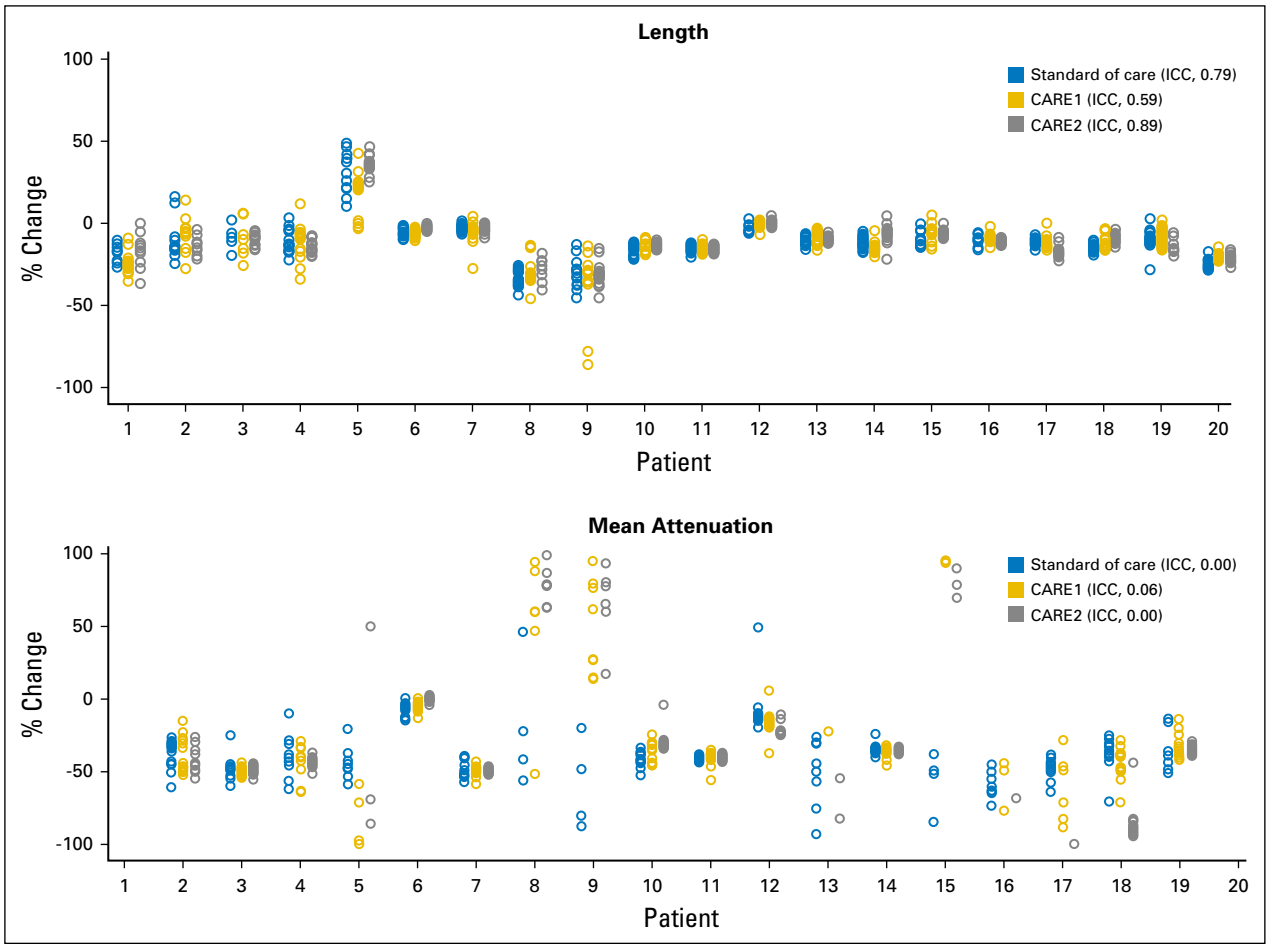

Fig 2. Scatter plots depicting the distributions of percent change in tumor length and mean attenuation for each patient according to tumor response evaluation method. Readers (N = 11) picked the target lesions for the standard-of-care and computer-assisted response evaluation (CARE1) methods, but images of preselected target lesions were used for CARE2. Interobserver agreement was good to very good for measuring percent change in tumor length. Patients 1 and 20 had nonenhanced computed tomography images, so mean attenuation was not measured in these patients. Interobserver agreement was poor for measuring percent change in mean attenuation by any of the methods tested. To facilitate a direct comparison, *y*-axis scales for each graph were restricted to the same range (−100% to +100%); however, for percent change in mean attenuation, there were 136 values outside of this range, with 13 having > 1000% change and three having > 5,000% change. ICC, intraclass correlation coefficient.

evaluation process was similar to methods that are used in prospective multi-institutional studies. Second, the high per-patient error rates in our study likely underestimate true error rates as readers only evaluated two imaging time points and did not perform longitudinal tumor assessments that are used in clinical trials and clinical practice and only one error type per patient was counted, despite the possibility of compounded errors—for example, a target lesion selection error that leads to measurement, calculation, and response categorization errors at the current time point or later time points. Third, the same set of patient images was evaluated twice by readers so that a direct comparison of two tumor response assessment methods could be made, potentially leading to recall bias. Efforts to reduce recall bias included the crossover study design, patient randomization and deidentification, and a 2-week washout period between reading sessions. Fourth, the study was not designed or powered to assess longitudinal differences in objective tumor response reclassification or changes in clinical management associated with use of the CARE method. Fifth, our study focused only on tumor assessments using CT images; however, the basic rule-setting structure is amenable to other imaging modalities.

In conclusion, the CARE method significantly reduces errors and time of evaluation while maintaining high interobserver agreement, which indicates better overall effectiveness than manual tumor response evaluation methods that are the current standard of care. As metastatic tumor response evaluation defines critical

end points in oncologic patient care, methods that reduce errors, reduce time of evaluation, and improve documentation while maintaining high interobserver agreement could lead to needed advancements in clinical trials and clinical care.

DOI: https://doi.org/10.1200/CCI.17.00026
Published online on ascopubs.org/journal/cci on August 30, 2017.

AUTHOR CONTRIBUTIONS

Conception and design: Brian C. Allen, Reza Sirous, Seth T. Lirette, Zhen J. Wang, Jacob E. Bieszczad, Atul B. Shinagare, Matthew S. Davenport, Brian Rini, Andrew D. Smith
Administrative support: Andrew D. Smith
Collection and assembly of data: Brian C. Allen, Edward Florez, Reza Sirous, Seth T. Lirette, Erick M. Remer, Zhen J. Wang, Jacob E. Bieszczad, Kelly L. Cox, Hyunseon C. Kang, Sadhna B. Nandwana, Rupan Sanyal, Atul B. Shinagare, Judd Storrs, Matthew S. Davenport, Brian Rini, Andrew D. Smith
Data analysis and interpretation: Brian C. Allen, Edward Florez, Reza Sirous, Michael Griswold, Erick M. Remer, Zhen J. Wang, Jacob E. Bieszczad, Kelly L. Cox, Ajit H. Goenka, Candace M. Howard-Claudio, Hyunseon C. Kang, Rupan Sanyal, J. Clark Henegan, Matthew S. Davenport, Balaji Ganeshan, Amit Vasanji, Brian Rini, Andrew D. Smith
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

Brian C. Allen
No relationship to disclose

Edward Florez
No relationship to disclose

Reza Sirous
No relationship to disclose

Seth T. Lirette
No relationship to disclose

Michael Griswold
No relationship to disclose

Erick M. Remer
No relationship to disclose

Zhen J. Wang
Stock and Other Ownership Interests: Nextrast
Patents, Royalties, Other Intellectual Property: Nextrast (I)

Jacob E. Bieszczad
No relationship to disclose

Kelly L. Cox
No relationship to disclose

Ajit H. Goenka
Stock and Other Ownership Interests: Bristol-Myers Squibb, Cellectar
Travel, Accommodations, Expenses: GE Healthcare

Candace M. Howard-Claudio
No relationship to disclose

Hyunseon C. Kang
Travel, Accommodations, Expenses: Siemens Healthineers

Sadhna B. Nandwana
No relationship to disclose

Rupan Sanyal
Honoraria: ABC Education

Atul B. Shinagare
Honoraria: Arog
Consulting or Advisory Role: Arog

J. Clark Henegan
No relationship to disclose

Judd Storrs
Consulting or Advisory Role: Absist
Patents, Royalties, Other Intellectual Property: Co-inventor US2011123078
Travel, Accommodations, Expenses: Absist

Matthew S. Davenport
No relationship to disclose

Balaji Ganeshan
Employment: TexRAD Ltd
Leadership: TexRAD Ltd, Feedback Plc, Stone Checker Software Ltd, Prostate Checker Ltd
Stock and Other Ownership Interests: Feedback Plc
Patents, Royalties, Other Intellectual Property: Inventor on a patent surrounding medical image texture analysis to quantify tissue/tumor heterogeneity
Travel, Accommodations, Expenses: TexRAD Ltd

Amit Vasanji
No relationship to disclose

Brian Rini
Consulting or Advisory Role: Pfizer
Research Funding: Pfizer (Inst), Genentech (Inst), Bristol-Myers Squibb
Travel, Accommodations, Expenses: Pfizer

Andrew D. Smith
Leadership: Radiostics, eMASS, Liver Nodularity, Color Enhanced Detection
Stock and Other Ownership Interests: Radiostics, eMASS, Liver Nodularity, Color Enhanced Detection
Research Funding: Pfizer (Inst)
Patents, Royalties, Other Intellectual Property: Patent pending "Computer-assisted tumor response assessment and evaluation of the vascular tumor burden," method for standardizing target lesion selection and tracking on medical images, method for standardization and color-enhanced detection for CT imaging, method for the detection and staging of liver fibrosis from image-acquired data

Affiliations

Brian C. Allen, Duke University Medical Center, Durham, NC; **Edward Florez**, **Reza Sirous**, **Seth T. Lirette**, **Michael Griswold**, **Candace M. Howard-Claudio**, **J. Clark Henegan**, **Judd Storrs**, and **Andrew D. Smith**, University of Mississippi Medical Center, Jackson, MS; **Erick M. Remer** and **Brian Rini**, The Cleveland Clinic; **Amit Vasanji**, ImageIQ, Cleveland; **Jacob E. Bieszczad**, University of Toledo Medical Center, Toledo, OH; **Zhen J. Wang**, University of California at San Francisco Medical Center, San Francisco, CA; **Kelly L. Cox** and **Sadhna B. Nandwana**, Emory University School of Medicine, Atlanta, GA; **Ajit H. Goenka**, The Mayo Clinic, Rochester, MN; **Hyunseon C. Kang**, University of Texas MD Anderson Cancer Center, Houston, TX; **Rupan Sanyal**, University of Alabama at Birmingham Medical Center, Birmingham, AL; **Atul B. Shinagare**, Dana-Farber Cancer Institute/Brigham and Women's Hospital, Harvard University, Boston, MA; **Matthew S. Davenport**, University of Michigan Health System, Ann Arbor, MI; and **Balaji Ganeshan**, University College of London, London, United Kingdom.

REFERENCES

1. Eisenhauer EA, Therasse P, Bogaerts J, et al: New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer 45:228-247, 2009
2. Therasse P, Arbuck SG, Eisenhauer EA, et al: New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst 92:205-216, 2000
3. Smith AD, Lieber ML, Shah SN: Assessing tumor response and detecting recurrence in metastatic renal cell carcinoma on targeted therapy: Importance of size and attenuation on contrast-enhanced CT. AJR Am J Roentgenol 194:157-165, 2010
4. Escudier B, Eisen T, Stadler WM, et al: Sorafenib in advanced clear-cell renal-cell carcinoma. N Engl J Med 356:125-134, 2007
5. Motzer RJ, Michaelson MD, Redman BG, et al: Activity of SU11248, a multitargeted inhibitor of vascular endothelial growth factor receptor and platelet-derived growth factor receptor, in patients with metastatic renal cell carcinoma. J Clin Oncol 24:16-24, 2006
6. Tuma RS: Sometimes size doesn't matter: Reevaluating RECIST and tumor response rate endpoints. J Natl Cancer Inst 98:1272-1274, 2006
7. Benjamin RS, Choi H, Macapinlac HA, et al: We should desist using RECIST, at least in GIST. J Clin Oncol 25:1760-1764, 2007
8. Choi H, Charnsangavej C, Faria SC, et al: Correlation of computed tomography and positron emission tomography in patients with metastatic gastrointestinal stromal tumor treated at a single institution with imatinib mesylate: Proposal of new computed tomography response criteria. J Clin Oncol 25:1753-1759, 2007
9. Smith AD, Shah SN, Rini BI, et al: Morphology, Attenuation, Size, and Structure (MASS) criteria: Assessing response and predicting clinical outcome in metastatic renal cell carcinoma on antiangiogenic targeted therapy. AJR Am J Roentgenol 194:1470-1478, 2010
10. Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al: Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 215:554-562, 2000
11. Rubin GD, Lyo JK, Paik DS, et al: Pulmonary nodules on multi-detector row CT scans: Performance comparison of radiologists and computer-aided detection. Radiology 234:274-283, 2005
12. Folio LR, Choi MM, Solomon JM, et al: Automated registration, segmentation, and measurement of metastatic melanoma tumors in serial CT scans. Acad Radiol 20:604-613, 2013
13. Folio LR, Sandouk A, Huang J, et al: Consistency and efficiency of CT analysis of metastatic disease: Semiautomated lesion management application within a PACS. AJR Am J Roentgenol 201:618-625, 2013
14. Folio LR, Turkbey EB, Steinberg SM, et al: Viable tumor volume: Volume of interest within segmented metastatic lesions, a pilot study of proposed computed tomography response criteria for urothelial cancer. Eur J Radiol 84:1708-1714, 2015
15. Goyal N, Apolo AB, Berman ED, et al: ENABLE (Exportable Notation and Bookmark List Engine): An interface to manage tumor measurement data from PACS to cancer databases. J Digit Imaging 30:275-286, 2017
16. Abramson RG, McGhee CR, Lakomkin N, et al: Pitfalls in RECIST data extraction for clinical trials: Beyond the basics. Acad Radiol 22:779-786, 2015
17. Andoh H, McNulty NJ, Lewis PJ: Improving accuracy in reporting CT scans of oncology patients: Assessing the effect of education and feedback interventions on the application of the Response Evaluation Criteria in Solid Tumors (RECIST) criteria. Acad Radiol 20:351-357, 2013
18. Kekelidze M: 10 Most frequently made mistakes with RECIST 1.1: How radiologist can fail–and how to avoid them. 2014 European Congress of Radiology, Vienna, Austria, March 6-10, 2014
19. Teslenko I, Belotserkovsky M: 717 Common pitfalls of RECIST 1.1 application in clinical trials. Eur J Cancer 51:S132, 2015

20. Nishino M, Jagannathan JP, Krajewski KM, et al: Personalized tumor response assessment in the era of molecular medicine: Cancer-specific and therapy-specific response criteria to complement pitfalls of RECIST. AJR Am J Roentgenol 198:737-745, 2012

21. Ronot M, Bouattour M, Wassermann J, et al: Alternative response criteria (Choi, European Association for the Study of the Liver, and modified Response Evaluation Criteria in Solid Tumors [RECIST]) versus RECIST 1.1 in patients with advanced hepatocellular carcinoma treated with sorafenib. Oncologist 19:394-402, 2014

22. Smith AD, Zhang X, Bryan J, et al: Vascular tumor burden as a new quantitative CT biomarker for predicting metastatic RCC response to antiangiogenic therapy. Radiology 281:484-498, 2016

23. Motzer RJ, Hutson TE, Tomczak P, et al: Sunitinib versus interferon alfa in metastatic renal-cell carcinoma. N Engl J Med 356:115-124, 2007

24. Brufau BP, Cerqueda CS, Villalba LB, et al: Metastatic renal cell carcinoma: Radiologic findings and assessment of response to targeted antiangiogenic therapy by using multidetector CT. Radiographics 33:1691-1716, 2013

25. Belton AL, Saini S, Liebermann K, et al: Tumour size measurement in an oncology clinical trial: Comparison between off-site and on-site measurements. Clin Radiol 58:311-314, 2003

26. Dodd LE, Korn EL, Freidlin B, et al: Blinded independent central review of progression-free survival in phase III clinical trials: Important design element or unnecessary expense? J Clin Oncol 26:3791-3796, 2008

27. Tang PA, Pond GR, Chen EX: Influence of an independent review committee on assessment of response rate and progression-free survival in phase III clinical trials. Ann Oncol 21:19-26, 2010

APPENDIX

Reading Sessions

All patients were evaluated once by using the standard-of-care method and twice by using the computer-assisted response evaluation (CARE) method, with differences in how target lesions were identified. A crossover design, patient random assignment, and 2-week washout period were used to reduce recall bias. In brief, a total of 11 readers participated. Two reader pools (1 and 2) and three reading sessions (A, B, and C) were used. For reading sessions A and B, readers were provided with images from baseline computed tomography exam and initial follow-up computed tomography exam, and readers independently picked their own target lesions as they evaluated the images. Reading session C used the CARE method only, was designed to eliminate interobserver variability as a result of target lesion selection, and used a set arrangement of target lesion images that were provided to all readers.

For reading session A, reader pool 1 evaluated participants 1 to 10 by using the standard-of-care method followed by participants 11 to 20 by using the CARE method. Reader pool 2 evaluated participants 11 to 20 by using the standard-of-care method followed by participants 1 to 10 by using the CARE method.

To reduce recall bias, readers waited at least 14 days before beginning session B. For reading session B, reader pool 1 evaluated participants 1 to 10 in a randomized order by using the CARE method followed by participants 11 to 20 in a randomized order by using the standard-of-care method. Reader pool 2 evaluated participants 11 to 20 in a randomized order by using the CARE method followed by participants 1 to 10 in a randomized order by using the standard-of-care method (Fig. A1).

Reading session C was designed to eliminate interobserver variability as a result of target lesion selection. Target lesions were identified by A.D.S. and single Digital Imaging and Communications in Medicine format images with each target lesion from baseline computed tomography exam and initial follow-up computed tomography exam were provided to readers. All readers used an imaging atlas—digital slides that contained pictures of each target lesion with an arrow pointing to the target—and evaluated participants 1 to 20 by using the CARE method. For the purposes of this article, CARE1 indicates that the readers picked their own target lesions, and CARE2 indicates that the target lesions were preselected for the readers.

Post-study Survey

The post-study survey indicated that 100% of readers (11 of 11) strongly preferred the CARE method and found CARE to be much easier to use than the standard-of-care method. Most readers (73%; eight of 11) indicated that CARE would improve overall productivity in the clinical setting, with 18% (two of 11) undecided and 9% (one of 11) indicating that productivity would not be improved. Most readers (82%; nine of 11) indicated that CARE would improve accuracy compared with the standard-of-care method, with 9% (one of 11) undecided, and 9% (one of 11) indicating that CARE would not improve accuracy. Most readers (82%; nine of 11) indicated that they would be very likely to use CARE in the future, with 9% (one of 11) indicating that they were somewhat likely and 9% (one of 11) undecided.

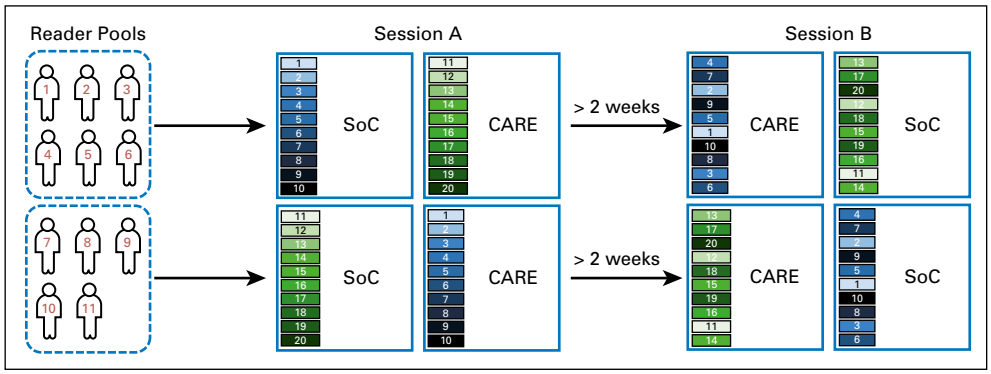

Fig A1. Reading session flowchart. CARE, computer-assisted response evaluation; SoC, standard of care.

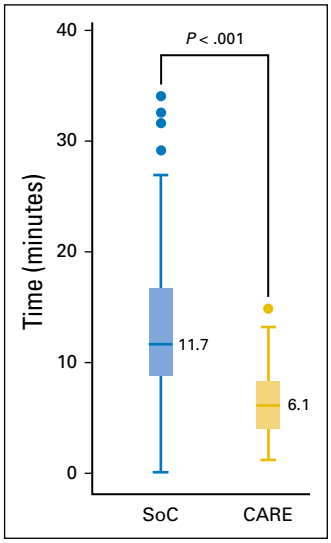

Fig A2. Box plots comparing patient tumor response evaluation time using the standard-of-care (SoC) method versus computer-assisted response evaluation (CARE) methods.

Table A1. Errors in Tumor Response Assessment According to Reading Session

Errors in Tumor Response Assessment	Standard of Care			CARE		
	Session 1	Session 2	*P*	Session 1	Session 2	*P*
Error in target lesion selection per RECIST version 1.1						
Selection of more than five total target lesions	0.0% (0.0/20)	0.0% (0.0/20)	—	0.0% (0.0/20)	0.0% (0.0/20)	—
Selection of more than two target lesions per organ system	2.5% (0.5/20)	0.0% (0.0/20)	.064	0.0% (0.0/20)	0.0% (0.0/20)	—
Selection of a metastasis measuring < 1.0 cm in long axis	0.0% (0.0/20)	0.0% (0.0/20)	—	0.0% (0.0/20)	0.0% (0.0/20)	—
Selection of a lymph node measuring < 1.5 cm in short axis	1.0% (0.2/20)	2.5% (0.5/20)	.312	0.0% (0.0/20)	0.0% (0.0/20)	—
Error in orientation of target lesion measurement						
Metastasis measured in short axis	0.0% (0.0/20)	0.0% (0.0/20)	—	0.0% (0.0/20)	0.0% (0.0/20)	—
Lymph node measured in long axis	0.0% (0.0/20)	3.0% (0.6/20)	.027	0.0% (0.0/20)	0.0% (0.0/20)	—
Error in data transfer						
> 0.5 cm in length between annotated image and database	0.0% (0.0/20)	2.5% (0.5/20)	.075	0.0% (0.0/20)	0.0% (0.0/20)	—
> 5 HU difference between annotated image and database	6.0% (1.2/20)	7.5% (1.5/20)	.786	0.0% (0.0/20)	0.0% (0.0/20)	—
Error in application of MDA* per MASS criteria						
Failure to identify MDA despite decreased attenuation ⩾ 40 HU	3.0% (0.6/20)	2.5% (0.5/20)	.700	0.0% (0.0/20)	0.0% (0.0/20)	—
MDA applied to a lung target lesion	2.0% (0.4/20)	3.0% (0.6/20)	.407	0.0% (0.0/20)	0.0% (0.0/20)	—
Error in calculation on the basis of tumor measurements						
Incorrect calculation of percent change in size	3.0% (0.6/20)	0.0% (0.0/20)	.041	0.0% (0.0/20)	0.0% (0.0/20)	—
Incorrect calculation of percent change in tumor attenuation	7.5% (1.5/20)	5.5% (1.1/20)	.579	0.0% (0.0/20)	0.0% (0.0/20)	—
Error in objective response categorization on the basis of calculations						
RECIST version 1.1 categorization error	2.0% (0.4/20)	0.0% (0.0/20)	.153	0.0% (0.0/20)	0.0% (0.0/20)	—
Choi criteria categorization error	3.0% (0.6/20)	2.5% (0.5/20)	.700	0.0% (0.0/20)	0.0% (0.0/20)	—
MASS criteria categorization error	4.0% (0.8/20)	5.5% (1.1/20)	.757	0.0% (0.0/20)	0.0% (0.0/20)	—
Error in image archiving						
Failure to store annotated images	4.0% (0.8/20)	1.0% (0.2/20)	.087	0.0% (0.0/20)	0.0% (0.0/20)	—
Total errors per response criteria						
RECIST version 1.1 errors	12.5% (2.5/20)	10.0% (2.0/20)	.509	0.0% (0.0/20)	0.0% (0.0/20)	—
Choi criteria errors	26.5% (5.3/20)	22.5% (4.5/20)	.527	0.0% (0.0/20)	0.0% (0.0/20)	—
MASS criteria errors	23.0% (4.6/20)	23.0% (4.6/20)	> .999	0.0% (0.0/20)	0.0% (0.0/20)	—
Total errors for complete assessment by three response criteria	32.5% (6.5/20)	29.0% (5.8/20)	.555	0.0% (0.0/20)	0.0% (0.0/20)	—

Abbreviations: CARE, computer-assisted response evaluation; HU, Hounsfield units; MDA, marked decreased attenuation; MASS criteria, Morphology, Attenuation, Size, and Structure criteria.

*MDA is defined as > 40 HU decrease in the mean attenuation of a target lesions compared with baseline. According to MASS criteria, MDA should not be applied to lung target lesions.