Intraobserver Variability in Bladder Cancer Treatment Response Assessment With and Without Computerized Decision Support

Lubomir M. Hadjiiski[1], Kenny H. Cha[1], Richard H. Cohan[1], Heang-Ping Chan[1], Elaine M. Caoili[1], Matthew S. Davenport[1,2], Ravi K. Samala[1], Alon Z. Weizer[2], Ajjai Alva[3], Galina Kirova-Nedyalkova[4], Kimberly Shampain[1], Nathaniel Meyer[1], Daniel Barkmeier[1], Sean A Woolen[5], Prasad R. Shankar[1], Isaac R. Francis[1], and Phillip L. Palmbos[3]

Departments of [1]Radiology; [2]Urology, Comprehensive Cancer Center; [3]Internal Medicine, Division of Hematology-Oncology, University of Michigan, Ann Arbor, MI; [4]Department of Radiology, Acibadem City Clinic Tokuda Hospital, Sofia, Bulgaria; and [5]Department of Radiology, University of California, San Francisco, Medical Center, San Francisco, CA

Corresponding Author:
Lubomir M. Hadjiiski, PhD
Department of Radiology, University of Michigan, 1500 E. Medical Center Drive, MIB C476, Ann Arbor, MI 48109-5842;
E-mail: lhadjisk@umich.edu

Key Words: Bladder cancer, treatment response assessment, intraobserver variability, radiomics, observer performance study, decision support systems
Abbreviations: computerized decision-support system for treatment response assessment (CDSS-T), deep-learning convolutional neural networks (DL-CNN), methotrexate vinblastine doxorubicin cisplatin (MVAC), institutional review board (IRB), volume of interest (VOI), graphical user interface (GUI), Auto-initialized cascaded level sets (AI-CALS), region of interest (ROI), receiver operating characteristic (ROC), computer-aided diagnosis (CAD), definite complete response (T0 disease), response evaluation criteria in solid tumors (RECIST), multireader multicase (MRMC), methodology for analysis of the not "fully-crossed" observer data (iMRMCMRMC), area under the curve (AUC)

ABSTRACT

We evaluated the intraobserver variability of physicians aided by a computerized decision-support system for treatment response assessment (CDSS-T) to identify patients who show complete response to neoadjuvant chemotherapy for bladder cancer, and the effects of the intraobserver variability on physicians' assessment accuracy. A CDSS-T tool was developed that uses a combination of deep learning neural network and radiomic features from computed tomography (CT) scans to detect bladder cancers that have fully responded to neoadjuvant treatment. Pre- and postchemotherapy CT scans of 157 bladder cancers from 123 patients were collected. In a multireader, multicase observer study, physician-observers estimated the likelihood of pathologic T0 disease by viewing paired pre/posttreatment CT scans placed side by side on an in-house-developed graphical user interface. Five abdominal radiologists, 4 diagnostic radiology residents, 2 oncologists, and 1 urologist participated as observers. They first provided an estimate without CDSS-T and then with CDSS-T. A subset of cases was evaluated twice to study the intraobserver variability and its effects on observer consistency. The mean areas under the curves for assessment of pathologic T0 disease were 0.85 for CDSS-T alone, 0.76 for physicians without CDSS-T and improved to 0.80 for physicians with CDSS-T ($P = .001$) in the original evaluation, and 0.78 for physicians without CDSS-T and improved to 0.81 for physicians with CDSS-T ($P = .010$) in the repeated evaluation. The intraobserver variability was significantly reduced with CDSS-T ($P < .0001$). The CDSS-T can significantly reduce physicians' variability and improve their accuracy for identifying complete response of muscle-invasive bladder cancer to neoadjuvant chemotherapy.

INTRODUCTION

Approximately 81,400 new cases of bladder cancer (62,100 in men, 19,300 in women) will be diagnosed in 2020, resulting in 17,980 deaths (13,050 male, 4,930 female) according to estimates by the American Cancer Society (1). Only 51% of bladder cancers are diagnosed at an early stage (stage T1 or less) when the cancer involves only the inner mucosal layer of the bladder wall (1) and is relatively easier to treat.

Improvement in patient survival and decrease in probability of metastatic disease is observed when a neoadjuvant chemotherapy was performed prior to radical cystectomy (2–4). However, neoadjuvant chemotherapy can cause significant toxicities, such

ISSN 2379-1381 https://doi.org/10.18383/j.tom.2020.00013

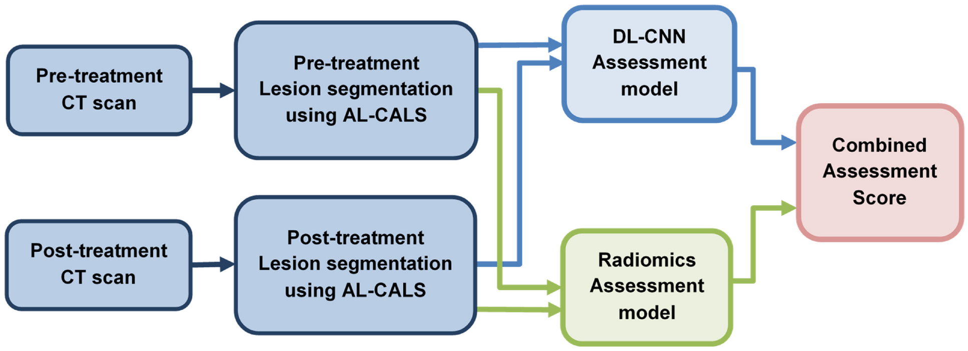

Pre-treatment CT scan
Pre-treatment Lesion segmentation using AL-CALS
Post-treatment CT scan
Post-treatment Lesion segmentation using AL-CALS
DL-CNN Assessment model
Radiomics Assessment model
Combined Assessment Score

Figure 1. Image analysis pipeline of the computerized decision-support system for treatment response assessment (CDSS-T system). AI-CALS: auto-initialized cascaded level sets. DL-CNN: deep-learning convolutional neural network.

as neutropenic fever, sepsis, mucositis, nausea, vomiting, malaise, and alopecia (5). Assessing response to neoadjuvant chemotherapy is not reliable at present, which may cause some patients to suffer adverse reactions to treatment with chemotherapy while gaining minimal benefit (6, 7). It is important to develop an accurate method for assessment of treatment response. Such a method could be very useful for personalizing therapy to patients in the neoadjuvant chemotherapy setting. It might also facilitate optimal selection of patients for bladder-sparing therapy (8), in which trimodal therapy (ie, transurethral resection, chemotherapy, radiation) can be used as a curative option for patients who do not wish to undergo the morbidity of radical cystectomy.

A computerized decision-support system for muscle-invasive bladder cancer treatment response assessment (CDSS-T) using imaging information from computed tomography (CT) examinations was developed in our laboratory. The CDSS-T tool estimates the likelihood that a patient has completely responded to neoadjuvant chemotherapy (9). It integrates deep-learning convolutional neural networks (DL-CNN) and radiomics features. We have used the CDSS-T as a physicians' aid in an observer study for assessment of the likelihood that a patient has completely responded to neoadjuvant chemotherapy (10). The physicians' assessment accuracy improved when CDSS-T was used than when CDSS-T was not used (10).

In this study, we evaluated the intraobserver variability in physicians' assessment aided by the CDSS-T of complete radiographic response to neoadjuvant chemotherapy, and the effects of that intraobserver variability on the physicians' assessment accuracy.

METHODS

Data Set

The study population consisted of 123 subjects with 157 muscle-invasive bladder cancers who had undergone CT scanning of the pelvis before and after neoadjuvant chemotherapy treatment before radical cystectomy. One hundred subjects were males with a mean age of 63 years (range, 43–84 years), and 23 were females with a mean age of 23 years (range, 37–82 years). The chemotherapy treatment was performed with MVAC (methotrexate, vinblastine, doxorubicin, cisplatin) or an alternative regimen (variably including carboplatin, paclitaxel, gemcitabine, etoposide). Three cycles of chemotherapy treatment were performed. Institutional Review Board (IRB) approval was obtained for this study.

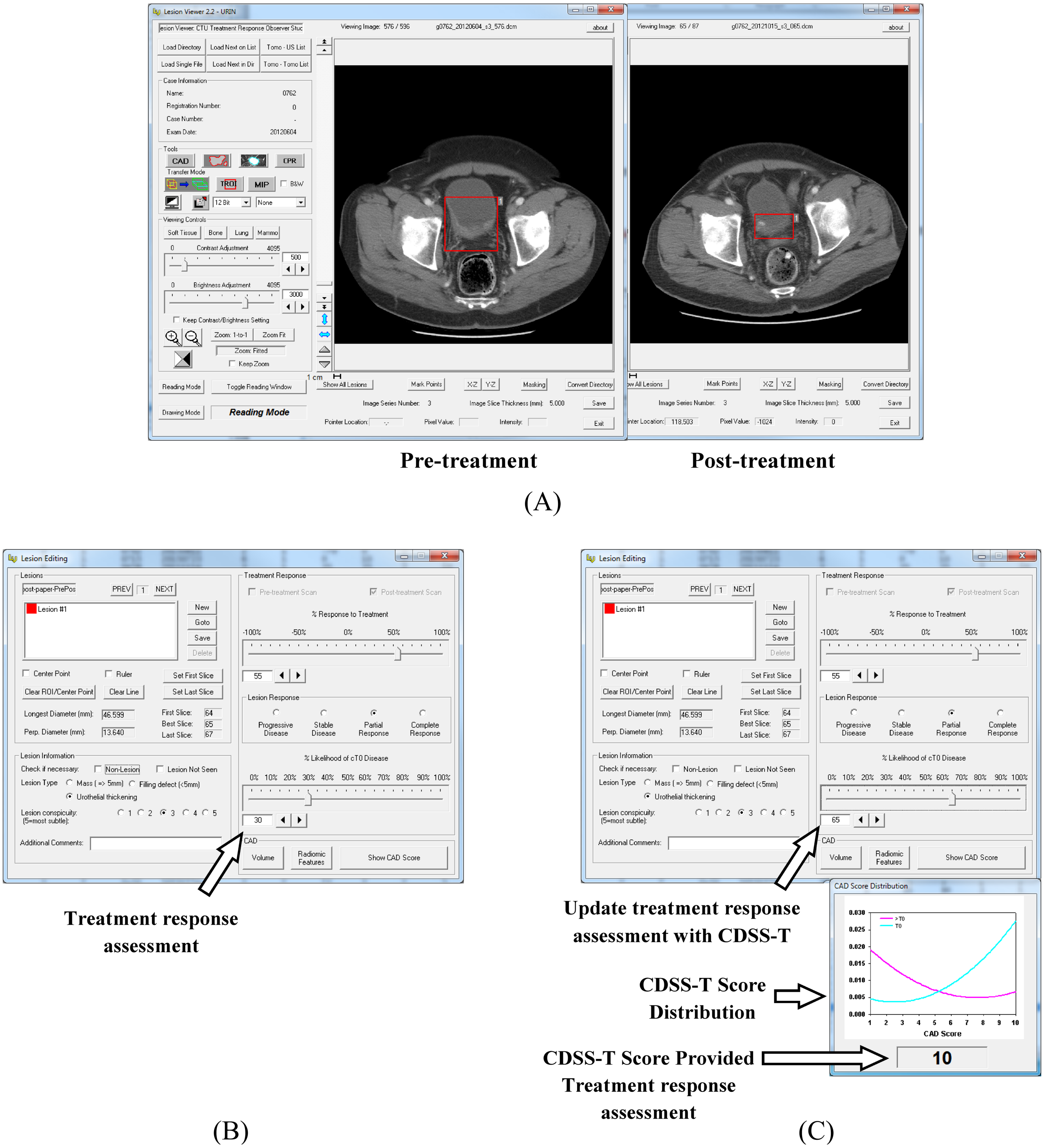
For all subjects, pretreatment and posttreatment CT scans of the pelvis with or without contrast material were acquired with GE Healthcare (WI) Lightspeed MDCT scanners using 120 kVp and 120–280 mA, at a pixel size range of 0.586–0.977 mm and a slice interval range of 0.625–7 mm. Pretreatment CT scans were acquired ∼1 month before the first cycle of chemotherapy. Posttreatment imaging was acquired at ∼1 month after completion of the therapy. The time interval between the pre- and posttreatment scans was 4 months on average. One to 2 months after completion of neoadjuvant chemotherapy, a radical cystectomy was performed. The final cancer stage and whether the subject had responded completely to neoadjuvant chemotherapy (ie, pathologic T0; the primary outcome measure) was determined on the basis of the pathology obtained from the bladder at the time of surgery. The pathology cancer stage was used as a reference standard. A radiologist (R.H.C) with over 30 years of experience reading abdominal CT marked all cancer locations on the pre- and postchemotherapy CT scans and defined a volume of interest (VOI) with a bounding box that enclosed the cancers using a custom graphical user interface (GUI), MiViewer, developed at the University of Michigan CAD-AI Research Laboratory. This reference radiologist did not participate as an observer in the treatment response assessment experiment.

Computerized Decision Support System for Treatment Response Assessment (CDSS-T)

Our CDSS-T system integrates DL-CNN and radiomic features to distinguish between bladder cancers that have fully responded to treatment (ie, pathologic stage T0) and those that have not (ie, pathologic stage T1–T4) (9). The CDSS-T system segments bladder cancers using our in-house-developed segmentation tool, autoinitialized cascaded level sets (AI-CALS). (11). Radiomic features were extracted from the segmented tumor. The image analysis pipeline of the CDSS-T system is shown in Figure 1.

DL-CNN Assessment Model

We trained a DL-CNN to distinguish complete responders from noncomplete-responders as described previously (9, 12). In brief,



Lesion Viewer: CTU Treatment Response Observer Stuc

Viewing Image: 576 / 596   g0762_20120604_s3_576.dcm

Case Information
Name: 0762
Registration Number: 0
Case Number: -
Exam Date: 20120604

12 Bit   None

Contrast Adjustment 0 – 4095: 500
Brightness Adjustment 0 – 4095: 3000
□ Keep Contrast/Brightness Setting
Zoom: Fitted
□ Keep Zoom

Reading Mode

Image Series Number: 3   Image Slice Thickness (mm): 5.000
Pointer Location: -,-   Pixel Value: Intensity:

Viewing Image: 65 / 87   g0762_20121015_s3_065.dcm

Image Series Number: 3   Image Slice Thickness (mm): 5.000
Pointer Location: 118.503   Pixel Value: -1024   Intensity: 0

(B) Lesion Editing

Lesions: post-paper-PrePos   1
🟥 Lesion #1
□ Center Point   □ Ruler
Longest Diameter (mm): 46.599   Perp. Diameter (mm): 13.640
First Slice: 64   Best Slice: 65   Last Slice: 67

Lesion Information
Check if necessary: □ Non-Lesion   □ Lesion Not Seen
Lesion Type: ○ Mass (=> 5mm)   ○ Filling defect (<5mm)   ● Urothelial thickening
Lesion conspicuity: (5=most subtle): ○ 1 ○ 2 ● 3 ○ 4 ○ 5
Additional Comments:

Treatment Response
□ Pre-treatment Scan   ☑ Post-treatment Scan
% Response to Treatment (-100% – 100%): 55
Lesion Response: ○ Progressive Disease   ○ Stable Disease   ● Partial Response   ○ Complete Response
% Likelihood of cT0 Disease (0% – 100%): 30

(C) Lesion Editing

Lesions: post-paper-PrePos   1
🟥 Lesion #1
□ Center Point   □ Ruler
Longest Diameter (mm): 46.599   Perp. Diameter (mm): 13.640
First Slice: 64   Best Slice: 65   Last Slice: 67
Check if necessary: □ Non-Lesion   □ Lesion Not Seen
Lesion Type: ○ Mass (=> 5mm)   ○ Filling defect (<5mm)   ● Urothelial thickening
Lesion conspicuity: (5=most subtle): ○ 1 ○ 2 ● 3 ○ 4 ○ 5
□ Pre-treatment Scan   ☑ Post-treatment Scan
% Response to Treatment: 55
Lesion Response: ○ Progressive Disease   ○ Stable Disease   ● Partial Response   ○ Complete Response
% Likelihood of cT0 Disease: 65

CAD Score Distribution: 10

Figure 2. Graphical user interface (GUI) for reading with and without the computer-aided diagnosis (CAD) system designed for supporting treatment response assessment (CDSS-T). (A) The pre- and posttreatment scans are shown side by side, and (B) the observer estimates the treatment response, recording the estimate in the interface indicated by the arrow. (C) The observer is shown the CDSS-T score and the score distribution of the 2 classes is displayed for reference, as indicated by the middle arrow and bottom arrow, respectively. The observer may revise their treatment response assessment after considering the CDSS-T score using the interface pointed to by the top arrow.

"hybrid" regions of interests (ROIs) were first generated from the pre- and post-treatment ROIs extracted from within the segmented cancers on the pre- and posttreatment CT scans. Each hybrid ROI was formed from a digitally concatenated side-by-side pair of the pre- and posttreatment ROIs. For each cancer, a large number of hybrid ROIs were generated by taking different combinations of the pre- and posttreatment ROIs. All hybrid ROIs from the same cancer were labeled as a complete responder (ie, pathologic stage T0) or a noncomplete-responder (ie, pathologic stages T1–T4) according to the postcystectomy-determined pathologic cancer stage. A leave-one-case-out cross-validation scheme was used for the training and testing of the DL-CNN model. For each leave-one-case-out partition, all hybrid ROIs except for those from the left-out case were used as a training set to train the DL-CNN. The hybrid ROIs from the left-out case were used as test and the trained DL-CNN was then deployed to these test hybrid ROIs. Therefore, a likelihood score of pathologic T0 disease for each of the test ROIs was obtained. Finally, by using the average of the likelihood scores among the ROIs associated with the specific cancer, a "per-cancer" summary score was obtained.

Radiomics Assessment Model

We also developed a radiomics-based model to distinguish complete responders from noncomplete-responders (9). In total, 91 radiomics features, which previously were shown to be useful in analyzing breast masses, lung nodules, and bladder cancer treatment response assessment, were extracted from every segmented cancer. Details of the radiomics features can be found in (9, 13, 14). The percent difference of each radiomic feature between the pre- and posttreatment tumor was calculated for every pre–post CT pair of a given bladder cancer. A 2-loop leave-one-case-out cross-validation scheme (15) was used to build this assessment model to separate the training procedure, which included feature selection and classifier training, from the testing cases. Within the inner loop, the subset of features was selected and the classifier weights were trained with a leave-one-case-out scheme by using the training partition. In the outer loop the trained classifier was deployed to the left-out test case. In such a way the test case is kept independent from the training process. An average of 4 features was selected, including 2 contrast features and 2 run-length statistics features.

CAD Score Generation

The final CDSS-T score was obtained by combining the test scores from both the DL-CNN and the radiomics assessment models. The CDSS-T combined score was generated by taking the maximum of the 2 scores. Receiver operating characteristic (ROC) analysis was performed on the CDSS-T scores. To communicate conveniently the CDSS-T scores to the physicians, the CDSS-T scores were linearly scaled within the interval between 1 and 10, rounding to the nearest whole integer. These rounded scores were referred to as computer-aided diagnosis (CAD) scores. A score of 1 corresponded to the lowest likelihood that the lesion pair was indicative of complete response. A score of 10 corresponded to the highest likelihood that the lesion pair was indicative of complete response. Fitted curves to the distributions of the linearly transformed scores for both the noncomplete-responders and the complete responders were obtained. The area under both of the fitted distribution curves was then normalized to a value of 1. The normalized fitted distribution curves (Figure 2C) were displayed on the GUI as a reference together with the cancer-specific CDSS-T likelihood score to be used as decision support in the computer-aided reading by the observer.

Table 1. Diagnostic Performance in Terms of AUC of Physicians Without and With the Use of CDSS-T for the Assessment of Complete Response to Neoadjuvant Chemotherapy on the Entire Data Set of 157 Bladder Cancers

Observer	AUC Without CDSS-T	AUC With CDSS-T
Physician 1	0.74 ± 0.04	0.77 ± 0.04
Physician 2	0.74 ± 0.04	0.77 ± 0.04
Physician 3	0.74 ± 0.04	0.76 ± 0.04
Physician 4	0.76 ± 0.04	0.79 ± 0.04
Physician 5	0.74 ± 0.04	0.74 ± 0.04
Physician 6	0.76 ± 0.04	0.77 ± 0.04
Physician 7	0.66 ± 0.05	0.73 ± 0.04
Physician 8	0.73 ± 0.04	0.75 ± 0.04
Physician 9	0.78 ± 0.04	0.81 ± 0.04
Physician 10	0.73 ± 0.04	0.76 ± 0.04
Physician 11	0.72 ± 0.04	0.76 ± 0.04
Physician 12	0.75 ± 0.04	0.78 ± 0.04
Mean AUC	0.74	0.77

Observer Performance Study

Twelve physicians participated as observers in this study including 5 abdominal-fellowship-trained attending radiologists (faculty experience, 2–36 years), 1 second-year radiology resident, 3 fourth-year radiology residents, 1 attending urologist (faculty experience, 11 years), and 2 attending oncologists (faculty experience, 3 and 10 years). Each observer reviewed each pre- and posttreatment CT pair displayed side by side on a specialized GUI that allows common interactive functions such as windowing, scrolling, and zooming (Figure 2). The observer was asked to provide an estimate of the likelihood of having complete response to treatment of the cancer by inspecting the pre- and posttreatment CT pair. The bladder tumor to be assessed was marked by a VOI box on both the pre- and posttreatment scans. In cases containing multiple cancers and therefore multiple VOIs, each VOI was analyzed separately (Figure 2A). Each observer was given unlimited time for the evaluation and was blinded to the reference standard and to the results of the other observers. To minimize bias related to fatigue or learning due to reading order, the sequence of cases in the reading list was randomized differently for each observer.

For each cancer, each observer provided an estimate of its likelihood of complete response on a scale of 0% to 100%, where 0% indicated definite residual viable neoplasm (>T0 disease) and 100% indicated definite complete response (T0 disease) (Figure 2B). Reader estimates were provided first without and then with access to the CAD likelihood score (Figure 2C). In this way, the

Table 2. Diagnostic Performance in Terms of AUC of Physicians Without and With the Use of CDSS-T for the Assessment of Complete Response to Neoadjuvant Chemotherapy on the First 51 Cases in Each Observer's Individually Randomized Reading List

	AUC Original Evaluation		AUC Repeated Evaluation	
Observer	**Without CDSS-T**	**With CDSS-T**	**Without CDSS-T**	**With CDSS-T**
Physician 1	0.76 ± 0.07	0.76 ± 0.07	0.79 ± 0.07	0.77 ± 0.07
Physician 2	0.88 ± 0.05	0.90 ± 0.04	0.88 ± 0.05	0.93 ± 0.04
Physician 3	0.69 ± 0.08	0.70 ± 0.08	0.66 ± 0.08	0.70 ± 0.08
Physician 4	0.70 ± 0.07	0.78 ± 0.06	0.83 ± 0.06	0.83 ± 0.06
Physician 5	0.83 ± 0.06	0.86 ± 0.06	0.81 ± 0.07	0.82 ± 0.07
Physician 6	0.75 ± 0.08	0.76 ± 0.08	0.83 ± 0.06	0.87 ± 0.05
Physician 7	0.65 ± 0.08	0.74 ± 0.07	0.73 ± 0.07	0.77 ± 0.06
Physician 8	0.75 ± 0.08	0.78 ± 0.08	0.81 ± 0.08	0.82 ± 0.07
Physician 9	0.81 ± 0.06	0.86 ± 0.05	0.79 ± 0.06	0.85 ± 0.05
Physician 10	0.80 ± 0.08	0.85 ± 0.07	0.81 ± 0.07	0.88 ± 0.06
Physician 11	0.65 ± 0.10	0.71 ± 0.09	0.65 ± 0.09	0.70 ± 0.09
Physician 12	0.82 ± 0.07	0.85 ± 0.06	0.76 ± 0.07	0.77 ± 0.07
Mean AUC	0.76	0.80	0.78	0.81
Mean standard deviation	0.073	0.069	0.069	0.064

In the 12 groups of 51 cases, each group contained different cases for each observer, were evaluated 2 times, shown as original evaluation and repeated evaluation.

Statistical significance in the difference:

AUC:

AUC(orig, with) versus AUC(orig, without): $P = .001$.
AUC(repeat, with) versus AUC(repeat, without): $P = .010$.
AUC(orig, without) versus AUC(repeat, without): $P = .083$.
AUC(orig, with) versus AUC(repeat, with): $P = .222$.

Standard deviation of AUC (SD_{AUC}):

SD_{AUC}(orig, with) versus SD_{AUC}(orig, without): $P < .0002$.
SD_{AUC}(repeat, with) versus SD_{AUC}(repeat, without): $P < .004$.
SD_{AUC}(orig, without) versus SD_{AUC}(repeat, without): $P = .112$.
SD_{AUC}(orig, with) versus SD_{AUC}(repeat, with): $P = .066$.

observers were given the opportunity to modify their estimate after being provided the CAD score, although they could leave it unchanged if they wished.

Each observer was also asked to estimate a percentage response of tumor to the neoadjuvant chemotherapy on a scale of −100% to +100% using RECIST 1.1 (16) measurement criteria, where 0% indicated no change between pre- and posttreatment CT scans, −100% indicated at least doubling of tumor size, and 100% indicated a complete response.

To study the intraobserver variability, each observer was asked to repeat the evaluation of the first 51 cases in the observer's individually randomized reading list after completing the evaluation of all cases in the list. Because each observer's list was randomized differently, the first 51 cases were different for each observer. We define the first reading of these first 51 cases for each observer as "original evaluation" to distinguish it from the repeated evaluation in the following discussion. The washout time between the original and repeated evaluations was ∼1 month to avoid potential memorization effects. The observers were not informed that they are repeating the evaluation of the cases. The observers were also blinded to the reference standard and to the results of the other observers.

Data Analysis

The observers' estimates were analyzed with multireader, multicase (MRMC) receiver operating characteristic (ROC) methodology using the radical cystectomy specimen as the reference standard (17). iMRMC methodology was also used for the analysis of the not "fully-crossed" intra-observer variability data, which were analyzed as an alternative design study (18, 19). The area under the curve (AUC) and the statistical significance of the difference in readings with and without CDSS-T were calculated. One outcome was a comparison of the diagnostic accuracy of the physicians in diagnosing T0 disease after treatment without CDSS-T and after the physicians had CDSS-T for decision support. Another outcome was an assessment of the intraobserver variability by comparing the results of the original and the repeated evaluation of the corresponding subsets of cases for each observer. The AUC and the statistical significance of the difference between the 2 evaluations were calculated.

An additional measure of the intraobserver variability was based on the standard deviation of the differences of the observer's original evaluation likelihood estimates and the observer's corresponding repeated evaluation likelihood estimates. The intraobserver variability assessments were performed for the observers' evaluations without and without CDSS-T and then compared.

The average standard deviation of the likelihood estimates by the observers per treatment pair was analyzed to study the effects of CDSS-T on inter- and intraobserver variability. The standard deviation of the observers' likelihood estimates of a given cancer was used as a measure of the level of difficulty, assuming that inter- and intraobserver variabilities would be smaller for easier cancers.

Pearson correlation was used to examine if there is relationship between the average level of difficulty of the case group and the AUC of reading the same case group by a given observer. The correlation was calculated for the AUCs of both readings with and without CDSS-T in both the original and repeated evaluations and the AUC of the CDSS-T alone. For all analyses, a *P*-value of <.05 was considered to indicate a significant difference.

RESULTS

Surgical histology revealed that 25% (40 / 157) of bladder cancers were determined to have a pathologic stage of T0 following neoadjuvant chemotherapy (ie, 40 complete responders). The average maximum diameter for these 40 completely responding lesions was 30.1 mm on pretreatment scans and 14.3 mm on posttreatment scans. Suspected lesions on posttreatment scans in these patients were found to represent an inflamed bladder wall or an entirely necrotic treated tumor. The average maximum diameter for the remaining 117 incompletely responding lesions was 43.0 mm on pretreatment scans and 31.2 mm on posttreatment scans.

Approximately 24% (12/51) of the bladder cancers were determined to be complete responders after neoadjuvant chemotherapy for each of the 12 subsets of 51 cases for the 12 observers used to study the intra-observer variability.

Overall Results for All Cancers

The overall results for all cancers (157 cancer pairs) are summarized in the following as a reference for the current study. A detailed analysis of the overall results can be found elsewhere (10).

The individual AUC values of the 12 observers are shown in Table 1. In general, the physicians' diagnostic accuracy significantly increased (*P* = .01) and physicians' diagnostic variability significantly decreased (*P* < .001) with the aid of CDSS-T. The average AUC for all of the physicians combined was 0.74 (range, 0.66–0.78) without CDSS-T, and it increased to 0.77 (range, 0.73–0.81) with CDSS-T. This difference was statistically significant (*P* = .01). In comparison, the AUC for assessment of complete response by CDSS-T alone was 0.80 ± 0.04.

Intraobserver Variability

The original and repeated evaluations of the first 51 cases in each observer's individually randomized reading list and estimation of the intraobserver variability are analyzed below. Twelve groups of 51 cases, each group contained the first 51 cases read by each observer, were evaluated 2 times, referred to as original evaluation and repeated evaluation.

The individual AUC values of the 12 observers for the original and repeated evaluations are shown in Table 2 and Figures 3 and 4. For the original evaluation, the average AUC of the 12 observers without the CDSS-T was 0.76 (range, 0.65–0.88) that increased to 0.80 (range, 0.70–0.90) with CDSS-T. The improvement was statistically significant (*P* = .001). For the repeated evaluation, the average AUC of the observers without the CDSS-T was 0.78 (range, 0.65–0.88) that increased to 0.81 (range, 0.70–0.93) with CDSS-T. The improvement was also statistically significant (*P* = .010).

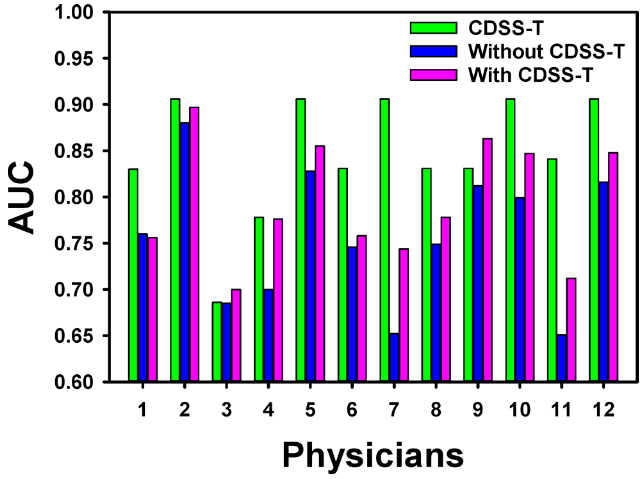


Figure 3. AUC values for the 12 observers with and without CDSS-T and the corresponding CDSS-T alone of the first 51 cases in each observer's individually randomized reading list for the original evaluation. The performance of all but 1 (physician 1) of the physicians increased using CDSS-T. Two physicians (physicians 3 and 9) with CDSS-T performed better than the CDSS-T alone.

However, there was no statistically significant difference between the average AUCs for the original and the repeated evaluations without CDSS-T (*P* = .083) or for the evaluations with CDSS-T (*P* = .222).

The standard deviations of the AUCs were smaller for both the original and the repeated evaluations with CDSS-T than for those without CDSS-T: an average of 0.073 without CDSS-T versus an average of 0.069 with CDSS-T (*P* < .0002) for the original evaluation, and an average of 0.069 without CDSS-T versus an average of 0.064 with CDSS-T (*P* < .004) for the repeated evaluation. In addition, for both without and with CDSS-T, the standard deviations of the AUCs were smaller for the repeated evaluation than for the original evaluation. However, the differences did not reach statistical significance (*P* > .07).

When evaluating with CDSS-T, 2 observers performed better than the CDSS-T alone in the original evaluation (Figure 3). In the repeated evaluation, 3 additional observers (5 in total) performed better than the CDSS-T alone when they evaluated with CDSS-T (Figure 4). The average AUC over the 12 groups of 51 cases for assessment of complete response by CDSS-T alone was 0.85 ± 0.06.

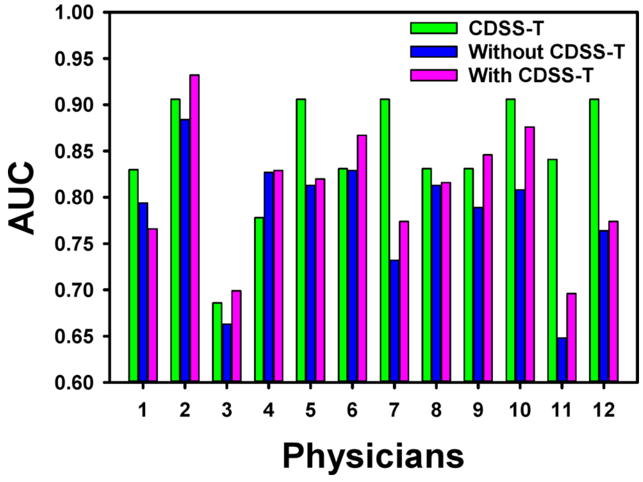


Figure 4. AUC values for the 12 observers with and without CDSS-T and the corresponding CDSS-T alone of the first 51 cases in each observer's individually randomized reading list for the repeated evaluation. The performance of all but one (physician 1) of the physicians increased using CDSS-T. Five physicians (physicians 2, 3, 4, 6, and 9) with CDSST performed better than the CDSS-T alone.

Table 3. Intraobserver Variability Assessment Based on the Standard Deviation (SD) of the Differences of the Observer's Original Evaluation Likelihood Estimates and the Observer's Corresponding Repeated Evaluation Likelihood Estimates

	SD of the differences in the observers' likelihood estimates between the original and repeated evaluation	
Observer	**Without CDSS-T**	**With CDSS-T**
Physician 1	26.34	20.70
Physician 2	15.35	13.89
Physician 3	23.75	20.95
Physician 4	18.87	14.46
Physician 5	33.59	28.24
Physician 6	32.28	22.92
Physician 7	30.46	21.57
Physician 8	15.99	13.40
Physician 9	25.42	18.99
Physician 10	24.01	17.95
Physician 11	27.84	24.09
Physician 12	44.42	41.92
Mean	26.53	21.59

The intraobserver variability assessments were performed for the observer's evaluations without CDSS-T and with CDSS-T.
Mean SD with CDSS-T (21.59) was significantly smaller than the mean SD without CDSS-T (26.53), ($P < .0001$).

The intraobserver variability estimated as the mean standard deviation of the corresponding observers' likelihood estimates differences between the original and the repeated evaluations was 26.53 without CDSS-T and was reduced significantly to 21.59 with CDSS-T ($P < .0001$) (Table 3).

Difficulty of Cancers as a Performance Factor

The level of difficulty for the 12 case groups estimated by the inter-reader standard deviation of the member cases within the groups was moderately negatively correlated (r = −0.64) with the corresponding AUC for CDSS-T alone for the 12 groups.

The level of difficulty was also negatively correlated (r = −0.31) with the corresponding physicians' AUCs with and without CDSS-T for the 12 groups. In the repeated evaluation, the physicians' AUCs with and without CDSS-T was less negatively correlated (r = −0.10 and −0.22, respectively) with the level of difficulty compared with the original evaluation with CDSS-T.

DISCUSSION

In this study, we evaluated the intraobserver variability of physicians' treatment response assessments of bladder cancer after neoadjuvant chemotherapy in CT examinations via CDSS-T.

We observed statistically significant improvement in physicians' average performance when they used CDSS-T for evaluation than when they did not use CDSS-T for evaluation. There was improvement in all experiments including the evaluation with the entire data set as well as the original and repeated evaluations of the individualized subsets. We have found that the interobserver variability was significantly reduced with the use of CDSS-T in the previous study (10), and that the intraobserver variability was also significantly reduced with CDSS-T in the current study. This is important, because the CDSS-T was able to consistently improve the accuracy of the observer evaluations and reduce the observer variability in the different experiments including repeated evaluations.

The level of difficulty of the cases has a stronger impact on the CDSS-T performance alone than on the observer performance. The observers were even less affected in the repeated evaluation with CDSS-T.

For both without and with CDSS-T evaluations, we have observed a slight improvement trend in the observers' performance (increased average AUCs and reduced average variability [standard deviations]) for the repeated evaluation compared with the original evaluation. However the improvement was not statistically significant. In addition, a larger number of observers with CDSS-T performed better than the CDSS-T alone in the repeated evaluation. The observed trends of improved performance for the repeated evaluation are interesting. These may be attributed to the fact that the observers were becoming more experienced using the decision-support tool and were using it more effectively for improving their assessment. The understanding of how a user may be influenced by their experience with and confidence on a decision-support tool is a topic of interest for future studies.

There are limitations in this study. First, the CDSS-T scores were obtained through the leave-one-case-out cross-validation owing to the lack of a large data set. Ideally, the system should have been evaluated on an independent test set (20). However, the leave-one-case-out cross-validation approach is well established in the machine learning literature and is a statistically valid technique for estimating classifier performance in an unknown population. In the future, as we collect a larger data set, we will evaluate our system on an independent test set.

Second, we used a sequential design for our observer study experiment (21–23). The main reason is that the Food and Drug Administration approved the use of CAD so far is in the sequential mode as a second reader.

Third, although the performance of CDSS-T alone was higher than that of the observers in this study, the AUCs under all conditions were still modest, probably because of the challenging nature of this classification task. It is possible that the imaging modality itself provides limited radiomics or physiological information that neither a physician nor machine learning will be able to overcome. We are now attempting to improve the CDSS-T by using improved cancer segmentation methods (24), more advanced DL-CNN models (12), and most importantly, by combining the imaging-based assessment with other available clinical biomarkers, including results from bimanual examinations under anesthesia, results from transuretheral resection of bladder cancer (25), and molecular biomarkers such as genomics and proteomics. Fourth, none of our observers was experienced in using a decision-support tool for bladder cancer, because such decision-support tools are not yet available in the clinic for abdominopelvic applications. This may have limited the observers' confidence in the CDSS-T system at the beginning. We expect that physicians will become more receptive to CDSS-T "advice" after gaining experience with the system as observed in the repeated evaluation results in the current study. The increased experience and improved confidence in CDSS-T may result in further improvements in diagnostic accuracy.

CONCLUSIONS

There exists an intraobserver variability for the physicians in the assessment of patients' response to neoadjuvant chemotherapy for muscle-invasive bladder cancer in CT. This study shows that our computerized decision-support system, CDSS-T, can significantly reduce physicians' variability and improve their accuracy in identifying the complete response of muscle-invasive bladder cancer to neoadjuvant chemotherapy. To validate the impact of the CDSS-T on clinical decision-making, a large-scale observer study should be conducted in an independent case set. A fully validated CDSS-T may have the potential of improving physicians' decision in the selection of patients with muscle-invasive bladder cancer for bladder-sparing therapy.

ACKNOWLEDGMENT

This work is supported by National Institutes of Health grant number U01CA232931.

Conflict of Interest: None reported.

Disclosures: No disclosures to report.

REFERENCES

1. American Cancer Society. Cancer Facts & Figures 2020. 2020. https://www.cancer.org/.
2. Fagg SL, Dawson-Edwards P, Hughes MA, Latief TN, Rolfe EB, Fielding JWL. Cis-diamminedichloroplatinum (DDP) as initial treatment of invasive bladder cancer. Br J Urol. 1984;56:296–300.
3. Raghavan D, Pearson B, Coorey G, Woods W, Arnold D, Smith J, Donovan J, Langdon P. Intravenous cis-platinum for invasive bladder cancer. Safety and feasibility of a new approach. Med J Aust. 1984;140:276–278.
4. Meeks JJ, Bellmunt J, Bochner BH, Clarke NW, Daneshmand S, Galsky MD, Hahn NM, Lerner SP, Mason M, Powles T, Sternberg CN, Sonpavde G. A systematic review of neoadjuvant and adjuvant chemotherapy for muscle-invasive bladder cancer. Eur Urol. 2012;62:523–533.
5. Witjes JA, Wullink M, Oosterhof GO, de Mulder P. Toxicity and results of MVAC (methotrexate, vinblastine, adriamycin and cisplatin) chemotherapy in advanced urothelial carcinoma. Eur Urol. 1997;31:414–419.
6. Seiler R. Predicting response to neoadjuvant chemotherapy in bladder cancer: controversies remain with genomic DNA sequencing. Transl Androl Urol. 2016;5:271–273.
7. Kukreja JB, Porten S, Golla V, Ho PL, Noguera-Gonzalez G, Navai N, Kamat AM, Dinney CPN, Shah JB. Absence of tumor on repeat transurethral resection of bladder tumor does not predict final pathologic T0 stage in bladder cancer treated with radical cystectomy. Eur Urol Focus. 2018;4:720–724.
8. Kulkarni GS, Hermanns T, Wei YL, Bhindi B, Satkunasivam R, Athanasopoulos P, Bostrom PJ, Kuk C, Li K, Templeton AJ, Sridhar SS, van der Kwast TH, Chung P, Bristow RG, Milosevic M, Warde P, Fleshner NE, Jewett MAS, Bashir S, Zlotta AR. Propensity score analysis of radical cystectomy versus bladder-sparing trimodal therapy in the setting of a multidisciplinary bladder cancer clinic. J Clin Oncol. 2017;35:2299–2305.
9. Cha KH, Hadjiiski L, Chan HP, Weizer AZ, Alva A, Cohan RH, Caoili EM, Paramagul C, Samala RK. Bladder cancer treatment response assessment in CT using radiomics with deep-learning, Sci Rep. 2017;7:1–12.
10. Cha KH, Hadjiiski LM, Cohan RH, Chan HP, Caoili EM, Davenport M, Samala RK, Weizer AZ, Alva A, Kirova-Nedyalkova G, Shampain K, Meyer N, Barkmeier D, Woolen S, Shankar PR, Francis IR, Palmbos P. Diagnostic accuracy of CT for prediction of bladder cancer treatment response with and without computerized decision support. Acad Radiol. 2019;26:1137–1145.
11. Hadjiiski LM, Chan H-P, Caoili EM, Cohan RH, Wei J, Zhou C. Auto-initialized cascaded level set (AI-CALS) segmentation of bladder lesions on multi-detector row CT urography. Acad Radiol. 2013;20:148–155.
12. Wu E, Hadjiiski LM, Samala RK, Chan H-P, Cha KH, Richter C, Cohan RH, Caoili EM, Paramagul C, Alva A, Weizer AZ. Deep learning approach for assessment of bladder cancer treatment response. Tomography. 2019;5:201–208.
13. Sahiner B, Chan H-P, Petrick N, Helvie MA, Hadjiiski LM. Improvement of mammographic mass characterization using spiculation measures and morphological features. Med Phys. 2001;28:1455–1465.
14. Way TW, Hadjiiski LM, Sahiner B, Chan H-P, Cascade PN, Kazerooni EA, Bogot N, Zhou C. Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours. Med Phys. 2006;33:2323–2337.
15. Way TW, Sahiner B, Chan H-P, Hadjiiski L, Cascade PN, Chughtai A, Bogot N, Kazerooni E. Computer aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. Med Phys. 2009;36:3086–3098.
16. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45:228–247.
17. Dorfman DD, Berbaum KS, Metz CE, Obuchowski NA, Rockette H. http://perception.radiology.uiowa.edu/Software/ReceiverOperatingCharacteristicROC/MRMCAnalysis/tabid/116/Default.aspx.
18. Gallas BD, Bandos A, Samuelson FW, Wagner RF. A framework for random-effects ROC analysis: biases with the bootstrap and other variance estimators. Commun Stat Theory Methods. 2009;38:2586–2603.

19. Gallas BD. One-shot estimate of MRMC variance: AUC. Acad Radiol. 2006;13:353–362.
20. Petrick N, Sahiner B, Armato SG, Bert A, Correale L, Delsanto S, Freedman MT, Fryd D, Gur D, Hadjiiski L, Huo ZM, Jiang YL, Morra L, Paquerault S, Raykar V, Samuelson F, Summers RM, Tourassi G, Yoshida H, Zheng B, Zhou C, Chan H-P. Evaluation of computer-aided detection and diagnosis systems. Med Phys. 2013;40.
21. Beiden SV, Wagner RF, Doi K, Nishikawa RM, Freedman M, Lo S-CB, Xu X-W. Independent versus sequential reading in ROC studies of computer-assist modalities: analysis of component of variance. Acad Radiol. 2002;9:1036–1043.
22. Hadjiiski LM, Chan H-P, Sahiner B, Helvie MA, Roubidoux M, Blane C, Paramagul C, Petrick N, Bailey J, Klein K, Foster M, Patterson S, Adler D, Nees A, Shen J. Improvement of radiologists' characterization of malignant and benign breast masses in serial mammograms by computer-aided diagnosis: an ROC study. Radiology. 2004;233:255–265.
23. Hadjiiski LM, Chan H-P, Sahiner B, Helvie MA, Roubidoux M, Blane C, Paramagul C, Petrick N, Bailey J, Klein K, Foster M, Patterson S, Adler D, Nees A, Shen J. Breast masses: computer-aided diagnosis with serial mammograms. Radiology. 2006;240:343–356.
24. Cha KH, Hadjiiski LM, Samala RK, Chan HP, Cohan RH, Caoili EM, Paramagul C, Alva A, Weizer AZ. Bladder cancer segmentation in CT for treatment response assessment: application of deep-learning convolution neural network-a pilot study. Tomography. 2016;2:421–429.
25. Gordon MN, Cha KH, Hadjiiski L, Chan HP, Cohan RH, Caoili EM, Paramagul C, Alva A, Weizer AZ. Bladder cancer treatment response assessment with radiomic, clinical and radiologist semantic features. Proc SPIE, Medical Imaging. 2018;10575 105751Y-1.