Bladder Cancer Staging in CT Urography: Estimation and Validation of Decision Thresholds for a Radiomics-Based Decision Support System

Dhanuj Gandikota[1], Lubomir Hadjiiski[1], Heang-Ping Chan[1], Kenny H. Cha[2], Ravi Samala[1], Elaine M. Caoili[1], Richard H. Cohan[1], Alon Weizer[4], Ajjai Alva[3], Chintana Paramagul[1], Jun Wei[1], Chuan Zhou[1]

[1]Department of Radiology, University of Michigan, Ann Arbor, MI
[2]Center for Devices and Radiological Health, U.S. FDA, Silver Spring, MD
[3]Department of Internal Medicine-Hematology/Oncology, University of Michigan, Ann Arbor, MI
[4]Department of Urology, University of Michigan, Ann Arbor, MI

Abstract 250 Words: Stage T2 is the clinical threshold when bladder cancer is treated with neoadjuvant chemotherapy. In this study we refined a radiomics-based decision support system (CDSS-S) to aid clinicians in staging of bladder cancer in CT Urography (CTU). To train the CDSS-S, we used a data set of 84 bladder cancers from 76 CTU clinically staged cases, 43 cancers were below stage T2, and 41 were stage T2 or above. An independent test set comprising of 82 bladder cancers from 80 CTU clinically staged cases that were staged as T2 or above were also collected. Our AI-CALS method was utilized to segment the lesions from which radiomics features were extracted. The training set was split on 2 balanced partitions. Four classifiers were studied: Linear Discriminant Analysis (LDA), Support vector machines (SVM), Back-propagation neural networks (BPNN), and Random Forest (RAF) classifiers. Based on the likelihood scores for a training set, the decision threshold providing the highest classification accuracy for each classifier was determined. The classifier with the fixed decision threshold was then applied to the test set and the performance evaluated. The test classification accuracy for the LDA, SVM, BPNN, and RAF trained on Partition 1 was 0.95, 0.98, 0.88, and 0.89, respectively, and was 0.88, 0.94, 0.88, and 0.93, respectively, when trained on Partition 2. The test classification accuracy for the LDA, SVM, BPNN, and RAF trained on the entire training set was 0.94, 0.94, 0.94, and 0.89, respectively. The results show the potential of CDSS-S in bladder cancer stage assessment.

Purpose: Stage T2 is the clinical threshold when a decision of administrating neoadjuvant chemotherapy in the treatment of bladder cancer was made. An estimated 30% of patients are over-staged or understaged in clinical staging. The purpose of this study is to refine our quantitative computerized decision support system (CDSS-S) that could be used to assist clinicians in the staging of bladder cancer when using CT Urography (CTU).

Methods: With IRB approval, we collected a training set comprising of 84 bladder cancers from 76 CTU clinically staged cases before treatment. An independent test set of CTU scans prior to undergoing chemotherapy treatment was also collected comprising of 82 bladder cancers from 80 clinically staged cases. The lesions were categorized into two classes: (A) staged at or above T2 or (B) staged below T2. Of the 84 bladder cancers within the training set, 43 of the lesions were below stage T2, and 41 were stage T2 or above. The 82 cancers in the test set were staged as T2 or higher and all 80 patients underwent chemotherapy treatment. Our previously developed auto-initialized cascaded level sets (AI-CALS) method was utilized to segment the lesions. A bounding box provided by a radiologist for each lesion was used as input for the AI-CALS. 89 radiomics features were extracted from each segmented cancer. The training set was randomly split into two partitions.

Partition 1 contained 22 cancers staged below T2 and 20 cancers staged T2 or higher. Partition 2 contained 21 cancers staged below T2 and 21 cancers staged T2 or higher. Four types of classifiers, Linear Discriminant Analysis (LDA), Support vector machine (SVM), Back-propagation neural network (BPNN) and Random Forest (RAF), were trained to merge the important radiomics features and estimate a likelihood score for prediction of T2 stage. The classifiers were trained on each training partition, as well as on the entire training set, using feature selection and then evaluated on the independent test set.
LDA with stepwise feature selection was trained on each partition to merge the best subset of features. Three features were selected for Partition 1 and seven features were selected for Partition 2. SVM classifiers were trained on each partition with the corresponding LDA selected features as input. Using the features selected from the LDA as input, a BPNN with a single output node and a single hidden layer was also trained on each partition. The RAF classifier, implemented by WEKA, was trained on each partition using all 89 features with 100 trees and 5 features per tree. In addition, each of the four classifiers was trained on the entire training set that combined Partition 1 and Partition 2. LDA classifier with stepwise feature selection selected 5 features when trained on the combined training set.
For a given classifier, based on the training likelihood scores obtained on a given training set, a decision threshold that maximized the classification accuracy was determined. The classification accuracy was defined as the sum of true positives and true negatives divided by the total number of cancers. The decision threshold was then applied to the independent test set and the performance of the CDSS-S was evaluated in terms of the classification accuracy on the test cases.

Results: For the CDSS-S utilizing LDA, the training classification accuracy on Partition 1 and Partition 2 was 0.88 and 1.0, respectively, while the test classification accuracy was 0.95 and 0.88, respectively. For the CDSS-S utilizing SVM, the training classification accuracy on Partition 1 and Partition 2 was 0.88 and 1.0, respectively, with corresponding test classification accuracy of 0.98 and 0.94, respectively. The training classification accuracy for the BPNN based CDSS-S trained on Partition 1 and Partition 2 was 0.93 and 1.0 respectively, while the test classification accuracy was 0.88 for both. For the RAF based CDSS-S, the training classification accuracy on Partition 1 and Partition 2 was 1.0 for both, while the test classification accuracy was 0.89 and 0.93, respectively. The training classification accuracy for the LDA, SVM, BPNN, and RAF based CDSS-Ss trained on the entire combined training set was 0.92, 0.94, 0.93, 1.0, respectively, while the corresponding test classification accuracy was 0.94, 0.94, 0.94, and 0.89, respectively.

New or Breakthrough Work: Accurate staging of bladder cancer is critical for the correct decision of administering neodjuvant chemotherapy to the patients. An estimated 30% of patients are understaged or over-staged. Staging inaccuracy may be ascribed to the variability and subjectivity of clinicians in using available diagnostic information. An objective decision support system may be useful for assisting clinicians in making more consistent and precise staging assessments. In this study, we evaluated the robustness of the trained CDSS-S and the decision thresholds using an independent test set. Our preliminary results demonstrate the feasibility of a radiomics-based CDSS-S that can assist with bladder cancer staging.

Conclusion: We demonstrate the potential of using automatically extracted radiomic features from CTU and decision thresholds to build a statistical predictive model for the staging of bladder cancer. The high classification accuracy in pre-treatment CTU of bladder cancer cases validated the

performance of the predictive models. Further work includes the collection of a larger data set and improvement in the predictive model accuracy through the inclusion of clinical and molecular data.

Key Words: Bladder Cancer Staging, Radiomics, Classification, Segmentation

WHETHER THE WORK IS BEING, OR HAS BEEN SUBMITTED FOR PUBLICATION OR PRESENTATION ELSEWHERE: No

Abstract 100 words: We are refining a computerized decision support system (CDSS-S) to aid clinicians in staging of bladder cancer in CT Urography. The CDSS-S were trained on 84 bladder cancers and tested on 82 independent bladder cancers. LDA, SVM, BPNN, and RAF classifiers utilizing radiomic features were studied as an integral part of the CDSS-S. Decision thresholds determined from the training set were applied to the test set to evaluate the CDSS-S performance. The test classification accuracy for the LDA, SVM, BPNN, and RAF trained on the entire training set was 0.94, 0.94, 0.94, and 0.89, respectively, demonstrating the potential of the CDSS-S in bladder cancer staging.

Tables 1-4: Training and testing results for each CDSS-S model. Table 1 used LDA, Table 2 used SVM, Table 3 used BPNN, and Table 4 used RAF. Within these tables, classification accuracy is denoted by CA, sensitivity is denoted by SE and specificity is denoted by SP. The classifier scores were normalized to between 1 and 100. The number of true positives, true negatives, false positives, and false negatives are denoted by TP, TN, FP, and FN, respectively.

LDA	Train AUC	Train CA	TP	TN	FP	FN	Train SE	Train SP	Thresh	Test CA(SE)	TP	FN
Part1	0.92	0.88	19	18	4	1	0.95	0.82	45.3	0.95	78	4
Part2	1	1	21	21	0	0	1	1	51.5	0.88	72	10
Comb	0.95	0.92	40	37	6	1	0.98	0.86	42.8	0.94	77	5

SVM	Train AUC	Train CA	TP	TN	FP	FN	Train SE	Train SP	Thresh	Test CA(SE)	TP	FN
Part1	0.92	0.88	19	18	4	1	0.95	0.81	31.0	0.98	80	2
Part2	1	1	21	21	0	0	1	1	37.0	0.94	77	5
Comb	0.99	0.94	40	39	4	1	0.98	0.91	42.0	0.94	77	5

BPNN	Train AUC	Train CA	TP	TN	FP	FN	Train SE	Train SP	Thresh	Test CA(SE)	TP	FN
Part1	1	0.93	20	19	3	0	1	0.86	41.8	0.88	72	10
Part2	1	1	21	21	0	0	1	1	41.8	0.88	72	10
Comb	0.95	0.93	40	38	5	1	0.98	0.88	41.1	0.94	77	5

RAF	Train AUC	Train CA	TP	TN	FP	FN	Train SE	Train SP	Thresh	Test CA(SE)	TP	FN
Part1	1	1	23	19	0	0	1	1	34.0	0.89	73	9
Part2	1	1	21	21	0	0	1	1	32.0	0.93	76	6
Comb	1	1	41	43	0	0	1	1	25.0	0.89	73	9

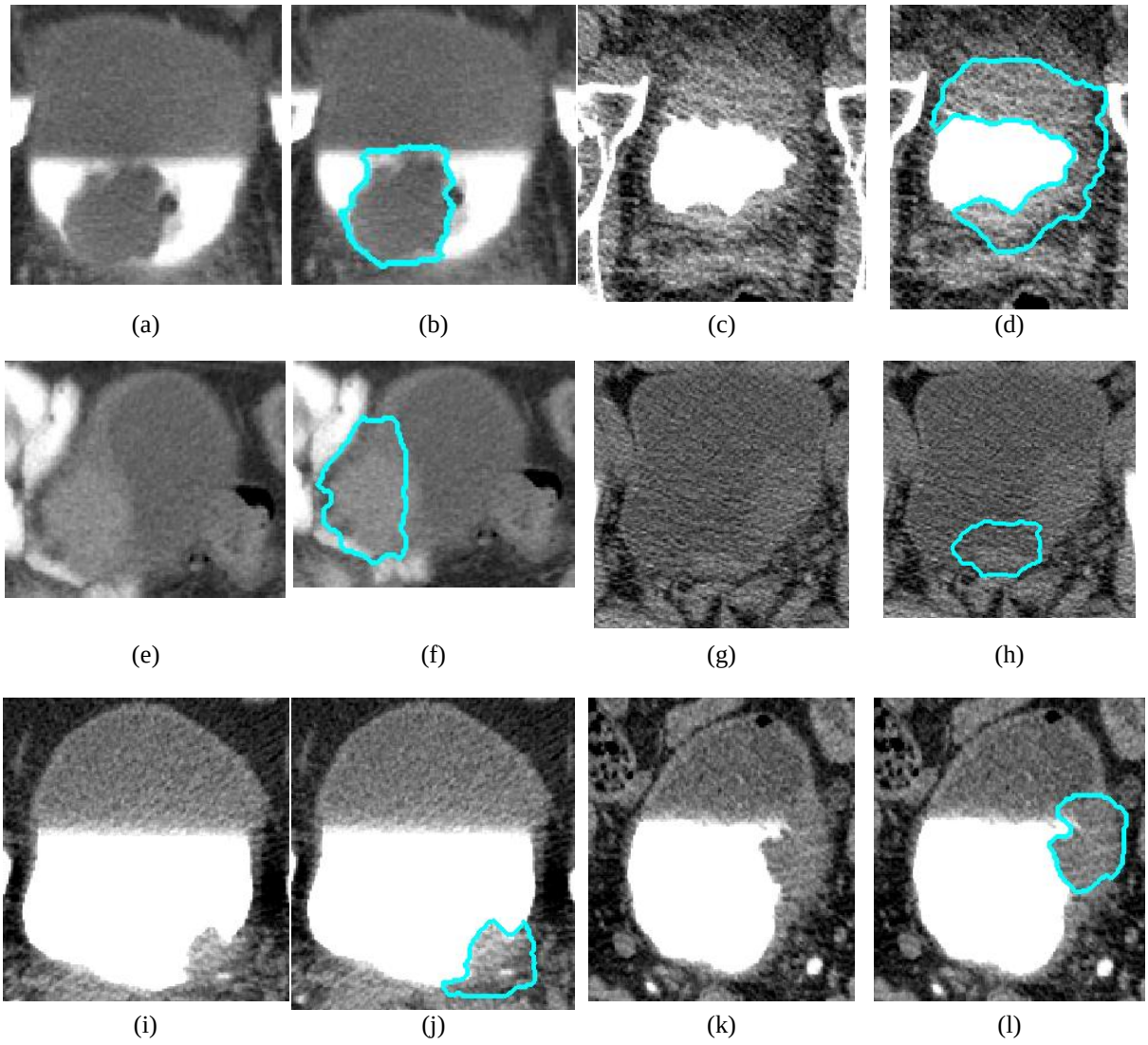

Fig. 1. Examples of CTU scans of bladder cancers from the test data set. The blue outlines are the AI-CALS segmentation. The following scores from the independent test set were obtained from training on the combined set of 84 bladder cancers. The stage T2 cancer in (a), (b) was classified accurately as T2 or above, with scores of LDA=100, SVM=43.7, BPNN=100, and RAF=70.8, above the threshold of each classifier. The stage T2 cancer in (c), (d) was classified accurately with scores of LDA=67.7, SVM=91, BPNN=96, RAF=70.8. The stage T2 cancer in (e), (f) was classified accurately with scores of LDA=62, SVM=84.3, BPNN=92.7, RAF=60.2. The stage T3 cancer in (g), (h) was misclassified, with scores of LDA=3.5, SVM=10.4, BPNN=0, and RAF=1.2, falling below the thresholds. The stage T2 cancer in (i), (j) was misclassified with scores of LDA=26.4, SVM=19.5, BPNN=23, and RAF=9.6. The stage T3 cancer in (k), (l) was misclassified with scores of LDA=26.4, SVM=19.5, BPNN=23, and classified accurately with a score of RAF=36.1.