Advancing Machine Learning for Precision Medicine and AI-Driven Clinical Decision Support for Bladder Cancer Treatment Response Assessment and Survival Prediction

by

Di Sun

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Nuclear Engineering and Radiological Sciences)
in the University of Michigan
2025

Doctoral Committee:

Professor Lubomir M. Hadjiyski, Co-Chair
Professor Martha M. Matuszak, Co-Chair
Professor Heang-Ping Chan
Professor Phillip L. Palmbos
Professor Sara A. Pozzi
Assistant Professor Liyue Shen

Di Sun

disun@umich.edu

ORCID iD: 0000-0001-6578-5142

© Di Sun 2025

Dedication

To my family, especially my parents – Shulan Xu and Xiguang Sun.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisors, Dr. Lubomir Hadjiyski and Dr. Martha Matuszak, for their invaluable guidance, patience, and encouragement throughout my PhD research. Their insights and expertise have been instrumental in shaping both my PhD journey and this dissertation. I am also profoundly grateful to my committee members, Dr. Heang-Ping Chan, Dr. Phillip Palmbos, Dr. Liyue Shen, and Dr. Sara Pozzi, for their constructive feedback and support, which significantly improved the quality of this work. I would also like to extend my sincere appreciation to Dr. Vikas Gulani and Dr. Chuan Zhou for their guidance and insightful discussions, which enriched my research in meaningful ways.

For an interdisciplinary study like this, encompassing engineering, physics, computer science, and medicine, collaboration with clinical experts has been indispensable. I am deeply thankful to my colleagues at Michigan Medicine—Dr. Richard Cohan, Dr. Phillip Palmbos, Dr. Elaine Caoili, Dr. Ajjai Alva, and Dr. Alon Weizer—for their professional expertise in radiology, oncology, urology, and clinical practice. Their advice served as a vital bridge between AI engineering and real-world medical applications. I am also grateful to my collaborators from the FDA, Dr. Kenny Cha and Dr. Ravi Samala, for providing novel perspectives that enhanced both the clinical relevance and technical depth of this research.

My time in the CAD-AI lab has been incredibly warm and memorable. The biweekly lab meetings with professors and research students — Mingjie Gao, Yifan Wang, Yahui Wang, and

Tao Chen—greatly enriched my understanding of other subfields, including breast and lung cancer research. I also have had the opportunity of working on fascinating projects and forming lasting friendships with John Gormley, Grace Bruno, Sagar Patil, and Rain Tarango. Together, we navigated the challenges of research and supported one another. The camaraderie within the CAD-AI lab made this journey not only more intellectually enriching but also truly enjoyable.

I would also like to acknowledge academic support from the Department of Nuclear Engineering and Radiological Sciences. Special thanks to the department chair, Todd Allen, and graduate student coordinator, Garnette Roberts, whose invaluable assistance with administrative matters ensured a smooth academic experience.

Beyond academia, my time in Ann Arbor was made unforgettable by the wonderful friends I met along the way. I am deeply grateful to Yawen, Can, and Yuling for their warm support and their exceptional cooking, which brought me a comforting taste of home whenever I felt homesick. Qiong, Yung-Wen, Shuhan, Emily, Chuwen, and Kexin, your liveliness and joyful company made my time here much more vibrant. Sally and Sophia, our countless shared moments, created a bond that made living abroad an adventure rather than a hardship.

I also hold immense gratitude for my lifelong friends, who have stood by me despite the years and distance. Their constant support and care have given me strength in moments of uncertainty. Thank you, Lan, Zhixin, Yan, Wuming, and Liya—you have been my anchor throughout this journey.

To my family—especially my parents and my brother—your unconditional love and support have been my greatest source of strength. This journey would not have been possible without you. I am also profoundly grateful to my uncles, who have provided invaluable guidance and support during pivotal moments in my life.

Finally, to my beloved Trung—your unwavering support, patience, and tenderness have been my rock. Our broad conversations about research, life, society, and universe have been distillated to another format of substance. Your warmth and fortitude have enriched my life.

This research was made possible through financial support from the National Institutes of Health (NIH): U01-CA232931. Additionally, portions of this work have been disseminated in various publications: Chapter 2 was included in *SPIE Proceeding Collections* [1]; work from Chapter 3 was published in Cancers [2, 3] and is currently under consideration at Scientific Reports (Nature family journal); Chapter 2.1 was published in Tomography [4]; and Chapter 5 will be featured in the upcoming 2025 *SPIE Proceeding Collections*.

Table of Contents

Dedication ii

Acknowledgements iii

List of Tables xi

List of Figures xiii

Abstract xviii

Chapter 1 Introduction 1

1.1 Motivation 1

1.2 Problem Definition 5

1.3 Contribution Summary 5

1.3.1 Additional Contributions 8

Chapter 2 Bladder Cancer Treatment Response Assessment to Neoadjuvant Chemotherapy 10

2.1 Bladder Cancer Treatment Response Assessment in CT Urography by Using Deep-learning and Radiomics 10

2.1.1 Abstract 10

2.1.2 Introduction 11

2.1.3 Methods and Materials 12

2.1.4 Results 20

2.1.5 Discussion 25

2.1.6 Conclusion 25

2.2 Multi-institutional Study of Radiomics Model for Assessment of Treatment Response to Neoadjuvant Chemotherapy of Bladder Cancer 26

2.2.1 Abstract 26

2.2.2 Introduction 26
2.2.3 Methods and Materials 26
2.2.4 Results 29
2.2.5 Conclusion 30
Chapter 3 Survival Prediction of Patients with Bladder Cancer After Cystectomy 31
3.1 Survival Prediction of Patients with Bladder Cancer after Cystectomy Based on Clinical, Radiomics, and Deep-learning Descriptors 31
3.1.1 Abstract 31
3.1.2 Introduction 31
3.1.3 Methods and Materials 33
3.1.4 Results 40
3.1.5 Discussion 46
3.1.6 Conclusion 48
3.2 Survival Prediction of Bladder Cancer Patients after Surgery based on Machine Learning Models with Radiomics and Genomic Descriptors 48
3.2.1 Abstract 48
3.2.2 Introduction 49
3.2.3 Methods and Materials 49
3.2.4 Results 50
3.2.5 Conclusion 53
3.3 Outcome Prediction Using Multi-modal Information: Integrating Large Language Model-extracted Clinical Information and Image Analysis 54
3.3.1 Abstract 54
3.3.2 Introduction 55
3.3.3 Methods and Materials 56
3.3.4 Results 65

3.3.5 Discussion 72

3.3.6 Conclusion 76

3.4 Evaluating the Reliability of Large Language Models for Clinical Data Extraction in Bladder Cancer Prognosis 76

3.4.1 Abstract 76

3.4.2 Introduction 77

3.4.3 Methods and Materials 78

3.4.4 Results 84

3.4.5 Discussion 94

3.4.6 Conclusion 97

Chapter 4 Computerized Decision Support for Bladder Cancer Treatment Response Assessment in CT Urography: Effect on Diagnostic Accuracy in Multi-institution Multi-specialty Observer Study 99

4.1 Abstract 99

4.2 Introduction 99

4.3 Methods and Materials 100

4.3.1 Data Set 100

4.3.2 Computerized AI-based Decision Support System for Treatment Response Assessment (CDSS-T) 101

4.3.3 CAD Score 102

4.3.4 Observer Performance Study 102

4.3.5 Statistical Analysis 104

4.4 Results 108

4.4.1 Overall Results for All Cancers 108

4.4.2 Easy vs. Difficult Cancer Subsets 110

4.4.3 Experienced vs. Inexperienced Observers 111

4.4.4 Multi-Specialty Observers 112

4.4.5 Multi-Institution Observers 113

4.4.6 Inter- and Intra-Observer Variability 114

4.5 Discussion 116

4.6 Conclusion 118

Chapter 5 Confidence Estimation of AI Model Output Affecting Decision Making for Bladder Cancer Treatment Response Assessment 119

5.1 Confidence Estimation of AI Model Output for Bladder Cancer Treatment Response Assessment in CT Urography 119

5.1.1 Abstract 119

5.1.2 Introduction 120

5.1.3 Methods and Materials 120

5.1.4 Results 124

5.1.5 Discussion 126

5.1.6 Conclusion 127

5.2 Observer Study: Impact of Case Complexities and Physician Characteristics on AI-Assisted Treatment Response Assessment in Bladder Cancer 127

5.2.1 Abstract 127

5.2.2 Introduction 128

5.2.3 Methods and Materials 129

5.2.4 Results 131

5.2.5 Discussion 132

5.2.6 Conclusion 134

Chapter 6 Summary and Future Work 135

6.1 Summary 135

6.2 Future Work 137

6.3 Conclusion 139

Bibliography 140

List of Tables

Table 2-1: Details of In-House Radiomics features. 17

Table 2-2: Correlation between In-House Radiomics and PyRadiomics features. 21

Table 2-3: Classification results based on In-House Radiomics model (In-House Radiomics features and Random Forest classifier) using the test set. 24

Table 2-4: CTU data collected from three institutions. 27

Table 3-1: Patient characteristics (n = 163). 41

Table 3-2: The *p*-values of survival prediction based on individual and combined descriptors. The critical α value for statistical significance was adjusted to be α = 0.017 (α = 0.05/3) according to the Bonferroni correction for multiple hypothesis testing. 45

Table 3-3: The average performances of the LLMs in clinical descriptor extraction with User1 prompt and User2 prompt, and the survival prediction performances of the C models (*n*=163). The entire set was a test set because none of the cases were used for training the LLMs or the C models. 66

Table 3-4: Performances of survival predictive models C, CR, CD, CRD on the independent test set (n=64), measured in AUC, alongside the accuracy of LLMs in extracting clinical descriptors. Clinical information was extracted manually (reference standard) and by LLMs with User1 prompt or User2 prompt. 68

Table 3-5: The direct survival predictions by GPT-4.0 with the SP prompt (survival prediction prompt). AUCs were calculated for the predictions of test sets of 163 patients and 64 patients. The LLM-cited sources, along with the predictions, were examined. Irrelevant sources cited by the LLM were defined as hallucinations. 72

Table 3-6: Fleiss' Kappa interpretation for evaluating agreement across experimental scenarios. 84

Table 3-7: LLM variability impacted by model types, model evolution over time, and input length and order. 86

Table 3-8: Summary of recommendations for the practice of applying LLMs in clinical information retrieval. 98

Table 4-1: Characteristics of the patients and cancers used in the observer study. 101

Table 4-2: Observers of different specialties and from different institutions. 104

Table 4-3: AUCs of observers with and without CDSS-T. The standard deviation of the AUCs of 17 observers was calculated for both the without CDSS-T and with CDSS-T conditions. ... 109

Table 4-4: Performance comparisons with and without CDSS-T for easy and difficult subsets. 111

Table 4-5: Performance comparison between experienced and inexperienced observers for the total of 157 lesion pairs. 111

Table 4-6: Performance comparison between observers from different specialties for the total of 157 lesion pairs. 113

Table 4-7: Performance comparison between observers from different institutions for the total of 157 lesion pairs. 114

Table 4-8: Diagnostic performance in terms of AUC of physicians without and with the CDSS-T aid for the assessment of complete response to neoadjuvant chemotherapy on the first 51 cases in each observer's individually randomized reading list. The iMRMC package provided standard deviation value along with each AUC. The mean and standard deviations of AUC values for the 17 observers with and without CDSS-T for both the original and the repeated evaluations were calculated. 115

Table 5-1: SVCE method. Variability was estimated from four feature sets extracted from four different lesion segmentation methods. Each feature set was input to the radiomics model for treatment response prediction. 125

Table 5-2: SVCE method. Variability was estimated from four hybrid ROI sets extracted from four different lesion segmentation methods. Each hybrid ROI set was input to the DL-CNN model for treatment response prediction. 125

Table 5-3: CMVCE method. The variability was estimated from the radiomics and the DL-CNN model scores when lesion segmentation was fixed to be AI-CALS_R1. 125

Table 5-4: CMVCE method. The variability was estimated from the radiomics and the DL-CNN model scores when lesion segmentation was fixed to be AI-CALS_R2. 126

Table 5-5: CMVCE method. The variability was estimated from the radiomics and the DL-CNN model scores when lesion segmentation was fixed to be AU-Net. 126

Table 5-6: CMVCE method. The variability was estimated from the radiomics and the DL-CNN model scores when lesion segmentation was fixed to be U-Net. 126

Table 5-7: AUCs of observers with and without CDSS-T for each subgroup. 133

List of Figures

Figure 1-1: Estimated number of cancer incidence worldwide in 2022, including all ages and both sexes. The total estimated incidence were 19,976,499, with bladder cancer ranking 9th [22]. 2

Figure 1-2: Geographical distribution of the age-standardized rate per 100, 000 incidence of bladder cancer across continents in 2022, including all ages and both sexes [22]. 3

Figure 1-3: Estimated number of bladder cancer diagnosed cases and deaths by sex in the U.S. in 2022. Bladder cancer ranks 4th among men [23]. 3

Figure 1-4: Overview of the dissertation's key contributions, core techniques, and the logical flow connecting individual studies into a cohesive body of work. 8

Figure 2-1: Examples of CTU scan. (a) A 3D volume rendering of CTU scan in which the bladder and ureters are visible. (b) An axial slice of a CTU scan in which the bladder is partially filled with intravenous (IV) contrast material. 14

Figure 2-2: Model design. The CTU was first segmented, and radiomics and deep-learning features were extracted from the segmented region. These features were then fused using a BPNN to generate a score indicating the likelihood of a complete response to NAC..... 15

Figure 2-3: Hybrid ROIs (hROIs) generated from a pre- and post-treatment lesion pair. In this example the cancer was stage T3 in pre-treatment, and stage T2 in post-treatment scan, resulting in a label of non-complete response (>T0 stage after treatment). 18

Figure 2-4: Subsets of hybrid ROIs. (a) A subset of complete responders (T0). (b) A subset of non-complete responders (>T0). 19

Figure 2-5: Schematic representation of the DL-CNN architecture used for image classification. The network consists of two convolutional layers, C1 and C2, each of them is followed by a max-pooling layer and a local response normalization layer, two locally connected layers, L3 and L4, and a fully connected layer, FC10. DL-CNN assessed the input hROIs and output the cancer treatment response score. 20

Figure 2-6: Heat map of the In-House radiomics features extracted from test set, visualizing the magnitude of extracted features. The left side is the features extracted from the pre-treatment scans, and the right side is features extracted from the post-treatment scans. 22

Figure 2-7: Examples of pre- and post- treatment lesion pairs and the response prediction by our trained models. The lesion of each scan was segmented by AI-CALS and was marked as

blue (pre-treatment scan) and yellow (post-treatment scan) outline. Cases (a) and (b) were complete responders. Both the DL-CNN and the radiomics models provided correct estimates. Cases (c) and (d) were non-complete responders. Both the DL-CNN and radiomics models provided the correct estimates. Case (e) was a complete responder. Both the DL-CNN and the radiomics model estimates were wrong. Case (f) is a non-complete responder. The DL-CNN model made the correct estimate but the radiomics model provided the wrong estimate. Case (g) is a non-complete responder. The estimate of the radiomics model was correct but the estimate of the DL-CNN model was wrong. 24

Figure 2-8: Study-population diagram. Out of 835 patients, 258 patients were identified for this study. Inclusion criteria include (1) patient had at least one pair of pre- and post-NAC CTU scans with acceptable image quality, (2) patient underwent neoadjuvant chemotherapy (NAC), and (3) patient underwent radical cystectomy, so that the ground truth of complete response or not to NAC can be determined..... 28

Figure 2-9: Examples of complete responders (CRs) and non-complete responders (NCRs). For each scan pair, the left side is the pre-treatment scan, the right side is the post-treatment scan. The two scan pairs on the first row are from Institution A, the two scan pairs on the second row are from Institution A, and the two scan pairs on the first row are from Institution C. 29

Figure 2-10: Classification results for treatment response assessment, including AUCs and ROC curves for three institutions. 30

Figure 3-1: Predictive model development process of this study. We collected clinical, histopathological information and CTU scans. We built and validated survival prediction models based on individual or combined descriptors. C = clinical descriptors; R = radiomics descriptors; D = deep-learning descriptor; CR = clinical + radiomics descriptors; CD = clinical + deep-learning descriptors; CRD = clinical + radiomics + deep-learning descriptors. DL-CNN = Deep learning–convolutional neural network..... 33

Figure 3-2: Study cohort diagram. Out of 781 patients, 163 patients (mean age 64±9 years; M:F 131:32) were identified for this study. Patients who (1) underwent neoadjuvant chemotherapy and had pre- and post-chemotherapy CTU scans, (2) underwent radical cystectomy, and (3) had follow-up data to determine their five-year survival status were eligible..... 35

Figure 3-3: CTU scan examples. (a) A 3D volume rendering CTU scan showing the ureter, bladder, and bladder lesion. (b) CTU scans that are potentially challenging for lesion segmentation. (b)-1: without contrast; (b)-2: delayed-contrast phase; (b)-3: bladder wall thickening; (b)-4: image artifact. (c) Examples of CTU scans in early contrast phase and without artifacts. The blue outlines are the AI-CALS segmentation outlines. 37

Figure 3-4: Structure of BPNN. The input x_i (i=1,2,3,...,n) are the descriptors. The hidden layer contains 13 nodes. The output y is the likelihood score assessing the survival of each patient..... 39

Figure 3-5: Histogram of patients' age at surgery. 42

Figure 3-6: Heatmap of Pearson correlation between any pair of the MI-selected radiomics features. Most of the correlation values were within ±0.3 which indicates a weak correlation. .. 43

Figure 3-7: ROC curves of survival prediction on the test set by using different descriptors. (a) ROC curves and AUC values for the individual descriptor. (b) ROC curves and AUC values for combined descriptors. (c) Direct comparison of the ROC curves of C and CRD descriptors. C = clinical descriptors; R = radiomics descriptors; D = deep-learning descriptor; CR = clinical + radiomics descriptors; CD = clinical + deep-learning descriptors; CRD = clinical + radiomics + deep-learning descriptors. 44

Figure 3-8: Kaplan–Meier survival probability curves for the alive and deceased groups stratified by C and CRD models. The survival probability of the two groups (deceased and alive) differed significantly based on either C ($p < 0.001$) or CRD ($p < 0.001$) models. The CRD model achieved better separation between the two groups compared to the C model. 46

Figure 3-9: Visualization of the magnitudes of 32 genomic markers obtained from collected tissue samples on (a) the training set, (b) the validation set, and (c) the test set. 50

Figure 3-10: Visualization of the magnitudes of 92 radiomics features extracted from the CTU sans on (a) the training set, (b) the validation set, and (c) the test set. 51

Figure 3-11: Visualization of the magnitudes of selected 5 genomic markers obtained from collected tissue samples on (a) the training set, (b) the validation set, and (c) the test set. 52

Figure 3-12: Visualization of the magnitudes of selected 8 radiomics features extracted from the CTU sans on (a) the training set, (b) the validation set, and (c) the test set. 52

Figure 3-13: Kaplan-Meier survival plots for (a) RF Model: The median survival times for the two groups (Group 1: deceased or Group 2: alive) are 1.25 and 2 years, respectively, with a significant difference in survival probability ($p = 0.014$). (b) GEN Model: The median survival times for the two groups (Group 1: deceased or Group 2: alive) are 0.195 and 2 years, respectively, with a significant difference in survival probability ($p < 0.001$), and (c) RF-GEN Model: The median survival times for the two groups (Group 1: deceased or Group 2: alive) are 0.205 and 2 years, respectively, with a significant difference in survival probability ($p < 0.001$). 53

Figure 3-14: Development and validation processes of predictive models C, R, D, CR, CD, and CRD, using clinical and radiomics descriptors as well as a deep learning assessment. 57

Figure 3-15: CTU images of a bladder cancer patient (a) pre- and (b) post-neoadjuvant chemotherapy, with the initial and residual cancer indicated by the green arrow. The CTU scans were interpreted by the radiologist (R.H.C.), who marked the volume of interest (VOIs) and the slice where the lesion had the most prominent appearance. The lesion was then automatically segmented with our in-house-developed AI-CALS model [46] within the VOI. .. 61

Figure 3-16: AUCs of LLM-aided models (C, CR, CD, CRD) compared to AUCs based on manually extracted clinical descriptors. Clinical information was extracted manually (reference standard) and by LLMs with User1 prompt or User2 prompt. 69

Figure 3-17: ROC curves and scatter plots showing that different models have the same AUC values but different predictions (survival likelihood scores). (a-1) ROC curves of predictions by model C-manual and model C-User2-GPT3.5-full-reports. Both had AUC=0.82±0.06. (a-2) Scatter plot of predictions (survival likelihood) by model C-manual and model C-User2-GPT3.5-full-reports. Pearson correlation coefficient is 0.90. (b-1) ROC curves of predictions by model C-User2-Dolly-snippet-report (accuracy=87%, AUC=0.77±0.04) and model C-User2-GPT4.0-snippet-reports (accuracy=96%, AUC=0.77±0.04). (b-2) Scatter plot of predictions (survival likelihood) by model C-User2-Dolly-snippet-report and model C-User2-GPT4.0-snippet-reports. Pearson correlation coefficient is 0.95. 70

Figure 3-18: Kaplan–Meier survival curves of C and CRD models (User1/User2-GPT4.0-snippet-reports) on the independent test set (n = 64). The clinical descriptors were extracted manually (reference standard) or by LLM through User1 prompt or User2 prompt. Stratification of the two groups (predicted deceased and predicted alive) by C and CRD (extracted by manual or GPT4.0-snippet-reports) achieved statistical significance ($p \leqslant 0.05$).. 71

Figure 3-19: Study design. 163 EMRs were input to various LLMs to perform information retrieval tasks. Each report was input 10 times independently to assess the reproducibility of each LLM. The extracted information was input to downstream models to perform survival prediction. One model was a nomogram that was solely based on clinical information and the other was a multi-modal model that was based on clinical information and image features. 79

Figure 3-20: Heatmap representing the frequency of extracted information for the three clinical variables (pT, pN, and LVI) across the 10 runs (N = 163) for the six LLMs: Dolly-v2-12b, Vicuna-v1.3-13b, Llama-2.0-13b, Llama-3.3-70b, GPT-4 Turbo, GPT-4o. The input for all the six models was the snippet reports and in their original sequence. The y-axis displays the categories for the specified clinical variables: pT (N/A, invalid, pT0, pTis, pTa, pT1, pT2, pT3, pT4), pN (N/A, invalid, N0, N1, N2, N3), and LVI (N/A, invalid, Yes, No), in which 'N/A' indicates that relevant information in the input report for LLMs was unavailable or absent, and 'invalid' signifies that the LLM's output could not be interpreted as meaningful or valid for the given variable. The x-axis corresponds to individual cases. Color intensity indicates the frequency of each category in the data. The color bar on the right provides a reference for the frequency scale. Red rectangles highlight the ground truth values for each case, enabling comparison with the LLM extractions. 90

Figure 3-21: Performance comparison (measured by AUC) of survival prediction across 10 runs from different conditions: (1) different model, (2) GPT-4o across 4 time points, (3) GPT-4 Turbo across snippet or full reports and original or random input sequence, (4) GPT-4o across snippet or full reports and original or random input sequence. The black dots marked the outliers that fall outside of the interquartile range (IQR) which is calculated as IQR=Q_3-Q_1, where Q_1 is the first quartile and Q_3 is the third quartile. The three horizontal dashed lines represent the AUCs of models based on the ground truth of clinical variables. 93

Figure 4-1: The graphical user interface for reading with and without our computerized decision support system (CDSS-T) for bladder cancer treatment response assessment. (a), (b) The pre- and post-treatment CTU scans are shown side-by-side. (c) The observer estimates the treatment response. (d) The observer is shown the CAD score and the score distribution of the

two classes as reference. (e) The observer may revise their treatment response assessment after considering the CAD score. 106

Figure 4-2: One step of observer study. Each observer performed treatment response assessment without CDSS-T. 107

Figure 4-3: CAD score indicating the treatment response likelihood of each case was provided to each observer. The observer can choose to update the treatment response assessment. 108

Figure 4-4: Results of the observer performance study. (a) AUCs of the 17 observers with and without CDSS-T. The performance of the CDSS-T system is shown with the dashed line. The performance of all but one observer (#6) increased with using CDSS-T. (b) Average ROC curves with and without CDSS-T. 110

Figure 4-5: ROC curves comparison for observers of different specialties. 112

Figure 4-6: ROC curves comparison for observers from different institutions. (a) Performance of observers from four institutions. (b) Performance of oncologists from three institutions. 113

Figure 5-1: Proposed methods to estimate the confidence of ML/AI model outputs. (a) Segmentation Variability-Based Confidence Estimation (SVCE). (b) Classification Model Variability-Based Confidence Estimation (CMVCE). 122

Figure 5-2: Segmentation of a cancerous lesion on CTU scan using AI-CALS_R1, AI-CALS_R2, AU-Net, and U-Net. 125

Figure 5-3: Paired violin plots for individual AUC w/o and w/ CDSS-T aid. (a) Experienced/inexperienced physicians (EP/IEP) on the easy and difficult subsets, w/o and w/ CDSS-T aid. (b) Radiologists and oncologists on the easy and difficult subsets, w/o and w/ CDSS-T aid. 134

Abstract

Bladder cancer is a common and clinically challenging malignancy with a substantial impact on global health, making it crucial to develop advanced methodologies for diagnosis, prognosis, and treatment planning. This dissertation explores the application of advanced machine learning (ML) and artificial intelligence (AI) techniques to improve clinical decision-making, enhance assessment of treatment response, and predict patient outcomes for bladder cancer. The research developed and evaluated several AI-driven models, focusing on integrating multi-modal data sources, including imaging, clinical, and genomic information, to build predictive tools that can guide personalized treatment strategies.

A key contribution of this work was the development of a deep-learning convolutional neural network model combined with radiomics features from CT urography, demonstrating the potential of AI in assessing bladder cancer treatment response to neoadjuvant chemotherapy. This model achieved an area under the receiver operating characteristic curve (AUC) of 0.77-0.80, highlighting its feasibility and potential for clinical implementation.

Additionally, the dissertation introduced a multi-modal model (termed CRD) for five-year survival prediction of bladder cancer patients post-cystectomy. This model, which integrated clinical, radiomics, and deep-learning features, achieved an AUC of 0.87, demonstrating the importance of combining diverse data sources for more accurate survival predictions. An enhanced CRD model was proposed by incorporating large language models (LLMs) into the modeling to improve efficiency by automating clinical information extraction from unstructured electronic medical records. The results showed the LLM-assisted model achieved survival prediction AUCs comparable to those based on manually curated data, emphasizing the potential of LLMs to automate clinical data extraction. LLM reliability and its impact on downstream prediction model were further investigated. The results showed GPT-4 and Llama held outstanding consistency in information retrieval for predictive modeling.

Furthermore, an observer study was conducted and its result demonstrated that AI-based decision support systems (CDSS-T) significantly improved diagnostic accuracy of 17 clinicians

when assessing bladder cancer treatment response. The use of AI tools to assist in clinical decision-making resulted in a notable increase in diagnostic performance, with AUC improving from 0.73 to 0.77 ($p = 0.002$). These findings highlight the practical benefits of integrating AI into clinical practice, suggesting that AI can complement and enhance clinician expertise, particularly in complex diagnostic scenarios.

The dissertation also explored methods for confidence estimation of AI outputs and examined the influence of case complexity and physician characteristics on AI-assisted assessments. This is crucial for providing the user with an estimate of the trustworthiness of AI output, allowing the user to properly weight the AI recommendation with their own assessment, thus maximizing the benefits of AI-assisted decision-making.

In conclusion, this work advances the integration of AI and ML in bladder cancer prognosis and treatment response assessment. By developing robust predictive models, evaluating AI in a pre-clinical multi-institutional observer study, and investigating the reliability and trustworthiness of ML models, this research underscores the transformative potential of AI in improving patient care. Future research will focus on integrating multi-institutional data to enhance model generalizability, developing longitudinal models to track disease progression, and conducting prospective study of the impact of AI tools in clinical settings. Ethical considerations and human-AI collaboration will be essential to ensure AI supports, rather than replaces, clinician judgment. This dissertation contributes to integrating AI into clinical workflows for more personalized, accurate, and timely interventions to improve outcomes for bladder cancer patients worldwide.

Chapter 1 Introduction

1.1 Motivation

The application of radiation in medicine has significantly shaped modern healthcare[5-7]. Medical imaging techniques such as X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, and positron emission tomography (PET) have become indispensable in diagnostics. These modalities enable non-invasive visualization of internal structures, facilitating early disease detection and treatment planning. By providing detailed anatomical and functional information, they enhance diagnostic accuracy and enable effective monitoring of various medical conditions, including cancer.

Beyond diagnostics, by integrating artificial intelligence (AI) into medical imaging, the developed methodologies and models have the potential to transform cancer diagnosis, prognosis and treatment [8-10]. AI models' ability to analyze vast amounts of data and differentiate abnormal from normal patterns has led to significant advancements in cancer diagnosis, prognosis, and treatment.

Deep learning (DL) models, for instance, have demonstrated proficiency in detecting and classifying tumors from imaging modalities [11], reducing interobserver variability [12] and improving early detection rates [13]. Machine learning (ML) models could enhance diagnostic accuracy by interpreting multi-modal data [14], including medical images, pathological slides, and genomic data. In prognosis, AI-powered models predict disease outcomes by analyzing clinical, pathological, and molecular data [15]. These models help clinicians stratify patients based on risk, facilitating personalized treatment strategies. AI also contributes to treatment

planning by predicting responses to therapies and identifying potential side effects [16, 17]. Machine learning models can analyze patient data to recommend optimal treatment regimens, enhancing intervention efficacy while minimizing adverse effects [18, 19].

Among various malignancies, bladder cancer is a significant health concern [20, 21], affecting the urinary tract and exhibiting high morbidity and mortality rates worldwide. It ranks as the 9th most common cancer globally (Figure 1-1). North America had the highest age-standardized rate among the continents (Figure 1-2). In the United States, it is the 4th most common cancer among men, with an estimated 61,700 diagnosed cases and an estimated 12,120 deaths (Figure 1-3). It is categorized into non-muscle-invasive and muscle-invasive types, each requiring distinct management approaches. Early detection and accurate staging are crucial for effective treatment and improved survival outcomes.

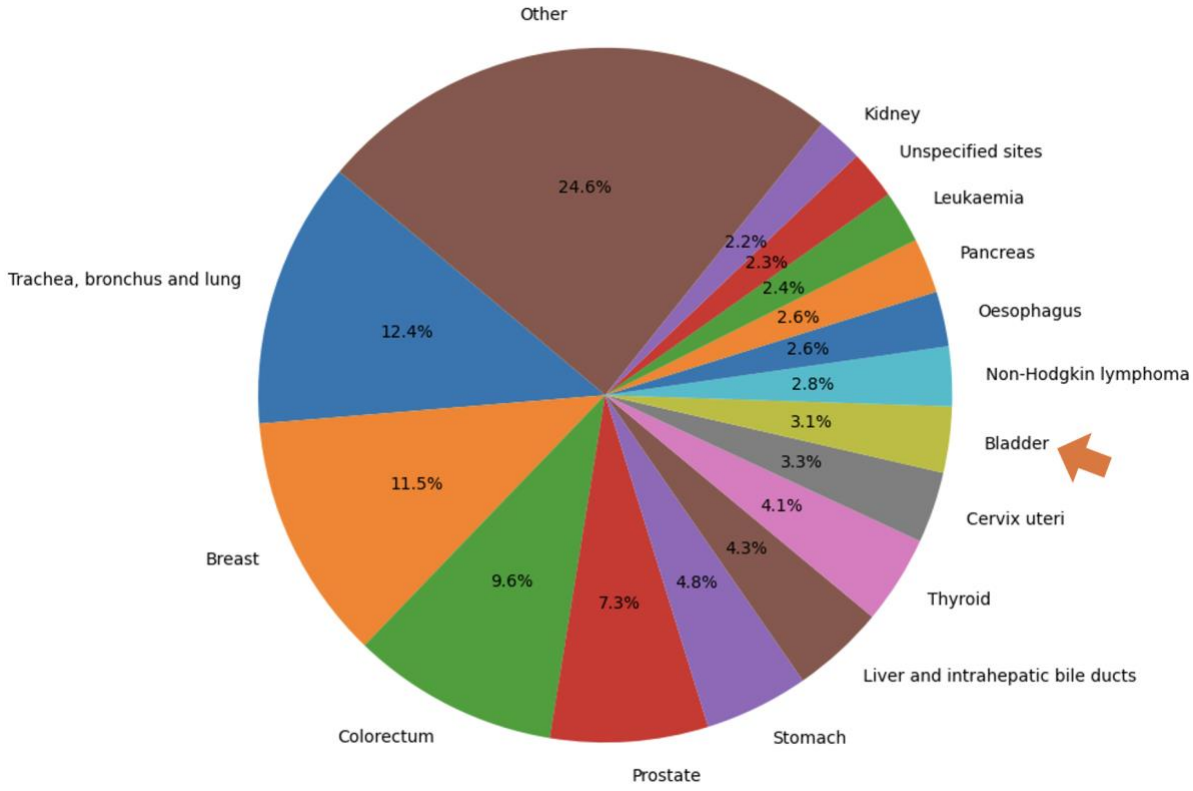


Figure 1-1: Estimated number of cancer incidence worldwide in 2022, including all ages and both sexes. The total estimated incidence were 19,976,499, with bladder cancer ranking 9th [22].

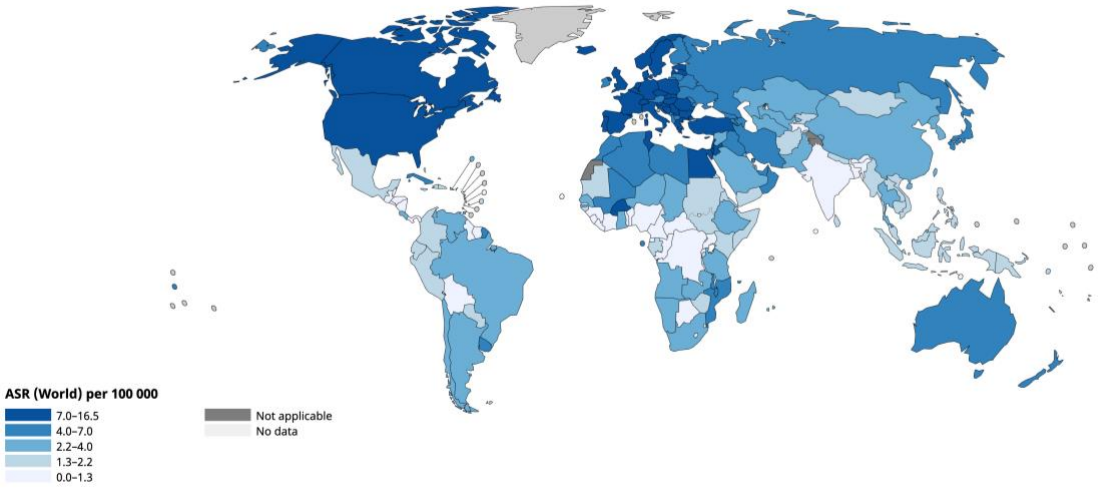


Figure 1-2: Geographical distribution of the age-standardized rate per 100, 000 incidence of bladder cancer across continents in 2022, including all ages and both sexes [22].



Figure 1-3: Estimated number of bladder cancer diagnosed cases and deaths by sex in the U.S. in 2022. Bladder cancer ranks 4th among men [23].

AI and DL models have been widely adopted to enhance diagnostic capabilities in bladder cancer. Six studies utilized DL-based algorithms had shown strong performance in analyzing cystoscopy, achieving high pooled sensitivity (89.7%) and specificity (96.1%) in detecting bladder tumors. The area under the receiver operating characteristic curve (AUC) ranged from 0.96 to 0.98 with an accuracy ranged from 85.6 to 96.9% [24-29]. The models

utilized included GoogleNet, ResNet-101, U-net, and LeNet. These ML models support clinicians by automation of lesion detection and by reducing oversight. In image-based assessment, radiomics has been combined with ML algorithms to predict muscle invasiveness of bladder tumors. For example, Sun et al. [30] developed a CT-based DL radiomics model that accurately predicted lymph node metastasis in bladder cancer patients, showing potential for aiding preoperative planning.

AI has also been applied in prognostic modeling. A review by Pinar et al. [31] emphasized that both ML and DL approaches have been utilized to predict recurrence and survival outcomes in bladder cancer, based on clinical, pathological, and imaging features. Such models help clinicians in tailoring personalized follow-up and treatment protocols.

In terms of treatment response, ML models have demonstrated their potential in predicting sensitivity to neoadjuvant chemotherapy (NAC) for muscle-invasive bladder cancer. A systematic review by Suartz et al. [32] reported that the models from 12 studies comprising 1523 patients achieved pooled sensitivities and specificities of 62% and 82%, respectively. Especially, the ML models with CT, genetic, and anatomopathologic data as input exhibited promising potential for predicting NAC response.

Despite these promising applications, the development of new AI/ML models with improved accuracy and robustness is still needed for the effective treatment response evaluation and survival prediction in bladder cancer. Moreover, comprehensive assessment of the impact of AI tools on clinical decision-making is also lacking. Further studies are needed to validate these technologies and to better understand their potential to support and enhance clinical workflows.

1.2 Problem Definition

This dissertation aims to develop predictive models that integrate advanced machine learning techniques for treatment response assessment, survival prediction, clinical decision-making in bladder cancer, and confidence estimation of AI model outputs.

Treatment response assessment can provide crucial information for optimizing therapeutic strategies for individual patients and contribute to the improvement of treatment protocols for broader patient populations. Survival prediction is a key aspect of cancer prognosis. Performing observer study using the AI-driven decision support system can evaluate the influence of AI tools on clinicians' decision-making processes. Confidence estimation of AI model outputs is crucial for determining the reliability of AI-generated assessments and their trustworthiness in clinical practice.

Together, these objectives not only explore the development of AI/ML models in bladder cancer treatment response monitoring and prognosis but also investigate their potential for adoption in clinical practice.

1.3 Contribution Summary

This dissertation presents a series of interconnected studies that collectively aim to enhance the assessment, prediction, and interpretation of treatment outcomes in bladder cancer, leveraging AI, radiomics, and multi-modal data integration. These contributions form a cohesive research trajectory that begins with model development for treatment response evaluation, advances toward robust survival prediction models, and culminates in the critical evaluation of AI-assisted decision-making in clinical tasks.

Figure 1-4 summarizes the dissertation's key contributions and core techniques, and illustrates the relationships between the different studies and their collective impact on addressing the target problem.

The key contributions of this dissertation are also summarized as follows:

- **Foundational Model Development**: A deep learning and radiomics-based model was developed for assessing bladder cancer lesions' response to NAC, establishing a baseline AI-driven approach for treatment response evaluation (Chapter 2.1) [1].
- **Generalizability Across Institutions**: Building on this, a multi-institutional validation study was conducted to assess the robustness and transferability of the developed radiomics model for NAC response prediction (Chapter 2.2).
- **Towards Long-Term Outcome Prediction:** Moving beyond treatment assessment, a multi-modal predictive model (CRD model) was designed to estimate five-year survival rates. This model integrates clinical (C), radiomics (R), and deep-learning descriptors (D) from both textual and imaging data (Chapter 3.1) [2].
- **Investigating Genomic Insights**: The research was extended to incorporate genomic descriptors with radiomics descriptors, enabling the prediction of two-year post-surgical survival (Chapter 3.2).
- **Efficiency Through Language Models**: To improve scalability and efficiency, large language models (LLMs) were incorporated into the CRD model pipeline to extract relevant information from unstructured clinical texts (Chapter 3.3) [3].
- **Evaluating AI Reliability**: The impact of LLM-derived information on model performance was critically assessed, offering insights into the reliability and limitations of LLMs in clinical information retrieval and their impact on the downstream CRD model (Chapter 3.4).

- **AI Impact on Physicians:** An observer study was conducted to evaluate how the AI model (Chapter 2.1) influences clinicians' decision-making in NAC response assessment, providing evidence on AI's practical utility in supporting clinical workflows (Chapter 4) [4].
- **Confidence Estimation in AI**: Methods were proposed to quantify the confidence of AI/ML model outputs, addressing a key concern in clinical adoption—trust and interpretability (Chapter 5.1).
- **Understanding Contextual Factors**: Finally, the study explored how case complexity and physician characteristics affect AI-assisted assessments, contributing to a nuanced understanding of human-AI collaboration in treatment response assessments in bladder cancer (Chapter 5.2).

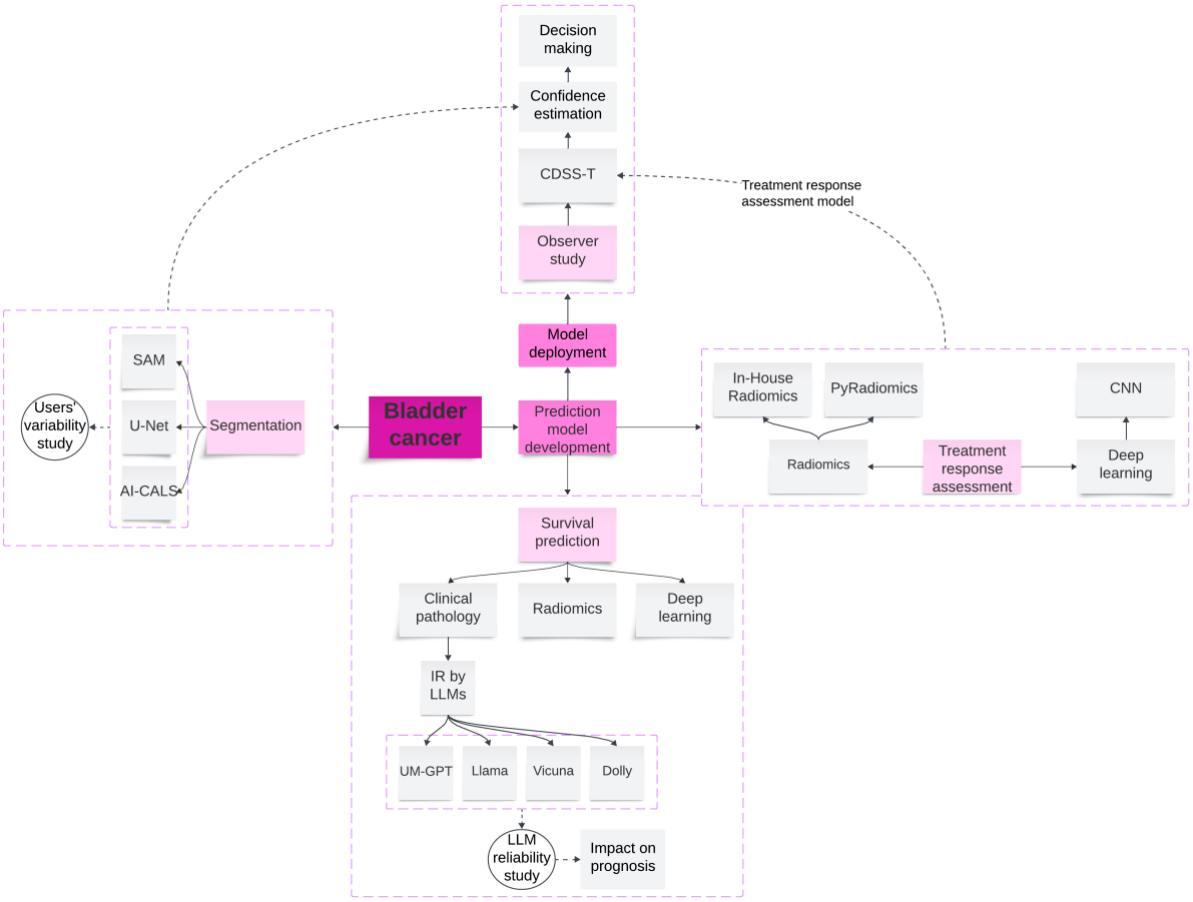


Figure 1-4: Overview of the dissertation's key contributions, core techniques, and the logical flow connecting individual studies into a cohesive body of work.

1.3.1 Additional Contributions

Several additional contributions, while not covered in detail within this dissertation, are presented here to provide a broader view of the PhD work. These studies were not elaborated upon for one of two primary reasons: either (1) they represent preliminary experiments and results or feasibility studies that have not yet been fully developed into comprehensive investigations, or (2) they involve collaborative efforts where the methodology or conceptual framework originated from this dissertation's work, but the studies were primarily conducted and led by other research teams.

These contributions provide additional evidence supporting the core research directions of this dissertation and demonstrate the broader applicability and value of the proposed approaches in different settings. Specifically, they include:

- Prediction of bladder cancer progression in immunotherapy treatment by long short-term memory (LSTM) model.
- Performing a multi-institutional observer study of AI-aided oncologists' performance on treatment response assessment of bladder cancer[33].
- Quantifying the impact of computer-aided diagnostic score on the diagnosis of functional seizures[34].
- Bladder cancer segmentation using U-Net based deep-learning.

Through these contributions, this dissertation seeks to advance the integration of AI and ML techniques in cancer research, specifically bladder cancer, enhancing treatment response monitoring, prognosis prediction, and the adoption of decision support systems into clinical use.

Chapter 2 Bladder Cancer Treatment Response Assessment to Neoadjuvant Chemotherapy

2.1 Bladder Cancer Treatment Response Assessment in CT Urography by Using Deep-learning and Radiomics

2.1.1 Abstract

We are developing deep-learning convolutional neural network (DL-CNN) and radiomics models to assist physicians in treatment response assessment in CT urography of bladder cancer to neoadjuvant chemotherapy (NAC). We collected a total of 264 pre- and post- treatment lesion pairs of 227 patients from University of Michigan hospital with IRB approval. The data were split into 3 sets by case: a training set including 35 complete responders (CRs) (T0 stage after treatment) and 113 non-complete responders (NCRs) (> T0 stage after treatment), a validation set including 5 CRs and 5 NCRs, and an independent test set including 19 CRs and 87 NCRs. The training set was used to train the models to classify CRs and NCRs and the selection of optimal models was guided by the validation set. The selected models were deployed on the test set to generate the likelihood score of CR of each pair. The classifying performance was evaluated by the area under ROC curve (AUC). The DL-CNN model achieved an AUC of 0.75 ± 0.06 on the test set. The radiomics model achieved an AUC of 0.76 ± 0.05. A combined DL-CNN and radiomics model increased the AUC to 0.77 ± 0.06. The results indicated the feasibility of using the DL-CNN model and radiomics model for assessing treatment response of bladder cancer.

2.1.2 Introduction

Neoadjuvant chemotherapy (NAC), administered prior to radical cystectomy, can significantly enhance the effectiveness of subsequent treatments for bladder carcinoma. NAC has been shown to achieve a complete response rate of approximately 30% and a 5-8% survival benefit at five years for patients with bladder cancer staged higher than T2 [35]. However, due to the heterogeneous genetic, epigenetic, and clinical nature of urothelial carcinoma, not all patients respond to NAC [36]. Additionally, the treatment can induce several toxicities, such as neutropenic fever, sepsis, mucositis, nausea, vomiting, malaise, and alopecia [37]. Therefore, early monitoring of bladder lesion response during the treatment cycle is essential to identify non-responding lesions, allowing patients to explore alternative treatments without enduring the toxicities of the entire chemotherapy course.

The assessment of NAC response can be approached from three main perspectives: clinical analysis, pathological biomarkers, and imaging evaluation. In clinical analysis, factors such as disease risk and smoking history can influence NAC response [38, 39]. Due to the histological heterogeneity of bladder cancer, investigating biomarkers is a complex process [40]. Recent studies have examined various biomarkers, including the pre-treatment neutrophil-to-lymphocyte ratio [41] and gut microbiome-based community signatures[42]. In terms of imaging, dynamic contrast-enhanced MRI has shown superior efficacy in distinguishing responders from non-responders compared to conventional MRI [43]. In one study, Fitoussi identified hypermetabolic residuals in the bladder and lymph nodes on FDG-PET/CT as predictors of response[44].

CT urography (CTU) plays a critical role in identifying urinary tract abnormalities, detecting enlarged lymph nodes that may harbor cancer, assessing tumor shape, size, and

location, and staging the disease [45]. Clinically, response to neoadjuvant chemotherapy (NAC) is commonly estimated using CTU by measuring changes in tumor size according to standardized criteria such as RECIST (Response Evaluation Criteria in Solid Tumors) or World Health Organization (WHO) tumor response criteria. However, these traditional size-based assessments can be limited in their ability to capture subtle morphological or biological changes, particularly in cases where tumors exhibit treatment-induced necrosis or fibrosis without significant size reduction. Moreover, inter-reader variability and delayed recognition of non-responders can hinder timely treatment adaptation.

To address these challenges, this study proposes a predictive model that integrates radiomics and deep-learning features extracted from CTU images, aiming to provide a more sensitive and objective tool for evaluating treatment response. By capturing high-dimensional imaging biomarkers beyond simple size metrics, the approach proposed in this study seeks to support physicians in making earlier and more accurate assessments of bladder cancer response to NAC.

2.1.3 Methods and Materials

2.1.3.1 Dataset

With approval from the Institutional Review Board (IRB), we collected CTU scans from 227 patients who underwent chemotherapy. Figure 2-1 shows an example of 3D CTU scan and axial slice of a CTU scan.

The CTU scans used in this study were acquired with GE Healthcare LightSpeed MDCT scanners, Issaquah, WA, USA, set at 120 kVp and 120–280 mA, with the slice interval ranging from 0.625 to 5 mm. The CTU series with the comparatively best quality were utilized regardless of whether they were performed after administration of contrast material or were non-contrast

images. The contrast-enhanced CTU scans were obtained 12 minutes after the initiation of the first bolus of a split-bolus IV contrast injection, and 2 minutes after the initiation of the second bolus of 175 ml of nonionic contrast material at a concentration of 300 mg iodine/ml.

Each patient had at least two CTU scans—one before NAC and another after NAC—allowing us to pair these scans as pre- and post-treatment scans. For patients with more than two scans, additional pre- and post-treatment pairs were obtained. In total, 264 pre- and post-treatment scan pairs were collected. The pathological cancer stage after radical cystectomy was used as the gold standard. The scan pairs were labeled as complete responders (CRs) or non-complete responders (NCRs).

We divided the 264 scan pairs by case into three sets: training, validation, and test sets. The training set consisted of 148 pairs (35 CRs, 113 NCRs), the validation set included 10 pairs (5 CRs, 5 NCRs), and the test set had remaining 106 scan pairs (19 CRs, 87 NCRs). The training and validation sets were used to obtain the optimal trained model, and the test set was kept independent to test the model. The majority of the test set cases were collected more recently than those in the training and validation sets, which was intended to simulate real-world clinical scenarios where the model must be robust and effective on new cases. This design reflects the assumption that the CAD model remains fixed in clinical applications and must be reliable for incoming new cases.

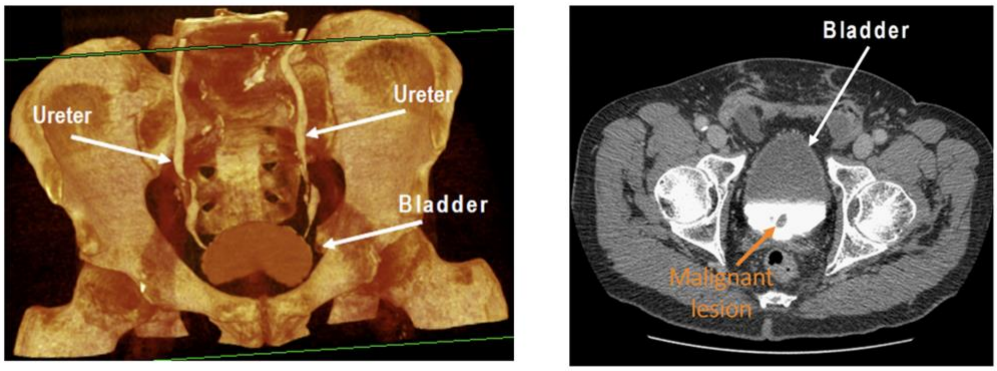


Figure 2-1: Examples of CTU scan. (a) A 3D volume rendering of CTU scan in which the bladder and ureters are visible. (b) An axial slice of a CTU scan in which the bladder is partially filled with intravenous (IV) contrast material.

2.1.3.2 Study Design

The model design is illustrated in Figure 2-2. Details of each model component are elaborated below.

With CTU as the input, we first performed segmentation, followed by extracting radiomics and deep-learning features from the segmented regions. Those features were used to classify CR and NCR. Radiomics and deep learning extract fundamentally different types of features from medical images. Radiomics relies on handcrafted features that capture predefined statistical, shape, and texture characteristics, offering high interpretability and suitability for smaller datasets, while deep learning automatically learns hierarchical and abstract features through neural networks, with limited transparency and usually requiring larger datasets. Combining both approaches leverages the strengths of each: radiomics contributes explainable, domain-relevant insights, while deep learning captures complex, non-linear patterns, resulting in more robust, accurate, and clinically informative models for medical imaging analysis.

The segmentation process requires as input a bounding box (region of interest (ROI) box) containing the cancer, which is manually marked. The manually marked bounding box may introduce variability in the extracted features and, consequently, affect the model's classification performance. In order to assess this variability, two users performed the segmentation to evaluate the model's robustness.

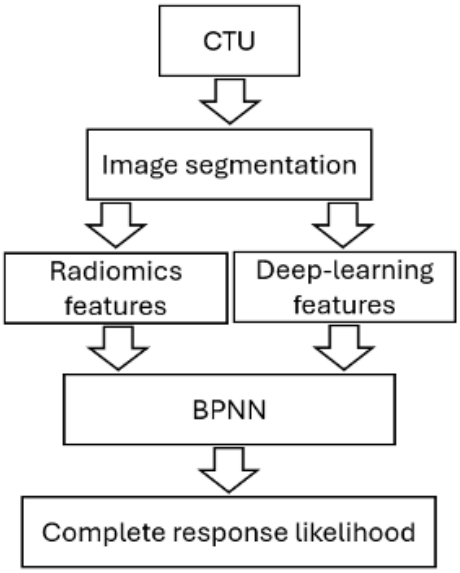


Figure 2-2: Model design. The CTU was first segmented, and radiomics and deep-learning features were extracted from the segmented region. These features were then fused using a BPNN to generate a score indicating the likelihood of a complete response to NAC.

2.1.3.3 Lesion segmentation

First, the user reviewed the scan and marked the lesions on the CTU scan using a ROI box. Then, the Auto-Initialized Cascaded Level Set (AI-CALS) algorithm [46] was applied to segment the lesions within the ROI box.

2.1.3.4 Radiomics Model

We proposed radiomics models based on the in-house radiomics package (In-House Radiomics model) developed in the CAD-AI Lab and the PyRadiomics package [47]. For both

models, we implemented a Random Forest [48] classifier for classification with the selected radiomics features.

By utilizing the In-House Radiomics Package to extract radiomics features from the segmented regions, a total of 91 radiomics features were extracted from each CTU scan, including morphological, texture, and intensity-based features. This set of features has proven useful in the analysis of various medical conditions, such as breast mass[49], lung nodules[50], and bladder cancer treatment response assessment[48]. These features were categorized in detail in Table 2-1.

We extracted 91 features from pre-treatment scans, denoted as pre-treatment features, f_{pre}, and 91 features from post-treatment scans, denoted as post-treatment features, f_{post}. We calculated the percentage difference of each feature between pre- and post-treatment scans for each pair using Equation 2-1, resulting in the creation of difference features, f_{diff}. Overall, a total of 273 features (f_{pre}, f_{post}, f_{diff},) was obtained from the pre- and post-treatment scans.

By using the publicly available PyRadiomics package [47], we extracted the radiomics features from the segmented image regions. This model enables the extraction of both morphological (size and shape) and gray-level features, for a total of 107 features. Unlike our In-House Radiomics Package, PyRadiomics also includes 3D morphological features. Similarly, we extracted 107 features from both pre-treatment and post-treatment scans, with percentage difference features calculated from the pre- and post-treatment features. As a result, we obtained a total of 321 features from PyRadiomics.

We evaluated the correlation of radiomics features between In-House Radiomics Package and PyRadiomics. This allowed us to comprehensively compare these two feature extraction methods and determine which one is more effective for analyzing medical images.

The high-dimensional nature of radiomics features can present challenges in terms of model interpretability, overfitting, and computational efficiency. In this study, we performed feature selection for dimensionality reduction using stepwise and brute-force methods. Brute-force selection conducts an exhaustive search over all possible feature subsets, while stepwise selection is a greedy, heuristic-based approach that incrementally adds or removes features based on model performance. The data-driven methods enabled us to identify effective feature subsets for optimizing model performance.

We also validated the selected set of the In-House Radiomics features from a previous study [48], which were selected based on different training set for the same task (treatment response assessment), on the test dataset of this study, to assess the robustness of the selected features.

$$f_{diff} = \frac{f_{pre} - f_{post}}{f_{pre}}$$

Equation 2-1

Table 2-1: Details of In-House Radiomics features.

Feature group (# of features)	**Sub-class (# of features)**
Morphological (41)	Size (6)
	Shape (5)
	Margin and surface roughness (18)
	Others (12)
Grey level (50)	Intensity (6)
	Contrast (10)
	Run length (32)
	Others (2)
Totally 91 features	

2.1.3.5 DL-CNN Model

As illustrated by Figure 2-3, from the segmented regions, we used a sliding window to extract one Region of Interest (ROI) of size 32×16 pixels from the pre-treatment scan and another 32×16 pixels ROI from the post-treatment scan. These two ROIs were then combined to form a hybrid ROI (hROI) of size 32×32 pixels. Within the segmented area, multiple ROIs can typically be extracted, meaning that one CTU scan pair can yield a number of hROIs. These hROIs served as the input units for the DL-CNN model. The ground truth for each hROI depends on the scan pair from which it was extracted. Figure 2-4 provides examples of a subset of hROIs.

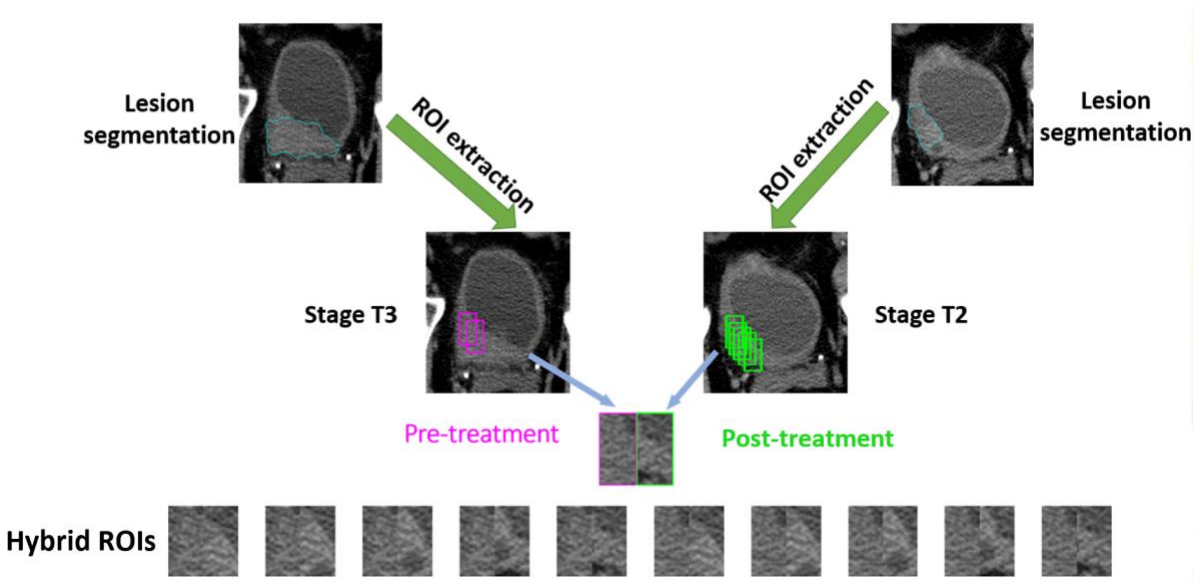


Figure 2-3: Hybrid ROIs (hROIs) generated from a pre- and post-treatment lesion pair. In this example the cancer was stage T3 in pre-treatment, and stage T2 in post-treatment scan, resulting in a label of non-complete response (>T0 stage after treatment).

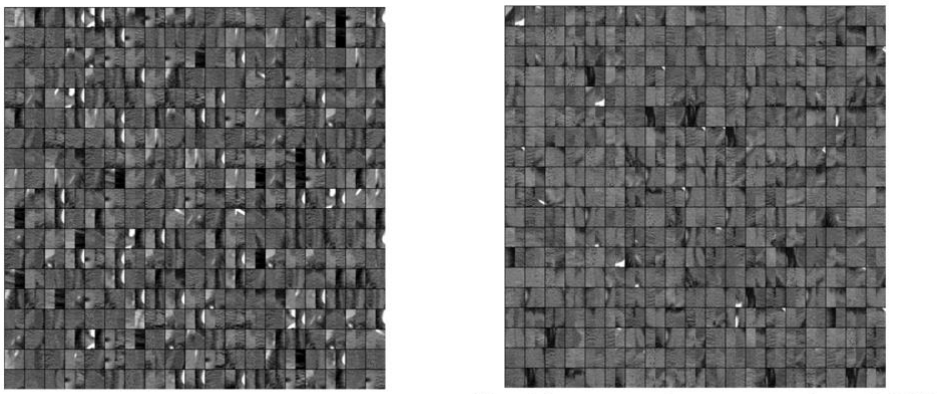


Figure 2-4: Subsets of hybrid ROIs. (a) A subset of complete responders (T0). (b) A subset of non-complete responders (>T0).

We developed the DL-CNN [51], a variation of AlexNet [52], using TensorFlow. The DL-CNN architecture, shown in Figure 2-5, consists of two convolutional layers (C1 and C2), two locally connected layers (L3 and L4), and a fully connected layer (FC10). The C1 layer uses 64 "5×5" kernels with a stride of 2, followed by local response normalization and max pooling with a "3×3" filter of stride 2. The C2 layer uses 64 "5×5" kernels with a stride of 1, followed by local response normalization and max pooling with a "2×2" filter of stride 1. The L3 layer contains 64 "3×3" kernels, while L4 consists of 32 "3×3" kernels. Finally, FC10 is a softmax layer that processes the output from L4 and generates a score representing the likelihood of a response to treatment, ranging from 0 to 1, where 0 indicates NCR and 1 indicates CR.

To train the DL-CNN model, we employed transfer learning [53]. First, the model was pre-trained using the CIFAR-10 image set. The pre-trained model was then fine-tuned on the bladder cancer training set through 5000 epochs of training. The validation set was used to identify the optimal training hyperparameters and determine the best-performing model. Once

the model was trained, it was deployed on an independent test set. The CR likelihood score for a scan pair was generated by taking the minimum score from the hROIs of that pair.

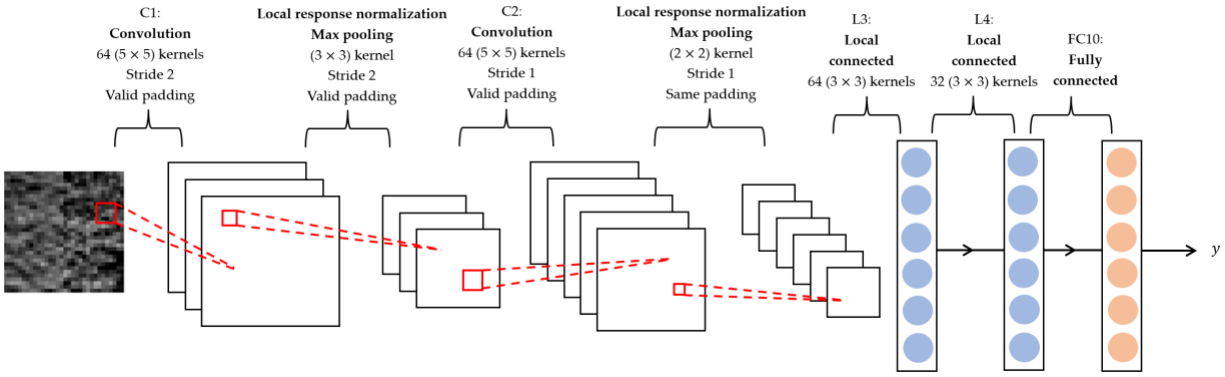


Figure 2-5: Schematic representation of the DL-CNN architecture used for image classification. The network consists of two convolutional layers, C1 and C2, each of them is followed by a max-pooling layer and a local response normalization layer, two locally connected layers, L3 and L4, and a fully connected layer, FC10. DL-CNN assessed the input hROIs and output the cancer treatment response score.

2.1.3.6 Combined model

We combined the selected In-House radiomics features with the prediction from the deep learning model through a backpropagation neural network (BPNN) [54].

2.1.3.7 Statistical Analysis

The area under the receiver operating characteristic curve (AUC) was used to evaluate the performance of the models. The correlation of the radiomics features from the two packages was calculated using Pearson's correlation coefficient. The performance variation between the two users was measured by the standard deviation.

2.1.4 Results

2.1.4.1 Radiomics Features

After comparing the In-House Radiomics features and PyRadiomics features, we found that five size features and three shape features had the same definition. Table 2-2 displays the

results of the correlation between the eight features. The majority of the matched feature pairs showed a high correlation, indicating the consistency of these quantitative imaging features developed by different institutions. However, some features, such as surface area, first-order mean, and first-order variance, exhibited low correlations (0.2–0.7). The possible differences for the "surface area" feature is because the PyRadiomics uses triangle mesh to calculate the surface area while the In-House package uses binary mask. The differences for the other two features are possibly due to the fact that the PyRadiomics features are extracted from a 3D scan of the lesion (multiple slices), while the In-House Radiomics features are derived from a 2D image (central slice) of the lesion. Overall, these findings highlight the reliability and agreement of the feature extraction methods, but also some differences due to feature implementation (as also discussed in this work [55]).

Figure 2-6 shows the heatmap of the In-House radiomics features, plotted using TreeView [56], reflecting the magnitude of extracted features from the test set.

Table 2-2: Correlation between In-House Radiomics and PyRadiomics features.

Matched features	**In-House**	**PyRadiomics**	**Correlation on pre-treatment CTU scans**	**Correlation on post-treatment CTU scans**
Feature 1	DIAM	original_shape_Maximum3DDiameter	0.98	0.98
Feature 2	SRFAREA	original_shape_SurfaceArea	0.68	0.69
Feature 3	VOL	original_shape_VoxelVolume	1	1
	VOL	original_shape_MeshVolume	1	1
Feature 4	MAX_CT	original_firstorder_Maximum	0.95	0.95
Feature 5	AVG	original_firstorder_Mean	0.52	0.46
Feature 6	MIN_CT	original_firstorder_Minimum	0.86	0.85
Feature 7	SKEW	original_firstorder_Skewness	0.94	0.92
Feature 8	VAR	original_firstorder_Variance	0.21	0.47

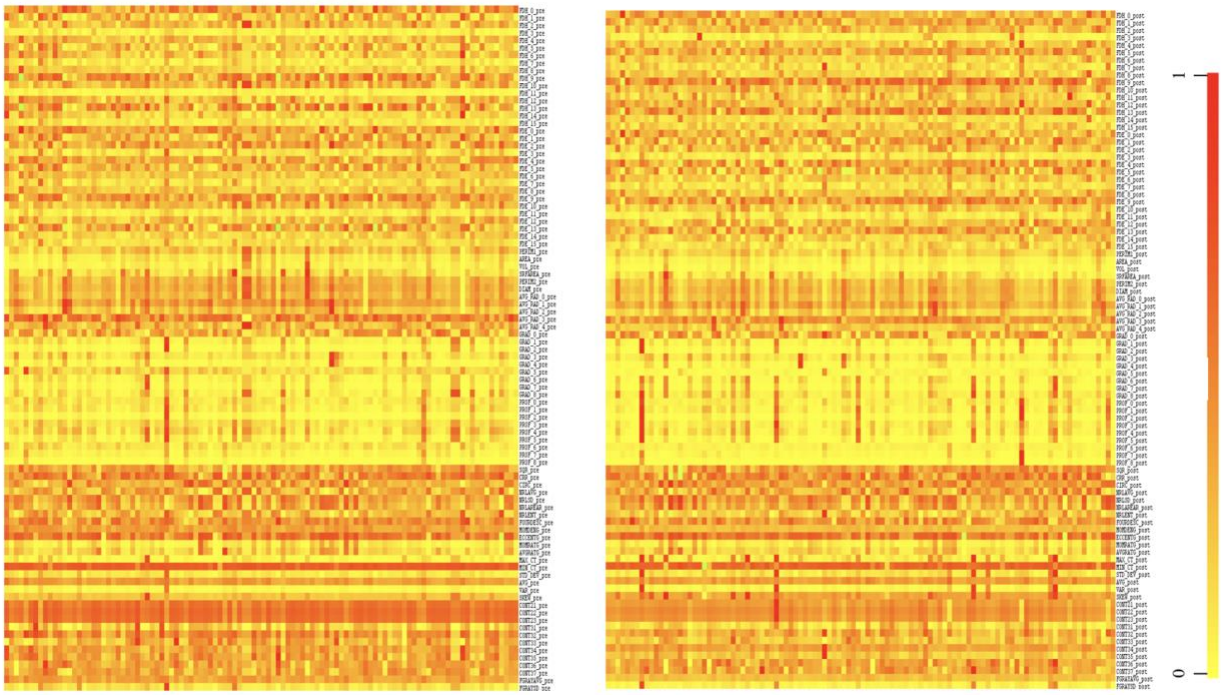


Figure 2-6: Heat map of the In-House radiomics features extracted from test set, visualizing the magnitude of extracted features. The left side is the features extracted from the pre-treatment scans, and the right side is features extracted from the post-treatment scans.

2.1.4.2 Model Classification

The performance of radiomics and deep-learning models are derived from User 1's segmentation. Feature selection and model weight optimization were conducted on the training/validation set. All the results presented here were obtained by deployment on the independent test set.

The performance of CR and NCR classification, measured by AUC, based on the In-House Radiomics model (In-House Radiomics features and random forest classifier) is shown in Table 2-3. For In-house Radiomics features, without feature selection, classification based on all f_{post} or all the f_{post} + f_{diff} achieved an AUC of 0.76 ± 0.05. The model achieved an AUC of 0.74 ± 0.06 using features selected from f_{pre} + f_{post} + f_{diff} by brute force, an AUC of 0.47 ±

0.07 using features selected by stepwise selection, and an AUC of 0.67 ± 0.06 based on selected features from a previous study. For PyRadiomics features, the best performance, with an AUC of 0.75 ± 0.05, was achieved using all extracted features without feature selection. With feature selection, the AUC performance was around 0.71. On average, seven features were selected. The selected features from the In-House Radiomics feature set and the PyRadiomics feature set were different.

Overall performance of the two radiomics packages in terms of AUC is comparable especially when all extracted features are used with Random Forest classifier.

The AUC for the DL-CNN model was 0.75 ± 0.06. Figure 2-7 shows some examples of pre- and post-treatment lesion pairs and the response predictions made by the DL-CNN and radiomics models.

The performance of the combined model, which fused the In-House Radiomics model with the DL-CNN, is 0.77 ± 0.06 for User 1 and 0.80 ± 0.05 for User 2. The combined model showed an improvement in performance compared to either the DL-CNN model or the radiomics model alone, indicating that the deep features and radiomics features provided complementary information, and combining them offered advantages. The difference in classification results of the combined model based on two users' segmentation was not statistically significant (p=0.58), which suggested that our classification models were robust in handling variations in user segmentation. This finding highlights the potential of our models to perform consistently well across diverse user groups.

Table 2-3: Classification results based on In-House Radiomics model (In-House Radiomics features and Random Forest classifier) using the test set.

Features set	**AUC (In-House Radiomics features)**	**AUC (PyRadiomics features)**
All pre-treatment features	0.56 ± 0.07	0.67 ± 0.06
All post-treatment features	0.76 ± 0.05	0.75 ± 0.05
All post and diff features	0.76 ± 0.05	0.75 ± 0.05
All features	0.75 ± 0.05	0.74 ± 0.06
Stepwise selected features	0.47 ± 0.07	0.70 ± 0.06
Brute force selected features	0.74 ± 0.06	0.72 ± 0.05
Selected features from previous study	0.67 ± 0.06	-

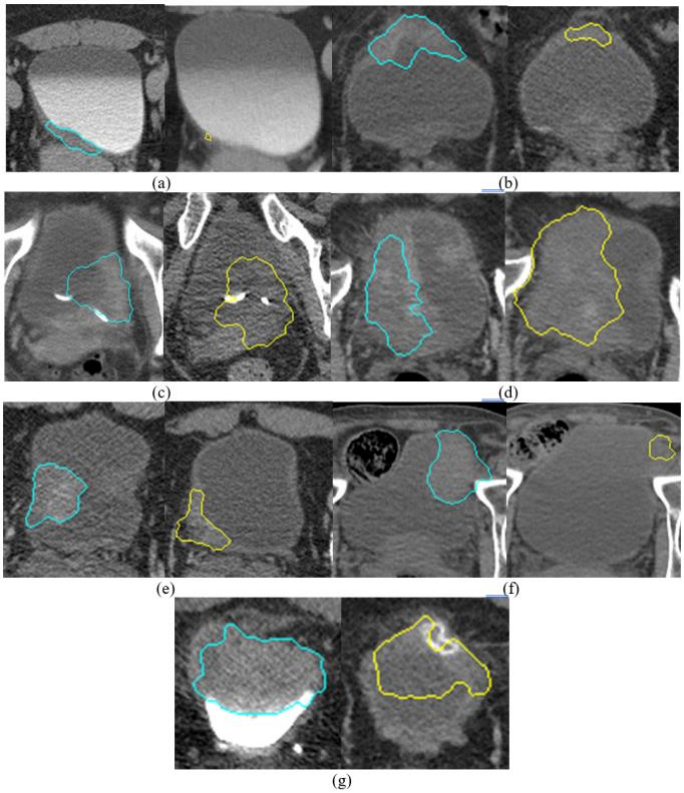


Figure 2-7: Examples of pre- and post- treatment lesion pairs and the response prediction by our trained models. The lesion of each scan was segmented by AI-CALS and was marked as blue (pre-treatment scan) and yellow (post-treatment scan) outline. Cases (a) and (b) were complete responders. Both the DL-CNN and the radiomics models provided correct estimates. Cases (c) and (d) were non-complete responders. Both the DL-CNN and radiomics models provided the correct estimates. Case (e) was a complete responder. Both the DL-CNN and the radiomics model estimates were wrong. Case (f) is a non-complete responder. The DL-CNN model made the correct estimate but the radiomics model provided the wrong estimate. Case (g) is a non-complete responder. The estimate of the radiomics model was correct but the estimate of the DL-CNN model was wrong.

2.1.5 Discussion

Compared to a previous study [4, 48] that reported an average AUC of 0.73 when 17 clinicians assessed the response of bladder cancer to NAC, our model has demonstrated a higher performance, achieving an AUC of 0.77 when classifying an independent test set. This indicates that our model has the potential to distinguish between CRs and NCRs with a level of accuracy comparable to physicians [4, 48]. In our previous studies [51], we applied the combined classification model to a relatively small test set of 42 lesion pairs, achieving an AUC of 0.79. With this study evaluating an enlarged independent test set of 106 lesion pairs, our results demonstrate the system's robustness in assessing the response to neoadjuvant chemotherapy for bladder cancer.

To further improve our classification model's performance, we need to address the overfitting issue in the machine learning models by augmenting the training dataset. Additionally, for the radiomics features, a more effective approach to feature selection is required to enhance generalizability while avoiding the heavy computation induced by greedy search algorithms. Nonetheless, the dataset in our study remains relatively small. We aim to collect more data from multi-institutional sources or explore federated learning to further test the model’s robustness.

2.1.6 Conclusion

In conclusion, the classification model based on radiomics and deep-learning features proposed in this study demonstrates the potential to classify complete responders and non-complete responders to neoadjuvant chemotherapy in bladder cancer patients. This could aid in treatment planning for patients who are not benefiting from neoadjuvant chemotherapy, helping optimize medical resource allocation and improving patient care.

2.2 Multi-institutional Study of Radiomics Model for Assessment of Treatment Response to Neoadjuvant Chemotherapy of Bladder Cancer

2.2.1 Abstract

A study using multi-institutional test sets was conducted to validate the generalizability of a radiomics-based machine learning model for assessing treatment response to neoadjuvant chemotherapy (NAC) in bladder cancer. Radiomics features were extracted from pre- and post-treatment CT urography scans of 295 patients across three institutions. A random forest classifier was trained on the radiomics features from Institution A and tested on Institution A test set and two sequestered external datasets. Model performance was evaluated using the area under the receiver operating characteristic curve (AUC). The model achieved AUCs of 0.77, 0.85, and 0.84 on internal and external test sets, demonstrating strong predictive performance. This study highlights the potential of a radiomics-based model for assessing NAC response in bladder cancer across institutions. Further validation with larger datasets is ongoing.

2.2.2 Introduction

The aim of this study is to validate the generalizability of the In-House Radiomics model based on radiologic imaging (CT urography) for assessing treatment response to neoadjuvant chemotherapy in bladder cancer patients using multi-institutional sequestered test sets.

2.2.3 Methods and Materials

With Institutional Review Board (IRB) approval, 835 patients with bladder cancer treated with neoadjuvant chemotherapy between 2003 and 2020 at three U.S. institutions were identified (Table 2-4). Patient selection criteria included having CTU scans before and after neoadjuvant chemotherapy treatment and a pathological cancer stage determined by post-radical cystectomy. The criteria outlined above were met by 258 patients (Figure 2-8) which generated 295 pre- and

post- CTU scan pairs (Table 2-4). The ground truth of treatment response was established based on pathological cancer stages: a stage of pT0 indicated a complete responder (CR), while any stage above pT0 was classified as a non-complete responder (NCR). Figure 2-9 shows some examples of CRs and NCRs from the three institutions.

The model development data set was sourced from Institution A and was split into three sets: training (56%: 35 CRs, 113 NCRs); validation (4%: 5 CRs, 5 NCRs); and a held-out internal test set (40%: 19 CRs, 87 NCRs). Institution B data (3 CRs, 16 NCRs) and Institution C data (4 CRs, 8 NCRs) were used as external test sets, which were sequestered from the developer until the model development was completed and frozen.

The In-House Radiomics model (Chapter 2.1.3.4) was utilized to perform the classification of CR and NCR. The area under the receiver operating characteristic curve (AUC) was used to evaluate the classification performance on the test sets.

Table 2-4: CTU data collected from three institutions.

	Institution name	**Initial patients #**	**Cases # used for model training and test**
Institution A	University of Michigan Health Center	781	264
Institution B	Penn State Health Milton S. Hershey Medical Center	34	19
Institution C	University of Iowa Health Care Medical Center	20	12
Total	-	835	295

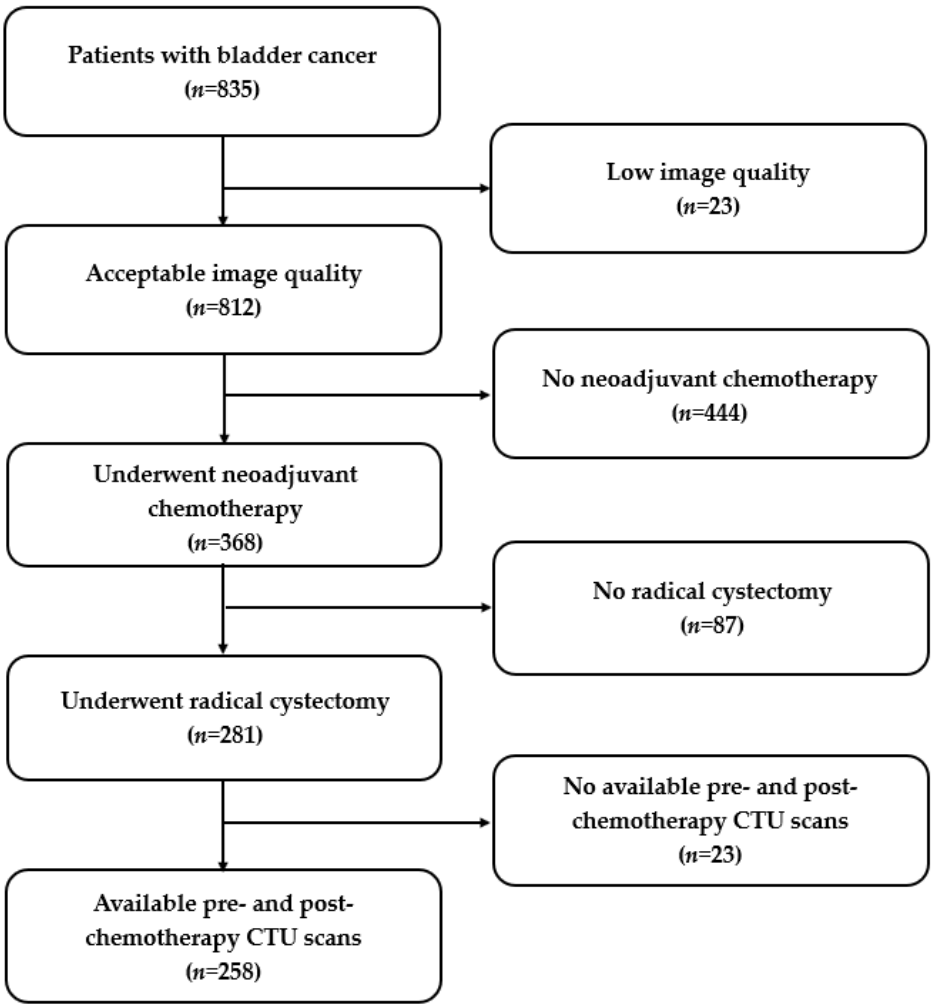


Figure 2-8: Study-population diagram. Out of 835 patients, 258 patients were identified for this study. Inclusion criteria include (1) patient had at least one pair of pre- and post-NAC CTU scans with acceptable image quality, (2) patient underwent neoadjuvant chemotherapy (NAC), and (3) patient underwent radical cystectomy, so that the ground truth of complete response or not to NAC can be determined.

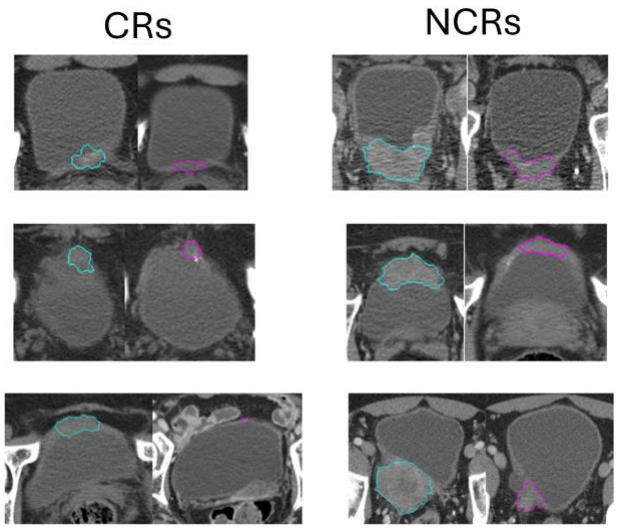


Figure 2-9: Examples of complete responders (CRs) and non-complete responders (NCRs). For each scan pair, the left side is the pre-treatment scan, the right side is the post-treatment scan. The two scan pairs on the first row are from Institution A, the two scan pairs on the second row are from Institution A, and the two scan pairs on the first row are from Institution C.

2.2.4 Results

The selected features included morphology features (lesion area and surface) and intensity-based features. The developed model achieved an AUC of 0.77±0.05, 0.85±0.13, and 0.84±0.12 on the test set of Institution A, B, and C, respectively (Figure 2-10).

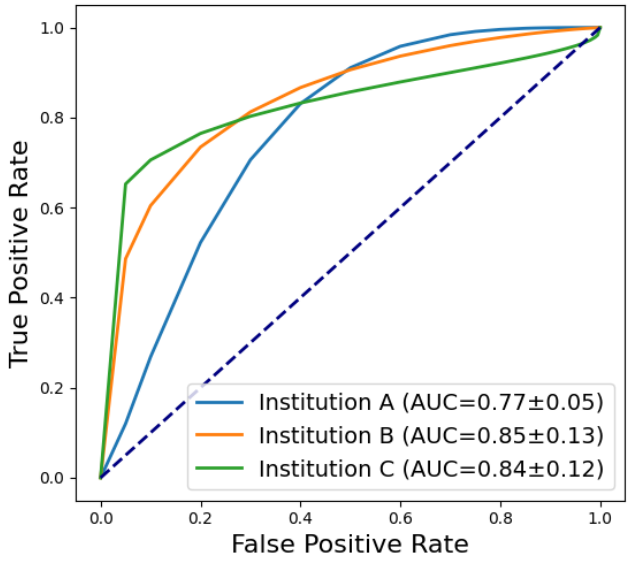


Figure 2-10: Classification results for treatment response assessment, including AUCs and ROC curves for three institutions.

2.2.5 Conclusion

While larger data sets are needed for further validation, this study demonstrates the potential generalizability of a developed radiomic model based on CTU images to assess the treatment response to neoadjuvant chemotherapy for bladder cancer patients across different institutions. Efforts to expand the external test sets are underway.

Chapter 3 Survival Prediction of Patients with Bladder Cancer After Cystectomy

3.1 Survival Prediction of Patients with Bladder Cancer after Cystectomy Based on Clinical, Radiomics, and Deep-learning Descriptors

3.1.1 Abstract

Accurate survival prediction for bladder cancer patients who have undergone radical cystectomy can improve their treatment management. However, the existing predictive models do not take advantage of both clinical and radiological imaging data. This study aimed to fill this gap by developing an approach that leveraged the strengths of clinical (C), radiomics (R), and deep-learning (D) descriptors to improve survival prediction. The dataset comprised 163 patients, including clinical, histopathological information, and CT urography scans. The data were divided by patient into training, validation, and test sets. We analyzed the clinical data by a nomogram and the image data by radiomics and deep-learning models. The descriptors were input into a BPNN model for survival prediction. The AUCs on the test set were (C): 0.82 ± 0.06, (R): 0.73 ± 0.07, (D): 0.71 ± 0.07, (CR): 0.86 ± 0.05, (CD): 0.86 ± 0.05, and (CRD): 0.87 ± 0.05. The predictions based on D and CRD descriptors showed a significant difference (p = 0.007). For Kaplan–Meier survival analysis, the deceased and alive groups were stratified successfully by C ($p < 0.001$) and CRD ($p < 0.001$), with CRD predicting the alive group more accurately. The results highlight the potential of combining C, R, and D descriptors to accurately predict the survival of bladder cancer patients after cystectomy.

3.1.2 Introduction

The five-year survival rate serves as a valuable indicator of treatment effectiveness, varying across different types of cancer. Testicular cancer (97%), melanoma of the skin (92.3%), and prostate cancer (88%) have the highest estimated five-year survival [57]. Conversely, lung and bronchial cancer exhibit the lowest survival rate at only 22% [57]. In the United States, the five-year relative survival rate for bladder cancer stands at 77% [57]. Specifically, for patients with bladder cancer who underwent radical cystectomy, the reported five-year survival rates range from 54.5% to 68% [58].

Survival prediction in bladder cancer can be performed based on data sources such as clinicopathological information [59], histological slides [60, 61], gene expression, or molecular markers [62]. The analysis methods include machine-learning models like support vector machines, logistic regression, and random forest [63, 64], as well as nomogram [65] and risk stratification [66]. Nomogram models, often developed based on large-size cohorts, can offer high-accuracy prediction capabilities.

CT urography (CTU) plays a significant role in bladder cancer diagnosis by providing 3D abdominal images that allow for the quantitative volumetric analysis of the bladder and lesion characteristics. The recent advances in deep-learning models offer new opportunities for extracting more comprehensive image features and the potential for improving the predictive models.

In this study, we combine clinical, histopathological features with CTU image features to harness the advantages of nomogram, radiomics, and deep-learning models. By introducing a hybrid approach combining clinical and imaging analytics, we aim to enhance the prediction of five-year survival for patients with bladder cancer after radical cystectomy, allowing better treatment planning and improving prognostic outcomes. The approach also has implications for

the assessment of the added value of imaging in patient management and for how the added predictive value could be used to select the utilization of imaging.

3.1.3 Methods and Materials

The model development process is depicted in Figure 3-1. First, we collected a comprehensive dataset comprising clinical, histopathological information and CTU scans. One pair of pre- and post-treatment CTU scans was obtained for each patient. We then processed the collected data and developed models incorporating nomogram, radiomics, and deep-learning descriptors. The models were built with a training set and optimized with a validation set. After the models were developed, they were deployed to an independent test set to evaluate their survival prediction ability on unseen cases.

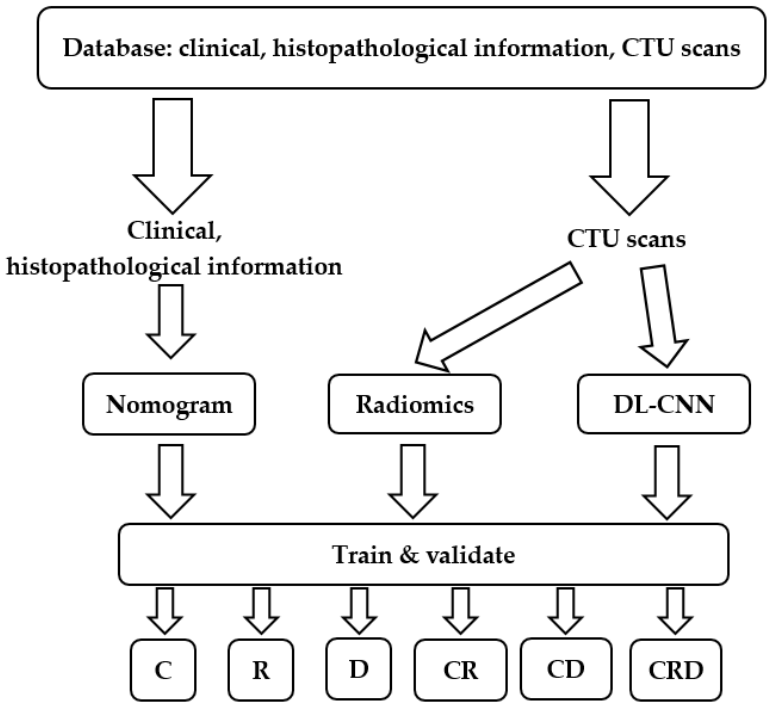


Figure 3-1: Predictive model development process of this study. We collected clinical, histopathological information and CTU scans. We built and validated survival prediction models based on individual or combined descriptors. C = clinical descriptors; R = radiomics descriptors; D = deep-learning descriptor; CR = clinical + radiomics descriptors; CD = clinical + deep-learning descriptors; CRD = clinical + radiomics + deep-learning descriptors. DL-CNN = Deep learning–convolutional neural network.

3.1.3.1 Patient Cohorts

This retrospective study is HIPAA compliant and Institutional Review Board (IRB) approved. We collected clinical reports and CTU scans from 781 bladder cancer patients spanning 2001 to 2020 at one institution.

Patients were selected if they met all of the following criteria: (1) having at least one set of pre- and post-chemotherapy CTU scans following neoadjuvant chemotherapy; (2) having undergone radical cystectomy; and (3) having available five-year follow-up data post-cystectomy to determine survival status. The reference standard for survival classification was determined by a research assistant (J.G.) using the patients' follow-up data.

Of the total of 781 bladder cancer patients, 163 patients satisfied the above criteria (Figure 3-2). Among these 163 patients, 48.5% (79/163) were alive five years after radical cystectomy, while 51.5% (84/163) died before reaching their fifth year post-cystectomy. We split the dataset into training, validation, and test sets: 56% (92/163) training (55 alive, 37 deceased); 4% (7/163) validation (4 alive, 3 deceased); and 40% (64/163) test (20 alive, 44 deceased), keeping a relatively large portion of the patients in both the training and test sets. The bladder cancer cases were collected chronologically, with medical records and exam dates in the training and validation sets dating from 2003 to 2016, and in the test set, dating from 2012 to 2020. This serial approach simulated, to some extent, the "real world" clinical situation, where the model is built based on previous cases and all weights and parameters are frozen after training, with the model then deployed to the new cases.

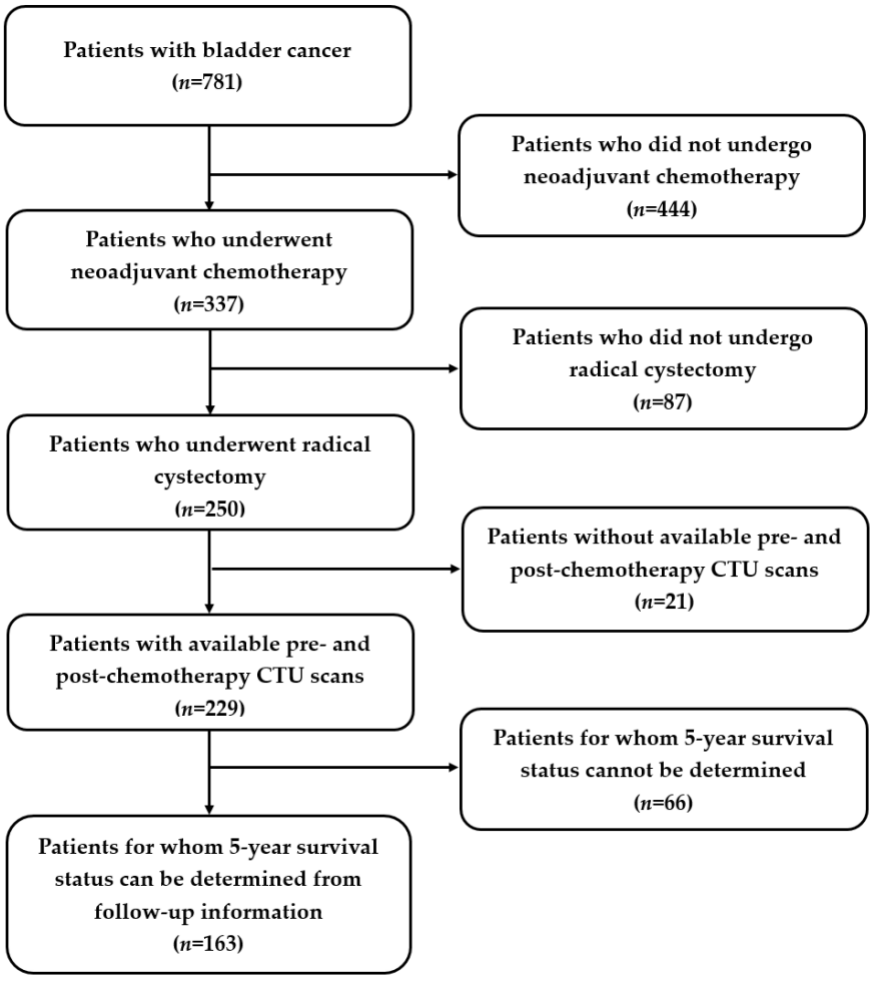


Figure 3-2: Study cohort diagram. Out of 781 patients, 163 patients (mean age 64±9 years; M:F 131:32) were identified for this study. Patients who (1) underwent neoadjuvant chemotherapy and had pre- and post-chemotherapy CTU scans, (2) underwent radical cystectomy, and (3) had follow-up data to determine their five-year survival status were eligible.

3.1.3.2 Nomogram

The histopathological information was analyzed using the nomogram developed by Shariat using a cohort of 731 patients [67]. The nomogram achieved a prediction accuracy of 0.791 in the study by Shariat [67], which was superior to that of the American Joint Committee on Cancer Staging [68] at 0.663 with $p = 0.001$.

Five clinical and histopathological indices were used as the input to the nomogram, including post-surgery pathologic stage, lymphovascular invasion (LVI), pathologic node stage, whether patients underwent neoadjuvant chemotherapy, and whether patients underwent adjuvant radiotherapy. For every patient, each of the five inputs was mapped to a "points" axis to determine how many points were contributed by each input. The sum of the points from the five inputs resulted in the estimation of the "total points" for the patient, which was then mapped to the *n*-year survival probability ($n = 5$ in the current study).

3.1.3.3 CTU Scans Processing

CTU scans for bladder cancer include a series of abdominal images, allowing for a comprehensive 3D inspection of the bladder. Figure 3-3a shows an example of a rendered 3D CTU scan. To extract radiomics and deep features from the CTU scans, the bladder cancer lesions were annotated by an experienced radiologist (R.H.C.) who has over 30 years of experience reading abdominal CT. The annotation process entailed several steps: (1) marking a volume of interest (VOI) for the lesion, (2) measuring the longest and perpendicular diameters of the tumor, (3) identifying the lesion location within the bladder, ureter, or other abdominal locations, and (4) evaluating the lesion type, edge characteristics, and likelihood of abnormality. While a patient could have multiple lesions, for this study, we limited the number of lesions to one per patient and selected the dominant lesion for feature extraction.

Lesion segmentation was obtained by our in-house developed semi-automatic algorithm AI-CALS [46] utilizing the marked VOIs. The segmentation provided 3D contours outlining the lesion, from which we performed feature extraction. In a previous study, we showed that the AI-CALS algorithm achieved an average volume error of $4.9 \pm 38.3\%$ when compared to the manual outlines drawn by an experienced radiologist [46]. Segmenting lesions in CTU scans can

be challenging due to various factors, such as patient positioning, image artifacts, and excreted contrast in the bladder. Ideally, since bladder cancer is expected to enhance in the early contrast phase, these images can help to distinguish neoplasm from the rest of the bladder tissue. However, some patients underwent CTU with different imaging protocols from referral sites including only non-contrast images (Figure 3-3b-1) for various reasons such as renal failure or patient refusal of contrast. Other protocols consisted of only delayed contrast phase imaging (Figure 3-3b-2) which could decrease the conspicuity of the lesion from surrounding structures. Figure 3-3b-3 shows a case with bladder-wall thickening after treatment, in which it is difficult to ascertain whether the lesion still existed. Additionally, artifacts in CTU scans can also complicate segmentation (Figure 3-3b-4). Despite these challenges, the performance of AI-CALS was satisfactory in most cases. Figure 3-3c illustrates CTU examples with early contrast phase and minimal artifacts, for which the segmentation was in good agreement with the manual reference.

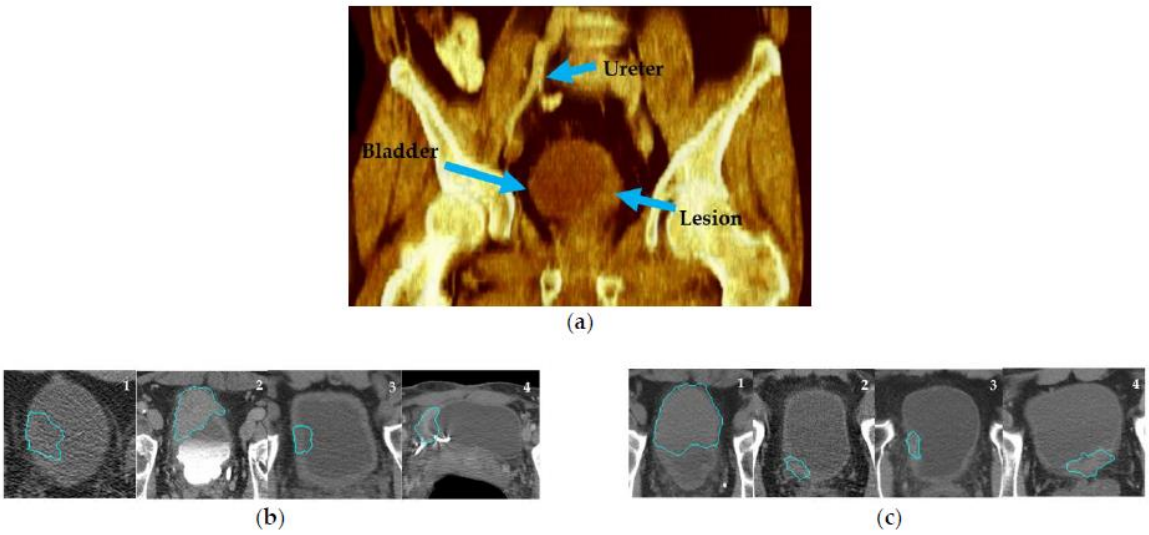


Figure 3-3: CTU scan examples. (a) A 3D volume rendering CTU scan showing the ureter, bladder, and bladder lesion. (b) CTU scans that are potentially challenging for lesion segmentation. (b)-1: without contrast; (b)-2: delayed-contrast phase; (b)-3: bladder wall thickening; (b)-4: image artifact. (c) Examples of CTU scans in early contrast phase and without artifacts. The blue outlines are the AI-CALS segmentation outlines.

3.1.3.4 Radiomics Model

We used the In-House Radiomics features (described in detail in Chapter 2.1.3.4) to build the radiomics model. After extracting 273 radiomics features from the pre- and post-treatment CTU scans, we employed a mutual information (MI) method for feature selection. Feature selection can eliminate the redundant variables, reduce the time and space requirements for data processing, and reduce the risk of the "curse of dimensionality" when the training data size is limited [69-71].

MI is capable of measuring the arbitrary dependencies between random variables and has been widely used for feature selection in machine learning [72]. One of the advantages of MI is its ability to assess the information content of features in nonlinear relations. Equation 3-2 shows that MI is formulated as the difference in the entropies between the variable itself (Equation 3-1) and the variable conditioned on the target value. The MI score ranges from 0 to ∞, where zero signifies complete independence between the feature and the target, while higher scores indicate greater relevance such that the corresponding feature may be useful for the learning task.

For a feature vector f, the entropy is:

$$H(F) = -\int P(f) log P(f) df$$

Equation 3-1

Given the condition of C, the MI between variables c and f is:

$$I(F;C) = H(F) - H(F|C)$$

Equation 3-2

In order to select the useful set of features, a threshold was imposed on the calculated MI. Pearson correlation was applied to estimate the feature correlation.

3.1.3.5 Deep-Learning Model

We utilized the DL-CNN model (detailed in Chapter 2.1.3.5) to predict the likelihood of five-year survival on a scale from 0 to 1, where a value of 1 indicates survival equal to or beyond five years, and 0 indicates less than five years.

3.1.3.6 Classification

The features from the three types of descriptors of each case, i.e., the selected radiomics features by the MI, the survival likelihood score from the deep-learning model, and the "points" derived from each of the five indices of the nomogram, were fed into a Back-Propagation Neural Network (BPNN) [54] to classify the likelihood of survival of each patient. The performance of the descriptors when they were used individually was compared to the utilization of the descriptors in combination. The structure of the BPNN is shown in Figure 3-4, where the parameters of the BPNN were selected as guided by the validation set.

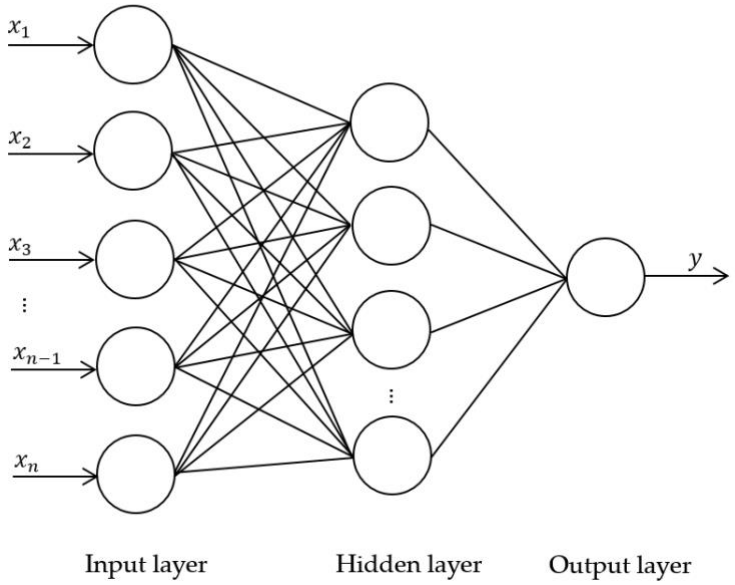


Figure 3-4: Structure of BPNN. The input x_i (i=1,2,3,...,n) are the descriptors. The hidden layer contains 13 nodes. The output y is the likelihood score assessing the survival of each patient.

3.1.3.7 Statistical Analysis

The classification performance of the descriptors was evaluated using the area under the receiver operating characteristics (ROC) curve (AUC). We compared three combinations of descriptors for survival prediction (C vs. CRD, R vs. CRD, and D vs. CRD); thus, the critical α value for statistical significance was adjusted to be $\alpha = 0.017$ ($\alpha = 0.05/3$) according to the Bonferroni correction for multiple hypothesis testing [73].

To further analyze the model performance on predicting survival outcomes, we also performed Kaplan–Meier analysis and a log-rank test [74] was used to evaluate if their survival probabilities differ significantly.

3.1.4 Results

3.1.4.1 Cohort Statistics

The demographic information and cancer stage of the 163 patients are shown in Table 3-1, including gender, race, tobacco use, clinical stage of cancer, and pathological stage of cancer after radical cystectomy. The distribution of patients' age at surgery is depicted in Figure 3-5. These statistics show that the dataset covered a wide range of patient demographics.

Table 3-1: Patient characteristics (n = 163).

Attributes	Sub-category	# of patients
Gender	Male	131
	Female	32
Age at surgery	Mean age ± standard deviation	64±9
Race	White	142
	Black/African Am.	13
	Am. Indian/Native	1
	Asian	2
	Unknown	5
Tobacco use	Current	40
	Former	88
	Never	34
	Unknown	1
Post-surgery stage	pT0	35
	pTa/pTi/pTis	15
	pT1	16
	pT2	36
	pT3	45
	pT4	16
Lymphovascular invasion (LVI)	Yes	61
	No	102
Pathologic node stage	N0	112
	N1	24
	N2	23
	N3	4
Neoadjuvant chemotherapy	Yes	163
	No	0
Adjuvant radiotherapy	Yes	0
	No	163

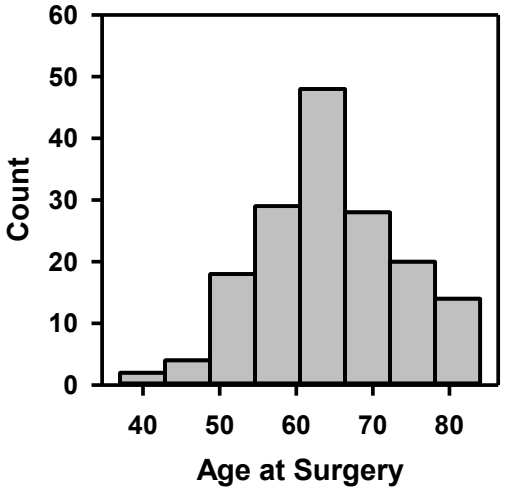


Figure 3-5: Histogram of patients' age at surgery.

3.1.4.2 Radiomics Features

The MI-selected radiomics features were evaluated by Pearson correlation on the training set. First, we compared the feature vector with the target vector. In our analysis, we aimed for a strong correlation between feature and target, as this would allow us to predict the target from the feature. Then we compared the feature vectors among themselves. It is preferable to have the selected features less correlated to make the model more robust. However, features with some correlation as a group may also be useful in the model design.

Out of the 12 selected features, we plotted a heatmap showing the correlation values between each pair of features (Figure 3-6). The majority of the correlation values fell within the range of ±0.3, indicating weak correlations. Weak correlations implied that the selected features were providing mostly distinct and independent information, which can contribute to the predictive power of the model with a low redundancy. On the other hand, the features with a certain level of correlation may contribute to the predictive power when combined with other features.

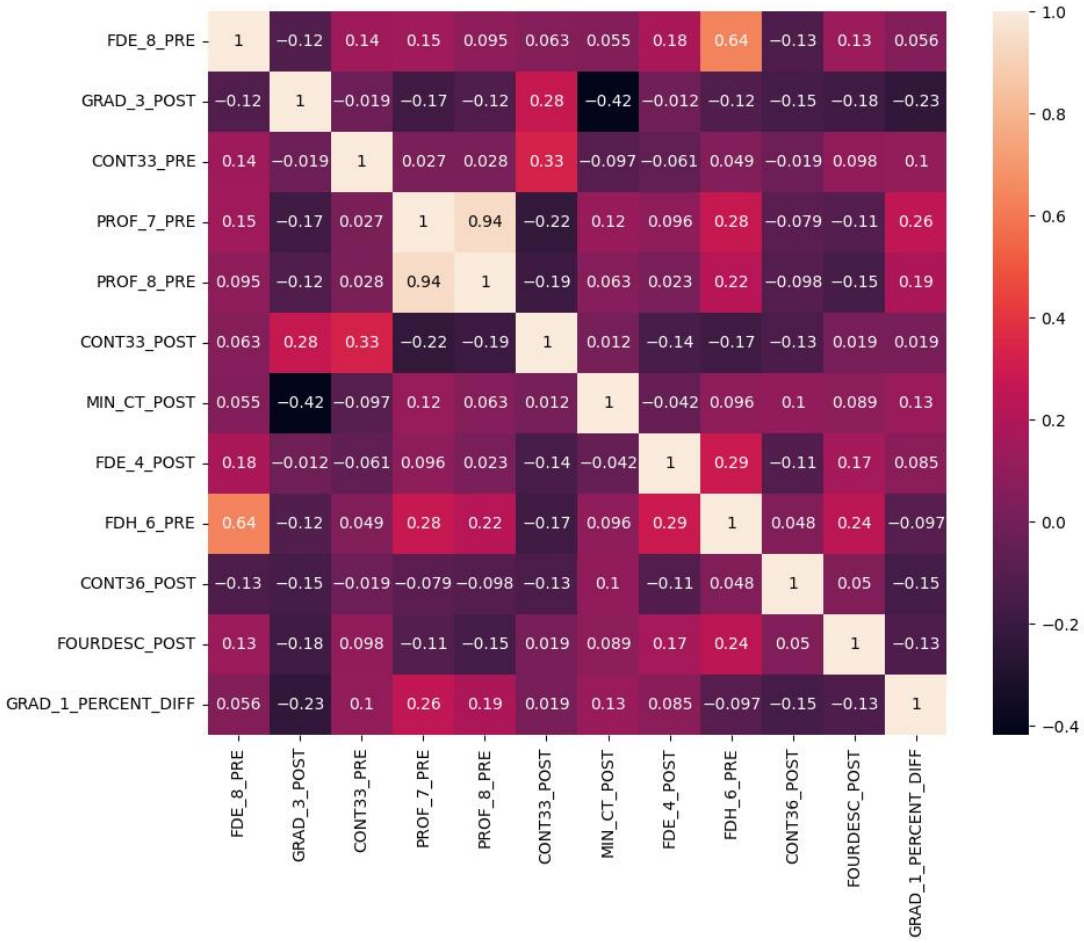


Figure 3-6: Heatmap of Pearson correlation between any pair of the MI-selected radiomics features. Most of the correlation values were within ±0.3 which indicates a weak correlation.

3.1.4.3 Survival Prediction

After the training of the models with the training set, we selected the best-performing clinical, radiomics, deep-learning, and combined models using the validation set. For the combined models, BPNN training required about 40 iterations. We deployed the selected models on the test set to evaluate their ability to predict patient survival. The ROC curves of the individual and combined descriptors are plotted in Figure 3-7. The clinical (C), radiomics (R), and deep-learning (D) descriptors achieved AUCs of 0.82 ± 0.06, 0.73 ± 0.07, and 0.71 ± 0.07,

respectively. By combining the R and D descriptors with the clinical C descriptors, the AUCs of the models were improved to 0.86 ± 0.05 for CR, 0.86 ± 0.05 for CD, and 0.87 ± 0.05 for CRD. The ROC curves of C and CRD are plotted in one figure (Figure 3-7c) to show their differences in classification performance. The classifications based on D and CRD descriptors achieved a significant level of difference (p = 0.007) (Table 3-2). Other pairs of comparisons did not reach statistical significance but showed strong trends towards improvement when used in combination.

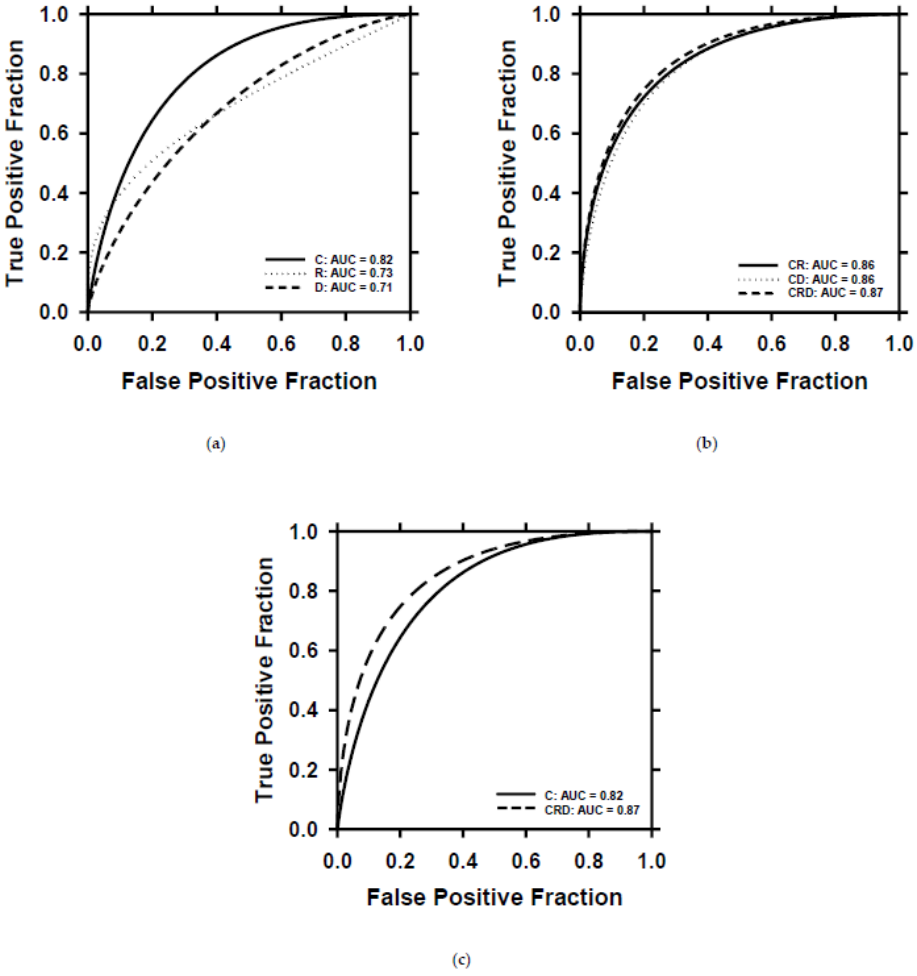


Figure 3-7: ROC curves of survival prediction on the test set by using different descriptors. (a) ROC curves and AUC values for the individual descriptor. (b) ROC curves and AUC values for combined descriptors. (c) Direct comparison of the ROC curves of C and CRD descriptors. C = clinical descriptors; R = radiomics descriptors; D = deep-learning descriptor; CR = clinical + radiomics descriptors; CD = clinical + deep-learning descriptors; CRD = clinical + radiomics + deep-learning descriptors.

Table 3-2: The *p*-values of survival prediction based on individual and combined descriptors. The critical α value for statistical significance was adjusted to be α = 0.017 (α = 0.05/3) according to the Bonferroni correction for multiple hypothesis testing.

Comparison	***p*-value (adjusted α = 0.017)**
C vs. CRD	0.153
R vs. CRD	0.056
D vs. CRD	0.007 *

* Achieved a significant difference.

To further investigate the impact of incorporating radiomics and deep-learning descriptors into clinical descriptors on survival prediction, we conducted Kaplan–Meier analysis on C and CRD descriptors. To categorize the patients in the test set into two groups (deceased and alive groups), we determined the cutoff values of the scores generated by BPNN for C and CRD descriptors by selecting the point of least misclassification for the C and CRD models separately based on the training and validation sets. Figure 3-8 displays the survival probability curves of the deceased and alive groups stratified by C and CRD models. The survival probability of the alive and deceased groups had significant difference ($p < 0.001$ for C and CRD models). The CRD model achieved better separation between the two groups compared to the C model.

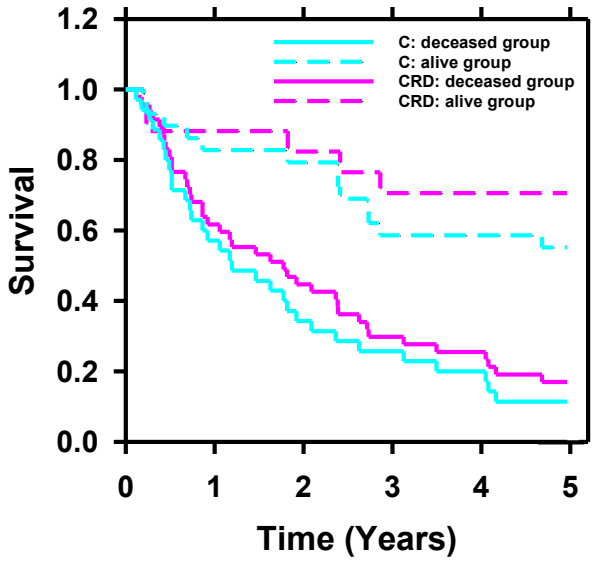


Figure 3-8: Kaplan–Meier survival probability curves for the alive and deceased groups stratified by C and CRD models. The survival probability of the two groups (deceased and alive) differed significantly based on either C ($p < 0.001$) or CRD ($p < 0.001$) models. The CRD model achieved better separation between the two groups compared to the C model.

3.1.5 Discussion

All three descriptors—C, R, and D—showed strength in classifying the test dataset. However, the clinical descriptors demonstrated a higher performance compared to the deep-learning and radiomics descriptors. One of the main factors contributing to this result is that the nomogram model utilized in the clinical descriptors was developed using a larger training set, while the deep-learning and radiomics models used here were trained on a smaller dataset. Deep-learning models are known to achieve a better generalization and robustness when they are trained with larger and more representative datasets [75].

With the integration of radiomics and deep-learning descriptors alongside the clinical descriptors, the CRD descriptors can stratify the deceased and alive groups more effectively. The lack of statistical significance in the difference between the AUCs of the C and CRD descriptors,

and the modest improvement in the stratification ability in the survival curves between the deceased and alive groups by CRD compared to that of the C descriptors, may be attributed to the small size of the test set, which could limit the statistical power to detect the small differences, and also the strong performance of the clinical descriptors alone. While this is a preliminary study and requires significant further investigation, the result has implications for the utilization of this approach in the future, as it could be used to assess whether an imaging test has added value over clinical data alone. If an imaging test is expected to add predictive value for the clinical task at hand, it will strengthen the utility of the imaging test for patient care.

The results also point to the importance of clinical information in the utilization of imaging, as this information can lead to the improved performance of imaging in answering the clinical question. This is well accepted in clinical radiology, but clinical information is not typically utilized in radiomics or machine-learning computational models for imaging tasks. This study demonstrated that the clinical data may contain important complementary information to improve the performance of these models.

Our study has some limitations. The dataset was relatively small and collected from a single site. We need to collect multisite cases and larger datasets to build more reliable and generalizable models. We also included non-contrast CT images which could affect the performance of the R and D models. On the other hand, the presence of non-contrast CT images represents the real clinical situation and the models may have been trained to be more robust with the heterogeneous image features.

Despite the limitations, this study demonstrated that CTU scans potentially can provide useful information to improve the survival prediction of patients with bladder cancer after radical cystectomy. Radiomics and deep-learning models can extract effective image features from CTU

scans. Combined with a nomogram, the prediction ability of the hybrid model can outperform the models using the individual types of descriptors.

3.1.6 Conclusion

While larger datasets are needed, this study demonstrates that combining radiomics and deep-learning descriptors with clinical descriptors holds promise for improving the prediction of the 5-year survival of bladder cancer patients after radical cystectomy. The accurate assessment of 5-year survival offers potential benefits with patient counseling and postoperative surveillance strategies.

3.2 Survival Prediction of Bladder Cancer Patients after Surgery based on Machine Learning Models with Radiomics and Genomic Descriptors

3.2.1 Abstract

This study explores the development of machine learning models to predict the two-year survival of bladder cancer patients following radical cystectomy using a combination of radiomics and genomic data. Data from a cohort of 85 bladder cancer patients were collected, which were split into training, validation, and test sets. Using radiomics features (RF) from pre-surgical CT urography scans and genomic markers (GEN) derived from tissue samples, we trained three models: a radiomics model (RFM), a genomic model (GENM), and a combined radiomics-genomics model (RF-GENM). Performance was assessed using the area under the ROC curve (AUC) and Kaplan-Meier analysis on the test set. The results showed that the RF-GENM achieved the highest AUC of 0.74±0.09, outperforming both the RFM (AUC=0.63±0.10) and GENM (AUC=0.69±0.10). These findings suggest that combining radiomics and genomic data can improve survival prediction and provide valuable insights for personalized treatment

and patient management. Further research with larger datasets is needed to validate these results and assess the generalizability of the model.

3.2.2 Introduction

The work in Chapter 3.1 demonstrated the feasibility of predicting five-year survival based on clinicopathological and radiology imaging data. In this study, we proposed using radiomics and genomic descriptors to develop machine learning models for predicting the two-year survival of bladder cancer patients after radical cystectomy. Accurate survival prediction can help with understanding disease heterogeneity, tailoring treatment and decision making, improving resource allocation, and counseling with patients and families.

3.2.3 Methods and Materials

With IRB approval, we retrospectively collected pre-surgical CT urography (CTU) exams from 2019 to 2021 for 85 patients with bladder cancer. The patients underwent procedures including primarily cystectomy and transurethral resection of a bladder tumor (TURBT) in the same timeframe, from which tissue samples were collected for genomic analysis. 2-year survival status was obtained as ground truth. We split the data into three sets: training (46 cases: 30 alive (A); 16 deceased (D)); validation (4 cases: 2 A; 2 D); and test (35 cases: 22 A; 13 D).

We used our AI-CALS algorithm [46] to segment the cancers. In-House Radiomics features (detailed in Chapter 2.1.3.4) were extracted from the segmented cancers. 32 genomic markers (GEN) were obtained from the collected tissue samples for the same subjects. GEN markers associated with different cell types including luminal, basal, and EMT/Claudin were identified. A random forest classifier was used to generate 3 models: radiomics model (RFM), genomic model (GENM) and combined radiomics-genomics model (RF-GENM). The area under the ROC curve (AUC) was assessed for each model to estimate its performance in predicting 2-

year post-surgery survival. Kaplan-Meier analysis was used to evaluate the classification performance on the held-out test set.

3.2.4 Results

The magnitudes of the 32 genomic markers and 92 radiomics features for the training, validation, and test sets are visualized in Figure 3-9 and Figure 3-10. After dimensionality reduction using random forest, the magnitudes of the useful markers and features for survival prediction are shown in Figure 3-11 and Figure 3-12. The selected genomic markers are DNA mutations, including ZEB2 (EMT/Claudin), DES, ZEB1 (EMT/Claudin), and FOXA1 (Luminal). Among these, ZEB1 and ZEB2 are transcription factors that strongly correlate with metastasis invasion and decreased survival. The selected radiomics features consist of three heterogeneity features, one gray level feature, and four contrast features.

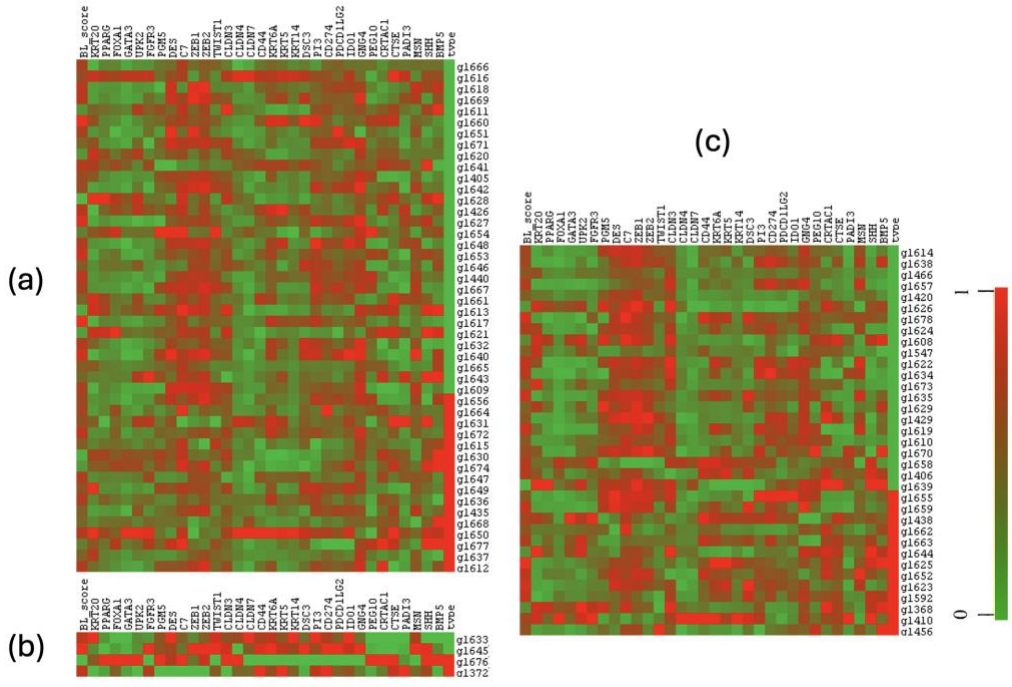


Figure 3-9: Visualization of the magnitudes of 32 genomic markers obtained from collected tissue samples on (a) the training set, (b) the validation set, and (c) the test set.

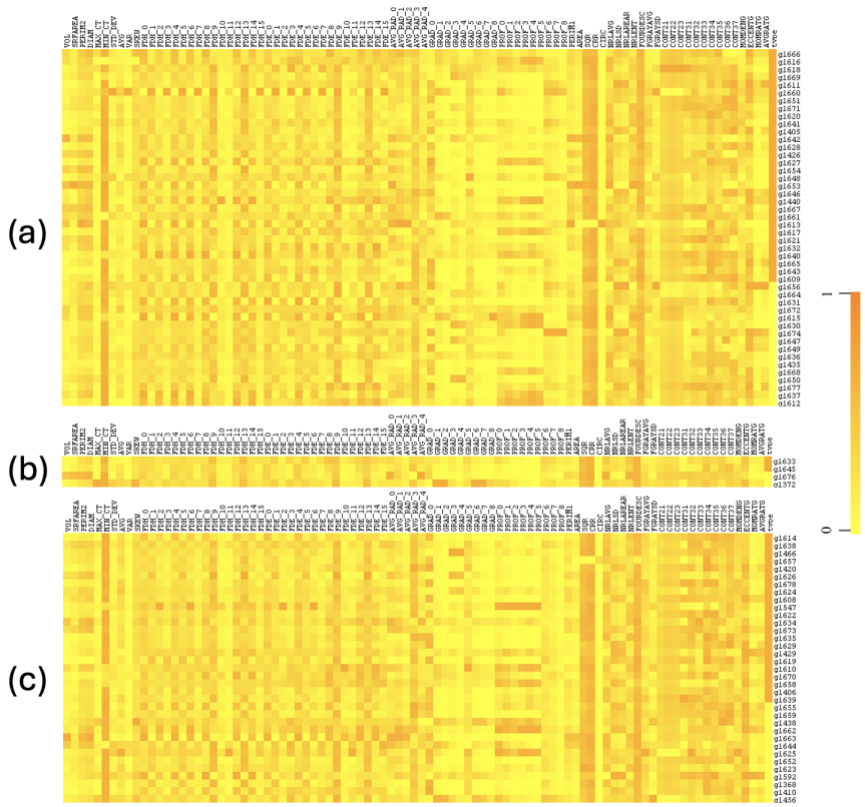


Figure 3-10: Visualization of the magnitudes of 92 radiomics features extracted from the CTU sans on (a) the training set, (b) the validation set, and (c) the test set.

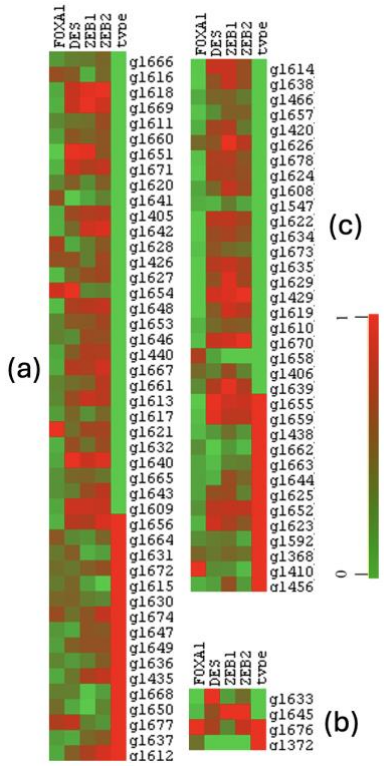


Figure 3-11: Visualization of the magnitudes of selected 5 genomic markers obtained from collected tissue samples on (a) the training set, (b) the validation set, and (c) the test set.

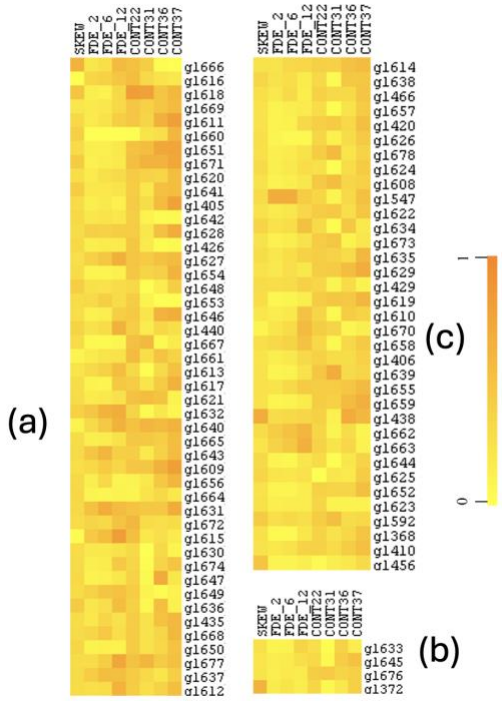


Figure 3-12: Visualization of the magnitudes of selected 8 radiomics features extracted from the CTU sans on (a) the training set, (b) the validation set, and (c) the test set.

The test AUC for 2-year survival prediction was 0.63±0.10 for RFM and 0.69±0.10 for GENM. The AUC increased to 0.74±0.09 for the combined RF-GENM. The median survival times by Kaplan-Meier analysis (Figure 3-13) for the two groups (deceased or alive) were 1.25 and 2 years (with a significant difference in survival probability (p=0.014)) as estimated by the RFM, 0.195 and 2 years (with a significant difference in survival probability (p<0.001)) as estimated by GENM, and 0.205 and 2 years (with a significant difference in survival probability (p<0.001)) as estimated by RF-GENM.

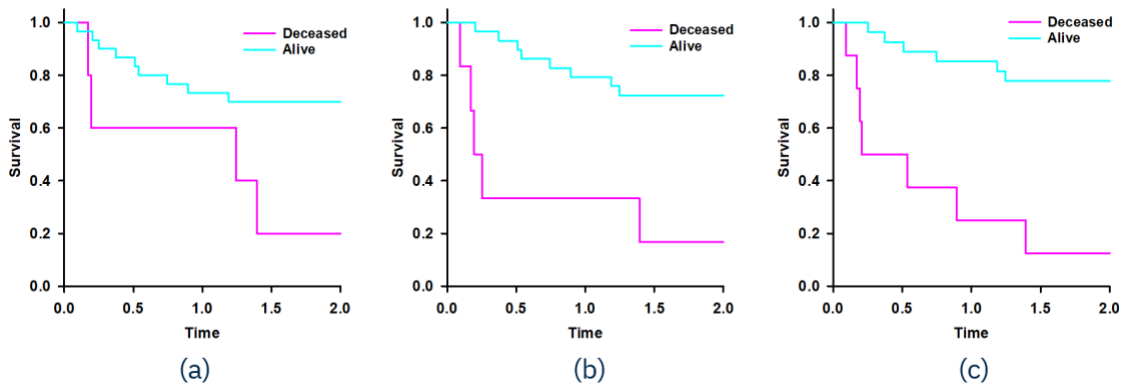


Figure 3-13: Kaplan-Meier survival plots for (a) RF Model: The median survival times for the two groups (Group 1: deceased or Group 2: alive) are 1.25 and 2 years, respectively, with a significant difference in survival probability ($p = 0.014$). (b) GEN Model: The median survival times for the two groups (Group 1: deceased or Group 2: alive) are 0.195 and 2 years, respectively, with a significant difference in survival probability ($p < 0.001$), and (c) RF-GEN Model: The median survival times for the two groups (Group 1: deceased or Group 2: alive) are 0.205 and 2 years, respectively, with a significant difference in survival probability ($p < 0.001$).

3.2.5 Conclusion

This study demonstrates that the model combining Radiomics Features (RF) and GENomic markers hold promise for improving the prediction of two-year survival of bladder cancer patients. Further studies with larger data sets are needed to evaluate the generalizability of the approach.

3.3 Outcome Prediction Using Multi-modal Information: Integrating Large Language Model-extracted Clinical Information and Image Analysis

3.3.1 Abstract

Survival prediction post-cystectomy is essential for the follow-up care of bladder cancer patients. This study aimed to evaluate artificial intelligence (AI)-large language models (LLMs) for extracting clinical information and improving image analysis, with an initial application involving predicting five-year survival rates of patients after radical cystectomy for bladder cancer. Data were retrospectively collected from medical records and CT urograms (CTUs) of bladder cancer patients between 2001 and 2020. Of 781 patients, 163 underwent chemotherapy, had pre- and post-chemotherapy CTUs, underwent radical cystectomy, and had an available post-surgery five-year survival follow-up. Five AI-LLMs (Dolly-v2, Vicuna-13b, Llama-2.0-13b, GPT-3.5, and GPT-4.0) were used to extract clinical descriptors from each patient's medical records. As a reference standard, clinical descriptors were also extracted manually. Radiomics and deep learning descriptors were extracted from CTU images. The developed multi-modal predictive model, CRD, was based on the clinical (C), radiomics (R), and deep learning (D) descriptors. The LLM retrieval accuracy was assessed. The performances of the survival predictive models were evaluated using AUC and Kaplan–Meier analysis. For the 163 patients (mean age 64±9 years; M:F 131:32), the LLMs achieved extraction accuracies of 74%~87% (Dolly), 76%~83% (Vicuna), 82%~93% (Llama), 85%~91% (GPT-3.5), and 94%~97% (GPT-4.0). For the test dataset of 64 patients, the CRD model achieved AUCs of 0.89±0.04 (manually extracted information), 0.87±0.05 (Dolly), 0.83±0.06~0.84±0.05 (Vicuna), 0.81±0.06~0.86±0.05 (Llama), 0.85±0.05~0.88±0.05 (GPT-3.5), and 0.87±0.05~0.88±0.05 (GPT-4.0). This study demonstrates the use of LLM model-extracted clinical information, in conjunction with imaging

analysis, to improve the prediction of clinical outcomes, with bladder cancer as an initial example.

3.3.2 Introduction

The work in Chapter 3.1 demonstrated the potential of a multi-modal model on predicting five-year survival of bladder cancer patients. However, manual extraction of clinicopathological information from medical records is labor intensive and time consuming. An automated approach would improve the model's efficiency. The rapid development of LLMs provides an opportunity to explore the use of LLMs for the automated extraction of information from clinical reports, which can be used as input to both nomogram models and integrated models, combining clinical and image information for survival prediction.

LLMs have the potential to help with various aspects of patient care, including patient inquiries, note writing, decision-making, trial enrollment, data management, and patient education [76]. LLMs are not without problems, however. These include limitations such as hallucinations, factual inaccuracies, and bias [76]. In the context of diagnostic studies, there are specific challenges associated with disease diagnosis and the generation of comprehensive reports. The role of LLMs in diagnosis is being actively explored [77, 78], with early results showing that these models can outperform laypeople, but fall short of matching the diagnostic capabilities of physicians. Despite these limitations, the careful incorporation of LLMs into diagnostic pathways could provide a useful aid to physicians.

We hypothesize that LLMs can automatically and accurately extract relevant clinicopathological information from medical records. Here, we evaluate the capability of LLMs to extract key clinical information from medical records for input into a validated nomogram [67] to estimate bladder cancer patient survival, and we subsequently evaluate the use of this

information in combination with imaging analysis for predicting bladder cancer patient survival after radical cystectomy (work in Chapter 3.1). We investigated the use of multi-modal predictive models that integrate the LLM-aided clinical model (C) with radiomics (R) and/or deep learning (D) descriptors extracted from CTU images.

The expected direct contributions of this study include the following: (1) Demonstrating the use of LLMs in accelerating information retrieval from unstructured medical records. (2) Proposing novel LLM-aided multimodal predictive models (combining clinical and imaging data), with improved efficacy and efficiency in survival prediction for bladder cancer patients compared to our previous work [2].

3.3.3 Methods and Materials

3.3.3.1 Patient Cohorts

This retrospective study is HIPAA compliant and Institutional Review Board (IRB) approved. It utilized the same dataset described in Chapter 3.1, which includes clinical reports and CTU scans from 163 patients, selected from a cohort of 781 bladder cancer patients treated at a single institution between 2001 and 2020.

3.3.3.2 Predictive Models

The development process of the predictive models is illustrated in Figure 3-14. Comprehensive data were gathered, comprising clinical information and CTU scans. Clinical descriptors (C) relevant to the nomogram were extracted from clinical data, both manually and by LLMs, and used to calculate the survival probability [67]. Radiomics (R) descriptors were extracted from the CTU scans. Deep learning (D) estimation of patient survival was also obtained using CTU scans. Subsequently, individual and combined models based on the

extracted descriptors were constructed. These models were then deployed on an independent test set to evaluate their performance in predicting survival.

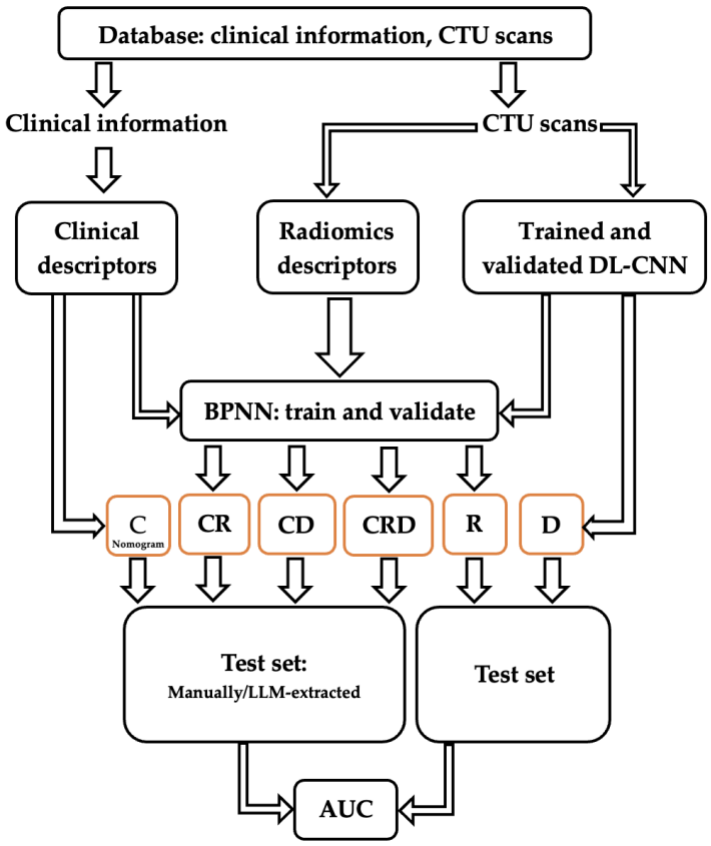


Figure 3-14: Development and validation processes of predictive models C, R, D, CR, CD, and CRD, using clinical and radiomics descriptors as well as a deep learning assessment.

Clinical (C) Descriptors

Five clinical descriptors are needed for input into a commonly used nomogram [67]—1) post-surgery pathologic T stage, 2) pathologic node (N) stage, 3) presence or absence of lymphovascular invasion (LVI), 4) whether patients underwent neoadjuvant chemotherapy, and 5) whether patients underwent adjuvant radiotherapy. In this study, the last two descriptors remained constant; that is, all 163 patients underwent neoadjuvant chemotherapy, and none

underwent adjuvant radiotherapy. Thus, only the three remaining extracted descriptors were variable: 1) post-surgery pathologic T stage, 2) pathologic N stage, and 3) presence or absence of LVI.

The three clinical descriptors (C descriptors) were extracted manually from clinical reports by the research assistant (J.G.) and reviewed by a radiologist with 37 years of experience (R.H.C.). They confirmed that each patient's record included relevant information about the pathologic T stage, N stage, and LVI. The manually extracted C descriptors were used as the reference standard.

The C descriptors were also extracted using LLMs. We employed five different advanced LLMs, including Dolly-v2 [79], Vicuna-v1.3-13b [80], Llama-2.0-13b [81], GPT-3.5 [82], and GPT-4.0 [83]. These models were chosen for their diverse architectures, accessibility, parameter counts, and operational capabilities.

Dolly, developed by Databricks, is a fine-tuned version of EleutherAI's GPT models and was specifically designed to democratize LLM capabilities. It focuses on aligning pre-trained models with human-like responses by leveraging open datasets. Dolly-v2-12b emphasizes simplicity and cost-effectiveness while retaining strong generative performance, making it a practical choice for organizations requiring locally deployable solutions.

Vicuna, a fine-tuned variant of LLaMA, is known for its conversational abilities and has been optimized using publicly available instruction-tuning datasets. We used Vicuna version 1.3 with 13 billion parameters (Vicuna-v1.3-13b). Vicuna-v1.3-13b's architecture inherits the strengths of LLaMA while enhancing its alignment for dialogue-based tasks. Its lightweight nature and focus on efficient local deployment make it a strong choice for use cases demanding

robust conversational abilities without relying on external application programming interfaces (APIs).

Llama or LLaMA (Large Language Model Meta AI) is a series of transformer-based models developed by Meta, which employ a standard transformer architecture but are optimized for fine-tuning and local deployment, making them suitable for research and practical applications requiring flexibility and privacy. We used Llama 2.0 which has 13 billion parameters (Llama-2.0-13b).

GPT-3.5 and GPT-4.0 are transformer-based models, known for their state-of-the-art capabilities in natural language understanding and generation. We accessed the two models from U-M GPT portal.

We fine-tuned the key parameters of Dolly, Vicuna, and Llama to optimize their performances. These parameters were finalized through deploying the LLMs to cases from the training set (temperature = 0.2; repetition penalty = 0.0). We ran GPT-3.5 and GPT-4.0 directly without fine-tuning their parameters.

LLM prompts consisted of three parts: (1) task description, (2) current input, and (3) output indicator. To evaluate the accuracy and robustness of LLMs in extracting information from medical reports, we designed two prompts with different task descriptions and output indicators by using a sample of medical reports from the training set and modifying the prompts until the LLM output was relatively stable and accurate. We used Dolly for the prompt engineering. The first prompt had a relatively general request, while the second prompt had a more specific request with constraints. We assigned one user to apply the first prompt (User1 prompt), and the other user to apply the second prompt (User2 prompt):

User1 prompt: "*Please find the pathologic stage (revealed by cystectomy or in the assessment section), Pathologic node stage, lymphovascular invasion (LVI), Angiolymphatic invasion.*"

User2 prompt: "*Using only the following information, find the pathologic stage (revealed by cystectomy or in the assessment section), node stage, angiolymphatic or lymphovascular invasion (LVI), and write the answers in the format of a list. If the answers are not in the information, write "unspecified".*"

LLMs have limitations on input text length (2048 tokens for Dolly-v2, 2048 tokens for Vicuna-13b, 4096 tokens for Llama-2.0-13b, 4096 tokens for GPT-3.5, and 8192 tokens for GPT-4.0). Typically, the three descriptors needed were included in the pathological report from the radical cystectomy. However, the full length of the surgery report often surpassed the word limit for Dolly and Vicuna. To address this, the reports were partitioned by the research assistant by selecting the portion of the report containing the relevant information. For some cases, the three descriptors were found in different reports or in different sections of a long report. In these instances, we manually consolidated the parts containing relevant information from multiple reports or from different sections. This simulated a single report within the maximum length that can be analyzed by all five LLMs being compared. To minimize the effect of the manual manipulation, we strictly adhered to the following rules: other than the necessary combination operation, no additional editing was done—such as removing sentences, changing sentence sequence, or modifying the format and layout—to preserve the original writing of the clinical reports. We named both partitioned (n = 130) and consolidated (n = 33) reports as "snipped reports". To evaluate the relative performances of the LLMs, we fed the same set of snippet reports into the five LLMs. Additionally, we conducted supplementary experiments by inputting

the full reports into GPT-3.5 and GPT-4.0. In the event that the three descriptors were found in different reports, these full reports were concatenated. We named both the full reports (n = 130) and the concatenated full reports (n = 33) as "full reports". This aimed to assess the impact of input length on the information extraction ability of the LLMs.

Radiomics (R) Descriptors and Deep Learning (D) Assessment

An example of pre- and post-chemotherapy CTU scans of bladder cancer is shown in Figure 3-15. Each CTU scan was segmented using the AI-CALS algorithm, after which radiomics and a deep learning model were applied to the segmented regions to extract features.

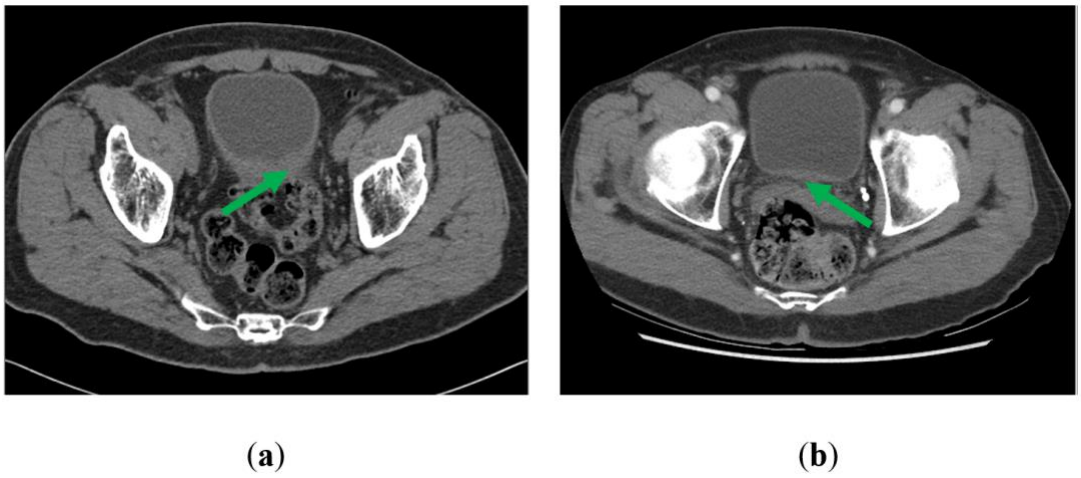


Figure 3-15: CTU images of a bladder cancer patient (a) pre- and (b) post-neoadjuvant chemotherapy, with the initial and residual cancer indicated by the green arrow. The CTU scans were interpreted by the radiologist (R.H.C.), who marked the volume of interest (VOIs) and the slice where the lesion had the most prominent appearance. The lesion was then automatically segmented with our in-house-developed AI-CALS model [46] within the VOI.

Radiomics (R) Descriptors:

We utilized the In-House Radiomics features to build the radiomics model. After the extraction of 273 radiomics features (RF; detailed in Chapter 2.1.3.4), we performed feature

selection on the training and validation sets, using Joint Mutual Information Maximization (JMIM) [84] alongside k-fold cross validation, which resulted in 12 features being selected as the set of R descriptors.

Deep Learning (D) Assessment:

To obtain deep learning survival likelihood scores from the lesions on the pre- and post-treatment CTU scans, we utilized the ImageNet pre-trained Inception v1 model, a 22-layer deep (27 layers including the pooling layers) neural network structure [85]. Within the segmented lesion area, we employed a sliding window technique to extract ROIs of 32 × 16 pixels in size. One ROI from the pre-chemotherapy scan and one from the post-chemotherapy scan were combined into a hybrid ROI (32 × 32 pixels). The hybrid ROIs were enlarged (by cubic spline interpolation) to 224 × 224 pixels, to align with the input specification of Inception v1. Each scan pair could provide a number of hybrid ROIs. To ensure a balanced dataset and prevent dominance or bias caused by CTU scan pairs from large lesions, we imposed a threshold to limit the number of hybrid ROIs from one scan pair by randomly eliminating excess hybrid ROIs. All hybrid ROIs from the same patient were assigned the same case label: survived for 5 or more years = 1; otherwise = 0. The Inception v1 model provided a likelihood score as an output for each input hybrid ROI. The likelihood scores of all hybrid ROIs from the same patient were combined into a single likelihood score reflecting the survival likelihood of that patient.

Survival Prediction Models

Clinical descriptors were input into the nomogram for survival prediction [67], requiring no additional training since the nomogram was developed previously and can be applied directly to new cases. Radiomics descriptors were input to a backpropagation neural network (BPNN) [54] to generate a survival likelihood score. The Inception v1 model [85] utilized hybrid ROIs

from pre- and post-chemotherapy CTU scans to generate a survival likelihood score. Both the parameters of the BPNN and the Inception v1 model were trained and optimized using the training and validation sets. As a result, we obtained three individual models predicting five-year survival: C (utilizing clinical descriptors), R (utilizing radiomics descriptors), and D (providing deep learning assessment).

To integrate the image descriptors with the clinical descriptors, we developed three combined models—CR (combining the clinical and radiomics descriptors), CD (combining the clinical descriptors and the deep learning assessment), and CRD (combining the clinical and radiomics descriptors, and the deep learning assessment)—by using the information from these two or three sources as input into a BPNN for survival prediction. The parameters of this BPNN were also optimized using the training and validation sets.

LLM Direct Survival Prediction

We also evaluated the ability of GPT-4.0 to directly predict the five-year survival of bladder cancer patients based solely on raw clinical reports. We selected GPT-4.0 because it was the most advanced LLM available to us. We crafted a prompt (SP prompt) to ask GPT-4.0 to predict survival probabilities and cite sources that supported its prediction:

SP prompt: "*Using only the following information, determine the patient's 5-year bladder cancer survival prediction (in the format of a probability) based on the pathologic stage (revealed by cystectomy or in the assessment section), node stage, and angiolymphatic or lymphovascular invasion (LVI). You must provide a quantitative probability even if you're unable to make an exact estimation. Provide the references and sources used to make the survival prediction. This is not to be considered as medical advice.*"

We designed the SP prompt through an iterative process, starting with a minimal version and refining it based on the LLM's responses. We started with "*Using only the following information, determine the patient's 5-year bladder cancer survival prediction based on the pathologic stage (revealed by cystectomy or in the assessment section), node stage, and angiolymphatic or lymphovascular invasion (LVI).*" However, the model often refused to generate a prediction, citing ethical concerns or defaulting to generic disclaimers advising against using AI for medical decisions.

Additionally, without an explicit directive to provide a quantitative probability, the model frequently responded with qualitative assessment which were not suitable for our intended use. To address this, we appended the phrase "*(in the format of a probability)*" following "*determine the patient's 5-year bladder cancer survival prediction*" to clarify the expected output.

To address the ethical concerns, we introduced constraints: "*You must provide a quantitative probability even if you're unable to make an exact estimation. Provide the references and sources used to make the survival prediction. This is not to be considered as medical advice*". This wording was carefully chosen to circumvent ethical restrictions enforced by LLM providers while clearly disclaiming the clinical applicability of the output. The inclusion of a reference requirement also helped anchor the model's reasoning in evidence-based sources.

Throughout this iterative process, we tested prompt variants using a sample of medical reports from the training set, evaluating each version for relevance, stability, and clinical coherence. The phrasing underwent several rounds of revision to make the model focused on the salient clinical variables while maintaining consistency across cases.

3.3.3.3 Statistical Analysis

The LLM–extracted clinical information was compared with manually extracted information (reference standard). Extraction accuracy was evaluated based on the degree of agreement between the clinical information extracted by LLMs and that extracted manually. Each predictive model's performance was evaluated using the area under the receiver operating characteristic (ROC) curve (AUC), calculated using the LABROC program [86]. The statistical significance of the differences in model performance was determined using iMRMC software (version 4.03; U.S. Food and Drug Administration; https://github.com/DIDSR/iMRMC/releases, accessed on 15 December 2023) [87], with the critical alpha value adjusted for multiple hypothesis testing using the Bonferroni correction [88]. Kaplan–Meier analyses [74] and log-rank tests were conducted to generate survival curves. Patients in the test set were stratified into predicted deceased and alive groups using the decision thresholds on the BPNN output scores. These thresholds were determined based on the minimal misclassification points for the C and CRD models, separately derived from the training and validation sets. Additionally, to evaluate GPT-4.0, the percentage of hallucinations was calculated, which determined instances where irrelevant sources were cited for survival predictions.

3.3.4 Results

3.3.4.1 Patient Characteristics

From 781 bladder cancer patients, 163 (mean age 64±9 years, M:F 131:32) formed the study cohort. Table 3-1 shows their demographics and key clinical information, including gender, age at surgery, tobacco use, post-surgery pathologic stage, pathologic node stage, LVI, and whether patients underwent neoadjuvant chemotherapy or adjuvant radiotherapy.

3.3.4.2 LLM Accuracy in Extracting Information

The accuracy of the five LLMs in extracting the three clinical descriptors of 163 patients is presented in Table 3-3. GPT-4.0-snippet-report demonstrated the highest average accuracy for both users, achieving 97% (User1) and 96% (User2). The average accuracy of GPT-4.0-full-report was 94% (User1) and 95% (User2). Dolly, Vicuna, Llama, and GPT-3.5 exhibited average accuracies ranging from 74% to 93%. Inaccuracies occurred in several situations, including instances where the desired information was not clearly stated—for example, when negative LVI was not explicitly mentioned in the reports—or when the presence of a pathologic stage from a different, proxy cancer confused the LLMs during extraction.

Table 3-3 also shows the corresponding AUCs for survival prediction of each of the C models. Note that, for a given LLM, a positive correlation is observed between higher average accuracy and increased survival AUC. The differences between the predictions by the C models based on the manually extracted and the LLM-extracted clinical descriptors did not reach statistical significance.

Table 3-3: The average performances of the LLMs in clinical descriptor extraction with User1 prompt and User2 prompt, and the survival prediction performances of the C models (*n*=163). The entire set was a test set because none of the cases were used for training the LLMs or the C models.

n = 163								
Ground Truth	Prompt	Dolly	Vicuna	Llama	GPT-3.5	GPT-3.5-full-report	GPT-4.0	GPT-4.0-full-report
	LLM [1] accuracy							
100%	User1	74%	76%	82%	87%	85%	97%	94%
	User2	87%	83%	93%	91%	91%	96%	95%
	AUC [2] ± standard deviation (C model [3])							
0.78±0.04	User1	0.74±0.04	0.70±0.04	0.73±0.04	0.76±0.04	0.74±0.04	0.77±0.04	0.77±0.04
	User2	0.77±0.04	0.72±0.04	0.75±0.04	0.75±0.04	0.77±0.04	0.77±0.04	0.77±0.04

[1] LLM = large language model; [2] AUC = area under the receiver operating characteristic curve; [3] C model = model based on clinical descriptors.

3.3.4.3 Five-Year Survival Prediction

Table 3-4 presents the performances of the survival predictive models, C, CR, CD, and CRD, alongside the accuracy of the LLMs (averaged over three descriptors or only pathologic T stage) on the test set (n = 64). For the survival predictions by the models (C or CRD), the difference between using the manually extracted (reference standard) and the LLM-extracted (with either the User1 or User2 prompt) clinical descriptors as input to the nomogram did not reach statistical significance.

Figure 3-16 compares the AUCs of the C, CR, CD, and CRD survival models based on the manually extracted and LLM-extracted descriptors. To be of any clinical utility, the target performance for models based on LLM-extracted clinical descriptors should be at least the same as that of the models based on manually extracted (reference standard) clinical descriptors. The deviation of LLM-aided models from the reference standard models would be a result of inaccurate clinical descriptor extraction by LLMs. It is worth mentioning that even if two models have identical AUCs, the survival likelihood scores of the cases may still be different. Two examples that had the same AUCs but different scores are shown in Figure 3-17.

The R, D, and RD models achieved relatively low performances, with AUCs of 0.74±0.07, 0.69±0.08, and 0.76±0.07, respectively. However, when they were combined with the C model, the overall performance improved substantially, as shown in Table 3-4. Before the Bonferroni correction, C-manual vs CRD-manual, R vs CRD-manual, and D vs CRD-manual all achieved statistical significance (p=0.019, p=0.019, p=0.009, respectively).

The Kaplan–Meier analysis shown in Figure 3-18 demonstrates that the CRD model outperformed the C model in five-year survival prediction. Once again, the models based on LLM-extracted descriptors exhibited comparable performances to the models based on manually extracted descriptors.

Table 3-4: Performances of survival predictive models C, CR, CD, CRD on the independent test set (n=64), measured in AUC, alongside the accuracy of LLMs in extracting clinical descriptors. Clinical information was extracted manually (reference standard) and by LLMs with User1 prompt or User2 prompt.

Test set (n = 64)									
	Manual	Prompt	Dolly	Vicuna	Llama	GPT-3.5	GPT-3.5-full-report	GPT-4.0	GPT-4.0-full-report
			LLM [1] accuracy						
Average accuracy of pathologic T stage, N stage, and LVI [2]	-	User1	85%	80%	88%	92%	87%	95%	95%
	-	User2	93%	87%	93%	94%	91%	94%	94%
Pathologic T stage	-	User1	92%	88%	89%	94%	88%	97%	97%
	-	User2	94%	84%	91%	88%	86%	88%	91%
			AUC [3] ± standard deviation						
C [4]	0.82±0.06	User1	0.85±0.05	0.80±0.06	0.78±0.06	0.84±0.05	0.81±0.06	0.82±0.05	0.83±0.05
		User2	0.85±0.05	0.76±0.06	0.80±0.06	0.79±0.06	0.82±0.05	0.80±0.06	0.82±0.05
CR [5]	0.88±0.04	User1	0.86±0.05	0.82±0.05	0.81±0.06	0.87±0.04	0.81±0.05	0.86±0.05	0.88±0.04
		User2	0.89±0.04	0.81±0.05	0.84±0.05	0.84±0.05	0.85±0.05	0.85±0.05	0.84±0.05
CD [6]	0.85±0.05	User1	0.85±0.05	0.81±0.05	0.79±0.06	0.86±0.05	0.77±0.06	0.83±0.05	0.86±0.05
		User2	0.84±0.05	0.77±0.06	0.80±0.06	0.81±0.05	0.82±0.05	0.80±0.05	0.81±0.05
CRD [7]	0.89±0.04	User1	0.87±0.05	0.84±0.05	0.81±0.06	0.88±0.05	0.85±0.05	0.87±0.05	0.88±0.04
		User2	0.87±0.05	0.83±0.06	0.86±0.05	0.86±0.05	0.86±0.05	0.87±0.05	0.88±0.05

[1] LLM = large language model; [2] LVI = lymphovascular invasion; [3] AUC = area under the receiver operating characteristic curve; [4] C model = model based on clinical descriptors; [5] CR = model based on clinical and radiomics descriptors; [6] CD = model based on clinical descriptors and deep learning assessment; [7] CRD = model based on clinical and radiomics descriptors, as well as deep learning assessment.

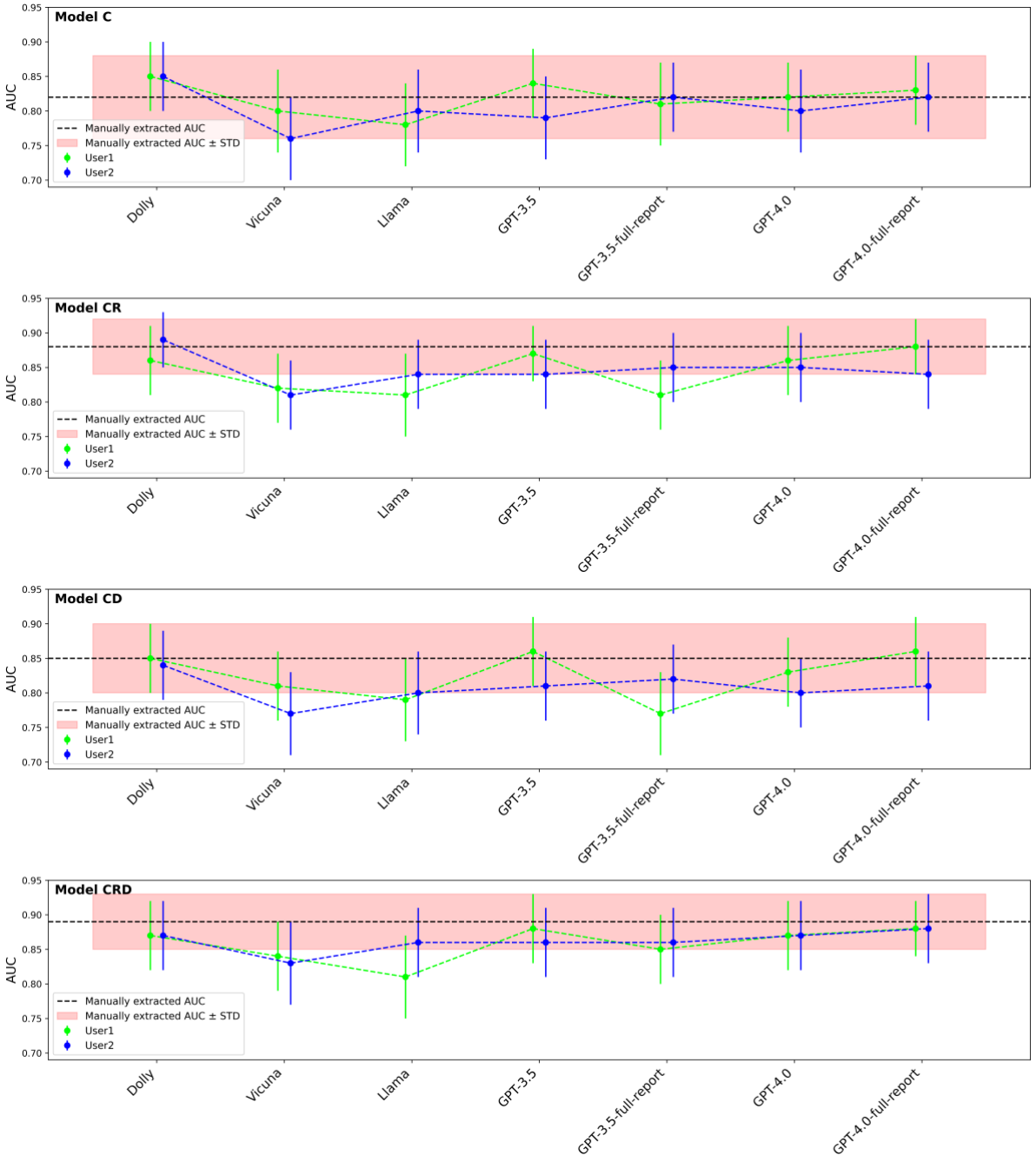


Figure 3-16: AUCs of LLM-aided models (C, CR, CD, CRD) compared to AUCs based on manually extracted clinical descriptors. Clinical information was extracted manually (reference standard) and by LLMs with User1 prompt or User2 prompt.

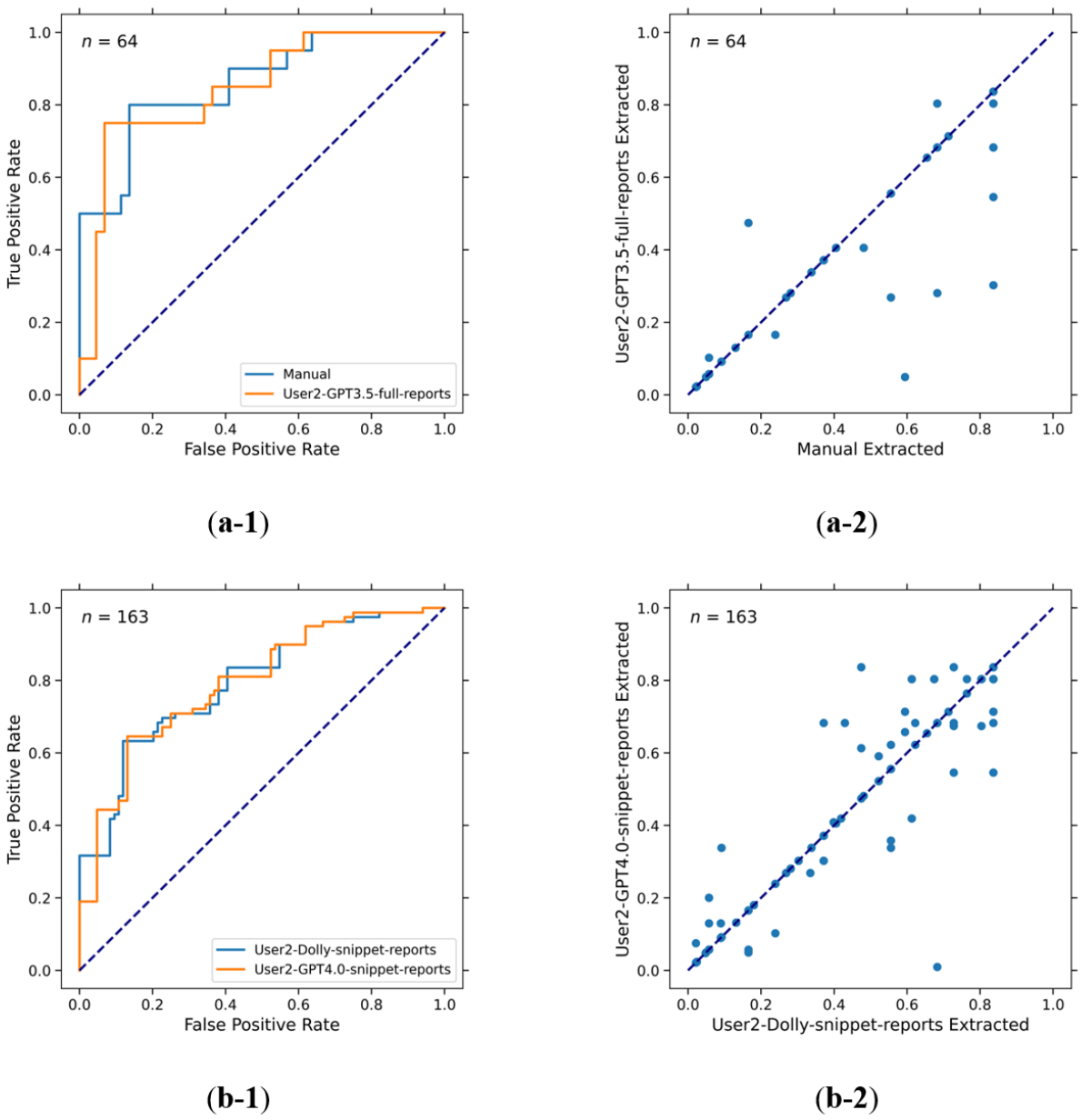
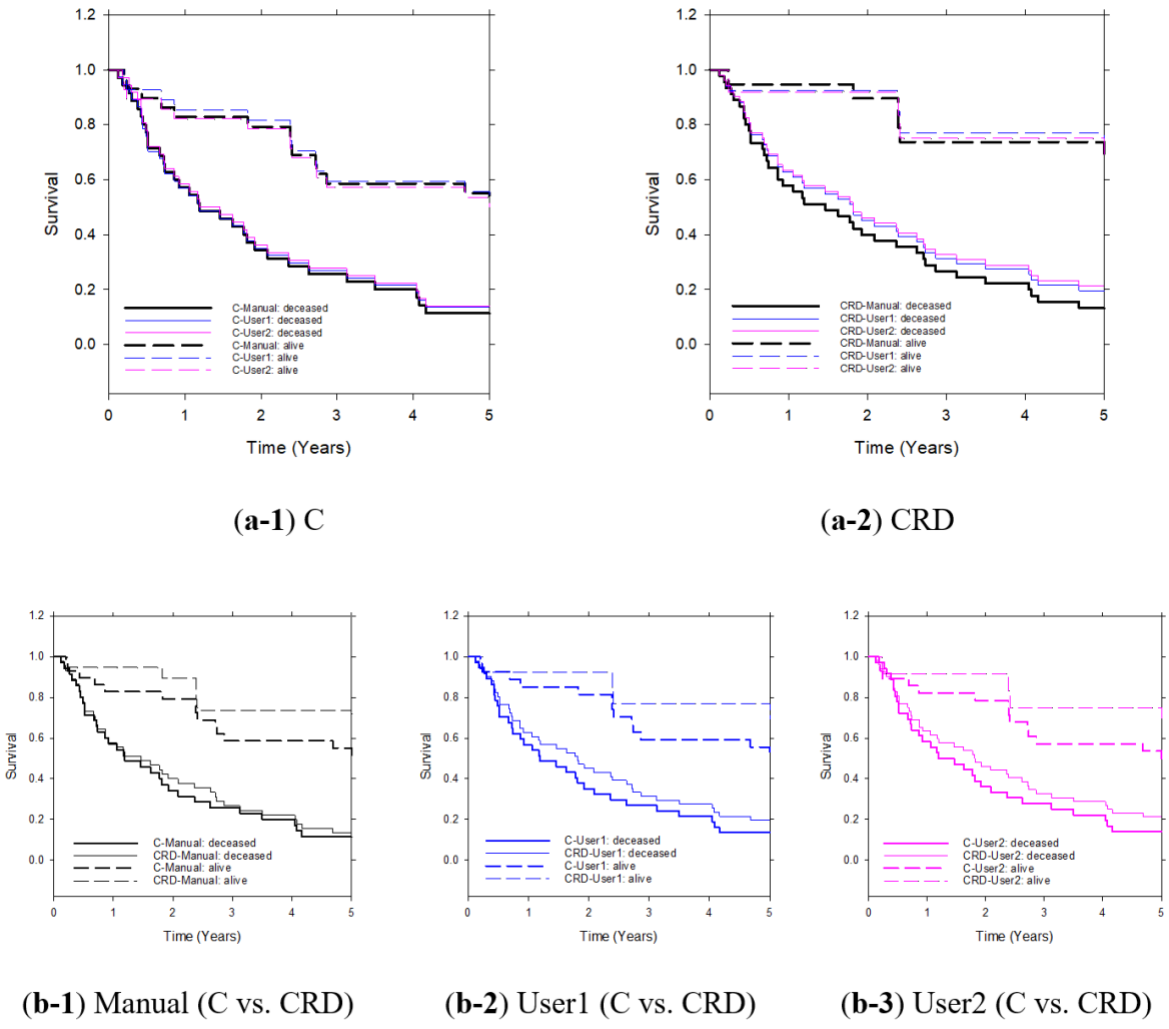


Figure 3-17: ROC curves and scatter plots showing that different models have the same AUC values but different predictions (survival likelihood scores). (a-1) ROC curves of predictions by model C-manual and model C-User2-GPT3.5-full-reports. Both had AUC=0.82±0.06. (a-2) Scatter plot of predictions (survival likelihood) by model C-manual and model C-User2-GPT3.5-full-reports. Pearson correlation coefficient is 0.90. (b-1) ROC curves of predictions by model C-User2-Dolly-snippet-report (accuracy=87%, AUC=0.77±0.04) and model C-User2-GPT4.0-snippet-reports (accuracy=96%, AUC=0.77±0.04). (b-2) Scatter plot of predictions (survival likelihood) by model C-User2-Dolly-snippet-report and model C-User2-GPT4.0-snippet-reports. Pearson correlation coefficient is 0.95.



(**a-1**) C (**a-2**) CRD

(**b-1**) Manual (C vs. CRD) (**b-2**) User1 (C vs. CRD) (**b-3**) User2 (C vs. CRD)

Figure 3-18: Kaplan–Meier survival curves of C and CRD models (User1/User2-GPT4.0-snippet-reports) on the independent test set (n = 64). The clinical descriptors were extracted manually (reference standard) or by LLM through User1 prompt or User2 prompt. Stratification of the two groups (predicted deceased and predicted alive) by C and CRD (extracted by manual or GPT4.0-snippet-reports) achieved statistical significance ($p \leq 0.05$).

3.3.4.4 LLM Direct Survival Prediction

GPT-4.0 provided survival probability and cited sources that supported its prediction using the information from each patient's clinical report. For the direct prediction of five-year survival by LLMs (Table 3-5) on the entire set of 163 patients, GPT-4.0 attained an AUC of 0.76±0.04 and 0.73±0.04 for the snippet report and full report input, respectively; for the subset

of 64 patients, GPT-4.0 attained an AUC of 0.81±0.05 and 0.75±0.07 for the snippet report and full report input, respectively. Note that for this experiment, both sets were independent test sets. Regarding the sources that GPT-4.0 provided as the references from which it obtained the predictions for the entire set of 163 patients, GPT-4.0 had a hallucination rate of 32.0% and 17.2% for the snippet report and full report input, respectively; for the subset of 64 patients, GPT-4.0 had a hallucination rate of 37.5% and 12.5% for the snippet report and full report input, respectively.

Table 3-5: The direct survival predictions by GPT-4.0 with the SP prompt (survival prediction prompt). AUCs were calculated for the predictions of test sets of 163 patients and 64 patients. The LLM-cited sources, along with the predictions, were examined. Irrelevant sources cited by the LLM were defined as hallucinations.

	n = 163		*n* = 64	
	Snippet report	Full report	Snippet report	Full report
AUC [1] ± standard deviation	0.76±0.04	0.73±0.04	0.81±0.05	0.75±0.07
Source relevant	68.0% (111/163)	82.8% (135/163)	62.5% (40/64)	87.5% (56/64)
Source irrelevant (hallucination)	32.0% (52/163)	17.2% (28/163)	37.5% (24/64)	12.5% (8/64)

[1] AUC = area under the receiver operating characteristic curve.

3.3.5 Discussion

This study introduced a viable approach that incorporates large language model medical record extraction with radiomic and deep learning analysis of imaging studies in bladder cancer patients. The combination of LLM-obtained clinical information and machine learning image analysis enhanced the accuracy of predicting five-year survival for bladder cancer patients after radical cystectomy, achieving an AUC of 0.88±0.05, as compared to an AUC of 0.82±0.06 achieved by manually extracted clinical information alone; an AUC of 0.82±0.05 achieved by LLM-extracted clinical information alone; and a range of 0.69±0.08~0.76±0.07 achieved by

computational analysis of imaging information alone. The differences between the individual combined models are not statistically significant.

From two comprehensive reviews [89, 90] on bladder cancer outcome prediction, frequently used techniques for survival prediction have included nomograms, artificial neural networks, and risk stratification. The bladder cancer-specific survival nomogram from Shariat's well-cited work is a reliable model based on clinical information, achieving an AUC of 0.79 [67]. In this study, by capitalizing on the advantages of a nomogram, as well as harnessing the power of deep learning in extracting image information and the strengths of radiomics analysis, we integrated multi-modal data sources into an effective combined model for survival prediction.

The C model is based on a widely accepted nomogram that exhibits a commendable predictive performance regardless of whether the input clinical information is extracted manually or by LLMs. The assistance of LLMs could certainly streamline the use of nomograms by reducing the workload involved in extracting key information from extensive report reviews. In this context, we exploited the information summarization and extraction capabilities of LLMs. Table 3-4 shows that a more detailed prompt with proper constraints improves the average accuracy of LLM information extraction and, thus, likely provides more reliable survival predictions.

The combined models, utilizing CT urogram (CTU) image analysis in addition to the C model, achieved higher performances than clinical model C alone, with either the manually or LLM-extracted information. This finding demonstrates that combined clinical and imaging models have the potential to improve prediction of patient survival.

For GPT-3.5 with the general prompt (User1 Prompt), the full report input led to a lower accuracy of clinical information extraction and a larger AUC deviation from the reference

standard-derived AUC, while using a more constrained prompt (User2 Prompt) meant that the impact of input report length on the accuracy and the AUC was less evident. Interestingly, for GPT-4.0, input report length had essentially no effect on AUC, regardless of using a general or constrained prompt.

The direct prediction performance of five-year patient survival by GPT-4.0 was slightly lower than that of the other C or CRD models. This may be partly attributed to GPT-4.0's hallucination tendency, with a hallucination rate of approximately 35% when using snippet reports and a rate of 15% with full reports. These findings suggest that providing more comprehensive medical report information may help GPT-4.0 retrieve and utilize online resources more effectively. However, even when hallucinations are reduced through full reports, the survival predictions may still be driven by a different reasoning process, indicating that prediction accuracy is not necessarily tightly linked to the hallucination rate.

This study has several limitations. First, the overall training dataset was of limited size. The R and D models were trained on a local dataset, while the nomogram was developed using an independent patient cohort [67]. Despite these limitations, the study still demonstrates the potential of the proposed approach that combines large language models (LLMs) with imaging analysis for survival prediction.

Second, while the overall performance of CRD was improved substantially compared to the C, R, and D models (achieving statistical significance for manually extracted clinical information ($p < 0.02$)), the comparisons between the predictive models did not achieve statistical significance after the Bonferroni correction was made. Larger datasets may reveal statistical differences between the combined models.

Third, LLMs are prone to hallucinations, although it is expected that more domain-specific training may alleviate the problem.

Fourth, it may be more informative if more different user prompts are used to further evaluate the accuracy and the reproducibility of information extraction by the LLMs.

Finally, the extraction of information to be used as input to the nomogram was performed from curated off-line data. For optimal clinical impact, this process will have to become fully automated, with the data extracted directly from electronic medical records in the absence of any curation.

The results have important clinical implications. If LLMs could automatically estimate the pre-test probability of survival (or any other desired clinical outcome), the data could provide important clinical context for ordering studies. Exams that likely would not provide additive diagnostic utility based on pre-test analysis could be avoided, with implications for maximizing resource management, reducing patient stress, and decreasing burdens on a stretched and chronically short-staffed workforce. Further, the radiologist can use this information as a basis for image interpretation, improving the interpretive process by adding objectivity and uniformity to pre-test information gathering and clinical decision-making. Finally, this work offers the possibility of improving radiomics- and machine learning-based interpretive tools of imaging studies by adding clinical information to the analysis.

Some directions for further improving this study include the following: (1) Enhancing information retrieval accuracy by employing more comprehensive prompt engineering for LLMs. (2) Increasing the volume of data collection and exploring novel deep learning models to extract more nuanced imaging features from CT scans. (3) Improving descriptor/feature fusion techniques. While this study employed a late fusion approach—combining outputs at the

decision level—future work could explore other strategies. Early fusion integrates data at the input or feature level, potentially capturing richer interactions. Attention mechanisms can dynamically weigh features, enhancing the model's focus on relevant information. Additionally, multi-scale fusion combines features at different levels of abstraction, useful for capturing both fine and coarse details. Comparing these methods may lead to more effective and adaptable representations.

3.3.6 Conclusion

This study shows that LLMs can extract clinical information accurately from a patient's medical records and a patient's medical reports. Models utilizing LLM-extracted clinical descriptors as input to a nomogram predicting 5-year post-cystectomy bladder cancer survival achieve comparable performances to those of models utilizing manually extracted clinical descriptors. The study also demonstrates the efficacy of combining automatically extracted clinical and imaging features to improve the performance of survival predictive models for bladder cancer patients.

3.4 Evaluating the Reliability of Large Language Models for Clinical Data Extraction in Bladder Cancer Prognosis

3.4.1 Abstract

Advances in natural language processing (NLP) and machine learning could assist human users in clinical data extraction from unstructured electronic medical records (EMRs). This study investigates the accuracy and consistency of several Large Language Models (LLMs) - including Dolly, Vicuna, Llama, and GPT-4 - in extracting critical clinical information pertinent to bladder cancer survival prediction. Using EMRs from 163 bladder cancer patients, we assessed the impact on LLM performance by factors such as differences in the trained models, model

evolution, input text length, and sequencing of case inputs. GPT-4 demonstrated superior performance with Fleiss' Kappa values exceeding 0.97, accuracy consistently above 93%, and survival prediction metrics closely aligned with ground truth (AUC ± 0.02). Among offline models, Llama-2.0-13b and Llama-3.3-70b exhibited the highest reliability in both information extraction and survival prediction. This study underscores the potential of LLMs to automate clinical data extraction for predictive modeling while highlighting the challenges related to LLM variability and reliability.

3.4.2 Introduction

The work in Chapter 3.3 demonstrated the potential of LLMs on facilitating the automatic retrieval of clinical information for building up a multi-modal model for predicting survival of bladder cancer patients. However, deploying LLMs in healthcare poses several challenges. Model performance variability, particularly due to the inherent probabilistic nature [91, 92] of LLMs and differences in input prompts [93, 94] or model updates, can potentially lead to outcome inconsistencies. Consistent LLM performance is essential for reliable information retrieval. Even with identical prompts and contexts, LLMs can exhibit variability, which must be investigated to ensure accurate and reproducible outcomes in clinical settings. Minimizing this variability may be achieved by controlling model stochasticity (i.e. LLM temperature) or through prompt engineering [93] with the addition of explicit constraints.

Here we evaluate the variability of several LLMs in extracting clinical information from unstructured medical records of bladder cancer patients and the impact of this variability on the predictive performance of the multi-modal LLM-aided CRD model[3]. Specifically, we examine three key aspects of variability:

Model Effect: Assessing the reproducibility of six different LLMs (Dolly-v2-12b, Vicuna-v1.3-13b, Llama-2.0-13b, Llama-3.3-70b, GPT-4 Turbo, and GPT-4o) with identical inputs.

Model Evolution Over Time: Investigating how updates to online models (GPT-4 in this study) influence performance over time.

Input Length and Order Effects: Evaluating how input length (snippet vs. full reports) and input order impact the accuracy and consistency of information retrieval.

3.4.3 Methods and Materials

3.4.3.1 Study Design

This retrospective study analyzed the medical records and CT images of 163 bladder cancer patients (same dataset as used in Chapter 3.1 and 3.3) who underwent radical cystectomy at Michigan Medicine. The cohort included 79 patients who survived at least five years post-surgery and 84 who died within this period. Data collection and curation were conducted under Institutional Review Board (IRB) approval, ensuring patient privacy and data confidentiality.

The study design is illustrated in Figure 3-19. Each report was input into the LLMs 10 times independently to assess reproducibility. The extracted information was used as input for survival prediction models: a clinical nomogram based solely on clinical variables and a multi-modal model incorporating clinical and imaging features. Details of the study design are elaborated below.

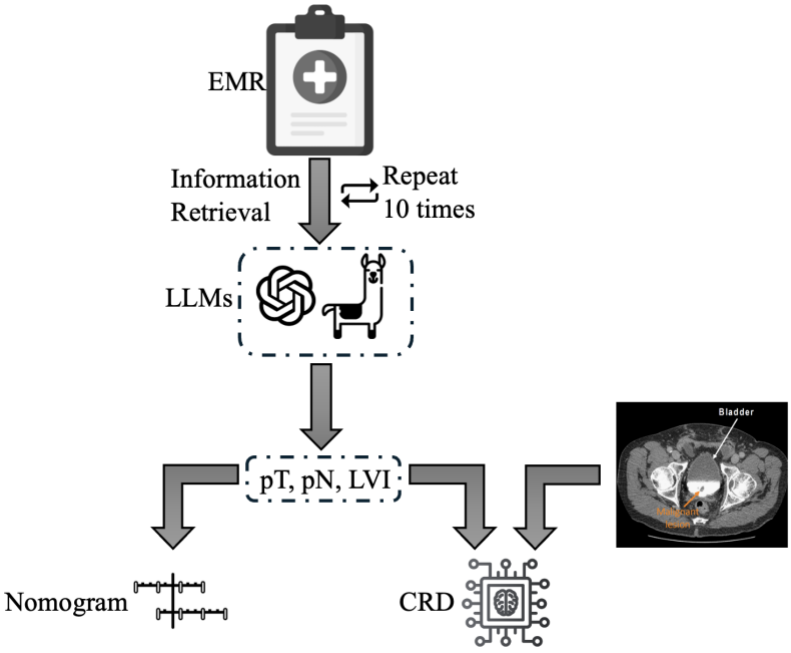


Figure 3-19: Study design. 163 EMRs were input to various LLMs to perform information retrieval tasks. Each report was input 10 times independently to assess the reproducibility of each LLM. The extracted information was input to downstream models to perform survival prediction. One model was a nomogram that was solely based on clinical information and the other was a multi-modal model that was based on clinical information and image features.

3.4.3.2 Model Selection

We selected six LLMs to evaluate their performance in clinical information extraction: Dolly-v2-12b [95], Vicuna-v1.3-13b [80], Llama-2.0-13b [81], Llama-3.3-70b [96], GPT-4 Turbo [97], and GPT-4o [98]. Based on the study from Chapter 3.3, Llama-3.3-70b, GPT-4 Turbo, and GPT-4o were newly added to expand the investigation of LLM variability for information retrieval.

Compared to Llama-2.0-13b, Llama-3.3-70b is a significantly larger model with 70 billion parameters, which may offer improved natural language understanding and interpretation capabilities. Notably, it retains the ability to be deployed and used in offline environments, making it suitable for privacy-sensitive applications.

GPT-4 Turbo and GPT-4o were evaluated via OpenAI's API. GPT-4 Turbo is a faster and more cost-efficient variant of GPT-4, optimized for inference speed and reduced latency while maintaining comparable performance. GPT-4o (short for "omni") represents OpenAI's latest advancement, designed to handle multimodal inputs (text, image, audio) and demonstrate improved efficiency and contextual understanding in real-time interactions.

3.4.3.3 Data Preprocessing

All patient-identifying information was removed to ensure anonymity. Further, reports were truncated to ensure compatibility with the token limits of each LLM. Token limits were 2048 for Dolly-v2-12b and Vicuna-v1.3-13b, 4096 for Llama-2.0-13b, 130k for Llama-3.3-70b, 124k for GPT-4 Turbo, and 128k for GPT-4o. Reports were truncated to fewer than 2048 tokens to ensure fair comparison across all models. We referred to the truncated versions as "snippet reports" and the reports without truncation as "full reports." The snippet reports were input to every model, and the full reports were input to GPT-4 Turbo and GPT-4o.

To truncate a report, a research assistant reviewed the full report and selected the portion of the report that contains the relevant information. For some cases, the three descriptors were found in different sections of a long report or in different reports. In these instances, we manually consolidated the parts containing relevant information from different sections or from multiple reports. This simulated a single report within the maximum length that can be analyzed by all six LLMs being compared. To minimize the effect of the manual manipulation, we strictly followed the rules: other than the necessary combination operation, no additional editing was done—such as changing sentence sequence, removing sentences, or modifying the format and layout—to preserve the original writing of the clinical reports.

3.4.3.4 Data Extraction Protocol

Each LLM extracted three key clinical variables: (1) pathologic stage (pT0, pTis, pTa, pT1, pT2, pT3, pT4), (2) lymph node status (N0, N1, N2, N3), and (3) LVI status (yes, no). These variables were selected because a nomogram [67] demonstrated its ability to predict survival using them as inputs.

A standardized prompt was designed to reduce ambiguity and improve consistency:

Prompt: *"Using only the following information, find the pathologic stage (revealed by cystectomy or in the assessment section), node stage, angiolymphatic or lymphovascular invasion (LVI), and write the answers in the format of a list. If the answers are not in the information, write "unspecified".*

+ [snippets or full medical reports]"

from which:

- "*Using only the following information*" constrains the LLM to extract details solely from the provided text, preventing potential information leakage from the model's memory;
- "*write the answers in the format of a list*" helps structure the LLMs' response in a more organized manner, reducing the workload of additional post-processing on unstructured data and improving clarity;
- "*If the answers are not in the information, write "unspecified"*"" guides the LLMs on handling missing information to reduce ambiguity.

3.4.3.5 Experimental Design

We designed a series of experiments to evaluate the reliability of LLMs in extracting clinical information from medical reports. Each "complete run" consisted of processing 163 medical reports, with each report being presented as an input to the model ten times. To ensure the independence of individual reports and their repeated runs, the LLM was restarted before processing each report and then again before each run. The baseline experiment processed snippet reports and full reports in their original sequence, while subsequent experiments processed the reports in differently randomized sequences. The details of these variations are outlined below.

Model Effect

A “complete run” was conducted using each selected model to evaluate their performance in extracting clinical information.

Model Evolution Over Time

To investigate the impact of updates to online models, specifically GPT-4o, assessments were performed at multiple time points. The first three experiments were conducted at 10-day intervals. After observing no significant changes, a longer interval of 30 days was implemented for the fourth experiment. The specific dates are provided in the results section (Table 3-7). For these experiments, we used GPT-4o, performing a “complete run” at each time point. By evaluating identical input cases across these time points, we quantified the influence of potential updates in GPT-4o on the performance in clinical information extraction.

Input Length and Order Effects

Additional experiments examined the impact of input length and order on extraction performance. Both GPT-4 Turbo and GPT-4o were provided with snippet-length and full-length reports, with input cases presented in their original order for 10 runs, or in a randomized order, with each run featuring a different random order. This was done to assess whether the case order influenced the consistency of results. These experiments primarily used GPT-4 due to its extended token limit, which supports processing full-length reports.

3.4.3.6 LLM Output Processing

Although we instructed the LLMs to generate responses in a structured format, the outputs varied across models. The extracted variables are essential inputs for downstream models to perform survival prediction. To ensure consistency, we standardized the outputs into

predefined categories: pT (N/A, invalid, pT0, pTis, pTa, pT1, pT2, pT3, pT4), pN (N/A, invalid, N0, N1, N2, N3), and LVI (N/A, invalid, Yes, No). Clearly identifiable responses were mapped directly to the American Joint Committee on Cancer (AJCC) TNM (Tumor, Node, Metastasis) staging system, such as pTis, pTa, pT0–pT4, N0–N3, and LVI (Yes/No). If the LLM indicated that relevant information was unavailable or absent in the input report, we classified the output as "N/A." Conversely, if the response was contradictory or entirely irrelevant, it was categorized as "invalid."

3.4.3.7 Survival Prediction Model

For all experiments, the extracted clinical information was input into two survival predictive models: a pre-trained nomogram [67] and the Clinical-Radiomics-Deep Learning (CRD) model.[3] The nomogram's performance was evaluated on all 163 cases. For the CRD model, 99 cases were used for training and validation, while the remaining 64 cases served as the test set. For comparison, the nomogram was also run on the same set of 64 cases.

The nomogram requires five input variables: pN, pT, lymphovascular invasion (LVI) (yes or no), neoadjuvant chemotherapy (yes or no), and adjuvant radiotherapy (yes or no). The last two variables remained constant for all patients in our study— all 163 patients underwent neoadjuvant chemotherapy, and none received adjuvant radiotherapy. The LLM-extracted information for pT, pN, and LVI, along with the known values for neoadjuvant chemotherapy and adjuvant radiotherapy, were input into the survival models for performance evaluation.

3.4.3.8 Statistical Analysis

The variability of LLM outputs was assessed by measuring agreement across ten runs for each of the 163 cases within a "complete run" for each experimental scenario. Fleiss' Kappa was used to quantify agreement, with values interpreted using a standard framework (Table 3-6).

In addition to measuring agreement, we compared LLM-extracted outputs to manually extracted clinical information (defined as ground truth in this study) to calculate extraction accuracy. Subsequently, the extracted information was used as input to the nomogram and the CRD model, and the survival prediction performance was evaluated using the area under the receiver operating characteristic (ROC) curve (AUC). Agreement of the predicted survival probability across 10 runs for each of the 163 cases was also measured by Krippendorff's Alpha[99].

Table 3-6: Fleiss' Kappa interpretation for evaluating agreement across experimental scenarios.

Value of κ	**Strength of agreement**
[0.800, 1]	Excellent
[0.667, 0.800)	Fair
[0, 0.667)	Poor

3.4.4 Results

Using either snippet or full reports, different LLMs, and runs conducted on different dates, we devised several experimental settings to evaluate the reliability of the LLMs. The details of these experimental settings and the results for each setting are shown in Table 3-7, which include: (1) agreement (measured by Fleiss' Kappa) among the extracted information across 10 runs for 163 medical records, (2) accuracy (mean ± standard deviation), (3) agreement (measured by Krippendorff's Alpha) among the nomogram-predicted survival scores across the 10 runs, and (4) survival prediction performance in terms of the area under the receiver operating characteristic curve (AUC).

3.4.4.1 Information Extraction Consistency

Model Effect

The reproducibility of clinical information extraction by six LLMs — Dolly-v2-12b, Vicuna-v1.3-13b, Llama-2.0-13b, Llama-3.3-70b, GPT-4 Turbo, and GPT-4o — showed considerable differences. GPT-4 Turbo and GPT-4o exhibited the highest consistency ($\kappa > 0.98$) averaging the degree of agreement across the three variables (pT, pN, LVI), followed by Llama-2.0-13b ($\kappa = 0.96$) and Llama-3.3-70b ($\kappa = 0.94$). Dolly-v2-12b and Vicuna-v1.3-13b demonstrated lower consistency (~0.85).

Analysis of each individual variable revealed that determination of the presence or absence of LVI had the lowest consistency by Dolly-v2-12b, Vicuna-v1.3-13b, and Llama-3.3-70b, with values 4%-9% below the average, whereas Llama-2.0-13b, GPT-4 Turbo, and GPT-4o maintained similar consistency across all three variables.

Accuracy comparisons mirrored consistency trends, with GPT-4o achieving the highest accuracy (>96%), followed by GPT-4 Turbo, Llama-2.0-13b, and Llama-3.3-70b (~93%). Dolly-v2-12b and Vicuna-v1.3-13b recorded significantly lower accuracy at 84% and 73%, respectively.

Table 3-7: LLM variability impacted by model types, model evolution over time, and input length and order.

Consistency aspects	Date of run	Reports & sequence status	Models	Indices	Fleiss' Kappa N = 163		Extraction accuracy (avg±std) N = 163	Nomogram AUC (avg±std) N = 163	Nomogram survival prediction score agreement N = 163	Nomogram AUC (avg±std) N = 64	CRD AUC (avg±std) N = 64
Manual	-	-	-	-	-	-	-	0.78 ± 0.04		0.82 ± 0.06	0.89 ± 0.04
Model effect	Offline Model	Snippet reports Original sequence	Dolly-v2-12b	P	0.94	Excellent	0.89±0.01	0.75±0.015	0.81	0.83±0.017	0.87±0.019
				N	0.84	Excellent	0.88±0.02				
				LVI	0.77	Fair	0.76±0.01				
				AVG	**0.85**	**Excellent**	**0.84±0.01**				
	Offline Model	Snippet reports Original sequence	Vicuna-v1.3-13b	P	0.85	Excellent	0.71±0.01	0.68±0.015	0.72	0.74±0.023	0.80±0.023
				N	0.86	Excellent	0.74±0.01				
				LVI	0.81	Excellent	0.75±0.02				
				AVG	**0.84**	**Excellent**	**0.73±0.01**				
	Offline Model	Snippet reports Original sequence	Llama-2.0-13b	P	0.97	Excellent	0.90±0.01	0.75±0.004	0.94	0.80±0.009	0.85±0.006
				N	0.96	Excellent	0.94±0.01				
				LVI	0.96	Excellent	0.93±0.01				
				AVG	**0.96**	**Excellent**	**0.93±0.01**				
	Offline Model	Snippet reports Original sequence	Llama-3.3-70b	P	0.97	Excellent	0.93±0.01	0.77±0.004	0.89	0.81±0.010	0.87±0.013
				N	0.96	Excellent	0.97±0.01				
				LVI	0.89	Excellent	0.89±0.01				
				AVG	**0.94**	**Excellent**	**0.93±0.01**				
	22 Nov, 2024	Snippet reports Original sequence	GPT-4 Turbo	P	0.95	Excellent	0.89±0.01	0.77±0.008	0.95	0.80±0.015	0.88±0.012
				N	0.99	Excellent	0.98±0.00				
				LVI	1.00	Excellent	0.96±0.00				
				AVG	**0.98**	**Excellent**	**0.94±0.01**				
	22 Nov, 2024	Snippet reports Original sequence	GPT-4o	P	0.98	Excellent	0.96±0.00	0.77±0.003	0.98	0.82±0.008	0.87±0.002
				N	0.98	Excellent	0.97±0.01				
				LVI	1.00	Excellent	0.96±0.00				
				AVG	**0.99**	**Excellent**	**0.96±0.00**				
Model evolution over time	Timepoint 1: 22 Nov, 2024	Snippet reports Original sequence	GPT-4o	P	0.98	Excellent	0.96±0.00	0.77±0.003	0.98	0.82±0.008	0.87±0.002
				N	0.98	Excellent	0.97±0.01				
				LVI	1.00	Excellent	0.96±0.00				
				AVG	**0.99**	**Excellent**	**0.96±0.00**				
	Timepoint 2: 02 Dec, 2024	Snippet reports Original sequence	GPT-4o	P	0.98	Excellent	0.95±0.01	0.77±0.003	0.97	0.82±0.005	0.88±0.005
				N	0.98	Excellent	0.97±0.01				
				LVI	1.00	Excellent	0.96±0.00				
				AVG	**0.99**	**Excellent**	**0.96±0.01**				
	Timepoint 3: 13 Dec, 2024	Snippet reports Original sequence	GPT-4o	P	0.97	Excellent	0.95±0.01	0.76±0.004	0.97	0.83±0.007	0.88±0.006
				N	0.97	Excellent	0.96±0.01				
				LVI	1.00	Excellent	0.96±0.00				
				AVG	**0.98**	**Excellent**	**0.95±0.01**				
	Timepoint 4: 11 Jan, 2025	Snippet reports Original sequence	GPT-4o	P	0.97	Excellent	0.93±0.00	0.78±0.005	0.97	0.83±0.008	0.88±0.005
				N	0.97	Excellent	0.97±0.01				
				LVI	1.00	Excellent	0.96±0.00				
				AVG	**0.98**	**Excellent**	**0.95±0.00**				
GPT-4 Turbo input length and order	23 Nov, 2024	Snippet reports Random sequences	GPT-4 Turbo	P	0.95	Excellent	0.89±0.01	0.76±0.007	0.95	0.80±0.010	0.88±0.008
				N	0.99	Excellent	0.98±0.00				
				LVI	1.00	Excellent	0.96±0.00				
				AVG	**0.98**	**Excellent**	**0.94±0.01**				
	23 Nov, 2024	Snippet reports Original sequence	GPT-4 Turbo	P	0.95	Excellent	0.89±0.01	0.77±0.008	0.95	0.80±0.015	0.88±0.012
				N	0.99	Excellent	0.98±0.00				
				LVI	1.00	Excellent	0.96±0.00				
				AVG	**0.98**	**Excellent**	**0.94±0.01**				
	23 Nov, 2024	Full reports Random sequences	GPT-4 Turbo	P	0.95	Excellent	0.92±0.02	0.77±0.005	0.95	0.83±0.008	0.88±0.012
				N	0.99	Excellent	0.97±0.00				
				LVI	0.99	Excellent	0.96±0.00				
				AVG	**0.98**	**Excellent**	**0.95±0.01**				
	23 Nov, 2024	Full reports Original sequence	GPT-4 Turbo	P	0.96	Excellent	0.92±0.01	0.77±0.004	0.96	0.83±0.008	0.89±0.005
				N	0.99	Excellent	0.97±0.00				
				LVI	0.99	Excellent	0.96±0.00				
				AVG	**0.98**	**Excellent**	**0.95±0.01**				
GPT-4o input length and order	22 Nov, 2024	Snippet reports Random sequences	GPT-4o	P	0.98	Excellent	0.95±0.00	0.77±0.004	0.97	0.82±0.004	0.87±0.000
				N	0.98	Excellent	0.97±0.01				
				LVI	1.00	Excellent	0.96±0.01				
				AVG	**0.99**	**Excellent**	**0.96±0.00**				
	22 Nov, 2024	Snippet reports Original sequence	GPT-4o	P	0.98	Excellent	0.96±0.00	0.77±0.003	0.98	0.82±0.008	0.87±0.002
				N	0.98	Excellent	0.97±0.01				
				LVI	1.00	Excellent	0.96±0.00				
				AVG	**0.99**	**Excellent**	**0.96±0.00**				
	22 Nov, 2024	Full reports Random sequences	GPT-4o	P	0.97	Excellent	0.95±0.01	0.77±0.004	0.97	0.84±0.009	0.89±0.003
				N	0.98	Excellent	0.95±0.01				
				LVI	0.99	Excellent	0.96±0.00				
				AVG	**0.98**	**Excellent**	**0.95±0.01**				
	22 Nov, 2024	Full reports Original sequence	GPT-4o	P	0.97	Excellent	0.95±0.01	0.77±0.004	0.96	0.84±0.012	0.89±0.008
				N	0.97	Excellent	0.95±0.01				
				LVI	0.99	Excellent	0.96±0.00				
				AVG	**0.98**	**Excellent**	**0.96±0.01**				

Heatmaps in Figure 3-20 illustrate 10-run agreement and accuracy. Ideal performance, indicated by yellow cells for 10-time agreement and alignment with the ground truth (red boxes), was most evident for GPT-4o. Among offline models, Llama-2.0-13b and Llama-3.3-70b displayed better consistency and accuracy, while Dolly-v2-12b and Vicuna-v1.3-13b lagged.

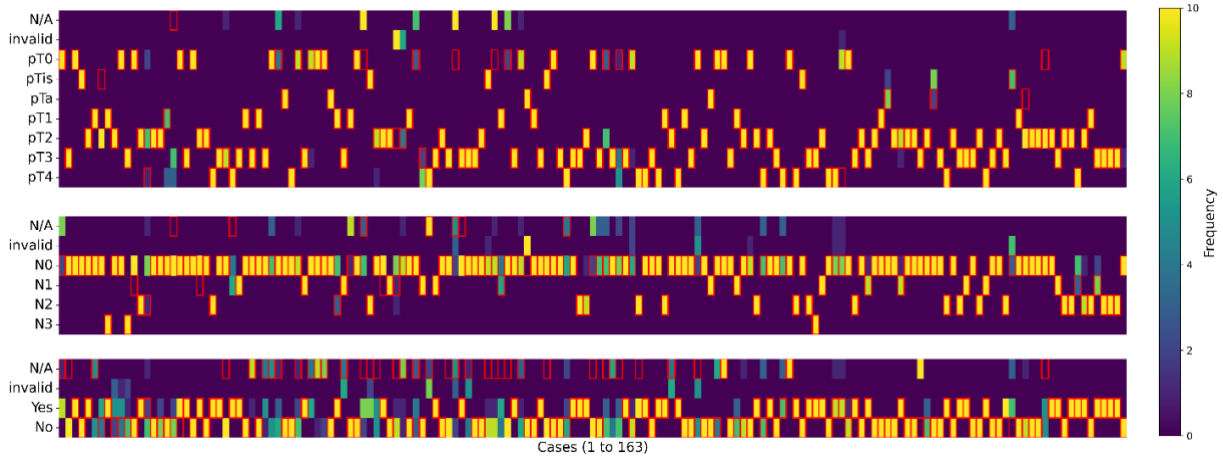
Model Evolution Over Time

GPT-4o was tested at four time points (22 Nov, 02 Dec, 13 Dec 2024, and 11 Jan 2025) to assess the impact of model updates on performance. Across the first three time points separated by 10-day intervals, the model maintained exceptional agreement (AVG $\kappa > 0.98$) and high accuracy (95%-97%). Extending the interval to 30 days resulted in similar agreement, though pT accuracy decreased slightly to 93%.

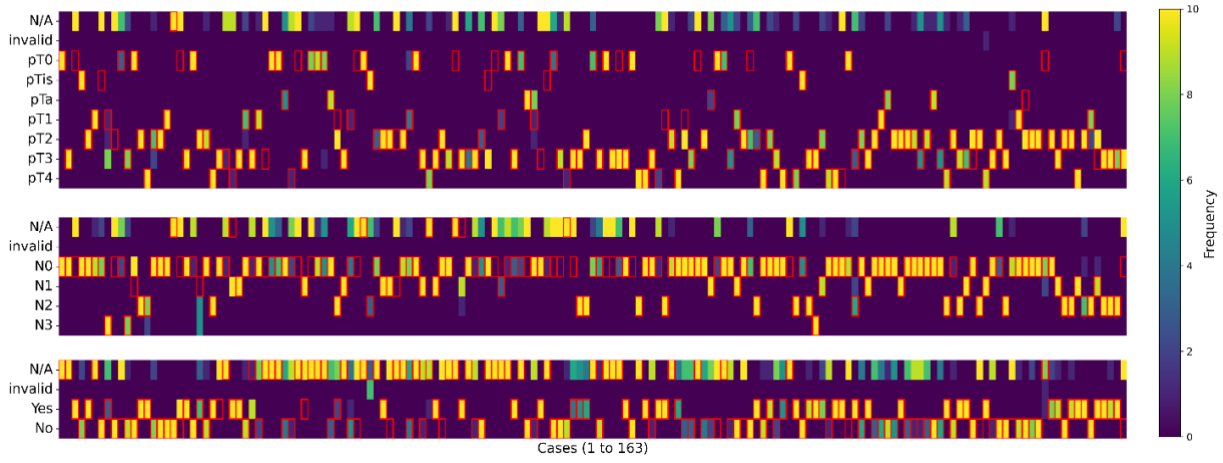
Input Length and Order Effects

The performance of GPT-4 Turbo and GPT-4o was evaluated using both snippet-length and full-length reports to determine the impact of input length on the accuracy of clinical information extraction. GPT-4 Turbo consistently achieved high Kappa values ($\geqslant 0.95$) across both input lengths. While there were slight differences in accuracy, snippet-length reports yielded a lower accuracy of 89% on pT, compared to 92% for full-length reports. These results indicate that GPT-4 Turbo effectively handles both brief and detailed inputs with comparable accuracy, provided the input remains within the token limit. Additionally, randomizing the order of input cases had no significant effect on either the 10-run agreement or the accuracy of the extracted information. This demonstrates that GPT-4 Turbo operates independently in each session, maintaining the integrity of extracted data and avoiding any carryover effects from prior interactions.

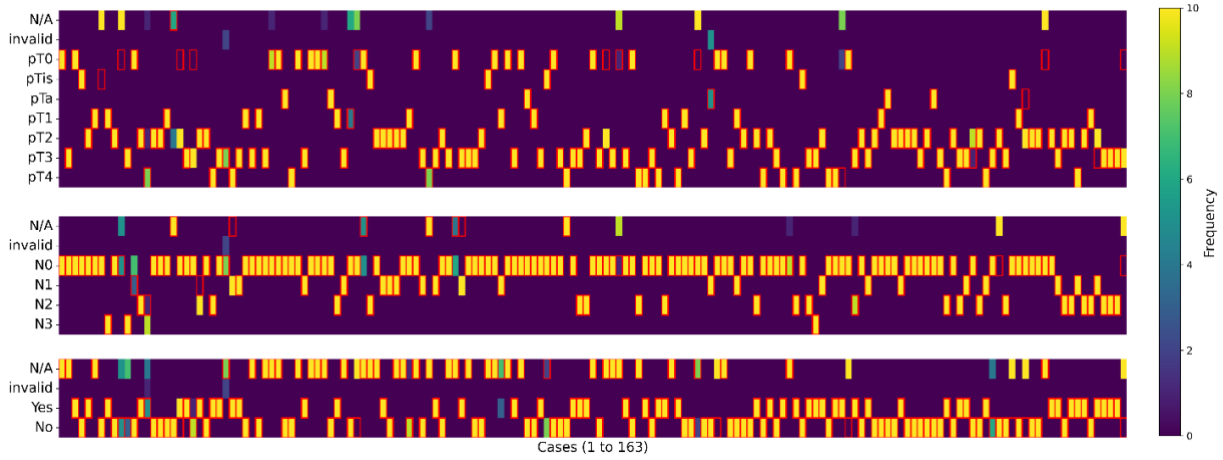
Similarly, GPT-4o exhibited minimal sensitivity to input length or sequence order. Kappa values and extraction accuracy remained consistently high across all experiments, with Kappa values exceeding 0.97 and accuracy consistently above 95%. Compared to GPT-4 Turbo, GPT-4o demonstrated slightly higher agreement and accuracy across all conditions, showing its stronger reliability and robustness in clinical information extraction tasks.



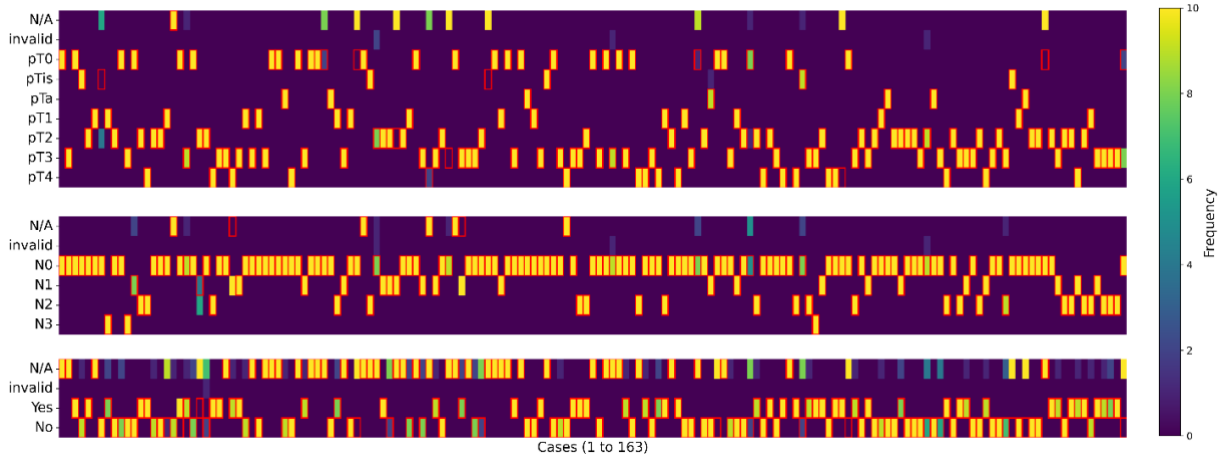
(a) Dolly-v2-12b



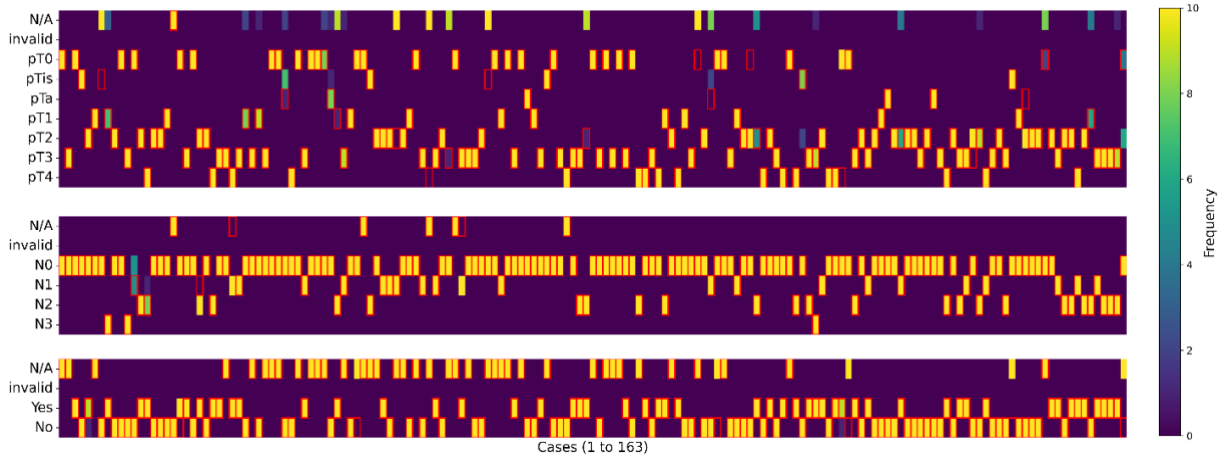
(b) Vicuna-v1.3-13b



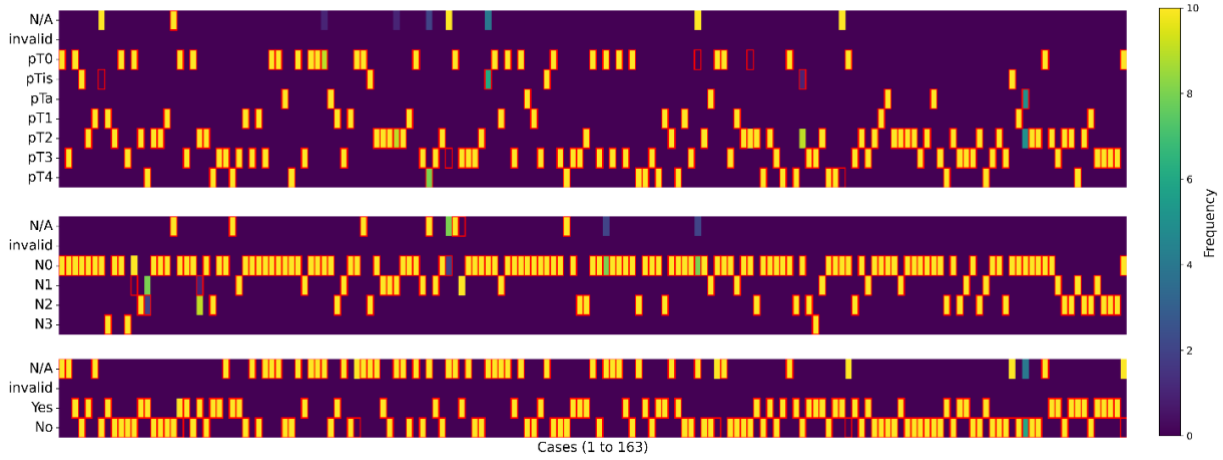
(c) Llama-2.0-13b



(d) Llama-3.3-70b



(e) GPT-4 Turbo. Run on 22 Nov, 2024.



(f) GPT-4o. Run on 22 Nov, 2024.

Figure 3-20: Heatmap representing the frequency of extracted information for the three clinical variables (pT, pN, and LVI) across the 10 runs (N = 163) for the six LLMs: Dolly-v2-12b, Vicuna-v1.3-13b, Llama-2.0-13b, Llama-3.3-70b, GPT-4 Turbo, GPT-4o. The input for all the six models was the snippet reports and in their original sequence. The y-axis displays the categories for the specified clinical variables: pT (N/A, invalid, pT0, pTis, pTa, pT1, pT2, pT3, pT4), pN (N/A, invalid, N0, N1, N2, N3), and LVI (N/A, invalid, Yes, No), in which 'N/A' indicates that relevant information in the input report for LLMs was unavailable or absent, and 'invalid' signifies that the LLM's output could not be interpreted as meaningful or valid for the given variable. The x-axis corresponds to individual cases. Color intensity indicates the frequency of each category in the data. The color bar on the right provides a reference for the frequency scale. Red rectangles highlight the ground truth values for each case, enabling comparison with the LLM extractions.

3.4.4.2 Survival Prediction Reliability

We evaluated survival prediction using two models: a nomogram based solely on clinical information [67] and a multi-modal model (CRD) [3] that incorporates clinical information, radiomics features, and deep-learning-derived features from CT images. Figure 3-21 illustrates the performance of the nomogram (evaluated on 163 and 64 cases) and CRD (evaluated on 64 cases). The ideal model, as indicated by the boxplot, would exhibit the highest AUC value with the narrowest distribution across the 10 runs. The results show that the CRD model achieves a higher AUC than the nomogram (N = 64), consistent with findings from our previous study [3].

Model Effect

In the nomogram analysis, survival predictions based on information extracted by Llama-3.3-70b, GPT-4 Turbo, and GPT-4o closely matched the ground truth prediction (AUC = 0.78), with each model achieving an average prediction of 0.77 among the 10 runs. Llama-2.0-13b and Dolly-v2-12b demonstrated similar average performance (AUC = 0.75); however, predictions from Dolly-v2-12b showed higher variability (±0.015) compared to those from Llama-2.0-13b (±0.004). Among the models, Llama-2.0-13b, Llama-3.3-70b, and GPT-4o exhibited the least variation (±0.004), followed by GPT-4 Turbo (±0.008), while Dolly-v2-12b and Vicuna-v1.3-13b showed the largest variations (±0.015).

For the CRD model, survival predictions from Dolly-v2-12b, Llama-3.3-70b, GPT-4 Turbo, and GPT-4o closely aligned with the ground truth prediction (AUC = 0.89), with an average around 0.87. Llama-2.0-13b had a slightly lower AUC of 0.85, and Vicuna-v1.3-13b yielded the lowest at 0.80. GPT-4o showed the smallest variation (±0.002), followed by Llama-2.0-13b (±0.006). GPT-4 Turbo, Dolly-v2-12b, and Llama-3.3-70b exhibited moderate variations (between ±0.01 and ±0.02), while Vicuna-v1.3-13b had the largest variation (±0.023).

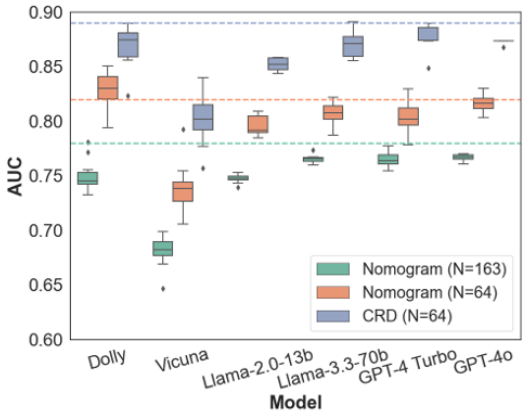
Model Evolution Over Time

With the extracted information by GPT-4o across four time points, survival prediction performance for both the nomogram and CRD models remained closely aligned with those using ground truth clinical information, with minimal variation (<±0.005 for nomogram; <±0.006 for CRD). This stability highlights a critical consideration for clinical applications of LLMs: while updates to LLMs are intended to enhance overall capabilities, they may inadvertently introduce variability that impacts performance in precision-dependent tasks like medical data extraction.

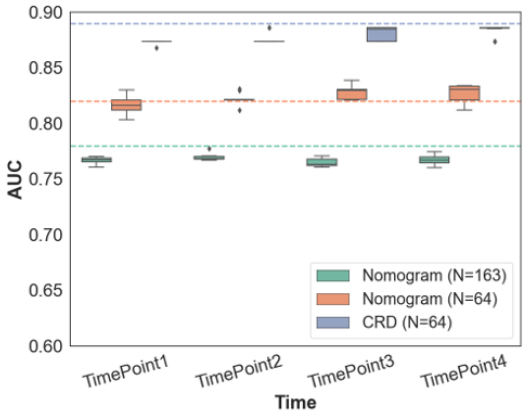
Ongoing monitoring and validation of LLM performance are essential to ensure updates do not negatively affect clinical decision making.

Input Length and Order Effects

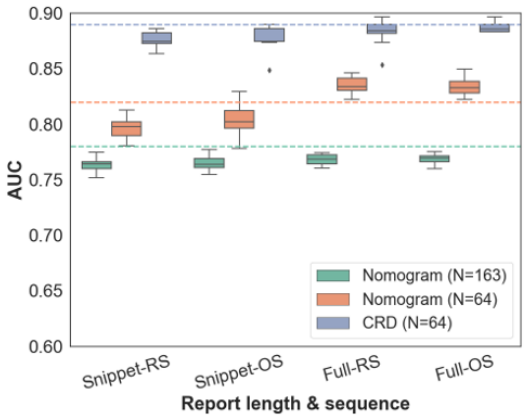
Variations in input length and sequential order had no notable impact on the performance of either the nomogram or CRD models in survival prediction. This robustness underscores the reliability of these LLMs for the task being studied.



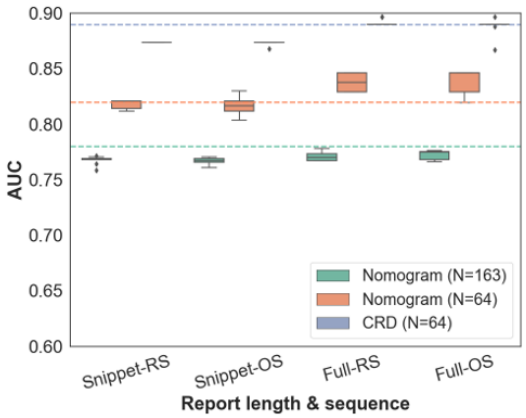
(a) Six different LLMs: Dolly-v2-12b, Vicuna-v1.3-13b, Llama-2.0-13b, Llama-3.3-70b, GPT-4 Turbo, and GPT-4o.



(b) Four time points running GPT-4o. The first three timepoints have an interval of 10 days, and the last interval is 30 days.



(c) GPT-4 Turbo tested by different length of reports (snippet or full reports) and original sequence (OS) or random sequence (RS).



(d) GPT-4o tested by different length of reports (snippet or full reports) and original sequence (OS) or random sequence (RS).

Figure 3-21: Performance comparison (measured by AUC) of survival prediction across 10 runs from different conditions: (1) different model, (2) GPT-4o across 4 time points, (3) GPT-4 Turbo across snippet or full reports and original or random input sequence, (4) GPT-4o across snippet or full reports and original or random input sequence. The black dots marked the outliers that fall outside of the interquartile range (IQR) which is calculated as IQR=Q_3-Q_1, where Q_1 is the first quartile and Q_3 is the third quartile. The three horizontal dashed lines represent the AUCs of models based on the ground truth of clinical variables.

3.4.5 Discussion

The findings of this study have significant implications for the application of LLMs in clinical practice, particularly in oncology, where timely, accurate, and reliable extraction of information from EMRs is critical.

3.4.5.1 Type of Information

Our results demonstrate that the type of information being extracted can impact the accuracy and consistency of LLMs. Specifically, we observed lower agreement among Dolly-v2-12b, Vicuna-v1.3-13b, and Llama-3.3-70b for lymphovascular invasion (LVI), compared to other data, such as pathological tumor (pT) and node (pN) staging. In contrast, GPT-4 Turbo and GPT-4o exhibited similar level of agreement among pT, pN, and LVI.

One possible explanation is that LVI is often documented in a less standardized and more implicit manner in pathology reports. When LVI is negative, it may not be explicitly mentioned at all, whereas a positive finding may be described in varied narrative forms (e.g., "lymphovascular invasion is present" or "evidence of vascular involvement"). This makes it more difficult for LLMs—especially those with fewer parameters or less instruction tuning, such as Dolly-v2-12b and Vicuna-v1.3-13b—to infer the correct label without clear textual cues. In contrast, pT and pN staging typically follow more structured and consistent formats (e.g., "pT1", "pN0", "N1 disease"), making them easier to extract accurately across models.

The inherent ambiguity and sparsity of explicit LVI mentions may require models to understand context more deeply or make subtle inferences, where larger models, such as GPT-4 Turbo and GPT-4o with more parameters and training data, appear to perform better. However, the relatively lower agreement for LVI seen with Llama-3.3-70b ($\kappa = 0.89$) compared to Llama-2.0-13b ($\kappa = 0.96$) warrants further investigation.

3.4.5.2 Implications for Model Selection

In comparing the performance between Llama-2.0-13b and Llama-3.3-70b, it is evident that the assumption that models with more parameters necessarily yield higher accuracy or reliability does not hold true for tasks like clinical information extraction. Larger models also require significantly more computational resources, which may not be feasible in many clinical settings. When selecting models for such tasks, the number of parameters should not be the sole determining factor. Instead, performance consistency and computational efficiency should be prioritized.

Notably, we found that survival prediction accuracy and overall information extraction accuracy were not linearly correlated. Furthermore, the propagation of inaccuracies varied among models depending on the type of information extracted, emphasizing the need for careful evaluation of downstream impacts.

Among the evaluated models, GPT-4o stood out for its high reproducibility and accuracy (Fleiss' Kappa > 0.97 and accuracy > 93%) across various conditions. This consistency underscores its potential to reduce manual data entry time significantly, allowing clinicians to focus more on patient care. By accurately extracting key clinical variables such as pathologic stage, lymph node status, and LVI, LLMs like GPT-4o facilitate the deployment of survival prediction models, improve clinical decision-making, and streamline workflows. In large healthcare systems, where vast amounts of unstructured data must be processed quickly and accurately, automation provided by LLMs can be transformative.

3.4.5.3 Considerations for Offline Models

For healthcare environments with strict data privacy, security, or resource constraints, offline models are often necessary. Among the offline models, Llama-2.0-13b and Llama-3.3-70b demonstrated excellent agreement across repeat runs and consistently high accuracy in

survival prediction, closely aligning with ground truth. However, for scenarios with limited computational resources, Llama-2.0-13b serves as a preferred choice with comparable performance. Dolly-v2-12b could be considered a secondary option, while Vicuna-v1.3-13b may require further investigation or technical fine-tuning to improve its suitability for clinical tasks like this study. The relatively lower performance of Dolly-v2-12b and Vicuna-v1.3-13b may stem from differences in training data quality, model architecture, or their ability to generalize across complex medical domains like bladder cancer prognosis. Fine-tuning these models with domain-specific datasets or employing transfer learning techniques could potentially improve their accuracy and consistency.

3.4.5.4 Multi-Modal Predictive Models

One critical implication of this study is the potential for integrating LLMs into multi-modal predictive models. Bladder cancer prognosis, similar to many other clinical assessment tasks, benefits from combining textual data (e.g., pathology reports) with imaging data (e.g., CT scans). The Clinical-Radiomics-Deep-Learning (CRD) model evaluated in this study integrates clinical data extracted by LLMs with radiomic and deep-learning imaging features to predict patient outcomes with high accuracy. The ability of models like GPT-4 and Llama to extract clinical data reliably enhances the utility of such multi-modal models by ensuring accuracy and consistency in the textual component. The complementarity of data from multiple modalities leads to more robust and generalizable predictive models, likely to improve patient care. Moreover, LLM-driven automation makes it feasible to scale these predictive models to larger patient populations, thereby amplifying their clinical impacts.

3.4.5.5 Future Directions

While this study provides valuable insights into the variability of LLMs in clinical data extraction, several areas warrant further exploration. First, additional research is needed to evaluate the performance of these models across a broader spectrum of medical records and specialties, covering patients with a variety of neoplastic and inflammatory processes. Although bladder cancer pathology reports were a suitable test case, LLMs are likely to encounter differing and potentially application-specific complexity when applied to other domains such as cardiology, neurology, or infectious disease. Second, efforts should focus on improving lower-performing models like Dolly-v2-12b and Vicuna-v1.3-13b through domain-specific fine-tuning or more sophisticated training datasets. Finally, healthcare institutions adopting LLMs must establish robust validation frameworks to ensure that model updates—whether online or offline—do not inadvertently compromise performance in critical applications. Regular quality assurance monitoring and calibration are essential to maintaining the reliability of these tools as they evolve.

3.4.6 Conclusion

This study highlights the degree of variability in LLM performance when applied to clinical information extraction. The ability of LLMs to automate the extraction of critical clinical variables from unstructured medical records has significant implications for improving predictive modeling and clinical decision-making in oncologic patients, as well as, potentially, in patients with other diseases. As can be seen in our study, LLMs can be used for expediting relevant data extraction from a patient's medical records, which is then utilized for accurate prediction of a patient's prognosis. However, the variability observed in certain models underscores the need for careful selection and continuous evaluation of LLMs in clinical applications.

A summary of the important findings and recommendations for the practice of utilizing LLMs in clinical information retrieval is presented in Table 3-8.

Future research should focus on strategies for improving the reliability of lower-performing models and validating the robustness of LLMs across the wide range of medical disciplines. As LLM technology advances, it holds immense potential for transforming healthcare by enabling more accurate, efficient, and scalable clinical data extraction and analysis.

Table 3-8: Summary of recommendations for the practice of applying LLMs in clinical information retrieval.

Topic	**Key findings and recommendations**
Model selection	1. Among Dolly-v2-12b, Vicuna-v1.3-13b, Llama-2.0-13b, Llama-3.3-70b, GPT-4 Turbo, and GPT-4o, GPT-4 provided the highest accuracy in information retrieval and best performance in the downstream survival prediction model. 2. If local model deployment is required for data protection, Llama models have comparable performance as GPT-4. One can choose Llama-3.3-70b when there are sufficient computing resources. In the absence of sufficient computing resources, Llama-2.0-13b is an excellent alternative. 3. Update of an online model (GPT-4) did not have an apparent impact on the performance of information retrieval.
EMRs consideration	Report length of EMR and case order does not have an apparent impact on the performance of information retrieval when using GPT-4.
Prompt consideration	1. Add context constraints to prevent information leakage. 2. Add instructions to write the output in a formatted way to reduce the workload of post-processing on unstructured data. 3. Provide explicit guidance for handling missing information (or other edge situations) in retrieval tasks to reduce ambiguity of LLM response.
Downstream tasks	Review the step-by-step propagation of errors and inaccuracies caused by LLMs in downstream tasks.

Chapter 4 Computerized Decision Support for Bladder Cancer Treatment Response Assessment in CT Urography: Effect on Diagnostic Accuracy in Multi-institution Multi-specialty Observer Study

4.1 Abstract

This observer study investigates the effect of computerized artificial intelligence (AI)-based decision support system (CDSS-T) on physicians' diagnostic accuracy in assessing bladder cancer treatment response. The performance of 17 observers was evaluated when assessing bladder cancer treatment response without and with CDSS-T using pre- and post-chemotherapy CTU scans in 123 patients having 157 pre- and post-treatment cancer pairs. The impact of cancer case difficulty, observers' clinical experience, institution affiliation, and specialty on the observers' diagnostic performance with and without using CDSS-T were analyzed. In addition, inter- and intra-observer variability were analyzed. It was found that the average performance of the 17 observers was significantly improved ($p = 0.002$) when aided by the CDSS-T. The cancer case difficulty, institution affiliation, and specialty influenced the observers' performance without CDSS-T. The AI-based decision support system has the potential to improve the diagnostic accuracy in assessing bladder cancer treatment response and result in more consistent performance among all physicians.

4.2 Introduction

We have developed a computerized artificial intelligence (AI)-based decision support system for muscle-invasive bladder cancer treatment response assessment (CDSS-T) to assist physicians to evaluate the response to treatment of these cancers on pre- and post-treatment CT

urography (CTU) scans [48] based on the combined radiomics and deep-learning model detailed in Chapter 2.1. It is critical to gain understanding of various factors that may affect the impact of CDSS-T on physician performance in identifying bladder cancers with complete response after neoadjuvant chemotherapy through observer studies that can guide the design of future clinical trials. Patients with complete response may be considered for organ preservation therapy instead of cystectomy (the removal of the bladder). Our previous studies showed that CDSS-T can improve physician performance from a single institution [100-102]. The goal of the current study is to further investigate the impact of CDSS-T on the performance of physicians from different specialties and different institutions.

4.3 Methods and Materials

4.3.1 Data Set

With Institutional Review Board (IRB) approval, we collected the pre- and post-chemotherapy CTU scans of 123 patients (157 pre- and post-treatment cancer pairs). The pathological cancer stage after treatment was collected as reference standard for determining if a patient had complete response to treatment. The reference standard indicated that 40 out of 157 lesion pairs had complete response (stage T0) after chemotherapy, and 117 lesion pairs had incomplete response (>T0) after chemotherapy. The patient and cancer characteristics are presented in Table 4-1.

Table 4-1: Characteristics of the patients and cancers used in the observer study.

Characteristics	Notes	Detail		Total number
Patient gender and age	100 males	Mean: 63 years Range: 43–84 years		123 patients
	23 females	Mean: 63 years Range: 37–82 years		
Average maximum diameter (mm)	Completely responding cancers (T0 stage)	Pre-treatment: 30.1 Post-treatment: 14.3		157 cancer pairs
	Incompletely responding cancers (>T0 stage)	Pre-treatment: 43.0 Post-treatment: 31.2		
Cancer stage		Pre-treatment	Post-treatment	157 cancer pairs
	T0	0	40	
	T1	8	37	
	T2	76	23	
	T3	63	38	
	T4	10	19	

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

CDSS-T combined a deep-learning convolutional neural networks model (DL-CNN; detailed in Chapter 2.1.3.5) and a radiomics model to estimate the likelihood of response to neoadjuvant chemotherapy treatment [48]. Each model output a discriminant score for each scan pair. A combined score was generated by taking the maximum of the two scores [48].

We utilized the In-House Radiomics features (detailed in Chapter 2.1.3.4) to build up the radiomics model, employing a two-loop leave-one-case-out cross-validation scheme [103] . For the inner loop, the subset of features was selected, and the random forest classifier [48] was trained by using a leave-one-case-out scheme within the training partition. For the outer loop, the trained classifier was applied to the left-out test case and a discriminant score was generated by

the radiomics model. An average of four features was selected, including two run-length statistics features and two contrast features.

4.3.3 CAD Score

For each pre-post-treatment scan pair, a combined score was obtained by taking the larger value of the two discriminant scores generated by the DL-CNN model and the radiomics model. Receiver operating characteristics (ROC) analysis was performed on the combined scores to estimate the CDSS-T performance. Linearly scaling the combined scores generated computer-aided diagnosis (CAD) scores in the range from 1 to 10. Smaller scores indicated lower probability that the post-treatment lesion had a complete response, and larger scores indicated higher probability that the post-treatment lesion had a complete response A curve was fitted to the distribution of the linearly transformed scores for the cancers in each class, i.e., the stage T0 and the stage > T0 cancers. The area under each fitted curve was normalized to 1. The CAD score for the cancer being read and the two fitted curves were displayed to the observers during the CDSS-T assisted reading. The fitted distributions provided a reference of the likelihood of complete response for the CAD score (Figure 4-1d). The CAD scores and the fitted curves were predetermined and were the same for all observers.

4.3.4 Observer Performance Study

Seventeen physicians from six specialties and four institutions, University of Michigan (UM), Pennsylvania State University (PSU), University of Iowa (UI), and Tokuda Hospital (TH), Sofia, Bulgaria, provided assessment of response by reading the pair of pre- and post-treatment CTU for each cancer. The specialty and institution information of observers is presented in Table 4-2. A graphical user interface developed for CDSS-T was used to assist physicians to make the estimations. Figure 4-1 shows the overall design for this observer study: one pair of CTU was

showed side-by-side to each observer. **First, without CDSS-T,** the physician provided the estimation of treatment response (Figure 4-2), consisting of: (1) percentage response to treatment on a scale of −100% to +100% using the Response Evaluation Criteria in Solid Tumors (RECIST) criteria [104], where 0% indicated no change between the pre- and post-treatment scan, −100% indicated at least doubling of tumor size, and 100% indicated a complete response; (2) tumor response category (based on the RECIST criteria) including progressive disease, stable disease, partial response, and complete response [104]; (3) likelihood of T0 stage (complete response) on a scale of 0 to 100%, where 0 indicated no chance of this post-treatment cancer being at T0 stage and 100% indicated 100% chance of this post-treatment cancer being at T0 stage; (4) a recommendation for the next treatment procedure including surgery or radiation. **Second, with CDSS-T** (Figure 4-3), a CAD score along with the CAD score distributions was shown to the physician. The physician might revise his/her estimates or leave them unchanged. For the analysis of intra-observer variability, a subset (mean N = 51) of the 157 lesion pairs was assessed for the second time by the 17 observers. Note that this subset or 51 lesions was randomly sampled so that it varied for each observer. The first read of the subset was called the original evaluation, and the second read of the subset was the repeated evaluation in the following discussion.

Table 4-2: Observers of different specialties and from different institutions.

Specialty	Observer number	Proficiency	Institution			
			UM	UI	PSU	TH
Abdominal radiologist	5	Experienced	4	-	-	1
Diagnostic radiology resident	4	Inexperienced	4	-	-	-
Urologist	1	Experienced	1	-	-	-
Oncologist	5	Experienced	2	2	1	-
Medical student	1	Inexperienced	-	-	1	-
Neurology fellow	1	Inexperienced	1	-	-	-

UM: University of Michigan, UI: University of Iowa, PSU: Pennsylvania State University, TH: Tokuda Hospital, Sofia, Bulgaria.

4.3.5 Statistical Analysis

The physicians' estimates of the likelihood of T0 stage (complete response) were analyzed with multi-reader, multi-case (MRMC) ROC methodology [105]. The MRMC analysis tool can provide values of area under curve (AUC) and numerical information which can be used to plot ROC curves. A *p*-value of less than 0.05 was considered to indicate a significant difference.

A number of comparisons were performed, including between observers with different degrees of clinical experience, observers from different institutions, observers from different specialties, and between two subsets of cases with different levels of diagnostic difficulty. The intra-observer variability was estimated from the repeated readings of 51 cases. As for the degree of experience, the radiologists, oncologists, and urologists were regarded as experienced physicians in assessing bladder cancer treatment response, and the radiology residents and the neurology fellow were regarded as inexperienced physicians. Note that the medical student was also categorized as an "inexperienced physician" for simplicity of discussion. For each cancer,

the standard deviation of the estimates of likelihood of T0 stage by nine radiologists was used to evaluate the level of difficulty in assessing the response of the cancer. The assessment was categorized as relatively easy when the standard deviation was less than 25, and the corresponding scan pair was grouped into the "easy cancer" subset, or relatively difficult when the standard deviation was larger than 25, and the corresponding scan pair was grouped into the "difficult cancer" subset. The threshold of 25 was selected by approximately balancing the number of complete responses (T0 after chemotherapy) in the easy and difficult cancer subsets for the purpose of performing a reliable ROC analysis [6]. There were 95 pre-post-treatment scan pairs in the easy subset, in which 77 cancers did not respond completely (>T0) and 18 had a complete response (T0). The remaining 62 pairs were grouped in the difficult subset, with 40 cancers not responding completely and 22 with complete response.

The physicians' estimates of the likelihood of T0 stage of the original and repeated evaluations were not fully-crossed due to the varied subset for different observers. We used the iMRMC package [106] for the analysis of the intra-observer variability which can handle the data and output the AUCs and the statistical significance of the difference in the AUCs [106]. The Bland–Altman method and the Krippendorff's alpha method were also employed to evaluate the intra-observer variability.

The Krippendorff's alpha method assesses the agreement between two response outcomes from multiple cases and multiple observers, i.e., inter-observer variability, and allows for missing responses for the cases. We applied the Krippendorff's alpha method to the estimates of the 17 observers on the 157 cases. Krippendorff suggests 0.8 as a customary threshold for satisfactory reliability, but if tentative conclusions are acceptable, 0.667 is the lowest conceivable limit [99].

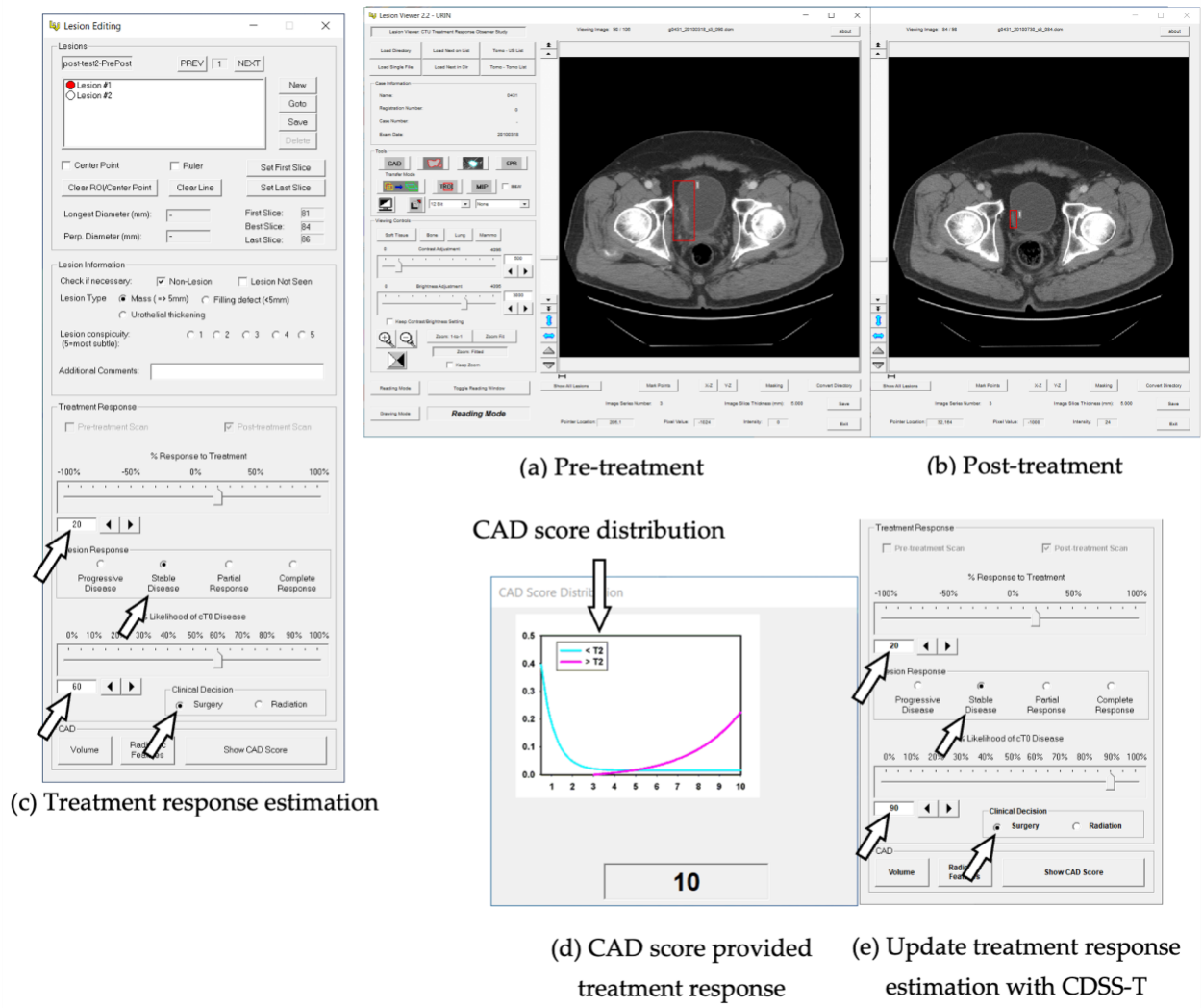
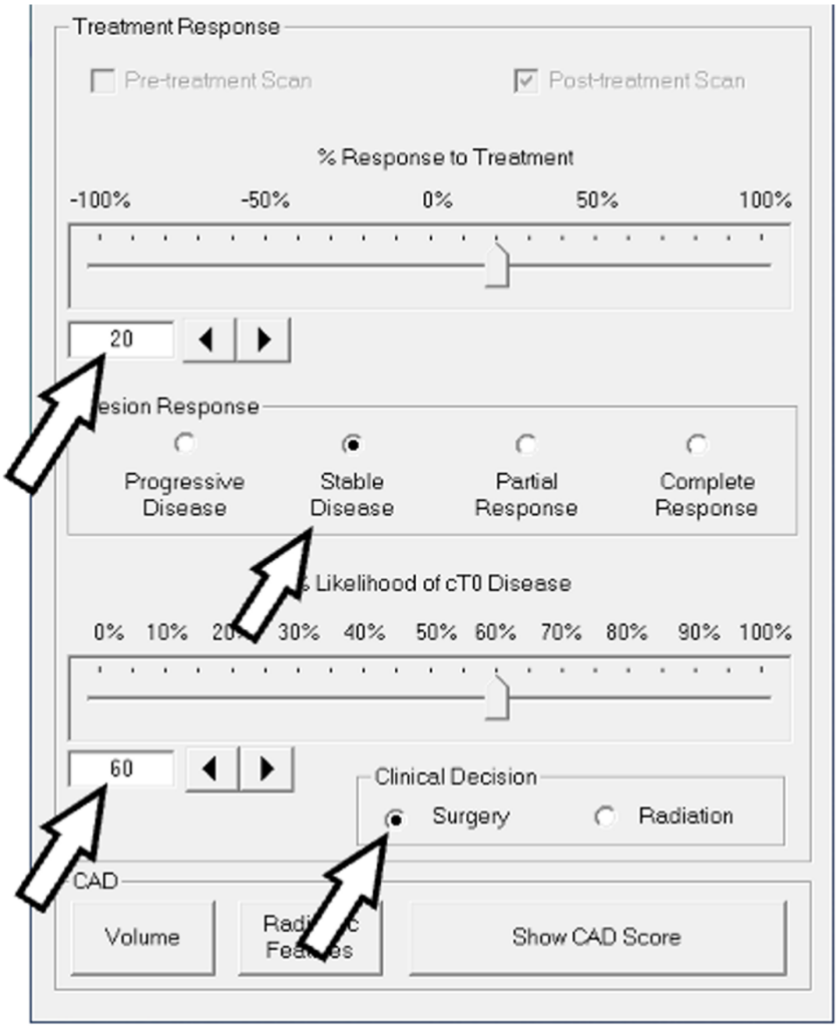
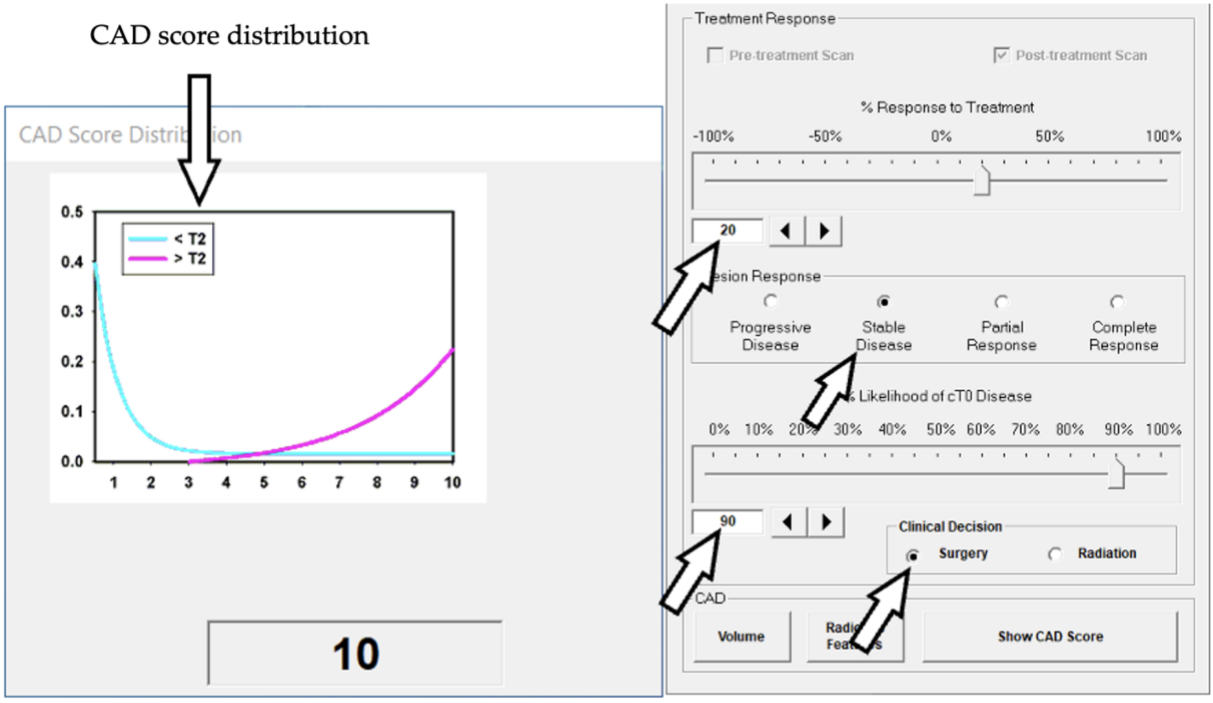


Figure 4-1: The graphical user interface for reading with and without our computerized decision support system (CDSS-T) for bladder cancer treatment response assessment. (a), (b) The pre- and post-treatment CTU scans are shown side-by-side. (c) The observer estimates the treatment response. (d) The observer is shown the CAD score and the score distribution of the two classes as reference. (e) The observer may revise their treatment response assessment after considering the CAD score.



(c) Treatment response estimation

Figure 4-2: One step of observer study. Each observer performed treatment response assessment without CDSS-T.



(d) CAD score provided treatment response

(e) Update treatment response estimation with CDSS-T

Figure 4-3: CAD score indicating the treatment response likelihood of each case was provided to each observer. The observer can choose to update the treatment response assessment.

4.4 Results

4.4.1 Overall Results for All Cancers

The AUC of the CDSS-T CAD scores was 0.80. The average ROC curves with and without CDSS-T aid for all observers together with the individual AUC values with and without CDSS-T aid for the 17 observers are presented in Figure 4-4. The individual *p* values are presented in Table 4-3. All observers except #6 had larger AUCs with the CDSS-T aid (Figure 4-4), but the improvement reached statistical significance only for 7 of the 17 observers (*p* values in Table 4-3). Observer #9 with CDSS-T aid had a larger AUC (0.81) than CDSS-T alone (0.80), but the difference did not achieve significance ($p = 0.776$). The average AUC over all observers

was 0.73 without CDSS-T and improved to 0.77 with CDSS-T. The difference was statistically significant ($p = 0.002$).

Table 4-3: AUCs of observers with and without CDSS-T. The standard deviation of the AUCs of 17 observers was calculated for both the without CDSS-T and with CDSS-T conditions.

Observer #	**AUC w/o CDSS-T**	**AUC w/ CDSS-T**	**Individual *p* value**
1	0.74	0.77	0.155
2	0.75	0.77	0.260
3	0.73	0.76	0.013 *
4	0.74	0.77	0.128
5	0.76	0.80	0.010 *
6	0.74	0.74	0.861
7	0.76	0.77	0.541
8	0.73	0.75	0.135
9	0.78	0.81	0.191
10	0.74	0.76	0.244
11	0.67	0.73	0.014 *
12	0.75	0.79	0.095
13	0.72	0.77	0.027 *
14	0.65	0.76	0.020 *
15	0.67	0.78	0.003 *
16	0.76	0.79	0.026 *
17	0.73	0.76	0.083
Mean AUC	0.73	0.77	0.002 *,$
Standard deviation	0.04	0.02	-

* Statistically significant difference at $p < 0.05$ level. $ Obtained from MRMC analysis.

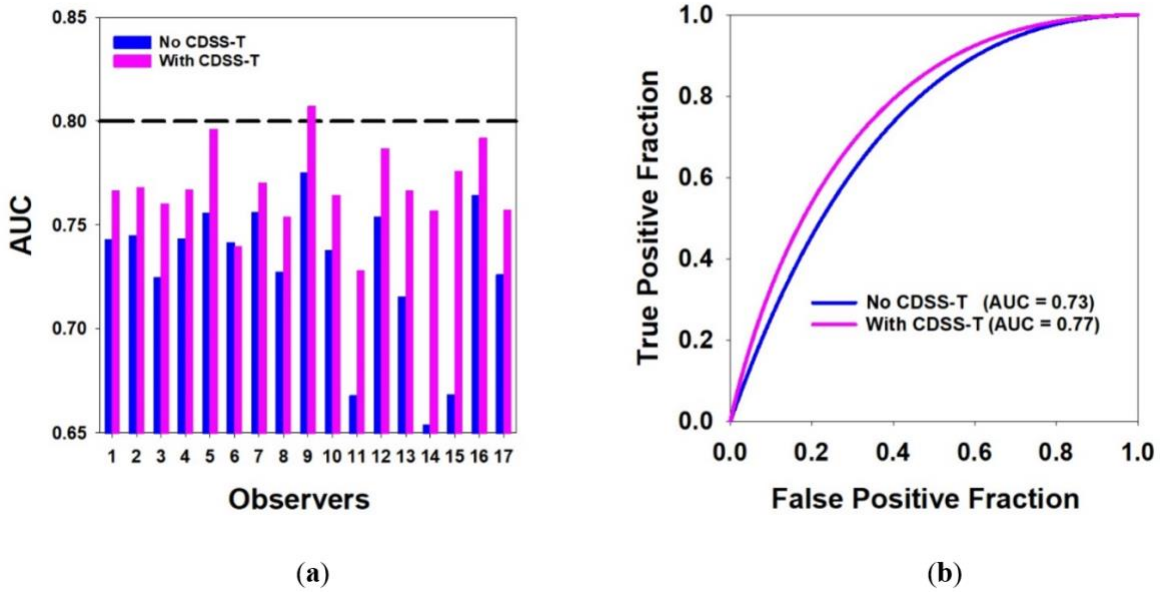


Figure 4-4: Results of the observer performance study. (a) AUCs of the 17 observers with and without CDSS-T. The performance of the CDSS-T system is shown with the dashed line. The performance of all but one observer (#6) increased with using CDSS-T. (b) Average ROC curves with and without CDSS-T.

4.4.2 Easy vs. Difficult Cancer Subsets

The AUC of the CDSS-T was 0.88 for the easy subset, and 0.67 for the difficult subset. The performance comparisons of observers for the easy subset and the difficult subset are shown in Table 4-4. All 17 physicians, and the 9 radiologists, achieved statistically significant improvement for the easy subset ($p < 0.035$), but had no significant improvement for the difficult subset ($p > 0.148$). The five oncologists had significant improvement for the difficult subset ($p = 0.009$), but a borderline significance for the easy subset ($p = 0.051$).

Table 4-4: Performance comparisons with and without CDSS-T for easy and difficult subsets.

	AUC of CDSS-T	Average AUC w/o CDSS-T	Average AUC w/ CDSS-T	*p* value	# of physicians
Easy subset	0.88	0.80	0.84	0.016 *	17 physicians
Difficult subset	0.67	0.58	0.62	0.148	
Easy subset	0.88	0.83	0.85	0.033 *	9 radiologists
Difficult subset	0.67	0.59	0.61	0.379	
Easy subset	0.88	0.78	0.84	0.051	5 oncologists
Difficult subset	0.67	0.57	0.63	0.009 *	

* Statistically significant difference at $p < 0.05$ level.

4.4.3 Experienced vs. Inexperienced Observers

The performance comparisons of experienced and inexperienced physicians are shown in Table 4-5. We can see there was no observable difference between their performances. The level of statistical significance of the inexperienced radiologists was slightly higher ($p = 0.007$) after using CDSS-T compared to that of experienced radiologists ($p = 0.06$). The use of CDSS-T resulted in more consistent performance among all subgroups of physicians (all AUC = 0.77).

Table 4-5: Performance comparison between experienced and inexperienced observers for the total of 157 lesion pairs.

	AUC of CDSS-T	Average AUC w/o CDSS-T	Average AUC w/ CDSS-T	*p* value	# of physicians
Experienced physicians	0.80	0.73	0.77	0.007 *	5 abdominal radiologists, 1 urologist, and 5 oncologists
Inexperienced physicians		0.73	0.77	0.019 *	5 residents and 1 medical student
Experienced radiologists		0.75	0.77	0.060	5 abdominal radiologists
Inexperienced radiologists		0.74	0.77	0.007 *	4 radiology residents
UM experienced radiologists		0.75	0.77	0.018 *	4 abdominal radiologists from UM
UM inexperienced radiologists		0.74	0.77	0.007 *	4 radiology residents from UM

* Statistically significant difference at $p < 0.05$ level.

4.4.4 Multi-Specialty Observers

The average AUC values and the corresponding ROC curves for the physicians from the different specialties are shown in Table 4-6 and Figure 4-5, respectively. Radiologists, oncologists, and the medical student had statistically significant improvement after using CDSS-T ($p < 0.020$). The improvement gained by the urologist and the neurology fellow did not reach statistical significance ($p = 0.244$ and $p = 0.083$, respectively). Note that the number of observers in each group was different so that the level of significance should not be directly compared. There was a difference between the average AUCs of observers from different specialties without CDSS-T. However, with CDSS-T the performance was improved to a similar level.

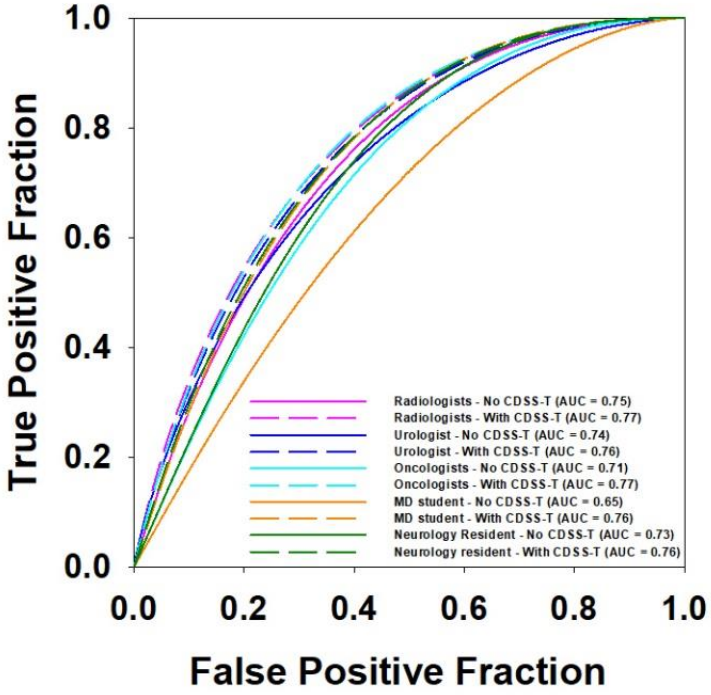


Figure 4-5: ROC curves comparison for observers of different specialties.

Table 4-6: Performance comparison between observers from different specialties for the total of 157 lesion pairs.

Specialty	AUC of CDSS-T	Average AUC w/o CDSS-T	Average AUC w/ CDSS-T	*p* value	# of physicians
Radiologists	0.80	0.75	0.77	0.014 *	9
Urologist		0.74	0.76	0.244	1
Oncologists		0.71	0.77	0.011 *	5
Medical student		0.65	0.76	0.020 *	1
Neurology fellow		0.73	0.76	0.083	1

* Statistically significant difference at $p < 0.05$ level.

4.4.5 Multi-Institution Observers

The average AUC values and the corresponding ROC curves of observers from different institutions are shown in Table 4-7 and Figure 4-6, respectively. The UM physicians (all specialties) and the PSU oncologist achieved statistically significant improvement with CDSS-T. Again, note the varied number of observers in the different groups so that the level of significance should not be compared directly.

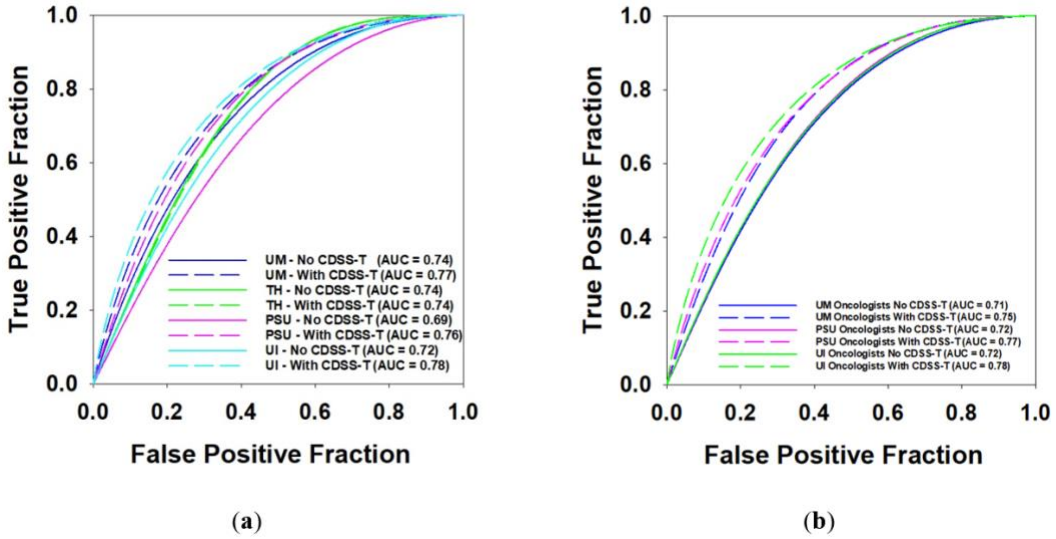


Figure 4-6: ROC curves comparison for observers from different institutions. (a) Performance of observers from four institutions. (b) Performance of oncologists from three institutions.

Table 4-7: Performance comparison between observers from different institutions for the total of 157 lesion pairs.

	AUC of CDSS-T	Average AUC w/o CDSS-T	Average AUC w/ CDSS-T	*p* value	# of physicians
UM physicians	0.8	0.74	0.77	0.002 *	12
TH physician		0.74	0.74	0.861	1
PSU physicians		0.69	0.76	0.117	2
UI physicians		0.72	0.78	0.326	2
UM oncologists		0.71	0.76	0.071	2
PSU oncologist		0.72	0.77	0.027 *	1
UI oncologists		0.72	0.78	0.326	2

* Statistically significant difference at $p < 0.05$ level.

4.4.6 Inter- and Intra-Observer Variability

The individual AUC values and the standard deviations of 17 observers for the original and the repeated evaluation are presented in Table 4-8. For the original evaluation, the average AUC increased significantly from 0.75 (without CDSS-T aid) to 0.81 (with CDSS-T aid) (p = 0.003). For the repeated evaluation, the average AUC also increased significantly from 0.77 (without CDSS-T aid) to 0.81 (with CDSS-T aid), (p = 0.006). There was no significant difference between the average AUC values for the original and repeated evaluations without CDSS-T (p = 0.217) or for the evaluations with CDSS-T (p = 0.692). The standard deviation of the observers' results was slightly smaller with CDSS-T.

4.4.6.1 Bland–Altman Analysis

The standard deviation of the original and repeated evaluations of each observer was calculated by the Bland–Altman method, without and with CDSS-T aid, respectively. The mean value of the standard deviations for the 17 observers was 25.47 without CDSS-T and was reduced to 19.72 with CDSS-T.

4.4.6.2 Krippendorff's Alpha Method

The agreement between the original and repeated evaluations of each of the observers without and with CDSS-T aid was calculated by the Krippendorff's alpha method. The average of the 17 Krippendorff's alpha reliability coefficients was 0.69 without CDSS-T and increased to 0.81 with CDSS-T.

The Krippendorff's alpha reliability coefficient of the 17 observers for the whole data set (N = 157) was 0.56 without CDSS-T and increased to 0.67 with CDSS-T.

Table 4-8: Diagnostic performance in terms of AUC of physicians without and with the CDSS-T aid for the assessment of complete response to neoadjuvant chemotherapy on the first 51 cases in each observer's individually randomized reading list. The iMRMC package provided standard deviation value along with each AUC. The mean and standard deviations of AUC values for the 17 observers with and without CDSS-T for both the original and the repeated evaluations were calculated.

Observer #	**AUC original evaluation**		**AUC repeated evaluation**	
	W/o CDSS-T	**W/ CDSS-T**	**W/o CDSS-T**	**W/ CDSS-T**
1	0.75 ± 0.08	0.76 ± 0.08	0.8 ± 0.07	0.79 ± 0.07
2	0.88 ± 0.05	0.91 ± 0.04	0.88 ± 0.05	0.92 ± 0.03
3	0.65 ± 0.10	0.72 ± 0.10	0.67 ± 0.10	0.72 ± 0.09
4	0.71 ± 0.09	0.71 ± 0.09	0.69 ± 0.09	0.71 ± 0.08
5	0.70 ± 0.07	0.78 ± 0.06	0.82 ± 0.06	0.83 ± 0.05
6	0.82 ± 0.07	0.85 ± 0.07	0.81 ± 0.07	0.81 ± 0.07
7	0.75 ± 0.08	0.77 ± 0.08	0.84 ± 0.05	0.87 ± 0.05
8	0.74 ± 0.09	0.77 ± 0.08	0.81 ± 0.08	0.8 ± 0.08
9	0.81 ± 0.06	0.85 ± 0.05	0.8 ± 0.06	0.85 ± 0.05
10	0.79 ± 0.08	0.84 ± 0.07	0.8 ± 0.07	0.87 ± 0.07
11	0.65 ± 0.08	0.75 ± 0.08	0.73 ± 0.07	0.78 ± 0.07
12	0.81 ± 0.07	0.85 ± 0.07	0.75 ± 0.08	0.76 ± 0.07
13	0.81 ± 0.06	0.89 ± 0.04	0.77 ± 0.07	0.83 ± 0.06
14	0.59 ± 0.10	0.82 ± 0.07	0.69 ± 0.10	0.81 ± 0.07
15	0.73 ± 0.07	0.88 ± 0.06	0.64 ± 0.10	0.83 ± 0.07
16	0.86 ± 0.05	0.93 ± 0.03	0.87 ± 0.05	0.87 ± 0.05
17	0.63 ± 0.08	0.69 ± 0.08	0.68 ± 0.10	0.76 ± 0.09
Mean AUC	0.75	0.81	0.77	0.81
Standard deviation	0.08	0.07	0.07	0.06
Statistical significance in the difference of AUC:				
AUC (orig.without) versus AUC (orig.with): $p = 0.003$ *				
AUC (repeat.without) versus AUC (repeat.with): $p = 0.006$ *				
AUC (orig.without) versus AUC (repeat.without): $p = 0.217$				
AUC (orig.with) versus AUC (repeat.with): $p = 0.692$				

* Statistically significant difference at $p < 0.05$ level.

4.5 Discussion

For the overall performance evaluation (17 observers and 157 cancers), the results showed that the CDSS-T aid can improve the observers' performance significantly in assessing bladder treatment response. The diagnostic difficulty of cancer cases can have an impact on the performance as seen in Table 4-4. However, the increases in AUC (w/o CDSS-T vs. w/ CDSS-T) between the easy and difficult subsets were essentially equal for all the groups in Table 4-4. While the other groups had significant improvement for easy subset but not for difficult subset, the five oncologists only had significant improvement for difficult subset, from AUC of 0.57 without CDSS-T to 0.63 with CDSS-T. This happened because only oncologists reached AUC of 0.63 with CDSS-T which was the best, while the other two groups were 0.61~0.62, and it increased from AUC of 0.57 which was the lowest AUC score without CDSS-T among the four groups. The level of physician experience only contributed to a small difference in the level of improvement (a difference of 0.01) in assessing cancer treatment response. Without CDSS-T, it may be expected that the performance of experienced observers would be better than that of inexperienced observers, as can be observed from the slightly higher AUCs of 0.74 and 0.75 in Table 4-5. With CDSS-T, the experienced and inexperienced observers reached the same average AUC level (0.77). The smaller gain in AUC by the experienced observers may reflect that they were more confident and persistent in their own assessments. More importantly, the CDSS-T aid was able to assist the inexperienced observers in making diagnosis at a level comparable to that of experienced observers.

The performance of observers from the different specialties without CDSS-T did differ. In this study, on average, oncologists had a larger gain in AUC than radiologists with the CDSS-T aid, suggesting that CDSS-T could be a potentially useful tool for non-radiology physicians. There was some difference in the improvement of observers from different institutions.

For both the original and repeated evaluations, the CDSS-T aid can improve the observers' performance significantly. The CDSS-T system can also narrow the performance gap between observers, as demonstrated by the Krippendorff's alpha reliability coefficients of the 17 observers without and with CDSS-T. With CDSS-T aid, the observers had less variability and better agreement for both the original and repeated evaluations.

There were limitations in our study. In some of the experiments, the lack of statistical significance may be due to the small number of observers. For example, we had only one urologist and one neurology resident, and their individual performance was not significantly improved with the CDSS-T aid although they achieved substantial improvement. Additionally, the number of cases was relatively small. Leave-one-case-out cross-validation method was used to generate the CAD score. The AUC of the CDSS-T system was 0.80 which was higher than the majority of the AUCs of the observers. Ideally, the system would be evaluated on an independent test set [107]. Nonetheless, the study did demonstrate that, if a CDSS-T aid could achieve this level of performance in the clinical cases, it would have the potential of improving the treatment response assessment of bladder cancer for a wide range of physicians of different specialties, varied levels of experience, and from different institutions.

In our future work, we will enlarge the dataset and use an independent dataset for both the CDSS-T system evaluation and the observer study. We will improve the performance of the CDSS-T system by integrating radiomics with clinical data and molecular and histopathology biomarkers and by incorporating different deep learning models. We will also perform observer study on a larger scale to further test the effectiveness of the computerized decision support system.

Although the results of this observer study shows that our CDSS-T system is promising as an aid to physicians for bladder cancer treatment response assessment, extensive work is still required before translation to the clinic. First, further improvement of its performance is needed as described above. Second, the processing pipeline has to be automated, including the AL-CALS system for auto tumor segmentation [46] and radiomic feature extraction. Third, the CDSS-T system will need to undergo rigorous validation of its generalizability with a wide range of patient cases from multiple sites. Finally, well-controlled clinical trial should be conducted by incorporating it into a real-world workflow to evaluate its effectiveness and efficiency as a preparatory step before clinical translation.

4.6 Conclusion

In conclusion, our study demonstrated that the computerized decision support system (CDSS-T) has the potential to improve the diagnostic accuracy in assessing the complete response of muscle-invasive bladder cancer to neoadjuvant chemotherapy prior to radical cystectomy. The use of CDSS-T aid has resulted in improved and more consistent diagnostic performance among the physicians from multiple institutions and multiple specialties.

Chapter 5 Confidence Estimation of AI Model Output Affecting Decision Making for Bladder Cancer Treatment Response Assessment

5.1 Confidence Estimation of AI Model Output for Bladder Cancer Treatment Response Assessment in CT Urography

5.1.1 Abstract

Confidence estimation of machine learning (ML) and artificial intelligence (AI) model outputs is crucial for guiding decision-making. This study proposes two confidence estimation methods to enhance the trustworthiness of ML and AI models for classifying complete and non-complete responders for bladder cancer following neoadjuvant chemotherapy using CT urography: Segmentation Variability-Based Confidence Estimation (SVCE) and Classification Model Variability-Based Confidence Estimation (CMVCE). SVCE assesses prediction confidence based on the variability in predictions from models with different tumor segmentations, and thus extracted feature values, but the same classification model, while CMVCE evaluates confidence based on the variability in predictions from models with the same segmentations but different classification models. Our results demonstrate that cases with lower prediction variability consistently achieve better classification performance, with area under the receiver operating characteristic curve (AUC) values exceeding those of cases with higher variability. This study highlights that our proposed variability-based methods have the potential to provide an indicator of trustworthiness of the ML/AI prediction, which may enhance clinicians' confidence in adopting ML/AI models for decision support.

5.1.2 Introduction

The trustworthiness and reliability of machine learning (ML) and artificial intelligence (AI) models are critical concerns when integrating them into physicians' workflows. A key challenge is ensuring that users understand how much to trust the model's output. Confidence estimation—the ability of a model to indicate the reliability of its predictions—can help clinicians make more informed decisions by supplementing their expertise with an assessment of uncertainty. In this study, we propose methods to estimate the variability of ML/AI models in classifying complete responders and non-complete responders following neoadjuvant chemotherapy (NAC) for bladder cancer based on CT urography (CTU), based on the work of CDSS-T (detailed in Chapter 2.1). We evaluate whether the variability can serve as an indicator of trustworthiness and reliability of the ML/AI model's prediction.

5.1.3 Methods and Materials

5.1.3.1 Dataset

With Institutional Review Board (IRB) approval, we retrospectively collected 234 pre- and post-NAC CTU scan pairs from 204 bladder cancer patients. Complete response (CR), also known as complete remission, refers to the disappearance of all signs of cancer after treatment. The ground truth for complete response in our study was determined based on radical cystectomy pathology: pT0 (complete response) and >pT0 (non-complete response). Among the dataset, 51 scan pairs correspond to complete responders (CRs), while 183 scan pairs correspond to non-complete responders (NCRs). The 234 scan pairs were split by case into independent training, validation, and test sets.

5.1.3.2 Confidence Estimation

In our ML/AI prediction framework, a segmentation algorithm first identifies the tumor lesion from the CTU scan, defining the region of interest (ROI). A classification model then extracts and analyzes the image features from the ROI, performs classification, and generates a prediction score. Given these two critical components—segmentation and classification—we propose two methods (Figure 5-1) to estimate the confidence of model predictions:

Segmentation Variability-Based Confidence Estimation (SVCE): This method estimates confidence by leveraging an ensemble of ML/AI models, each using a different segmentation algorithm to delineate the lesion while employing the same classification model to generate the prediction score. Variability in segmentation impacts the extracted feature values and, consequently, the model's prediction scores. We hypothesize that greater variability in predictions indicates higher uncertainty and lower confidence, resulting from the case being more challenging to analyze or being more "difficult."

Classification Model Variability-Based Confidence Estimation (CMVCE): This method assesses confidence by using an ensemble of ML/AI models, each applying the same segmentation algorithm but different classification models. Since different classification models may emphasize different aspects of image information, they produce varying predictions, contributing to confidence estimation. In analogy to above, we hypothesize that greater variability in predictions indicates higher uncertainty and lower confidence, resulting from the case being more "difficult."

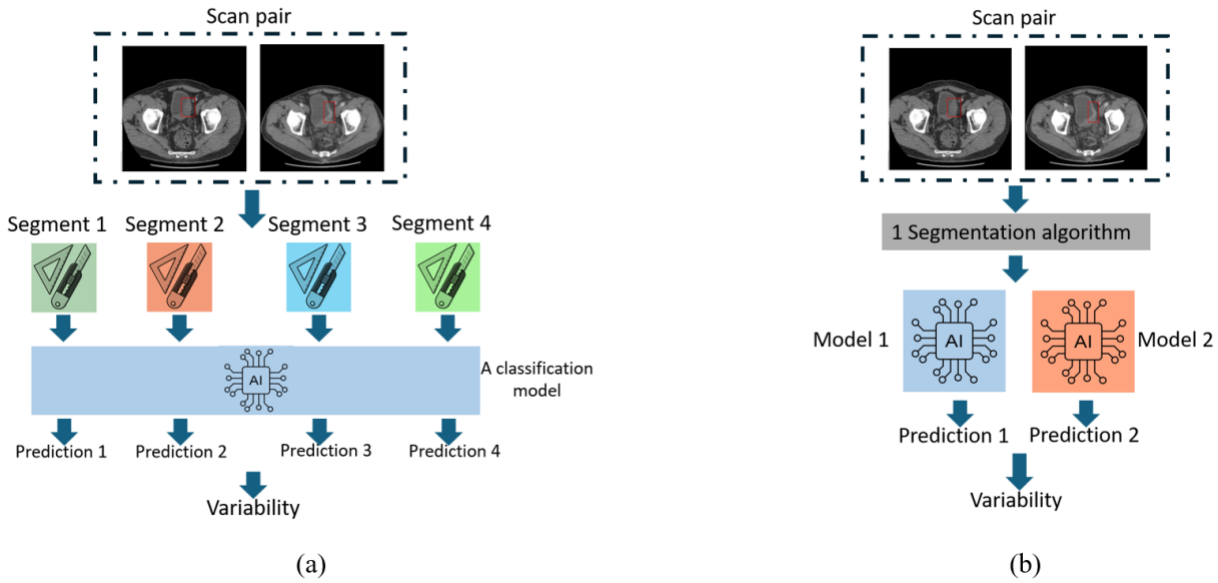


Figure 5-1: Proposed methods to estimate the confidence of ML/AI model outputs. (a) Segmentation Variability-Based Confidence Estimation (SVCE). (b) Classification Model Variability-Based Confidence Estimation (CMVCE).

Segmentation Algorithms

Three different segmentation algorithms were utilized: auto-initialized cascaded level sets system (AI-CALS) [46], U-Net model (U-Net) [108-110], and Attention U-Net segmentation model (AU-Net) [109, 110]. Among these, AI-CALS relies on a reader-defined ROI box, within which automated segmentation is performed. We have two readers to perform the segmentation by AI-CALS. In total, we obtained four sets of segmentations for all cases: AI-CALS_R1, AI-CALS_R2, U-Net, and AU-Net.

Classification Models

After obtaining the segmented tumor lesions, classification models were used to extract and analyze image features. We used two classification models [1].

The first classification model is a deep-learning convolutional neural network (DL-CNN; detailed in Chapter 2.1.3.5), the DL-CNN was initially trained to classify CRs and NCRs using

hybrid ROIs from the training set. The trained models were then deployed on the validation set to select the optimal model. Finally, the optimal model was applied to the hybrid ROIs of the test set, where the likelihood score of a complete response for each lesion pair was obtained by taking the minimum score among all hybrid ROIs from the same lesion pair.

The second classification model is based on the In-House Radiomics model (detailed in Chapter 2.1.3.4). After training the model on the training set, parameter optimization was guided by the validation set. The model with the optimal parameter settings was then deployed on the test set to perform classification, generating a likelihood score of a complete response for each lesion pair.

Confidence Estimation

We developed prediction models based on different combinations of segmentation algorithms and classification models. Here is a specific example of how confidence estimation was performed. Using four different segmentation algorithms combined with the same classification model (SVCE method) resulted in four prediction models: AI-CALS_R1-DL-CNN, AI-CALS_R2-DL-CNN, U-Net-DL-CNN, and AU-Net-DL-CNN.

For each case in the dataset, we calculated the standard deviation of the outputs from these four models. The variability scale across the training set was then determined, and a threshold was specified to categorize the variability level. Cases with variability below the threshold were classified as “easy,” while those with variability above the threshold were classified as “difficult.” The terms “easy” and “difficult” represent the confidence levels assigned to the predictions for each test case.

5.1.3.3 Statistical Analysis

To validate our hypothesis, we evaluated the classification performance for each ensemble of models in the "easy" and "difficult" groups using the area under the receiver operating characteristic curve (AUC). AUCs were computed using the LABROC method, which also estimates the standard error (SE) of the AUC based on Taylor series expansion and asymptotic approximations.

If the AUC for the "easy" group was higher than that for the "difficult" group, this was evidence supporting our hypothesis that low variability in the ML/AI score for a case indicated that the prediction was more accurate and trustworthy. Thus, the variability-based estimations would provide useful information for clinicians to judge if they could accept the ML/AI model's recommendation for decision support.

5.1.4 Results

Figure 5-2 shows an example of a cancerous lesion on CTU scan segmented by the AI-CALS_R1, AI-CALS_R2, AU-Net, and U-Net. The results obtained using the SVCE method are presented in Table 5-1 and Table 5-2. A consistent trend was observed, indicating that the AUC for the "difficult" cases was lower (AUC range: 0.58–0.80) compared to the "easy" cases (AUC range: 0.85–0.92) for the radiomics model. A similar trend was also observed for the DL-CNN model.

The results based on the CMVCE method are shown in Table 5-3–Table 5-6, where a higher AUC was consistently observed for the "easy" group compared to the "difficult" group.

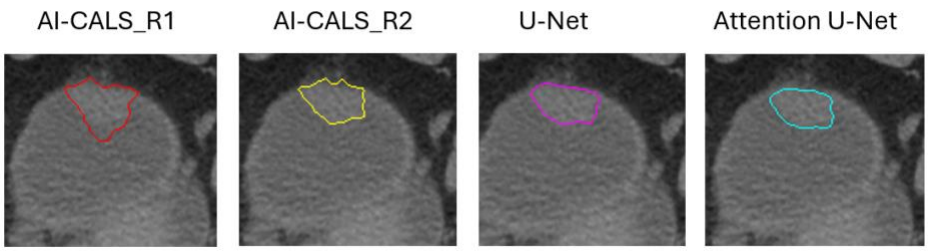


Figure 5-2: Segmentation of a cancerous lesion on CTU scan using AI-CALS_R1, AI-CALS_R2, AU-Net, and U-Net.

Table 5-1: SVCE method. Variability was estimated from four feature sets extracted from four different lesion segmentation methods. Each feature set was input to the radiomics model for treatment response prediction.

Classification model	**Segmentation algorithm**	**Overall (AUC)**	**"Difficult" (AUC ± SE)**	**"Easy" (AUC ± SE)**
Radiomics	AI-CALS_R1	0.787 ± 0.057	0.771 ± 0.079	0.850 ± 0.078
	AI-CALS_R2	0.811 ± 0.053	0.798 ± 0.068	0.852 ± 0.075
	AU-Net	0.703 ± 0.079	0.606 ± 0.129	0.893 ± 0.054
	U-Net	0.685 ± 0.084	0.583 ± 0.131	0.920 ± 0.046

Table 5-2: SVCE method. Variability was estimated from four hybrid ROI sets extracted from four different lesion segmentation methods. Each hybrid ROI set was input to the DL-CNN model for treatment response prediction.

Classification model	**Segmentation algorithm**	**Overall (AUC)**	**"Difficult" (AUC ± SE)**	**"Easy" (AUC ± SE)**
Dl-CNN	AI-CALS_R1	0.805 ± 0.062	0.527 ± 0.166	0.821 ± 0.089
	AI-CALS_R2	0.778 ± 0.072	0.540 ± 0.140	0.718 ± 0.110
	AU-Net	0.721 ± 0.071	0.330 ± 0.132	0.638 ± 0.119
	U-Net	0.631 ± 0.089	0.397 ± 0.140	0.527 ± 0.114

Table 5-3: CMVCE method. The variability was estimated from the radiomics and the DL-CNN model scores when lesion segmentation was fixed to be AI-CALS_R1.

Segmentation algorithm	**Classification model**	**Overall (AUC)**	**"Difficult" (AUC ± SE)**	**"Easy" (AUC ± SE)**
AI-CALS_R1	Radiomics	0.787 ± 0.057	0.603 ± 0.148	0.798 ± 0.076
	DL-CNN	0.805 ± 0.062	0.739 ± 0.132	0.747 ± 0.083

Table 5-4: CMVCE method. The variability was estimated from the radiomics and the DL-CNN model scores when lesion segmentation was fixed to be AI-CALS_R2.

Segmentation algorithm	Classification model	Overall (AUC)	"Difficult" (AUC ± SE)	"Easy" (AUC ± SE)
AI-CALS_R2	Radiomics	0.811 ± 0.053	0.481 ± 0.126	0.877 ± 0.063
	DL-CNN	0.778 ± 0.072	0.585 ± 0.123	0.813 ± 0.120

Table 5-5: CMVCE method. The variability was estimated from the radiomics and the DL-CNN model scores when lesion segmentation was fixed to be AU-Net.

Segmentation algorithm	Classification model	Overall (AUC)	"Difficult" (AUC ± SE)	"Easy" (AUC ± SE)
AU-Net	Radiomics	0.703 ± 0.079	0.454 ± 0.160	0.733 ± 0.102
	DL-CNN	0.721 ± 0.071	0.308 ± 0.169	0.800 ± 0.082

Table 5-6: CMVCE method. The variability was estimated from the radiomics and the DL-CNN model scores when lesion segmentation was fixed to be U-Net.

Segmentation algorithm	Classification model	Overall (AUC)	"Difficult" (AUC ± SE)	"Easy" (AUC)
U-Net	Radiomics	0.685 ± 0.084	0.268 ± 0.130	0.642 ± 0.144
	DL-CNN	0.631 ± 0.089	0.298 ± 0.155	0.764 ± 0.101

5.1.5 Discussion

For this study, the ML/AI prediction models are subject to two primary types of uncertainty: epistemic and aleatoric.

One contributing factor to epistemic uncertainty is the model's insufficient knowledge of the correct segmentation. Aleatoric uncertainty, on the other hand, arises when certain cases inherently contain ambiguous features, making them difficult to classify, regardless of segmentation quality.

If the segmentation remains constant but the predictions vary significantly, this suggests that the image features are challenging for the model to interpret or learn in relation to the

classification task. This scenario highlights the difficult cases, a concept central to the CMVCE method.

In contrast, for cases that are truly "easy," classification should remain relatively stable even with variations in segmentation. However, for "difficult" cases, even slight changes in segmentation may result in substantially different predictions. This aligns with the principles of the SVCE method.

Together, these methods provide a framework for investigating both epistemic and aleatoric uncertainties in model predictions.

Future directions for this research include refining segmentation techniques to improve accuracy, exploring additional methods for uncertainty quantification, and expanding the study to multi-center datasets to enhance the generalizability of the findings.

5.1.6 Conclusion

This study demonstrated the feasibility of estimating the confidence of ML/AI model outputs using the SVCE and CMVCE methods, highlighting the relationship between prediction confidence and accuracy. The results consistently showed that the AUC for "difficult" cases was lower compared to that for "easy" cases, indicating that predictions for highly variable cases tend to be less reliable. By incorporating confidence estimation into medical decision-making, users could gain a better understanding of how much to trust the model's output in conjunction with their own clinical judgment.

5.2 Observer Study: Impact of Case Complexities and Physician Characteristics on AI-Assisted Treatment Response Assessment in Bladder Cancer

5.2.1 Abstract

This study explores the impact of physician experience, specialty, and institutional background on the performance of AI-assisted assessment of treatment responses in bladder cancer patients. Utilizing pre- and post-chemotherapy CTU scans from 123 patients, 17 physicians with varying levels of experience and from different specialties and institutions assessed 157 lesion pairs. The lesion pairs were divided into easy and difficult cases to evaluate the AI system's effectiveness in different scenarios. The study revealed that AI assistance significantly improved diagnostic accuracy in easy cases for both experienced and inexperienced physicians, with a great benefit observed in radiologists and oncologists. In difficult cases, the AI's impact was present but less pronounced, indicating that while AI can enhance performance in challenging situations, its effectiveness is more limited in complex cases. Additionally, the study found that institutional background influenced the effectiveness of AI assistance, suggesting that certain training or cultural factors may affect physicians' trust in AI recommendations. The findings underscore the potential of AI to support clinical decision-making in bladder cancer treatment response assessment, particularly in less complex cases. However, they also highlight the need for tailored implementation and user training of AI systems to maximize their effectiveness across different medical specialties and institutions. By aligning AI tools with the specific needs and expertise of physicians, their confidence and efficacy in using AI in complex medical scenarios can be enhanced.

5.2.2 Introduction

Based on the work of CDSS-T (detailed in Chapter 4), and confidence estimation of AI/ML models' output (detailed in Chapter 5.1), we further investigate how different physician characteristics affect their interaction with and reliance on AI tools, which is crucial for optimizing AI's adoption in medical decision-making. The primary goal of this study is to assess

how physician experience level, specialty, and institutional background impact the effectiveness of AI-assisted assessment of treatment responses for bladder cancer patients. Specifically, the study focuses on evaluating these effects in both easy and difficult cancer cases.

5.2.3 Methods and Materials

5.2.3.1 Data

With Institutional Review Board (IRB) approval, we collected CTU scans from 123 patients who had undergone neoadjuvant chemotherapy (NAC). In total, we identified 157 pairs of pre- and post-NAC cancerous lesions. The pathological cancer stage following NAC served as the ground truth for determining whether a patient had achieved a complete response to treatment. Among the 157 lesion pairs, 40 (25%) showed a complete response (stage T0) after chemotherapy, while 117 (75%) exhibited an incomplete response (>T0). The lesion pairs were divided into easy and difficult subsets, allowing for an analysis of how the impact of the AI-aid on physicians' decision varied based on case complexity.

5.2.3.2 Observers Study Design

The study involved 17 physicians [4] (same as Chapter 4) with varying levels of experience, specialties, and institutional affiliations. Observer characteristics can be found in Table 4-2.The observers included both experienced and inexperienced radiologists, oncologists, a urologist, a neurology resident, and a medical student. These observers were affiliated with four different institutions: the University of Michigan (UM), Pennsylvania State University (PSU), the University of Iowa (UI), and Tokuda Hospital (TH) in Sofia, Bulgaria.

We have integrated our AI model into an interactive platform, referred to as a decision support system in treatment response assessments of bladder cancer patients (CDSS-T). This

platform allows observers to review CT images, examine the AI's prediction, and provide their evaluations before and after being shown the AI's prediction.

The observer study followed a structured process (Figure 4-1; detailed in Chapter 4.3.4). Based on the ratings among the nine radiologists, the 157 lesion pairs were divided into "easy" and "difficult" groups, where the easy case was when the standard deviation of the ratings was smaller than 25, and the difficult case was when the standard deviation of the ratings was larger than 25. The threshold of 25 was selected by approximately balancing the number of complete responses (T0 after chemotherapy) in the easy and difficult cancer subsets for the purpose of performing a reliable ROC analysis [4].

5.2.3.3 Evaluation Metrics

The performance of the treatment response assessment was evaluated using the Area Under the Curve (AUC) metric. The study compared the average AUC scores achieved by the observers with and without CDSS-T assistance across different observer characteristics and case complexities. For each subgroup, the effect of AI-aid was evaluated using multi-reader multi-case (MRMC) [105] statistical analysis.

5.2.3.4 Subgroup Analysis

The impact of the AI-aid on observer's decision was evaluated based on the experience level, specialty, and institution of the observers through subgroup analyses:

Experience Level: Comparisons were made between experienced and inexperienced physicians across both easy and difficult cases.

Specialty: The study analyzed how the AI-aid influenced physicians or trainees depend on their specialties, including radiology, oncology, urology, neurology, and trainees.

Institutional Background: The study also examined how institutional affiliation (UM, PSU, UI, and TH) might affect the observers' performance with AI assistance.

5.2.4 Results

The results are summarized in Table 5-7, which presents the observer's performance based on experience level, specialty, and institutional background. Figure 5-3 offers a more intuitive illustration of the change in performance.

5.2.4.1 Experience Level

Easy Cases: Experienced physicians (N=11) exhibited a significant improvement in AUC with AI assistance, increasing from 0.80 to 0.84 ($p = 0.013$). Inexperienced physicians (N=6) also demonstrated a noticeable improvement, with AUC rising from 0.79 to 0.84 ($p = 0.190$), indicating that AI assistance benefited physicians across experience levels in easy cases.

Difficult Cases: For difficult cases, the AI's effect was moderate. Experienced physicians obtained an AUC increased from 0.59 to 0.62 ($p = 0.138$), while inexperienced physicians improved from 0.58 to 0.61 ($p = 0.300$). These results suggested that while AI enhanced performance in challenging cases, the extent of improvement was less pronounced compared to easy cases.

5.2.4.2 Specialty

Radiologists: In easy cases, experienced radiologists (N=9) benefited significantly from AI assistance, with AUC improving from 0.83 to 0.85 ($p = 0.033$). In difficult cases, the improvement was moderate, with AUC increasing from 0.59 to 0.61 ($p = 0.379$).

Oncologists: Oncologists (N=5) showed notable gains with AI in easy cases, with AUC rising from 0.78 to 0.84 ($p = 0.051$). In difficult cases, they gained statistically significant improvement, with AUC increasing from 0.57 to 0.63 ($p = 0.009$).

Other Specialties: The AI's impact varied across other specialties. Urologist (N=1) and neurology resident (N=1) showed improvements, but these were less consistent and generally moderate, particularly in difficult cases.

5.2.4.3 Institutional Background

UM: Physicians from UM (N=12) showed significant improvement in easy cases with AI assistance, with AUC increasing from 0.81 to 0.84 ($p = 0.020$). In difficult cases, the improvement was moderate, with AUC rising from 0.59 to 0.62 ($p = 0.199$), reflecting a benefit from AI across case complexities.

UI, TH, and PSU: Physicians from these institutions (N=2, 1, 2, respectively) experienced improvements in both easy and difficult cases with AI assistance, except the TH physician on the difficult cases.

5.2.5 Discussion

It is important to note that the significance of the performance difference between the assessment with and without CDSS-T is strongly influenced on the number of observers in each subgroup. In some subgroups, although performance differences reach statistical significance, the imbalance in observer numbers limits the ability to attribute these statistical differences solely to observer characteristics. To more reliably assess the effectiveness of the model, a larger-scale observer study with balanced subgroup representation is warranted.

Table 5-7: AUCs of observers with and without CDSS-T for each subgroup.

Categories	Subgroups	Subset	AUC of CDSS-T	Average AUC w/o CDSS-T	Average AUC w/ CDSS-T	*p* value	# of observers
Experience level	Experienced physicians	Easy	0.88	0.80	0.84	0.013*	11
		Difficult	0.67	0.59	0.62	0.138	
	Inexperienced physicians	Easy	0.88	0.79	0.84	0.190	6
		Difficult	0.67	0.58	0.61	0.300	
	Experienced radiologists	Easy	0.88	0.83	0.86	0.002*	5
		Difficult	0.67	0.59	0.60	0.669	
	Inexperienced radiologists	Easy	0.88	0.82	0.84	0.144	4
		Difficult	0.67	0.58	0.61	0.273	
	UM experienced radiologists	Easy	0.88	0.84	0.87	<0.001*	4
		Difficult	0.67	0.58	0.59	0.483	
	UM inexperienced radiologists	Easy	0.88	0.82	0.84	0.144	4
		Difficult	0.67	0.58	0.61	0.273	
Specialties	Oncologists	Easy	0.88	0.78	0.84	0.051	5
		Difficult	0.67	0.57	0.63	0.009*	
	Radiologists	Easy	0.88	0.83	0.85	0.033*	9
		Difficult	0.67	0.59	0.61	0.379	
	Urologists	Easy	0.88	0.76	0.79	0.382	1
		Difficult	0.67	0.64	0.67	0.595	
	Medical student	Easy	0.88	0.65	0.86	0.004*	1
		Difficult	0.67	0.60	0.60	0.978	
	Neurology resident	Easy	0.88	0.81	0.82	0.337	1
		Difficult	0.67	0.55	0.60	0.395	
Institutions	UM	Easy	0.88	0.81	0.84	0.020*	12
		Difficult	0.67	0.59	0.62	0.199	
	UI	Easy	0.88	0.78	0.86	0.407	2
		Difficult	0.67	0.56	0.63	0.094	
	TH	Easy	0.88	0.80	0.81	0.441	1
		Difficult	0.67	0.64	0.62	0.389	
	PSU	Easy	0.88	0.72	0.85	0.341	2
		Difficult	0.67	0.57	0.60	0.512	

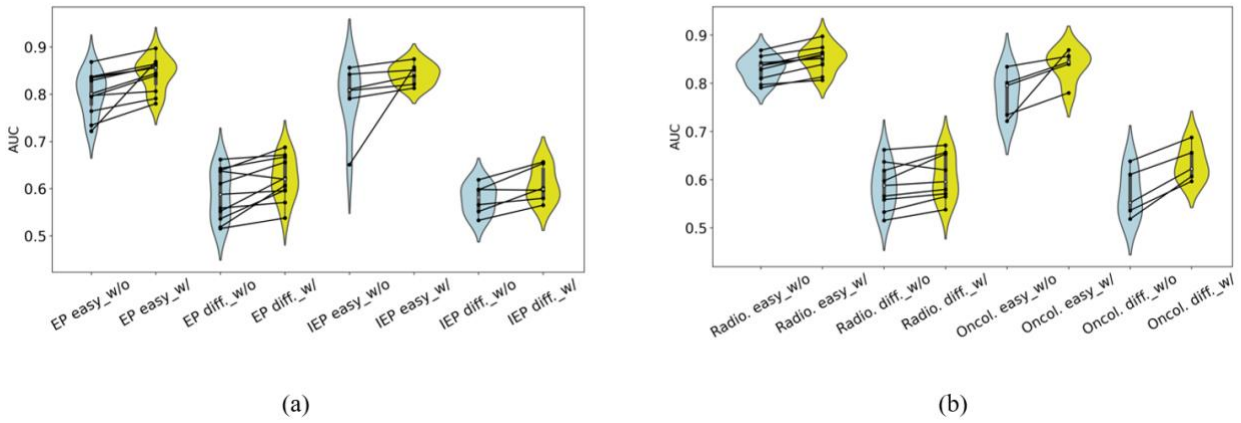


Figure 5-3: Paired violin plots for individual AUC w/o and w/ CDSS-T aid. (a) Experienced/inexperienced physicians (EP/IEP) on the easy and difficult subsets, w/o and w/ CDSS-T aid. (b) Radiologists and oncologists on the easy and difficult subsets, w/o and w/ CDSS-T aid.

5.2.6 Conclusion

This study demonstrates that AI could enhance physicians' performance in assessing bladder cancer treatment responses in both easy and difficult cases. The AI system improved diagnostic accuracy in easy cases, irrespective of the physician's experience level, specialty, or institutional background. However, in challenging cases, the AI's impact was present but more moderate. These findings highlight the potential of AI to support clinical decision-making across a range of cases, with varying degrees of effectiveness depending on the complexity of the case and AI's performance on the case. The study also indicated that institutional background might influence the effectiveness of AI assistance, suggesting that certain training or cultural factors may affect physicians' trust in AI recommendations. This underscores the need for tailored implementation and user training of AI systems to maximize the benefits of AI aid across different medical specialties and institutions.

Chapter 6 Summary and Future Work

This dissertation explored the application of advanced machine learning (ML) and artificial intelligence (AI) techniques in improving the treatment response assessment and prognosis for bladder cancer. Through a series of models and studies, this work demonstrates the potential of AI-driven tools to enhance clinical decision-making and patient outcomes. In this chapter, the key findings are summarized, and potential directions for future research are proposed.

6.1 Summary

The primary aims of this dissertation were to 1) develop AI/ML models that assist in the assessment of bladder cancer treatment response and prediction of patient survival, and 2) evaluate their impact and reliability in clinical decision-making. Each chapter contributes to the overarching goals in distinct ways:

Chapter 1 provided an introduction, outlining motivation, problem definition, and a summary of the main contributions of this dissertation.

Chapter 2 presented machine learning models to access the treatment response of bladder cancer to neoadjuvant chemotherapy.

Chapter 2.1 studied a deep-learning convolutional neural network (DL-CNN) and radiomics model for assessing bladder cancer treatment response to neoadjuvant chemotherapy using CT urography. The combined DL-CNN and radiomics model achieved an area under the

ROC curve (AUC) of 0.77-0.80, showing the feasibility of using AI to assess treatment response in bladder cancer.

Chapter 2.2 presented a multi-institutional study of radiomics model for assessment of treatment response to neoadjuvant chemotherapy of bladder cancer was presented.

Chapter 3 explored a multi-modal survival prediction model for estimating the five-year survival of bladder cancer patients post-cystectomy.

Chapter 3.1 proposed the CRD model, which integrated clinical, radiomics, and deep-learning descriptors, achieving an AUC of 0.87, highlighting the importance of combining multiple data sources to improve survival prediction.

Chapter 3.2 explored the use of machine learning models based on radiomics and genomic data (RF-GENM) to predict the two-year survival of bladder cancer patients following radical cystectomy, achieving the highest AUC of 0.74±0.09, outperforming both the radiomics model (AUC=0.63±0.10) and genomic model (AUC=0.69±0.10).

Chapter 3.3 evaluated the use of large language models (LLMs) in improving the survival prediction model by extracting clinical information from unstructured textual data in electronic medical records (EMRs). The LLM-aided CRD model achieved AUCs comparable to those based on ground truth textual data, demonstrating the potential of LLMs to automate clinical data extraction and enhance the efficiency of predictive modeling.

Chapter 3.4 assessed the reliability of various LLMs in clinical data extraction for bladder cancer prognosis. The results revealed that GPT-4 exhibited outstanding performance, with information extraction accuracy consistently exceeding 93% and survival prediction metrics closely aligned with ground truth (AUC ± 0.02). Llama showed the highest reliability among the offline LLMs being evaluated. This chapter underscores the reliability of LLMs in extracting

clinically relevant information for AI-driven prediction tasks, while also addressing challenges related to model variability.

Chapter 4 investigated the impact of AI-based decision support systems (CDSS-T) on physicians' diagnostic accuracy in assessing bladder cancer treatment response. The study found that the use of CDSS-T significantly improved the diagnostic performance of 17 observers. This suggested that AI-based systems can enhance decision-making in clinical settings, particularly when integrated with the clinical expertise of healthcare providers.

Chapter 5 explored methods for estimating the confidence of AI/ML model outputs, as well as the influence of case complexity and physician characteristics on AI-assisted assessments. This work is critical in determining the trustworthiness of AI tools in clinical practice and ensuring that AI-driven insights are both accurate and actionable.

6.2 Future Work

While the results presented in this dissertation demonstrate the potential of AI in treatment response assessment and prognosis of bladder cancer, there remain several opportunities for future research to further enhance the utility of these models and extend their application in clinical practice.

Integration of Multi-Institutional Data: The models developed in this dissertation were based on data from a single institution. Future work should explore the integration of multi-institutional data to increase the generalizability and robustness of the models. This would require harmonizing data from diverse sources and refining the models to ensure consistent performance across different clinical settings.

Longitudinal Model Development: Bladder cancer progression is dynamic, and treatment responses and patient outcomes could evolve over time. Developing longitudinal

models that incorporate longer sequential data—beyond the two time points used in this study—could further improve the prediction of treatment responses and survival outcomes. This would involve leveraging extended temporal data to more accurately capture disease progression and therapeutic effects.

Expanding the use of Large Language Models: The promising results of LLMs in clinical data extraction suggest further exploration of their application in other areas of healthcare. Future work could focus on fine-tuning LLMs for more specific medical tasks, such as personalized treatment recommendation, automated clinical documentation, and enhanced disease classification. Fine-tuning these models with domain-specific datasets could improve their accuracy and reliability in specialized medical contexts. Additionally, LLMs can be utilized for imaging data analysis to enhance diagnostic capabilities. The potential benefits of applying LLMs to process imaging data include automated annotation, identification of subtle imaging biomarkers, and improved consistency in image interpretation. These advancements could lead to more efficient workflows, reduced clinician workload, and ultimately, better patient outcomes.

Prospective Study to Evaluating AI Systems in Clinical Practice: The observer study in Chapter 4 highlighted the potential of AI to enhance physician decision-making. Future research could extend this work by conducting large-scale prospective studies that evaluate the impact of AI on clinical outcomes, such as patient survival and treatment-related complications.

Translation to Broader Clinical Applications: While this work focused on bladder cancer, the developed methodology for conducting observer study has been applied to other disease in one of the collaborative projects, functional seizures [34], demonstrating its potential broad applicability. Additionally, other methodologies, such as the use of LLMs for clinical data processing and interpretation, and the integration of radiologic imaging with clinicopathological

information, hold strong potential for widespread application across various disease domains. These approaches could be adapted to study other cancers, such as lung or breast cancer, as well as chronic conditions like cardiovascular disease or neurodegenerative disorders. Future research could further explore the generalizability of these methods to different datasets and clinical contexts, supporting the development of versatile, disease-agnostic AI tools that enhance diagnosis, prognosis, and treatment planning across diverse medical specialties.

Ethical Considerations and Human-AI Collaboration: As AI becomes increasingly integrated into clinical decision-making, it is essential to explore the ethical implications of AI systems in healthcare, including issues such as bias, transparency, accountability, and patient autonomy. It is also crucial to examine how AI can be designed to meaningfully collaborate with clinicians, including preventing overreliance on AI, and ensuring AI tools support rather than undermine physician judgment.

6.3 Conclusion

This dissertation contributes to advancing the integration of AI and machine learning in bladder cancer prognosis and treatment response assessment. By developing predictive models, evaluating AI tools in observer study, and investigating the reliability of machine learning models, this work paves the way for the widespread adoption of AI-driven clinical decision support systems. The promising results highlight the transformative potential of AI in improving patient care, while the future directions outlined in this chapter provide a roadmap for continued progress in the field. As AI and machine learning continue to evolve, their integration into healthcare will undoubtedly lead to more accurate, personalized, and timely interventions, improving outcomes for bladder cancer patients worldwide.

Bibliography

1. Sun, D., L. Hadjiiski, H.-P. Chan, R. H. Cohan, E. M. Caoili, A. Alva, K. H. Cha, R. K. Samala and C. Zhou. "Bladder cancer treatment response assessment in CT urography by using deep-learning and radiomics." *SPIE Medical Imaging 2023: Computer-Aided Diagnosis* 12465 (2023): 738-42.
2. Sun, D., L. Hadjiiski, J. Gormley, H.-P. Chan, E. M. Caoili, R. H. Cohan, A. Alva, V. Gulani and C. Zhou. "Survival prediction of patients with bladder cancer after cystectomy based on clinical, radiomics, and deep-learning descriptors." *Cancers* 15 (2023): 4372.
3. Sun, D., L. Hadjiiski, J. Gormley, H.-P. Chan, E. Caoili, R. Cohan, A. Alva, G. Bruno, R. Mihalcea and C. Zhou. "Outcome prediction using multi-modal information: Integrating large language model-extracted clinical information and image analysis." *Cancers* 16 (2024): 2402.
4. Sun, D., L. Hadjiiski, A. Alva, Y. Zakharia, M. Joshi, H.-P. Chan, R. Garje, L. Pomerantz, D. Elhag and R. H. Cohan. "Computerized decision support for bladder cancer treatment response assessment in CT urography: Effect on diagnostic accuracy in multi-institution multi-specialty study." *Tomography* 8 (2022): 644-56.
5. Smith-Bindman, R., M. L. Kwan, E. C. Marlow, M. K. Theis, W. Bolch, S. Y. Cheng, E. J. Bowles, J. R. Duncan, R. T. Greenlee and L. H. Kushi. "Trends in use of medical imaging in us health care systems and in Ontario, Canada, 2000-2016." *Jama* 322 (2019): 843-56.
6. Smith-Bindman, R., D. L. Miglioretti and E. B. Larson. "Rising use of diagnostic medical imaging in a large integrated health system." *Health affairs* 27 (2008): 1491-502.
7. Hendee, W. R., G. J. Becker, J. P. Borgstede, J. Bosma, W. J. Casarella, B. A. Erickson, C. D. Maynard, J. H. Thrall and P. E. Wallner. "Addressing overutilization in medical imaging." *Radiology* 257 (2010): 240-45.
8. Chan, H.-P., R. K. Samala, L. M. Hadjiiski and C. Zhou. "Deep learning in medical image analysis." *Deep learning in medical image analysis: challenges and applications* (2020): 3-21.
9. Ting, D. S., Y. Liu, P. Burlina, X. Xu, N. M. Bressler and T. Y. Wong. "AI for medical imaging goes deep." *Nature medicine* 24 (2018): 539-40.
10. Pesapane, F., M. Codari and F. Sardanelli. "Artificial intelligence in medical imaging: Threat or opportunity? Radiologists again at the forefront of innovation in medicine." *European radiology experimental* 2 (2018): 1-10.
11. Samala, R. K., H. P. Chan, L. Hadjiiski, M. A. Helvie, J. Wei and K. Cha. "Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography." *Medical physics* 43 (2016): 6654-66.
12. Krönke, M., C. Eilers, D. Dimova, M. Köhler, G. Buschner, L. Schweiger, L. Konstantinidou, M. Makowski, J. Nagarajah and N. Navab. "Tracked 3D ultrasound and

deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry." *Plos one* 17 (2022): e0268550.
13. Helaly, H. A., M. Badawy and A. Y. Haikal. "Deep learning approach for early detection of alzheimer's disease." *Cognitive computation* 14 (2022): 1711-27.
14. Madabhushi, A. and G. Lee. "Image analysis and machine learning in digital pathology: Challenges and opportunities." *Medical image analysis* 33 (2016): 170-75.
15. Bera, K., K. A. Schalper, D. L. Rimm, V. Velcheti and A. Madabhushi. "Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology." *Nature reviews Clinical oncology* 16 (2019): 703-15.
16. Wang, C., X. Zhu, J. C. Hong and D. Zheng. "Artificial intelligence in radiotherapy treatment planning: Present and future." *Technology in cancer research & treatment* 18 (2019): 1533033819873922.
17. Netherton, T. J., C. E. Cardenas, D. J. Rhee, L. E. Court and B. M. Beadle. "The emergence of artificial intelligence within radiation oncology treatment planning." *Oncology* 99 (2021): 124-34.
18. Bica, I., A. M. Alaa, C. Lambert and M. Van Der Schaar. "From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges." *Clinical Pharmacology & Therapeutics* 109 (2021): 87-100.
19. Liu, Y., B. Logan, N. Liu, Z. Xu, J. Tang and Y. Wang. "Deep reinforcement learning for dynamic treatment regimes on medical registry data." *IEEE international conference on healthcare informatics (ICHI)* 2017: 380-85.
20. Dyrskjøt, L., D. E. Hansel, J. A. Efstathiou, M. A. Knowles, M. D. Galsky, J. Teoh and D. Theodorescu. "Bladder cancer." *Nature Reviews Disease Primers* 9 (2023): 58.
21. Lenis, A. T., P. M. Lec, K. Chamie and M. Mshs. "Bladder cancer: A review." *Jama* 324 (2020): 1980-91.
22. IARC (International Agency for Research on Cancer). 2024. *Cancer today*. Available online: https://gco.iarc.who.int/today (accessed on 8 February 2024.)
23. American Cancer Society. 2022. *Cancer facts & figures 2022*. Atlanta, Ga: American Cancer Society.
24. Ikeda, A., H. Nosato, Y. Kochi, T. Kojima, K. Kawai, H. Sakanashi, M. Murakawa and H. Nishiyama. "Support system of cystoscopic diagnosis for bladder cancer based on artificial intelligence." *Journal of endourology* 34 (2020): 352-58.
25. Teoh, J., E. Chan, A. Cheuk, R. Chan, J. Qin and C. Ng. "A newly developed computer-aided endoscopic diagnostic system for bladder cancer detection." *European Urology Open Science* 19 (2020): e1364-e65.
26. Shkolyar, E., X. Jia, T. C. Chang, D. Trivedi, K. E. Mach, M. Q.-H. Meng, L. Xing and J. C. Liao. "Augmented bladder tumor detection using deep learning." *European urology* 76 (2019): 714-18.
27. Suarez-Ibarrola, R., M. Kriegmair, F. Waldbillig, B. Grüne, M. Negassi, U. Parupalli, A. Schmitt, A. Reiterer, C. Müller and A. Scheurer. "A novel endoimaging system for endoscopic 3D reconstruction in bladder cancer patients." *Minimally Invasive Therapy & Allied Technologies* 31 (2022): 34-41.
28. Yang, R., Y. Du, X. Weng, Z. Chen, S. Wang and X. Liu. "Automatic recognition of bladder tumours using deep learning technology and its clinical application." *The*

International Journal of Medical Robotics and Computer Assisted Surgery 17 (2021): e2194.

29. Chan, E. O.-T., B. Pradere and J. Y.-C. Teoh. "The use of artificial intelligence for the diagnosis of bladder cancer: A review and perspectives." *Current Opinion in Urology* 31 (2021): 397-403.
30. Sun, R., M. Zhang, L. Yang, S. Yang, N. Li, Y. Huang, H. Song, B. Wang, C. Huang and F. Hou. "Preoperative ct-based deep learning radiomics model to predict lymph node metastasis and patient prognosis in bladder cancer: A two-center study." *Insights into Imaging* 15 (2024): 21.
31. Pinar, U., B. Pradere and M. Roupret. "Artificial intelligence in bladder cancer prognosis: A pathway for personalized medicine." *Current opinion in urology* 31 (2021): 404-08.
32. Suartz, C. V., L. M. Martinez, M. D. Cordeiro, H. A. Flores, S. Kodama, L. Cardili, J. M. Mota, F. M. A. Coelho, J. de Bessa Junior and C. P. Camargo. "Artificial intelligence for predicting response to neoadjuvant chemotherapy for bladder cancer: A comprehensive systematic review and meta-analysis." *Canadian Urological Association Journal* 18 (2024): E276.
33. Joshi, M., D. Sun, A. S. Alva, P. Palmbos, Y. Zakharia, R. Garje, H.-P. Chan, R. H. Cohan and L. Hadjiiski. "A multi-institutional observer study of ai-aided oncologists' performance on treatment response assessment of bladder cancer." *American Society of Clinical Oncology* (2024): 4599.
34. McFarlane, K., C. Allas, S. Terman, S. H. Seo, A. Karimi, S. Sreenivasan, J. Grauer, D. Sun, M. Watson and E. Patterson. "Quantifying the impact of computer-aided diagnostic score on the diagnosis of functional seizures (s29. 009)." *Neurology* 102 (2024): 6152.
35. Grossman, H. B., R. B. Natale, C. M. Tangen, V. Speights, N. J. Vogelzang, D. L. Trump, R. W. d. White, M. F. Sarosdy, D. P. Wood Jr and D. Raghavan. "Neoadjuvant chemotherapy plus cystectomy compared with cystectomy alone for locally advanced bladder cancer." *New England Journal of Medicine* 349 (2003): 859-66.
36. ABCAM-a, C. "Neoadjuvant chemotherapy in invasive bladder cancer: Update of a systematic review and meta-analysis of individual patient data; advanced bladder cancer (abc) meta-analysis collaboration." *Eur Urol* 48 (2005): 202-05.
37. Witjes, J., M. Wullink, G. Oosterhof and P. De Mulder. "Toxicity and results of mvac (methotrexate, vinblastine, adriamycin and cisplatin) chemotherapy in advanced urothelial carcinoma." *European urology* 31 (1997): 414-19.
38. Lyon, T. D., I. Frank, V. Sharma, P. H. Shah, M. K. Tollefson, R. H. Thompson, R. J. Karnes, P. Thapa, J. C. Cheville and S. A. Boorjian. "A risk-stratified approach to neoadjuvant chemotherapy in muscle-invasive bladder cancer: Implications for patients classified with low-risk disease." *World Journal of Urology* 37 (2019): 1605-13.
39. Boeri, L., M. Soligo, I. Frank, S. A. Boorjian, R. H. Thompson, M. Tollefson, F. J. Quevedo, J. C. Cheville and R. J. Karnes. "Cigarette smoking is associated with adverse pathological response and increased disease recurrence amongst patients with muscle-invasive bladder cancer treated with cisplatin-based neoadjuvant chemotherapy and radical cystectomy: A single-centre experience." *BJU international* 123 (2019): 1011-19.
40. Motterle, G., J. R. Andrews, A. Morlacco and R. J. Karnes. "Predicting response to neoadjuvant chemotherapy in bladder cancer." *European Urology Focus* 6 (2020): 642-49.

41. von Deimling, M., V. M. Schuettfort, D. D'Andrea, B. Pradere, N. C. Grossmann, T. Kawada, T. Yanagisawa, M. Majdoub, E. Laukhtina and P. Rajwa. "Predictive and prognostic role of the neutrophil-to-lymphocyte ratio in muscle invasive bladder cancer treated with neoadjuvant chemotherapy and radical cystectomy." *Clinical Genitourinary Cancer* 21 (2023): 430-441.
42. Bukavina, L., R. Ginwala, M. Sindhani, M. Prunty, D. Geynisman, G. Pooja, H. Valentine, A. Calaway, J. R. Brown and A. Correa. "Role of gut microbiome in neoadjuvant chemotherapy response in urothelial carcinoma: A multi-institutional prospective cohort evaluation." *Cancer Research Communications* 4 (2024): 1505-1516.
43. Barentsz, J. O., O. Berger-Hartog, J. A. Witjes, C. Hulsbergen-van der Kaa, G. Oosterhof, J. VanderLaak, H. Kondacki and S. Ruijs. "Evaluation of chemotherapy in advanced urinary bladder cancer with fast dynamic contrast-enhanced MR imaging." *Radiology* 207 (1998): 791-97.
44. Fitoussi, O., J. B. Roche, J. Riviere, H. Wallerand, J. E. Poulain, P. Gordien, S. Galland, B. Henriques, C. Dupin and M. Vincent. "Accuracy of FDG-PET/CT for response evaluation of muscle-invasive bladder cancer following neoadjuvant or induction chemotherapy." *Urologia Internationalis* 107 (2023): 1-7.
45. Razavi, S. A., G. Sadigh, A. M. Kelly and P. Cronin. "Comparative effectiveness of imaging modalities for the diagnosis of upper and lower urinary tract malignancy: A critically appraised topic." *Academic Radiology* 19 (2012): 1134-40.
46. Hadjiiski, L., H.-P. Chan, E. M. Caoili, R. H. Cohan, J. Wei and C. Zhou. "Auto-initialized cascaded level set (ai-cals) segmentation of bladder lesions on multidetector row CT urography." *Academic radiology* 20 (2013): 148-55.
47. Van Griethuysen, J. J., A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper and H. J. Aerts. "Computational radiomics system to decode the radiographic phenotype." *Cancer research* 77 (2017): e104-e07.
48. Cha, K. H., L. Hadjiiski, H.-P. Chan, A. Z. Weizer, A. Alva, R. H. Cohan, E. M. Caoili, C. Paramagul and R. K. Samala. "Bladder cancer treatment response assessment in CT using radiomics with deep-learning." *Scientific reports* 7 (2017): 1-12.
49. Sahiner, B., H. P. Chan, N. Petrick, M. A. Helvie and L. M. Hadjiiski. "Improvement of mammographic mass characterization using spiculation measures and morphological features." *Medical physics* 28 (2001): 1455-65.
50. Way, T. W., L. M. Hadjiiski, B. Sahiner, H. P. Chan, P. N. Cascade, E. A. Kazerooni, N. Bogot and C. Zhou. "Computer‐aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours." *Medical physics* 33 (2006): 2323-37.
51. Wu, E., L. M. Hadjiiski, R. K. Samala, H.-P. Chan, K. H. Cha, C. Richter, R. H. Cohan, E. M. Caoili, C. Paramagul and A. Alva. "Deep learning approach for assessment of bladder cancer treatment response." *Tomography* 5 (2019): 201-08.
52. Krizhevsky, A., I. Sutskever and G. E. Hinton. "ImageNet classification with deep convolutional neural networks." *Communications of the ACM* 60 (2017): 84-90.
53. Krizhevsky, A. and G. Hinton. "Learning multiple layers of features from tiny images." *Computer Science University of Toronto, Canada* (2009).
54. Goh, A. T. "Back-propagation neural networks for modeling complex systems." *Artificial intelligence in engineering* 9 (1995): 143-51.

55. McNitt-Gray, M., S. Napel, A. Jaggi, S. Mattonen, L. Hadjiiski, M. Muzi, D. Goldgof, Y. Balagurunathan, L. Pierce and P. Kinahan. "Standardization in quantitative imaging: A multicenter comparison of radiomic features from different software packages on digital reference objects and patient data sets." *Tomography* 6 (2020): 118-128.
56. Saldanha, A. J. "Java treeview—extensible visualization of microarray data." *Bioinformatics* 20 (2004): 3246-48.
57. American Cancer Society. 2023. *Cancer facts & figures 2023*. Atlanta, Ga: American Cancer Society.
58. Stein, J. P., G. Lieskovsky, R. Cote, S. Groshen, A.-C. Feng, S. Boyd, E. Skinner, B. Bochner, D. Thangathurai and M. Mikhail. "Radical cystectomy in the treatment of invasive bladder cancer: Long-term results in 1,054 patients." *Journal of clinical oncology* 19 (2001): 666-75.
59. Eisenberg, M. S., S. A. Boorjian, J. C. Cheville, R. H. Thompson, P. Thapa, D. Kaushik and I. Frank. "The sparc score: A multifactorial outcome prediction model for patients undergoing radical cystectomy for bladder cancer." *The Journal of urology* 190 (2013): 2005-10.
60. Gavriel, C. G., N. Dimitriou, N. Brieu, I. P. Nearchou, O. Arandjelović, G. Schmidt, D. J. Harrison and P. D. Caie. "Assessment of immunological features in muscle-invasive bladder cancer prognosis using ensemble learning." *Cancers* 13 (2021): 1624.
61. Zheng, Q., R. Yang, X. Ni, S. Yang, L. Xiong, D. Yan, L. Xia, J. Yuan, J. Wang and P. Jiao. "Accurate diagnosis and survival prediction of bladder cancer using deep learning on histological slides." *Cancers* 14 (2022): 5807.
62. Seiler, R., H. A. D. Ashab, N. Erho, B. W. van Rhijn, B. Winters, J. Douglas, K. E. Van Kessel, E. E. F. van de Putte, M. Sommerlad and N. Q. Wang. "Impact of molecular subtypes in muscle-invasive bladder cancer on predicting response and survival after neoadjuvant chemotherapy." *European urology* 72 (2017): 544-54.
63. Bhambhvani, H. P., A. Zamora, E. Shkolyar, K. Prado, D. R. Greenberg, A. M. Kasman, J. Liao, S. Shah, S. Srinivas and E. C. Skinner. "Development of robust artificial neural networks for prediction of 5-year survival in bladder cancer." *Urologic Oncology: Seminars and Original Investigations* 39 (2021): 193-e7.
64. Gavriel, C. G., N. Dimitriou, N. Brieu, I. P. Nearchou, O. Arandjelović, G. Schmidt, D. J. Harrison and P. D. Caie. "Identification of immunological features enables survival prediction of muscle-invasive bladder cancer patients using machine learning." *BioRxiv* (2020): 2020.02.
65. Riester, M., J. M. Taylor, A. Feifer, T. Koppie, J. E. Rosenberg, R. J. Downey, B. H. Bochner and F. Michor. "Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer." *Clinical cancer research* 18 (2012): 1323-33.
66. Shariat, S. F., C. Bolenz, G. Godoy, Y. Fradet, R. Ashfaq, P. I. Karakiewicz, H. Isbarn, C. Jeldres, J. Rigaud and A. I. Sagalowsky. "Predictive value of combined immunohistochemical markers in patients with pt1 urothelial carcinoma at radical cystectomy." *The Journal of urology* 182 (2009): 78-84.
67. Shariat, S. F., P. I. Karakiewicz, G. S. Palapattu, G. E. Amiel, Y. Lotan, C. G. Rogers, A. Vazina, P. J. Bastian, A. Gupta and A. I. Sagalowsky. "Nomograms provide improved accuracy for predicting survival after radical cystectomy." *Clinical Cancer Research* 12 (2006): 6663-76.

68. Edge, S. B. and C. C. Compton. "The american joint committee on cancer: The 7th edition of the ajcc cancer staging manual and the future of tnm." *Annals of surgical oncology* 17 (2010): 1471-74.
69. Chan, H. P., B. Sahiner, R. F. Wagner and N. Petrick. "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers." *Medical physics* 26 (1999): 2654-68.
70. Sahiner, B., H. P. Chan, N. Petrick, R. F. Wagner and L. Hadjiiski. "Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size." *Medical physics* 27 (2000): 1509-22.
71. Way, T. W., B. Sahiner, L. M. Hadjiiski and H. P. Chan. "Effect of finite sample size on feature selection and classification: A simulation study." *Medical physics* 37 (2010): 907-20.
72. Battiti, R. "Using mutual information for selecting features in supervised neural net learning." *IEEE Transactions on neural networks* 5 (1994): 537-50.
73. Chen, S.-Y., Z. Feng and X. Yi. "A general introduction to adjustment for multiple comparisons." *Journal of thoracic disease* 9 (2017): 1725.
74. Altman, D. G. 1990. *Practical statistics for medical research*. Boca Raton, FL: Chapman and Hall/CRC.
75. Samala, R. K., H.-P. Chan, L. Hadjiiski, M. A. Helvie, C. D. Richter and K. H. Cha. "Breast cancer diagnosis in digital breast tomosynthesis: Effects of training sample size on multi-stage transfer learning using deep neural nets." *IEEE transactions on medical imaging* 38 (2018): 686-96.
76. Garg, R. K., V. L. Urs, A. A. Agarwal, S. K. Chaudhary, V. Paliwal and S. K. Kar. "Exploring the role of chatgpt in patient care (diagnosis and treatment) and medical research: A systematic review." *Health Promotion Perspectives* 13 (2023): 183.
77. Levine, D. M., R. Tuwani, B. Kompa, A. Varma, S. G. Finlayson, A. Mehrotra and A. Beam. "The diagnostic and triage accuracy of the GPT-3 artificial intelligence model." *medRxiv* (2023).
78. Eriksen, A. V., S. Möller and J. Ryg. "Use of GPT-4 to diagnose complex clinical cases." *NEJM-AI* 1 (2023): AIp2300031.
79. Conover, M., M. Hayes, A. Mathur, X. Meng, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell and M. Zaharia. *Free dolly: Introducing the world's first truly open instruction-tuned llm*. 2023,
80. Chiang, W.-L., Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang and J. E. Gonzalez. "Vicuna: An open-source chatbot impressing GPT-4 with 90%* chatgpt quality." Available online: https://vicuna (accessed 14 April 2023).
81. Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava and S. Bhosale. "Llama 2: Open foundation and fine-tuned chat models." *arXiv preprint* (2023): 2307.09288.
82. Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry and A. Askell. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-901.
83. OpenAI. "GPT-4 technical report." *arXiv preprint* (2023): *2303.08774*.
84. Bennasar, M., Y. Hicks and R. Setchi. "Feature selection using joint mutual information maximisation." *Expert Systems with Applications* 42 (2015): 8520-32.

85. Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich. "Going deeper with convolutions." *IEEE conference on computer vision and pattern recognition* (2015): 1-9.
86. Metz, C. E., B. A. Herman and J. H. Shen. "Maximum likelihood estimation of receiver operating characteristic (roc) curves from continuously-distributed data." *Statistics in medicine* 17 (1998): 1033-53.
87. Gallas, B. D. "Imrmc v4. 0: Application for analyzing and sizing mrmc reader studies." *Division of Imaging, Diagnostics, and Software Reliability, OSEL/CDRH/FDA, Silver Spring, Maryland* (2017).
88. Bland, J. M. and D. G. Altman. "Multiple significance tests: The bonferroni method." *Bmj* 310 (1995): 170.
89. Kluth, L. A., P. C. Black, B. H. Bochner, J. Catto, S. P. Lerner, A. Stenzl, R. Sylvester, A. J. Vickers, E. Xylinas and S. F. Shariat. "Prognostic and prediction tools in bladder cancer: A comprehensive review of the literature." *European urology* 68 (2015): 238-53.
90. Borhani, S., R. Borhani and A. Kajdacsy-Balla. "Artificial intelligence: A promising frontier in bladder cancer diagnosis and outcome prediction." *Critical reviews in oncology/hematology* 171 (2022): 103601.
91. Gu, B., R. J. Desai, K. J. Lin and J. Yang. "Probabilistic medical predictions of large language models." *npj Digital Medicine* 7 (2024): 367.
92. Abbasi Yadkori, Y., I. Kuzborskij, A. György and C. Szepesvari. "To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty." *Advances in Neural Information Processing Systems* 37 (2025): 58077-117.
93. White, J., Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith and D. C. Schmidt. "A prompt pattern catalog to enhance prompt engineering with chatgpt." *arXiv preprint* (2023): 2302.11382.
94. Wang, L., X. Chen, X. Deng, H. Wen, M. You, W. Liu, Q. Li and J. Li. "Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs." *NPJ digital medicine* 7 (2024): 41.
95. Conover, M., M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia and R. Xin. 2023. *Free dolly: Introducing the world's first truly open instruction-tuned LLM*.
96. Face, H. 2024. Available online: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct (accessed on 19 December 2024).
97. AI, O. 2023. *New models and developer products announced at devday*. Available online: https://openai.com/blog/new-models-and-developer-products-announced-at-devday (accessed on 15 October 2024).
98. AI, O. 2024. *Hello GPT-4o*. Available online: https://openai.com/index/hello-gpt-4o/ (accessed on 10 October 2024).
99. Krippendorff, K. "Measuring the reliability of qualitative text analysis data." *Quality and quantity* 38 (2004): 787-800.
100. Cha, K. H., L. M. Hadjiiski, R. H. Cohan, H.-P. Chan, E. M. Caoili, M. S. Davenport, R. K. Samala, A. Z. Weizer, A. Alva and G. Kirova-Nedyalkova. "Diagnostic accuracy of CT for prediction of bladder cancer treatment response with and without computerized decision support." *Academic radiology* 26 (2019): 1137-45.
101. Hadjiiski, L. M., K. H. Cha, R. H. Cohan, H.-P. Chan, E. M. Caoili, M. S. Davenport, R. K. Samala, A. Z. Weizer, A. Alva and G. Kirova-Nedyalkova. "Intraobserver variability

in bladder cancer treatment response assessment with and without computerized decision support." *Tomography* 6 (2020): 194-202.
102. Hadjiiski, L. M., M. Joshi, A. Alva, H.-P. Chan, R. H. Cohan, E. M. Caoili, G. Kirova-Nedyalkova, M. S. Davenport, P. R. Shankar and I. R. Francis. "Multi-institutional observer performance study for bladder cancer treatment response assessment in CT urography with and without computerized decision support." *SPIE Medical Imaging 2021: Computer-Aided Diagnosis* 11597 (2021): 369-375.
103. Way, T. W., B. Sahiner, H. P. Chan, L. Hadjiiski, P. N. Cascade, A. Chughtai, N. Bogot and E. Kazerooni. "Computer-aided diagnosis of pulmonary nodules on CT scans: Improvement of classification performance with nodule surface features." *Medical physics* 36 (2009): 3086-98.
104. Eisenhauer, E. A., P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther and M. Mooney. "New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1)." *European journal of cancer* 45 (2009): 228-47.
105. Berbaum, K. S., C. E. Metz, L. L. Pesce and K. M. Schartz. *DBM MRMC 2.2*. Available online: http://perception.radiology.uiowa.edu (accessed on 28 September 2021).
106. Gallas, B. D., A. Bandos, F. W. Samuelson and R. F. Wagner. "A framework for random-effects roc analysis: Biases with the bootstrap and other variance estimators." *Communications in Statistics—Theory and Methods* 38 (2009): 2586-603.
107. Petrick, N., B. Sahiner, S. G. Armato III, A. Bert, L. Correale, S. Delsanto, M. T. Freedman, D. Fryd, D. Gur and L. Hadjiiski. "Evaluation of computer-aided detection and diagnosis systems[a)]." *Medical physics* 40 (2013): 087001.
108. Clever, J., L. Hadjiiski, H.-P. Chan, R. H. Cohan, E. M. Caoili, K. Cha, R. Samala and C. Zhou. "Bladder cancer segmentation using U-Net-based deep-learning." *SPIE Medical Imaging 2023: Computer-Aided Diagnosis* 12465 (2023): 732-37.
109. Kushwaha, A., R. F. Mourad, K. Heist, H. Tariq, H.-P. Chan, B. D. Ross, T. L. Chenevert, D. Malyarenko and L. M. Hadjiiski. "Improved repeatability of mouse tibia volume segmentation in murine myelofibrosis model using deep learning." *Tomography* 9 (2023): 589-602.
110. Kushwaha, A., R. F. Mourad, K. Heist, D. Malyarenko, H.-P. Chan, T. L. Chenevert and L. M. Hadjiiski. "Segmentation of mouse tibia on mri using deep learning U-Net models." *SPIE Medical Imaging 2023: Computer-Aided Diagnosis* 12465 (2023): 638-43.