medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

# Intricacies of Human-AI Interaction in Dynamic Decision-Making for Precision Oncology: A Case Study in Response-Adaptive Radiotherapy

Dipesh Niraula*[1*], Kyle C Cuneo[2], Ivo D Dinov[3], Brian D Gonzalez[4], Jamalina B Jamaluddin[5], Jionghua (Judy) Jin[6], Yi Luo[1], Martha M Matuszak[2], Randall K Ten Haken[2], Alex K Bryant[2], Thomas J Dilling[7], Michael P Dykstra[2], Jessica M Frakes[7], Casey L Liveringhouse[7], Sean R Miller[2], Matthew N Mills[7], Russell F Palm[7], Samuel N Regan[2], Anupam Rishi[7], Javier F Torres-Roca[7], Hsiang-Hsuan Michael Yu[7], and Issam El Naqa[1]

[1]Department of Machine Learning, Moffitt Cancer Center, Tampa, FL, USA
[2]Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA
[3]Department of Health Behavior and Biological Sciences, University of Michigan, Ann Arbor, MI, USA
[4]Department of Health Outcomes and Behavior, Moffitt Cancer Center, Tampa, FL, USA
[5]Department of Nuclear Engineering and Radiological Sciences, University of Michigan, Ann Arbor, MI, USA
[6]Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, USA
[7]Department of Radiation Oncology, Moffitt Cancer Center, Tampa, FL, USA

## Abstract

Background: Adaptive treatment strategies that can dynamically react to individual cancer progression can provide effective personalized care. Longitudinal multi-omics information, paired with an artificially intelligent clinical decision support system (AI-CDSS) can assist clinicians in determining optimal therapeutic options and treatment adaptations. However, AI-CDSS is not perfectly accurate, as such, clinicians' over/under reliance on AI may lead to unintended consequences, ultimately failing to develop optimal strategies. To investigate such collaborative decision-making process, we conducted a Human-AI interaction case study on response-adaptive radiotherapy (RT).

Methods: We designed and conducted a two-phase study for two disease sites and two treatment modalities—adaptive RT for non-small cell lung cancer (NSCLC) and adaptive stereotactic body RT for hepatocellular carcinoma (HCC)—in which clinicians were asked to consider mid-treatment modification of the dose per fraction for a number of retrospective cancer patients without AI-support (Unassisted Phase) and with AI-assistance (AI-assisted Phase). The AI-CDSS graphically presented trade-offs in tumor control and the likelihood of toxicity to organs at risk, provided an optimal recommendation, and associated model uncertainties. In addition, we asked for clinicians' decision confidence level and trust level in individual AI recommendations and encouraged them to provide written remarks. We enrolled 13 evaluators (radiation oncology physicians and residents) from two medical institutions located in two different states, out of which, 4 evaluators volunteered in both NSCLC and HCC studies, resulting in a total of 17 completed evaluations (9 NSCLC, and 8 HCC). To limit the evaluation time to under an hour, we selected 8 treated patients for NSCLC and 9 for HCC, resulting in a total of 144 sets of evaluations (72 from NSCLC and 72 from HCC). Evaluation for each patient consisted of 8 required inputs and 2 optional remarks, resulting in up to a total of 1440 data points.

Results: AI-assistance did not homogeneously influence all experts and clinical decisions. From NSCLC cohort, 41 (57%) decisions and from HCC cohort, 34 (47%) decisions were adjusted after AI assistance. Two evaluations (12%) from the NSCLC cohort had zero decision adjustments, while the remaining 15 (88%) evaluations resulted in at least two decision adjustments. Decision adjustment level positively correlated with dissimilarity in decision-making with AI [NSCLC: $\rho = 0.53\,(p < 0.001)$; HCC: $\rho = 0.60\,(p < 0.001)$] indicating that evaluators adjusted their decision closer towards AI recommendation. Agreement with AI-recommendation positively correlated with AI Trust Level [NSCLC: $\rho = 0.59\,(p < 0.001)$; HCC: $\rho = 0.7\,(p < 0.001)$] indicating that evaluators followed AI's recommendation if they agreed with that recommendation. The correlation between decision confidence changes and decision adjustment level showed

*Corresponding Author: Dipesh.Niraula@moffitt.org

**NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.**

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

an opposite trend [NSCLC: $\rho = -0.24\,(p = 0.045)$, HCC: $\rho = 0.28\,(p = 0.017)$] reflecting the difference in behavior due to underlying differences in disease type and treatment modality. Decision confidence positively correlated with the closeness of decisions to the standard of care (NSCLC: 2 Gy/fx; HCC: 10 Gy/fx) indicating that evaluators were generally more confident in prescribing dose fractionations more similar to those used in standard clinical practice. Inter-evaluator agreement increased with AI-assistance indicating that AI-assistance can decrease inter-physician variability. The majority of decisions were adjusted to achieve higher tumor control in NSCLC and lower normal tissue complications in HCC. Analysis of evaluators' remarks indicated concerns for organs at risk and RT outcome estimates as important decision-making factors.

Conclusions: Human-AI interaction depends on the complex interrelationship between expert's prior knowledge and preferences, patient's state, disease site, treatment modality, model transparency, and AI's learned behavior and biases. The collaborative decision-making process can be summarized as follows: (i) some clinicians may not believe in an AI system, completely disregarding its recommendation, (ii) some clinicians may believe in the AI system but will critically analyze its recommendations on a case-by-case basis; (iii) when a clinician finds that the AI recommendation indicates the possibility for better outcomes they will adjust their decisions accordingly; and (iv) When a clinician finds that the AI recommendation indicate a worse possible outcome they will disregard it and seek their own alternative approach.

Keywords. Human-AI Interaction, Collaborative Decision-Making, Adaptive Treatment Strategy, Dynamic Decision-Making, ARCliDS, Deep Reinforcement Learning, Graph Neural Network, Response-Adaptive Radiotherapy, Precision Oncology, NSCLC, HCC.

# 1 Introduction

Development of novel cancer therapies[1,2] and increasing availability of longitudinal multi-omics data[3,4] have improved our ability to prescribe personalized, adaptive treatment strategies[5–7] capable of dynamically reacting to individual cancer progression while improving efficacy and minimizing side effects. However, the dynamic nature of adaptive strategies compounded by a wide range of clinical options, high data dimensionality, uncertainty in assessing treatment response, and the uncertainty in the future course of disease, present challenges in tailoring optimal strategies.[7] Assistance from artificial intelligence (AI) decision-support tools designed for precision oncology[8–12] that can provide individual treatment response assessments, outcome predictions, and optimal treatment recommendations can overcome such challenges. However, AI tools are not perfectly accurate, have inherent biases, and are limited by the quality of their training data;[13,14] as such, over/under reliance on AI[15] can result in sub-optimal therapeutics. Therefore, investigating collaborative Human-AI decision-making behavior[16,17] is crucial before treating advanced diseases with a relatively narrow therapeutic window and tighter margin of error, such as cancer.[9,16–20]

We conducted a Human-AI interaction study to investigate clinicians' (physicians and residents) collaborative decision-making behavior in knowledge-based response-adaptive radiotherapy (KBR-ART).[21–24] KBR-ART is a single-modality and single-intervention adaptive treatment strategy consisting of three phases: pre-treatment assessment, response evaluation, and adaptation. In the response evaluation phase, patients' treatment response is assessed by comparing pre and during-treatment multi-omics information and in the adaptation phase, a treatment plan is adapted (dose escalation/de-escalation). In this study, the clinicians collaborated with ARCliDS[22,25,26]—a software for dynamic decision-making developed using a model-based deep reinforcement learning algorithm. For application in KBR-ART, ARCliDS uses a graphical neural network-based model of radiotherapy environment which defines a patient's state via a graph of multi-omics features and is capable of assessing treatment response and predicting treatment outcomes. Prior to this study, we had developed ARCliDS modules for two modalities that were trained on two retrospective cohorts: non-small cell lung cancer (NSCLC) patients who had received adaptive radiotherapy (RT)[27] and hepatocellular carcinoma (HCC) patients who had received adaptive stereotactic body radiotherapy (SBRT).[28] In this study, we designed a two-phase Human-AI interaction study for each of the two modules, in which clinicians were asked to prescribe mid-treatment dose adaptation without (Unassisted phase) and with AI-assistance (AI-assisted phase) for a number of retrospective patients. In addition, they were asked to input their decision confidence level, and trust level on AI recommendation, and were encouraged to provide text remarks.

We designed web modules that closely simulate KBR-ART's decision-making process, as summarized in **Figure 1**. In the Unassisted phase, we presented the RT treatment plan (dose volume histogram, and three-dimensional dose distribution)

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

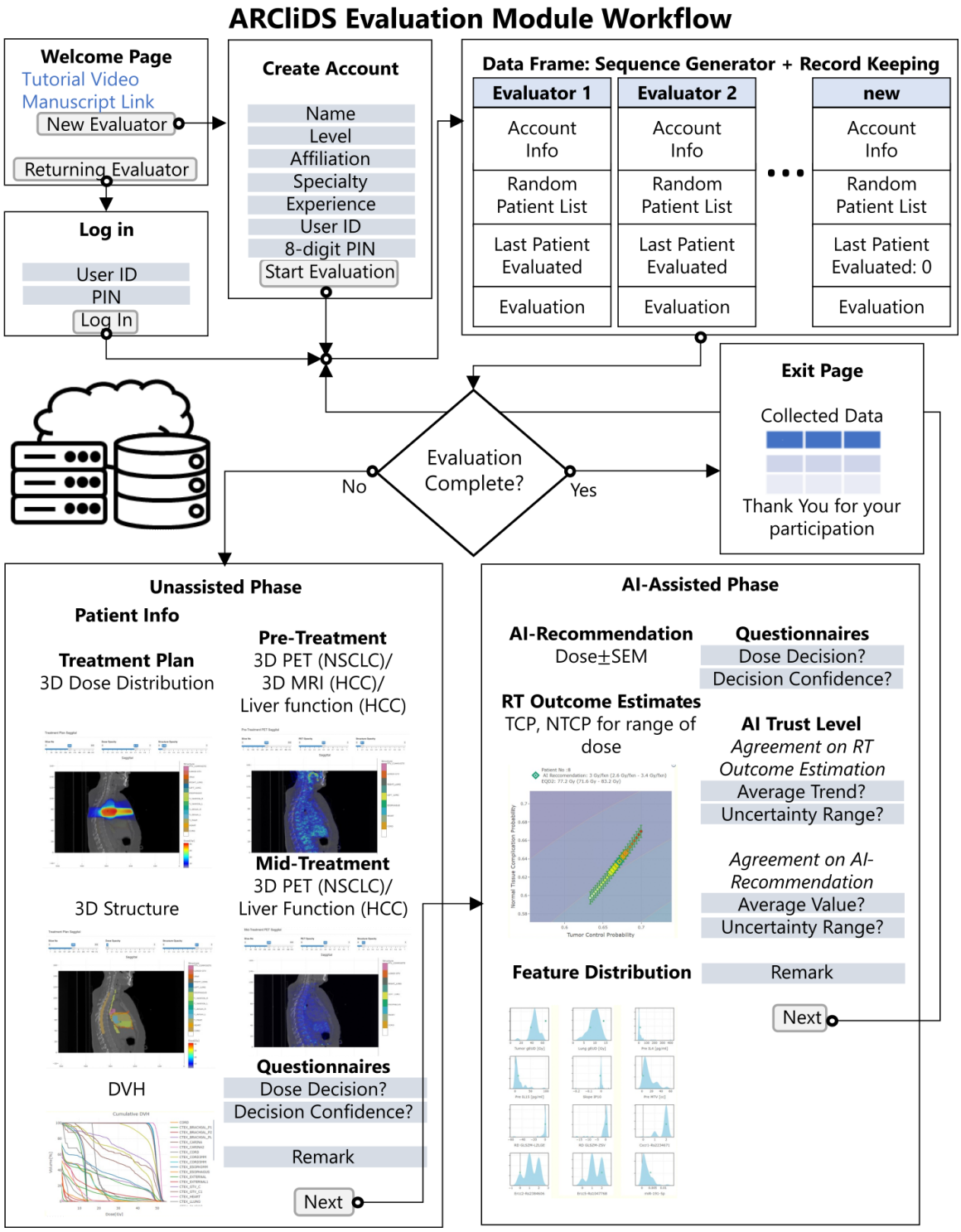



**Figure 1: ARCliDS Evaluation Module workflow**. The modules were deployed on cloud in shinyapps.io server and used google sheets as data storage. **Welcome Page** contains links to tutorial video hosted in YouTube and manuscript; from which a new evaluator can create a new account and returning user can login back to complete their evaluation. **Create Account** page consists of a series of input prompts including a unique username and 8-digit PIN so that they could log back in if disconnected unexpectedly or if needed to step away. **Data Frame** hosted as a google sheet automatically saves login info and evaluation input. Additionally, it saves randomized list of patients, and the last patient evaluated for each user account to check if evaluation is completed. **Unassisted Phase** page presents patient's relevant info, treatment plans including 3D dose distribution and structure, cumulative DVH, pre and mid treatment 3D PET scans for NSCLC, pre and mid treatment liver functions along with pre-treatment 3D MRI scans for HCC; and contains input prompts for dose decision, decision confidence, and a textbox for remark. **AI-Assisted Phase** page presents AI-recommendation, outcome estimation for a range of dose show in TCP vs NTCP outcome space, and Distribution plots for all the feature variables; and contains input prompts for dose decision and decision confidence, a series of multiple-choice questions to access the user trust level, and a textbox for remark. **Exit Page** marks the end of the evaluation.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

and CT/PET/MRI imaging from KBR-ART's response evaluation phase, and asked clinicians to assess mid-treatment response and input their dose decision for the remaining of treatment period along with their confidence level (0-5, 5 being the highest level). In the AI-assisted phase, we provided ARCliDS recommendation based on its assessment of treatment response and asked the evaluators to re-enter their decision and decision confidence level. Besides AI recommendation we included additional graphs to improve model transparency and explainability for establishing and maximizing trust on AI.[9,16,17,20,29] We included a two-dimensional outcome space for quantifying tradeoffs between tumor control probability (TCP) and normal tissue complication probability (NTCP)[30] across a range of RT dose fractionation options; graphically demonstrated the model uncertainty for both AI recommendation and outcome prediction; and included feature distribution plots with feature value marked in the foreground for presenting "whereabouts" of the individual patient with respect to the rest of the population. To assess evaluators' level of trust in the individual AI recommendation (0-5, 5 being the highest level; AI Trust Level), we included four multiple-choice questions. Lastly, we included text boxes for remarks to gain insights into the collaborative decision-making process. A detailed description of the evaluation modules is presented in **supplementary sections S3**, **S4**, and **S5**.

The main objectives of this study were twofold: broader model-agnostic investigation of a collaborative decision-making process, and model-specific application-grounded evaluation[31] of ARCliDS by domain expert and end-users. We included diverse study elements for attaining both levels of our objectives. We incorporated AIs for two diseases and two treatment modalities and enrolled evaluators from two medical institutions located in different states, with different experience levels, sub-specialties, and professional backgrounds. Considering the fact that evaluators' mental model of AI can affect human-AI interaction,[32] we took the following steps to homogenize evaluators' first impression of AI: (i) created two 10-minute tutorial videos;[33,34] (ii) conducted a pre-evaluation information session with the evaluators, in which, we played the training video and demonstrated sample evaluation followed by a question-and-answer round; and (iii) included a web-link to the tutorial video and original ARCliDS manuscript[35] in the evaluation modules. Furthermore, to optimize evaluator's time utilization, we designed the evaluation to take no more than an hour, added a user account system so the evaluation could be completed in multiple sessions if needed, and enabled automatic saving of all user inputs in the cloud. More details are presented in **supplementary sections S1**, **S2** and **S6**.

Treatment options (dose/fraction) in KBR-ART are continuous variables (NSCLC: 1.5-4 Gy/fx; HCC: 1-15 Gy/fx) and fundamentally differ from categorical decisions. In addition, retrospective clinical endpoints (local control and toxicity) are available for only one decision point and thus lack ground truth for validation in general. As such, our study is unique in comparison to other Human-AI interaction studies such as collaborative decision-making with image-based visual diagnostic AI in skin cancer,[36] with diagnostic AI for lesion detection and categorization from colonoscopy,[15] with diagnostic AI for malignant nodules detection from chest radiographs,[37] and with image-based AI for chemotherapy response assessment in bladder cancer from CT urography;[38] in which either decision option was categorical (e.g. multi-class image classification), or ground truth was available (e.g. labeled image or retrospective post-treatment tumor stage) or both. Moreover, whereas categorical decision with the availability of ground truth simplifies evaluation, continuous decisions provide opportunities for conducting high-resolution correlation analysis. Furthermore, dose options have a natural ordering property with respect to the radiobiological principle, i.e., higher dose corresponds to higher TCP as well as higher NTCP compared to lower doses. Thus, in this study along with analyzing the observed variables—decisions, decision confidences, and AI trust level—we derived a number of quantities to investigate human-AI interaction and collaborative decision-making behavior. In particular, we investigated decision adjustment frequency, decision adjustment level, dissimilarity in decision-making with AI, agreement with AI recommendation, and closeness of decision with standard of care.

# 2 Results

## 2.1 AI-assistance did not homogeneously influence all experts and all clinical decisions.

Our study accumulated 72 evaluations for NSCLC cohort (9 evaluators $\times$ 8 patients) and 72 evaluations for HCC cohort (8 evaluators $\times$ 9 patients). First, we analyzed AI-influence on a distribution level by performing a matched paired randomization t-test[39,40] between unassisted decision ($un$) and AI-assisted decision ($aia$). As shown in **Figures 2A**, **2B**, **2E**, and **2F**, we failed to reject the null hypothesis in most of the tests, grouped by both evaluators and patients, indicating absence of significant AI influence. However, on an individual level, we observed that AI assistance resulted in the adjustment of about half of the decisions [41 (57%) for NSCLC, and 34 (47%) for HCC], which is a considerable

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

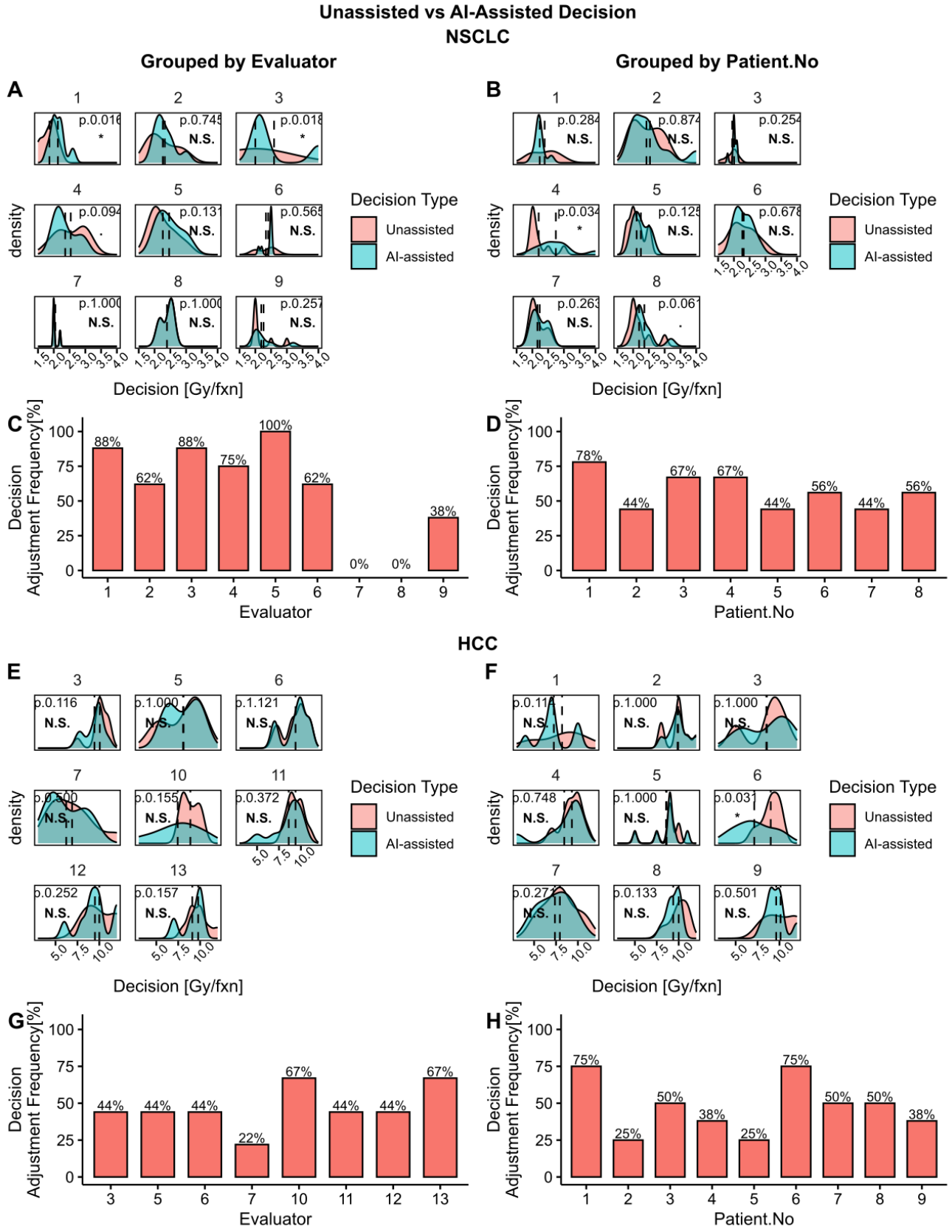



**Figure 2: Unassisted vs AI-assisted Decision analysis grouped by Evaluators and Patient Number**. Plots **A**, **B**, **C**, and **D** summarizes data from NSCLC and plots **E**, **F**, **G**, and **H** from HCC. Density plots **A**, **B**, **E**, and **F** compare the unassisted ($un$) and AI-assisted ($aia$) distribution and includes the p-values from a matched pair randomization t-test with $un - aia = 0$ as the null hypothesis and standard significance code: ******* ($p < 0.001$), ****** ($p < 0.01$), ***** ($p < 0.05$), **.** ($p < 0.1$), **N.S** ($p \geq 0.1$). Bar plots **C**, **D**, **G**, and **H** shows the frequency of decision adjustment ($un \neq aia$) after AI-assistance in percentages.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

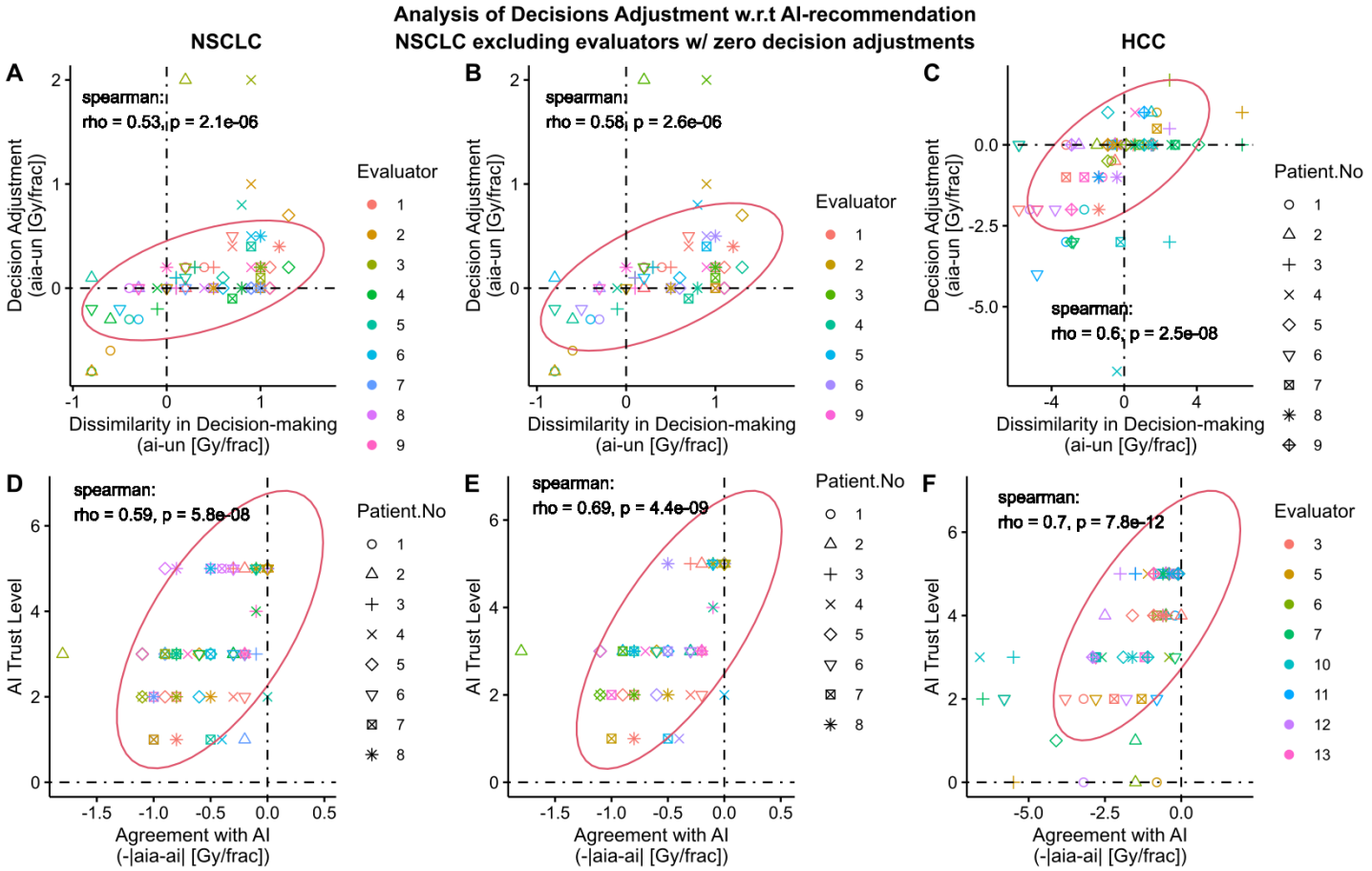



**Figure 3: Analysis of Decision adjustment with respect to AI recommendation.** The first, second, and third column of figures correspond to NSCLC, NSCLC excluding evaluators with zero decision adjustment, and HCC, respectively, and all the plots are colored by evaluators and marked coded patient number. All evaluators in HCC adjusted at least one of their decisions. Scatter plots **A**, **B**, and **C** shows relationship between the level of decision adjustment ($aia - un$) and dissimilarity with AI recommendation ($ai - un$). Scatter plots **D**, **E**, and **F** show relationship between AI Trust Level (0-5, 5 being the highest level) and agreement with AI-recommendation ($-|aia - ai|$). The $aia - ai = 0$ line corresponds to absolute agreement with the AI-recommendation. All plots include the Spearman correlation coefficients, p-values, and co-variance ellipse (95% confidence). Covariance ellipses are included for a visual insight about the data distribution.

proportion. Moreover, we found only 2 out of 17 evaluations contained zero decision adjustments i.e.,$un = aia$ for all cases, while the remaining 15 evaluations contained at least 2 decision adjustments [88% (15/17)]. In particular, NSCLC evaluators #7 and #8 made zero decision adjustments, however, all HCC evaluators made at least two decision adjustments. As shown in **Figures 2C** and **2D**, number of NSCLC decision adjustments ranged from 0% to 100% among evaluators and 44% to 78% among patients and as shown in **Figures 2G**, and **2H**, number of HCC decision adjustments ranged from 22% to 67% among evaluators and 25% to 75% among patients. Decision adjustment level (Gy/fx) is summarized in **supplementary Figure S5**. In addition, to investigate the relationship between unassisted and AI-assisted decisions, we performed a correlation analysis on clinical decisions, as presented in **supplementary section S9**. We found that a statistically significant positive correlation existed between $un$ and $aia$ but the correlation coefficient was not strong enough (i.e. near 1) to dismiss AI influence, further corroborating the observation that AI influence exists but not all experts and not all clinical decisions are influenced homogeneously.

## 2.2 Decision adjustment level positively correlated with dissimilarity in decision-making with AI.

Under a positive AI influence, we expect evaluators to adjust their decision closer towards AI recommendation ($ai$). In such a case, we expect higher level of decision adjustment ($aia - un$) for higher level of dissimilarity between AI recommendation and unassisted decision ($ai - un$) and vice versa. Conversely, we would expect no decision adjustment, $aia = un$, when, $ai = un$. As expected, we found an overall positive correlation between decision adjustment level and dissimilarity in decision-making with AI as shown in **Figures 3A**, **3B**, and **3C**. Specifically, for NSCLC, $\rho = 0.53\,(p <$

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

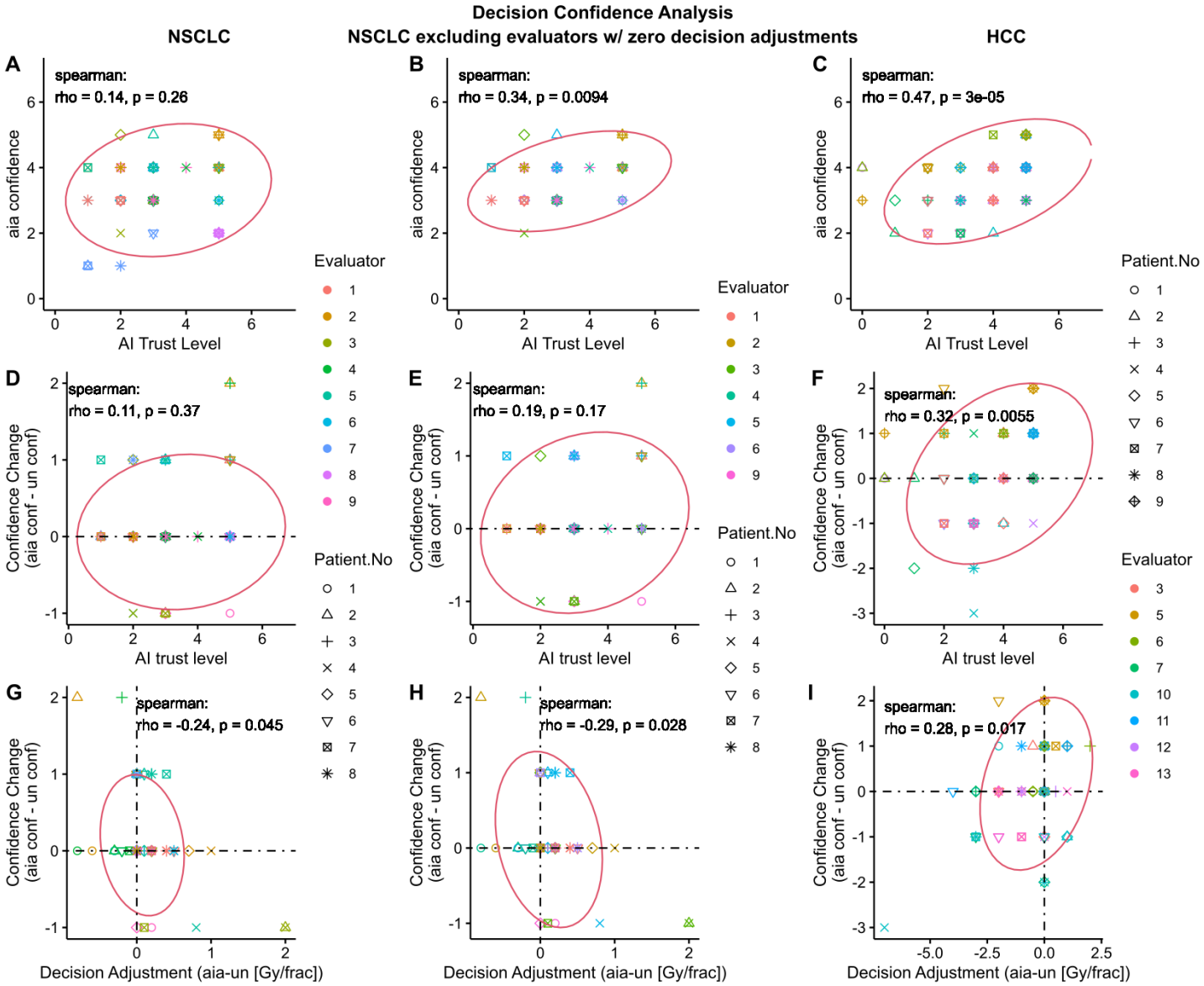



**Figure 4: Analysis of evaluator's decision confidence.** The first, second, and third column of figures correspond to NSCLC, NSCLC excluding Evaluators with zero decision adjustment, and HCC, respectively. All evaluators in HCC adjusted at least one of their decisions. 2D Scatter plots **A**, **B**, and **C** show the relationship between evaluators' AI-assisted decision ($aia$) confidence (0-5, 5 being the highest level) and AI Trust level (0-5, 5 being the highest level). 2D Scatter plots **D**, **E**, and **F** show the relationship between evaluators' change in confidence level ($aia\,conf - un\,conf$) and AI trust level. 2D Scatter plots **G**, **H** and **I** show the relationship between evaluators' change in confidence level and level of decision adjustment with AI-assistance. All plots include the Spearman correlation coefficients, p-values, and co-variance ellipse (95% confidence). Covariance ellipses are included for a visual insight about the data distribution.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

$0.001)$ , for NSCLC when excluding evaluators with zero decision adjustment, $\rho = 0.58\,(p < 0.001)$, and for HCC, $\rho = 0.60\,(p < 0.001)$. The positive spearman correlation coefficient indicates an increasing monotonic relationship between $ai - un$ and $aia - un$. Here we separately investigated the NSCLC cohort excluding zero decision adjustment to examine the contribution of strictly positive AI influence. Additionally, we observed that the majority of the data points reside in the first and third quadrant, which represents decision adjustments in alignment with AI's advice, whereas very few data points reside in the second and fourth quadrants which represents decision adjustment against AI's advice. These results further corroborate a positive AI influence on clinicians' decisions.

## 2.3 Agreement with AI-recommendation positively correlated with AI Trust Level.

Under a positive AI influence, we expect evaluators to follow AI's recommendation if they agree with that recommendation. In such a scenario, the level of agreement with AI-recommendation should correlate with their level of trust on that particular recommendation. To investigate such a relationship, we defined agreement as the additive inverse of the absolute difference between AI-assisted Decision and AI recommendation, i.e. $-|aia - ai|$. Based on this definition, the level of agreement peaks at 0 when $aia = ai$, and decreases when the difference between $aia$ and $ai$ increases in either direction. As expected, we found a positive correlation between evaluators' self-reported AI trust level and their agreement with AI, as shown in **Figures 3D**, **3E**, and **3F**. For NSCLC, $\rho = 0.59\,(p < 0.001)$, for NSCLC when excluding evaluators with zero decision adjustment, $\rho = 0.69\,(p < 0.001)$, and for HCC, $\rho = 0.7\,(p < 0.001)$. The positive Spearman correlation coefficient indicates an increasing monotonic relationship between AI trust level and agreement with AI. Note that the self-reported AI trust level comprises of 4 components as described in **supplementary section S3.4.4** and as shown in **supplementary Figure S7**.

## 2.4 Decision confidence analysis showed a mixed trend between NSCLC and HCC

We investigated various relationships for decision confidence levels. First, we investigated the relationship between evaluators' self-reported AI-assisted decision confidence and AI trust level as shown in **Figures 4A**, **4B**, and **4C**. Under positive AI influence, we would expect a positive correlation between aia conf and AI trust level. However, we only found positive correlation for two cases: for NSCLC excluding evaluators with zero decision adjustment, $\rho = 0.34\,(p = 0.0094)$, and for HCC, $\rho = 0.47\,(p < 0.001)$. For NSCLC, we found a much weaker correlation with no significance: $\rho = 0.14\,(p = 0.26)$, which must be due to the two evaluators with zero decision adjustment. Then, to examine if higher AI trust level corresponds to increase in decision confidence, we investigated the relationship between the change in confidence level $(aia\,conf - un\,conf)$ and AI-trust level as shown in **Figures 4D**, **4E**, and **4F**. We only found a statistically significant positive correlation for HCC i.e. $\rho = 0.32\,(p = 0.0055)$. For NSCLC, $\rho = 0.11\,(p = 0.37)$, and for NSCLC excluding evaluators with zero decision adjustment, $\rho = 0.19\,(p = 0.17)$. In addition, we investigated the relationship between change in confidence level and decision adjustment level, and interestingly, found opposite trends between NSCLC and HCC as shown in **Figures 4G**, **4H**, and **4I**. For NSCLC, $\rho = -0.24\,(p = 0.045)$, for NSCLC excluding evaluators with zero decision adjustment, $\rho = -0.29\,(p = 0.028)$, whereas, for HCC, $\rho = 0.28\,(p = 0.017)$. This indicates that, overall, the evaluators were more confident in reducing dose fractionation amount $(aia < un)$ for NSCLC patients and increasing dose fractionation amount $(aia > un)$ for HCC patients. Such behavior must have originated from the difference between the two diseases and treatment modalities as seen from **section 2.7**.

## 2.5 Decision confidence positively correlated with the closeness of decisions to the Standard of Care

We expect a higher confidence level for decisions close to the Standard of Care (SOC) Dose (NSCLC: 2 Gy/fx; HCC: 10 Gy/fx, fixed throughout the treatment period). As such, we investigated the relationship between decision confidence and the closeness of their decision to SOC. For this purpose, first we defined closeness as the additive inverse of the absolute difference between decision and SOC, i.e. $-|d - SOC|$, where $d$ is $un$ or $aia$. Based on this definition, closeness peaks at 0 when $d = SOC$ and decreases when the difference between decision and SOC increases in either direction. We found a positive correlation between $un\,conf$ and closeness to SOC as shown in **Figures 5A**, **5B**, and **5C**. For NSCLC, $\rho = 0.15\,(p = 0.21)$, for NSCLC excluding evaluators with zero AI-influence, $\rho = 0.25(p = 0.059)$, and for HCC, $\rho = 0.32(p = 0.0056)$. Similarly, we found a positive correlation between $aia\,conf$ and closeness to SOC as shown in **Figures 5D**, **5E**, and **5F**. For NSCLC, $\rho = 0.092(p = 0.44)$, for NSCLC excluding evaluators with zero decision adjustment, $\rho = 0.29\,(p = 0.029)$, and for HCC, $\rho = 0.48(p < 0.001)$. For NSCLC the p-value was not

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

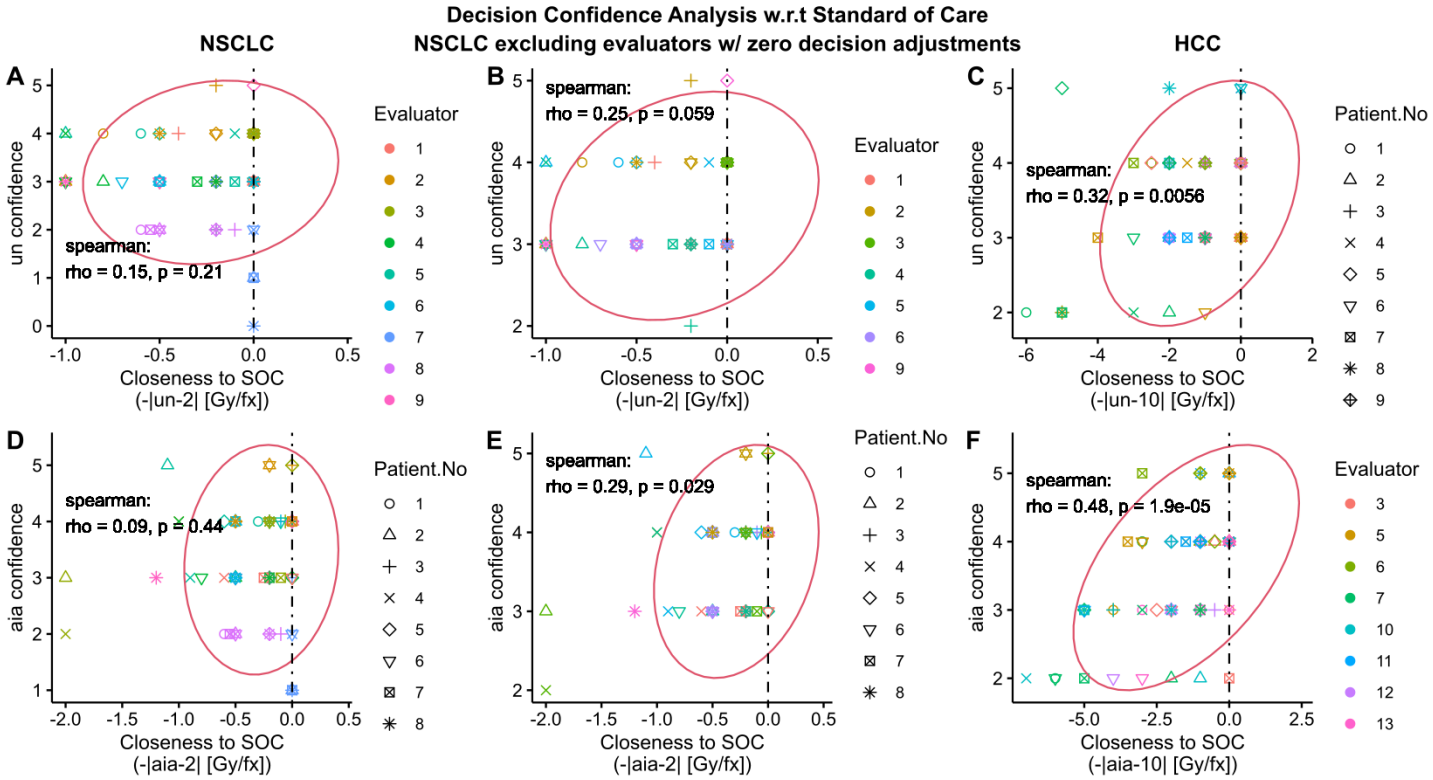



**Figure 5: Analysis of Evaluator's decision confidence with respect to the Standard of Care Dose fractionation (2 Gy/fx for NSCLC RT; 10 Gy/fx for HCC SBRT)**. The first, second, and third column of figures corresponds to NSCLC, NSCLC excluding evaluators with zero decision adjustment, and HCC, respectively. All evaluators in HCC adjusted at least one of their decisions. 2D Scatter plots **A**, **B**, and **C** show the relationship between unassisted decision ($un$) confidence (0-5, 5 being the highest level) and closeness of $un$ to the standard of care dose decision values (NSCLC: $-|un-2|$ Gy/fx); HCC: $-|un-10|$ Gy/fx). 2D Scatter plots **D**, **E**, and **F** show the relationship between AI-assisted decision ($aia$) confidence and closeness of $un$ to the standard of care dose decision values (NSCLC: $-|aia-2|$ Gy/fx); HCC: $-|aia-10|$ Gy/fx). All plots include the Spearman correlation coefficients, p-values, and co-variance ellipse (95% confidence). Covariance ellipses are included for a visual insight about the data distribution.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

significant to reject the null, i.e. $\rho = 0$ (**Figure 5A** and **5D**), but as expected, once we excluded evaluators with zero decision adjustments, we found a higher correlation value with increased significance. (**Figures 5B** and **5E**).

## 2.6 Inter-evaluator agreement increased with AI-assistance

We performed a concordance analysis on the decisions with intraclass correlation coefficient (ICC)[41–43] as listed in **Table 1**. We compared four types of ICCs between unassisted and AI-assisted decision for both NSCLC cohort and HCC cohort as shown in **Figure 6A** and **6B**, respectively. We found that on average the ICC between evaluators increased with AI-assistance for both cohorts. The higher ICC for AI-assisted decision indicates that AI-assistance resulted in decrease in inter-evaluator (inter-physicians) variability by reducing the uncertainty in decision-making.

## 2.7 The majority of decisions were adjusted to achieve higher tumor control in NSCLC and lower toxicity in HCC.

In the AI-assisted Phase, along with AI recommendation, evaluators were also provided with RT outcome estimates (RTOE). As such, we analyzed decision adjustment behavior with respect to outcome estimates as shown in **Figure 7** for NSCLC and **Figure 8** for HCC. Outcome estimates were provided in a two-dimensional outcome space spanned by TCP and NTCP. **Figures 7A** and **8A** summarize outcome estimates for unassisted decision and **Figures 7B** and **8B** for AI-assisted decision, grouped by individual patient, and color-coded and marked by individual evaluator. To investigate the effect of outcome estimates on decision-making, we isolated the adjusted decisions ($un \neq aia$) and analyzed the frequency of increase/decrease in TCP (**Figures 7C** and **8C**) and NTCP (**Figures 7D** and **8D**). We found that for NSCLC out of 41 total decision adjustments, 31 (76%) increased and 10 (24%) decreased both TCP and NTCP, whereas for HCC out of 34 total decision adjustments, 9 (26%) increased and 25 (74%) decreased both TCP and NTCP. Note that the simultaneous increase/decrease in TCP and NTCP is by design[22] which was added into the modeling to follow the radiobiological principle that states that increasing/decreasing radiation dose increases/decreases both TCP and NTCP. Since the clinical goal of RT is to maximize TCP while minimizing NTCP, the observation indicates that most of the decisions were adjusted to achieve higher TCP in NSCLC patients and lower NTCP in HCC patients. Such behavior could be attributed to the differences in TCP/NTCP slopes between NSCLC and HCC as can be seen from **Figures 7A**, **8A**, **7B**, and **8B**. Here the slope refers to the slope made by tuple $(tcp, ntcp)$ for a range of dose fractionation in the outcome space, clearly visible in **supplementary Figure S4**.

In the absence of ground truth, we analyzed the adjusted decision based on a scoring schema of toxicity free local control, i.e., $TCP(1 - NTCP)$, which reflects the clinical goal of RT. The scoring schema has a maximum value of 1 for ideal outcome of $(tcp, ntcp) = (1, 0)$ and minimum value of 0 for the dose-limiting factor $ntcp = 1$. The higher value indicates a higher TCP and a lower NTCP. **Figures 7E** and **8E** summarize the change in scores between unassisted and AI-assisted decision, grouped by patient. We observed both increase and decrease in both average and dispersion. To get a sense of the overall trend, we calculated the pairwise difference in the scoring $(aia\,score - un\,score)$ and summary statistics as shown in **Figures 7F** and **8F**, respectively. For NSCLC, we found a conflicting mean and median, where the mean showed an increase in scoring (desirable) while the median showed a decrease, whereas for HCC both mean and median showed improvement in decisions. However, in both cases, we found a right-skewed distribution which indicates an overall improvement in decision.

## 2.8 Analysis of evaluators' remarks indicated the concern for organs at risk and RT outcome estimates as important decision-making factors.

As a final step, we analyzed evaluators' self-reported text remarks as listed in **supplementary Tables S3** and **S4**. First, we summarized the remarks and then selected and counted the keywords, as shown in **Figures 9** and **10**. As the remarks were optional, we received only 48 total remarks: 14 from NSCLC Unassisted Phase, 14 from NSCLC AI-assisted Phase, 9 from HCC Unassisted Phase, and 11 from HCC AI-assisted Phase.

For NSCLC Unassisted Phase, we found that controlling the radiation-induced toxicity for organs at risk, especially for the esophagus, was the main factor for making dose escalation or de-escalation decision, which was mentioned 9 times, followed by patients' KBR-ART evaluation phase dose response (3 times). Other remarks pertained to disagreement with KBR-ART evaluation phase dose fractionation (1 time), disagreement with dose-volume histogram curve (2 times), preference for anatomical adaptation instead of dose adaptation (1 time), and trouble with PET treatment planner viewer

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

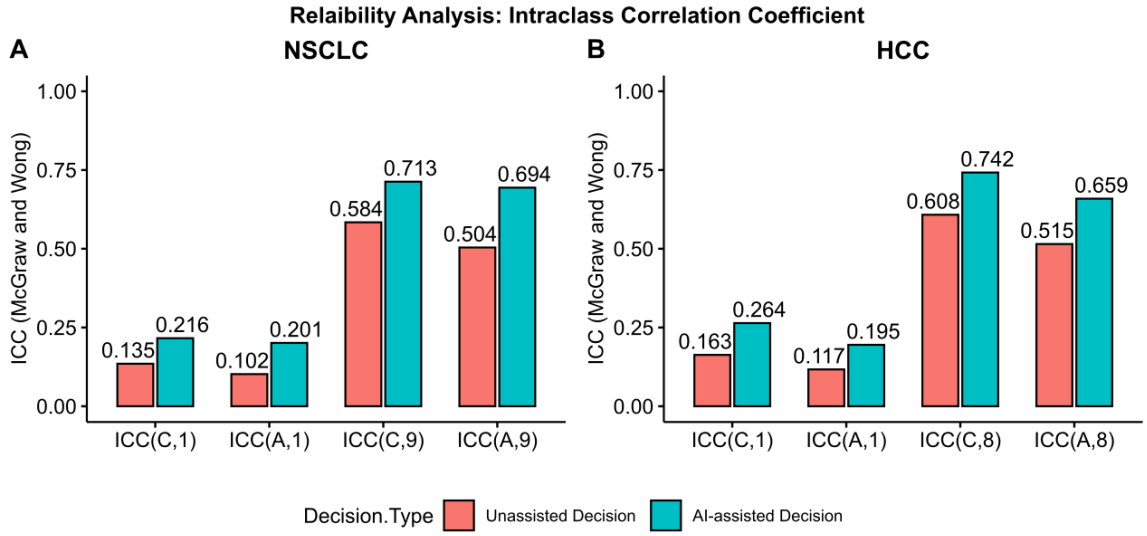



**Table 1: Intraclass Correlation Coefficient (McGraw and Wong)**

| **NSCLC** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Patient($n$) = 8, Evaluator($k$) = 9 two-way random model** | | | **Unassisted Decision** (F-stat = 2.405) | | | **AI-Assisted Decision** (F-stat = 3.482) | | |
| **Name** | **Type** | **Unit** | **Value** | **95% Conf Interval** | **P-value** | **Value** | **95% Conf Interval** | **P-value** |
| ICC(C,1) | consistency | single | 0.135 | [-0.005, 0.507] | 0.032 | 0.216 | [0.04, 0.606] | 0.004 |
| ICC(A,1) | agreement | single | 0.102 | [-0.003, 0.426] | 0.031 | 0.201 | [0.039, 0.584] | 0.003 |
| ICC(C,9) | consistency | average | 0.584 | [-0.049, 0.902] | 0.032 | 0.713 | [0.275, 0.933] | 0.004 |
| ICC(A,9) | agreement | average | 0.504 | [-0.053 ,0.871] | 0.035 | 0.694 | [0.264, 0.927] | 0.003 |
| **HCC** | | | | | | | | |
| **Patient($n$) = 9, Evaluator($k$) = 8 two-way random model** | | | **Unassisted Decision** (F-stat = 2.554) | | | **AI-Assisted Decision** (F-stat = 3.87) | | |
| **Name** | **Type** | **Unit** | **Value** | **95% Conf Interval** | **P-value** | **Value** | **95% Conf Interval** | **P-value** |
| ICC(C,1) | consistency | single | 0.163 | [0.006, 0.521] | 0.019 | 0.264 | [0.069, 0.631] | 0.001 |
| ICC(A,1) | agreement | single | 0.117 | [0.004, 0.427] | 0.019 | 0.195 | [0.045, 0.537] | 0.001 |
| ICC(C,8) | consistency | average | 0.608 | [0.049, 0.897] | 0.019 | 0.742 | [0.373, 0.932] | 0.001 |
| ICC(A,8) | agreement | average | 0.515 | [0.003, 0.858] | 0.024 | 0.659 | [0.254, 0.904] | 0.002 |

**Figure 6: Reliability analysis via Intraclass Correlation Coefficient (ICC)**. Bar plots **A** and **B** compare the McGraw and Wong's ICC between unassisted decision and AI-assisted decision for NSCLC and HCC, respectively and **Table 1** presents the summary including F statistics, 95% confidence interval, and p-value. We applied two-way random effects model to calculate the ICC for our $n \times k$ data structure where $n$ and $k$ are the number of patients and evaluators, respectively, which were both chosen randomly from a larger pool of patients and evaluators. ICC type Consistency measures the symmetric differences between the decisions of the $k$ evaluators, whereas ICC type Absolute Agreement measures the absolute differences. ICC unit Single rater corresponds to using the decision from a single evaluator as the basis for measurement and ICC unit Average corresponds to using the average decision from all evaluators.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

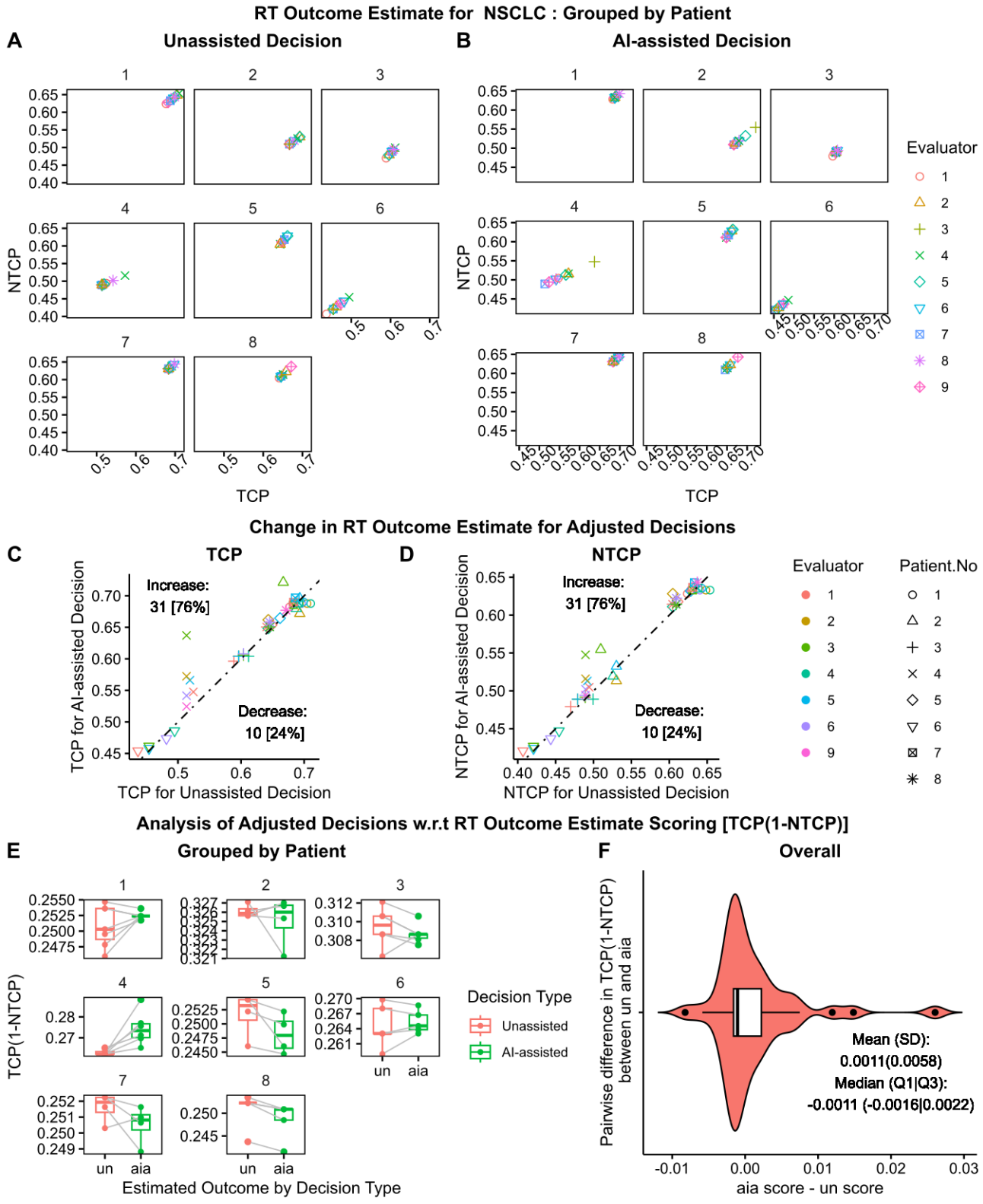



**Figure 7: Analysis of Decision Adjustment with respect to RT Outcome Estimate for NSCLC**. Scatter plots **A** and **B**, grouped by patient number, show the RT outcome estimate (RTOE) in the space spanned by tumor control probability (TCP) and normal tissue complication probability (NTCP) for Unassisted ($un$) and AI-assisted ($aia$) decision, respectively. Scatter plots **C** and **D** show the change in RTOE for adjusted decisions in $un$ vs $aia$ TCP space and $un$ vs $aia$ NTCP space, respectively, including the $45^\circ$ null dashed line. Out of 41 decision adjustment, 32 (76%) increased both TCP and NTCP while 10 (24%) decreased TCP and NTCP. Paired plot **E** and violin plot **F**, present analysis of adjusted decision based on RTOE scoring schema $TCP(1-NTCP)$ [1 for $(tcp, ntcp) = (1, 0)$, 0 for $ntcp = 1$]. Paired plots **E** compare the change in score for $un$ and $aia$ for each patient. Violin plot **F** presents the overall summary statistics for the pairwise difference in score between $aia$ and $un$: mean(sd)= 0.0011(0.0058); median(Q1|Q3)=-0.0011(-0.0016|0.0022).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

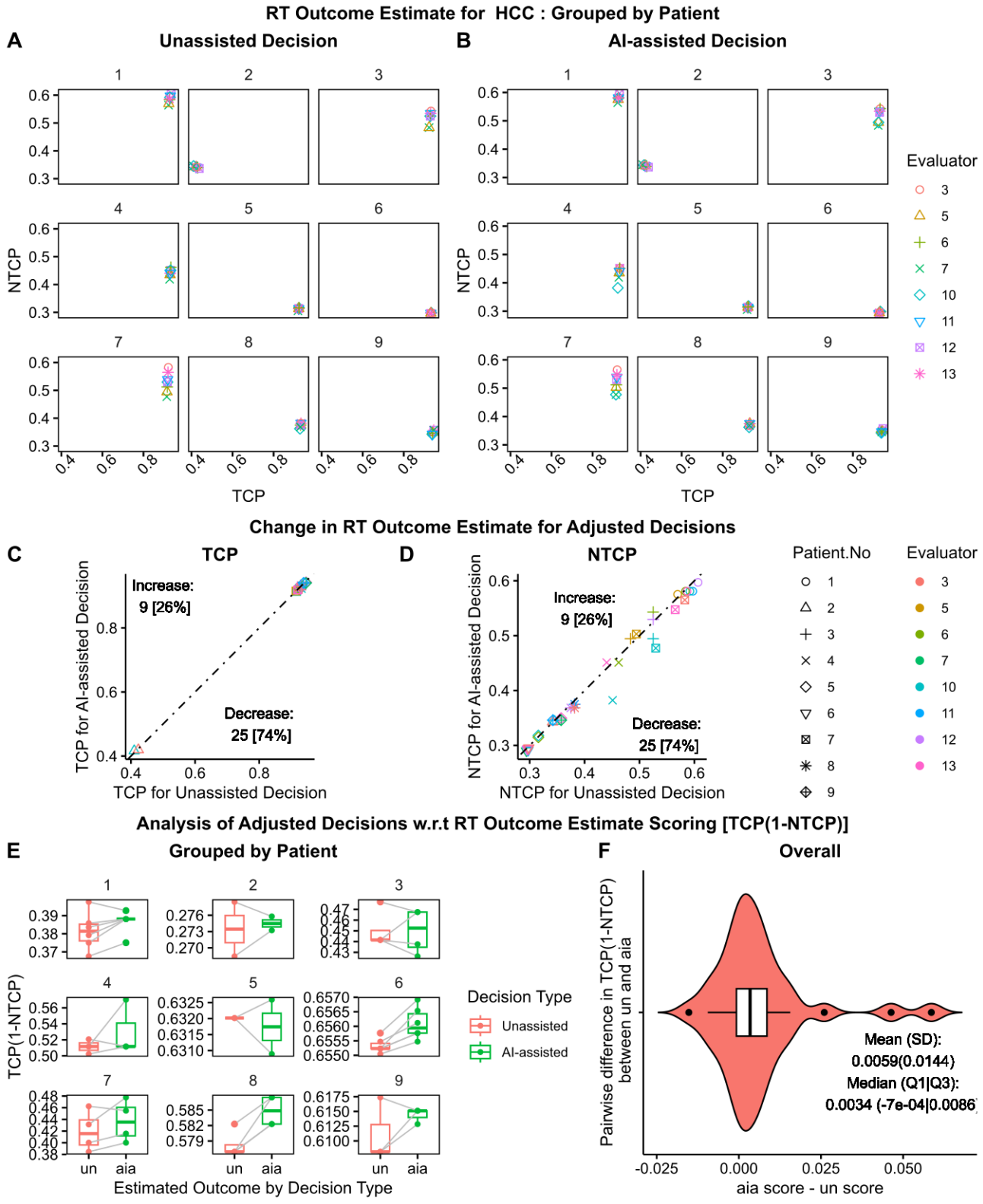



**Figure 8: Analysis of Decision Adjustment with respect to RT Outcome Estimate for HCC**. Scatter plots **A** and **B**, grouped by patient number, show the RT outcome estimate (RTOE) in the space spanned by tumor control probability (TCP) and normal tissue complication probability (NTCP) for Unassisted ($un$) and AI-assisted ($aia$) decision, respectively. Scatter plots **C** and **D** show the change in RTOE for adjusted decisions in $un$ vs $aia$ TCP space and $un$ vs $aia$ NTCP space, respectively, including the $45^\circ$ null dashed line. Out of 34 decision adjustment, 9 (26%) increased both TCP and NTCP while 25 (74%) decreased TCP and NTCP. Paired plot **E** and violin plot **F**, present analysis of adjusted decision based on RTOE scoring schema $TCP(1-NTCP)$ [1 for $(tcp, ntcp) = (1, 0)$, 0 for $ntcp = 1$]. Paired plots **E** compare the change in score for $un$ and $aia$ for each patient. Violin plot **F** presents the overall summary statistics for the pairwise difference in score between $aia$ and $un$: mean(sd)= 0.0059(0.0144); median(Q1|Q3)=0.0034(-7E-4|0.0086).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

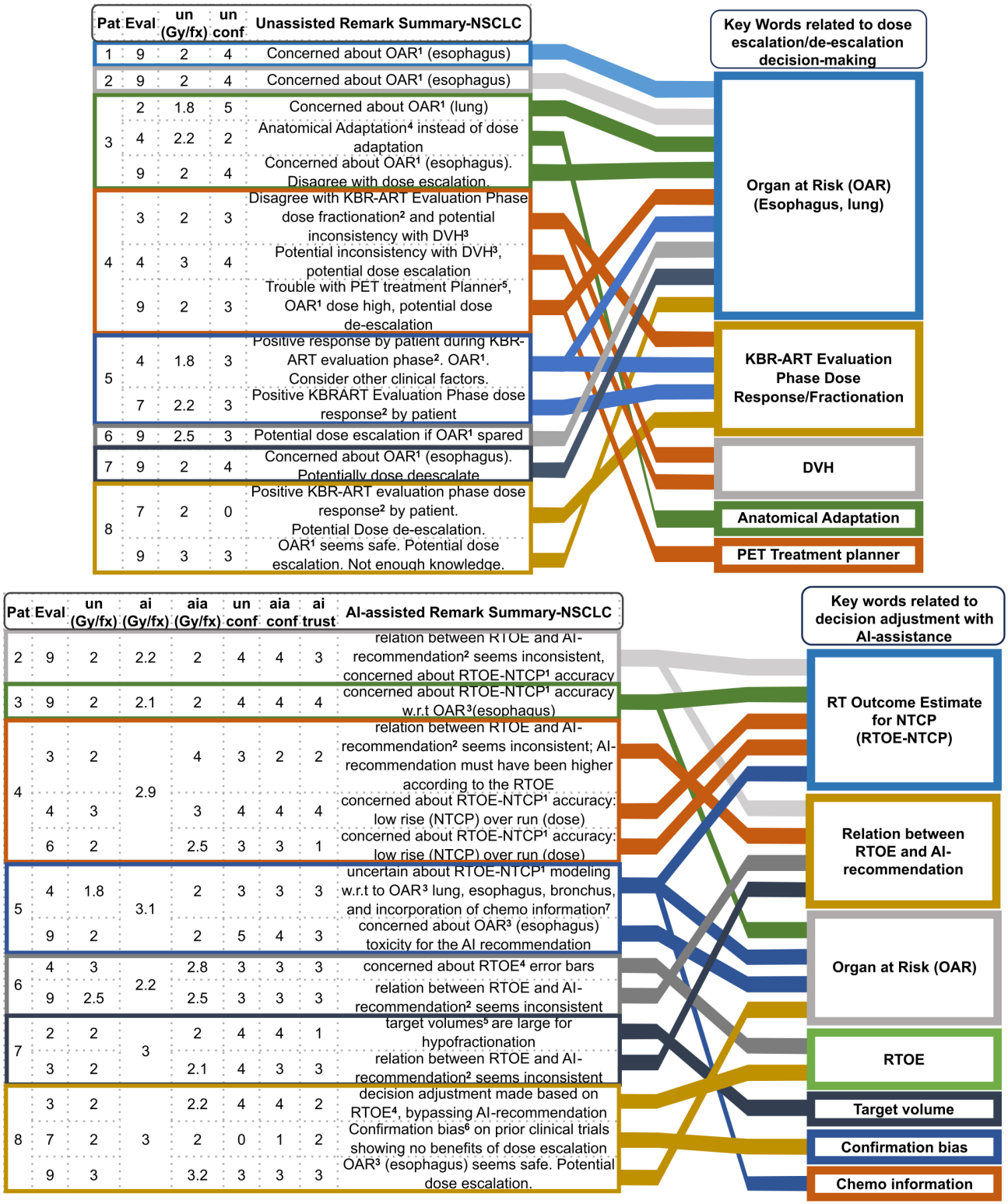
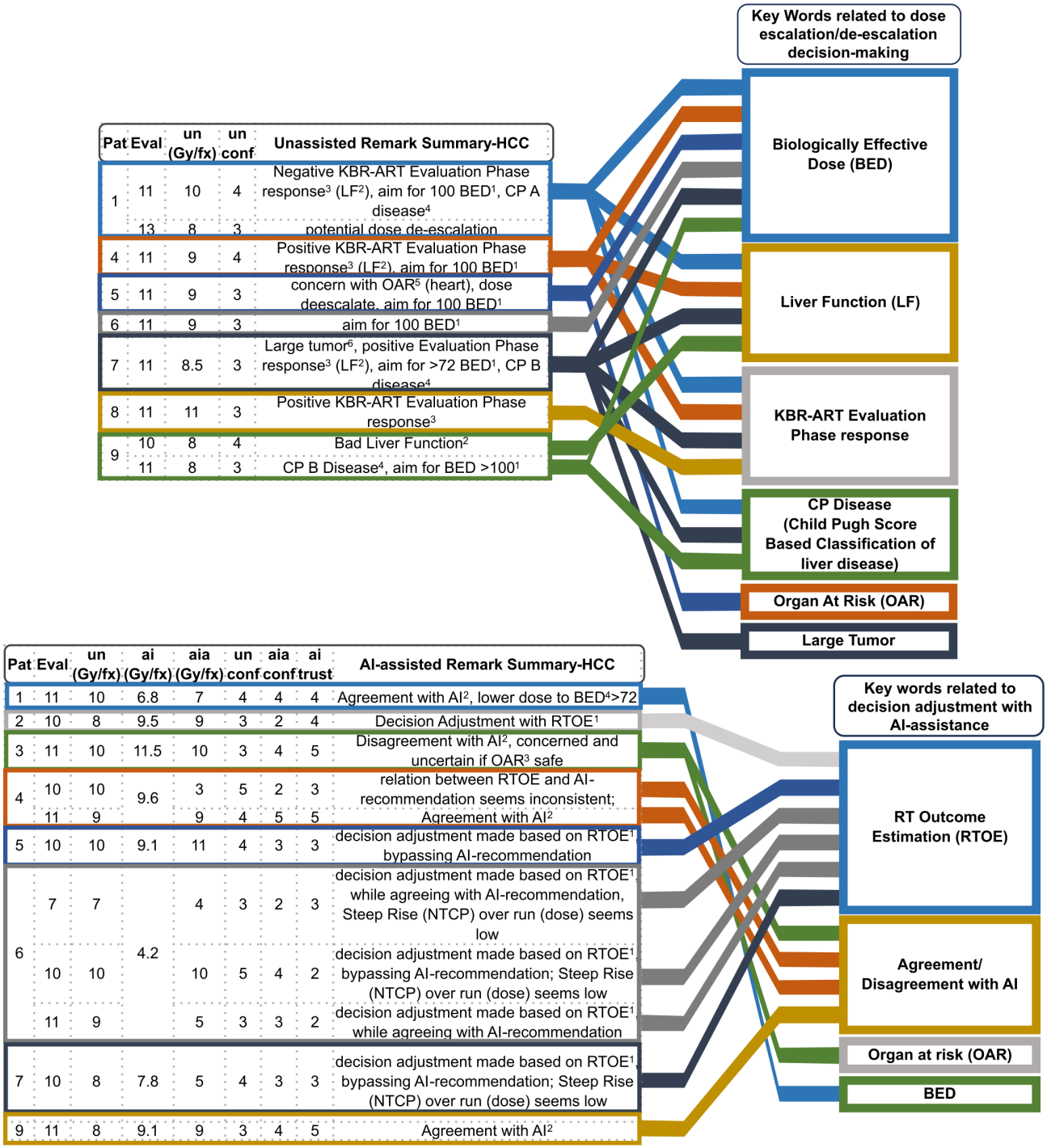



**Figure 9: Authors summary of Evaluators' Remark for NSCLC**. The tables on the left summarize Evaluators' Remark and present corresponding decision, decision confidence, and AI trust level for Un-assisted and AI-assisted Phase. To present a statistic the repeated keywords are identified and displayed in boxes in the right columns. The box height is proportional to keyword frequency, and they are sorted from the most frequent to least frequent. Evaluators' full remarks are included in the supplementary material. **Abbreviation**- **Pat**: Patient No; **Eval**: Evaluator; **un**: Unassisted Decision; **ai**: AI recommendation; **aia**: AI-assisted Decision; **un conf**: Unassisted Decision Confidence; **aia conf**: AI-assisted Decision Confidence; **ai trust**: AI recommendation trust level

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .




| Pat | Eval | un (Gy/fx) | un conf | Unassisted Remark Summary-HCC |
|---|---|---|---|---|
| 1 | 11 | 10 | 4 | Negative KBR-ART Evaluation Phase response[3] (LF[2]), aim for 100 BED[1], CP A disease[4] |
| | 13 | 8 | 3 | potential dose de-escalation |
| 4 | 11 | 9 | 4 | Positive KBR-ART Evaluation Phase response[3] (LF[2]), aim for 100 BED[1] |
| 5 | 11 | 9 | 3 | concern with OAR[5] (heart), dose deescalate, aim for 100 BED[1] |
| 6 | 11 | 9 | 3 | aim for 100 BED[1] |
| 7 | 11 | 8.5 | 3 | Large tumor[6], positive Evaluation Phase response[3] (LF[2]), aim for >72 BED[1], CP B disease[4] |
| 8 | 11 | 11 | 3 | Positive KBR-ART Evaluation Phase response[3] |
| 9 | 10 | 8 | 4 | Bad Liver Function[2] |
| | 11 | 8 | 3 | CP B Disease[4], aim for BED >100[1] |

| Pat | Eval | un (Gy/fx) | ai (Gy/fx) | aia (Gy/fx) | un conf | aia conf | ai trust | AI-assisted Remark Summary-HCC |
|---|---|---|---|---|---|---|---|---|
| 1 | 11 | 10 | 6.8 | 7 | 4 | 4 | 4 | Agreement with AI[2], lower dose to BED[4]>72 |
| 2 | 10 | 8 | 9.5 | 9 | 3 | 2 | 4 | Decision Adjustment with RTOE[1] |
| 3 | 11 | 10 | 11.5 | 10 | 3 | 4 | 5 | Disagreement with AI[2], concerned and uncertain if OAR[3] safe |
| 4 | 10 | 10 | 9.6 | 3 | 5 | 2 | 3 | relation between RTOE and AI-recommendation seems inconsistent; |
| | 11 | 9 | | 9 | 4 | 5 | 5 | Agreement with AI[2] |
| 5 | 10 | 10 | 9.1 | 11 | 4 | 3 | 3 | decision adjustment made based on RTOE[1] bypassing AI-recommendation |
| 6 | 7 | 7 | 4.2 | 4 | 3 | 2 | 3 | decision adjustment made based on RTOE[1] while agreeing with AI-recommendation, Steep Rise (NTCP) over run (dose) seems low |
| | 10 | 10 | | 10 | 5 | 4 | 2 | decision adjustment made based on RTOE[1] bypassing AI-recommendation; Steep Rise (NTCP) over run (dose) seems low |
| | 11 | 9 | | 5 | 3 | 3 | 2 | decision adjustment made based on RTOE[1] while agreeing with AI-recommendation |
| 7 | 10 | 8 | 7.8 | 5 | 4 | 3 | 3 | decision adjustment made based on RTOE[1], bypassing AI-recommendation; Steep Rise (NTCP) over run (dose) seems low |
| 9 | 11 | 8 | 9.1 | 9 | 3 | 4 | 5 | Agreement with AI[2] |

**Figure 10: Authors summary of Evaluators' Remark for HCC**. The tables on the left summarize Evaluators' Remark and present corresponding decision, decision confidence, and AI trust level for Un-assisted and AI-assisted Phase. To present a statistic the repeated keywords are identified and displayed in boxes in the right columns. The box height is proportional to keyword frequency, and they are sorted from the most frequent to least frequent. Evaluators' full remarks are included in the supplementary material. **Abbreviation**- **Pa**t: Patient No; **Eval**: Evaluator; **un**: Unassisted Decision; **ai**: AI recommendation; **aia**: AI-assisted Decision; **un conf**: Unassisted Decision Confidence; **aia conf**: AI-assisted Decision Confidence; **ai trust**: AI recommendation trust level

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

(1 time). We note that anatomical adaptation is a complementary adaptive strategy where a treatment plan is adapted to physiological changes such as weight loss or evolving tumor shape and size. For HCC Unassisted Phase, the summary of the remark indicated that the main decision-making scheme was to escalate/de-escalate dose over a certain amount of biologically effective dose (6 times) based on the KBR-ART evaluation phase response (4 times) in terms of liver function (4 times) and Child Pugh Score (3 times). Other remarks pertained to organs at risk (1 time) and large tumor size (1 time).

For NSCLC AI-assisted Phase, we found that the RT outcome estimation curve (2 times), especially the estimation for NTCP (5 times) played an important role in the decision adjustment. The main concern pertained to the lack of steep change in NTCP for a large change in the dose fractionation value and small error bars. There were concerns over the relation between RT outcome estimation and AI recommendation (4 times) and organs at risk toxicity (4 times). Other remarks pertained to large tumor volume (1 time) and whether chemotherapy information was included (1 time). The remark from evaluator #7 (one of the two evaluators with zero decision adjustment) stood out as it stated that prior clinical trials showed no benefits of dose escalation providing insight into their collaborative decision-making process. For HCC AI-assisted Phase, we again found that RT outcome estimation (6 times) played an important role in decision-making. In contrast to NSCLC, one of the concerns was the steep rise in NTCP for increasing dose fractionation. We also found that evaluators directly agreed/disagreed with the AI (4 times). Other remarks pertained to organs at risk (1 time) and biologically effective dose (1 time).

# 3 Discussion

The combination of the following two levels of seemingly conflicting observations suggests that AI-assistance does not uniformly influence an expert's decision-making process: (i) statistically non-significant difference between unassisted and AI-assisted decisions, yet considerable decision adjustment frequency and (ii) absence of decision adjustment from evaluators #7 and #8 for NSCLC, yet two decision adjustments from evaluator #7 for HCC. In particular, upon further investigation of evaluator #7's remark, we found that evaluator #7 may have a preference against dose escalation in NSCLC but is open to adaptation in HCC. The above conclusion comes from comparing the following two remarks: (i) in the AI-assisted Phase of NSCLC module, evaluator #7 left a remark for patient 8, "with multiple level 1 trials showing no benefit of dose escalation, it's hard to have any confidence in these recommendations"; (ii) in the AI-assisted Phase of HCC module, evaluator #7 left a remark for patient 6, "with TCP near identical per dose, it seems like omitting last two fractions should be an option" and adjusted their $un = 7$ Gy/fx to $aia = 4$ Gy/fx, seemingly agreeing with $ai = 4.2$ Gy/fx. Thus, we deduce that collaborative decision-making is a highly dependent process which in our case depended on clinician's prior knowledge, patient's condition, disease type, and treatment modality.

The observation of a positive correlation between decision adjustment level ($aia - un$) and dissimilarity in decision-making between evaluator and AI ($ai - un$); and between AI-trust level and agreement with AI ($-|aia - ai|$) makes it clear that understanding AI-influence in collaborative decision-making needs beyond binary considerations. The simplest of such considerations involves two conditions: first, whether the clinical experts believe in the benefits of AI system and second, if they do believe in the AI, the level of agreement with AI's recommendations and outcome prediction. In a discretized scenario, there can be four possibilities:

| AI Influence truth table | | |
|---|---|---|
| | $un = aia$ | $un \neq aia$ |
| $un = ai$ | Not Sure | Yes (based on RTOE) |
| $un \neq ai$ | No | Yes |

Clinicians that do not believe in an AI system (e.g. evaluators #7 and #8 for ARCliDS-NSCLC), are generally not influenced by any AI recommendations and tend not to adjust their decision ($un = aia$). However, even when a clinician believes in the AI system, if their unassisted decision happens to be the same or close to AI recommendation ($un = ai$), then they will again not adjust their decision ($un = aia$), and we wouldn't be able to determine AI-influence from such cases. Conversely, when clinicians make decision adjustments ($un \neq aia$) then we can consider AI-influence regardless of their agreement/disagreement with the AI recommendation. However, when evaluators adjust their decision even when their $un = ai$, then we can conclude that they must have made adjustments based on the outcome prediction (TCP/NTCP) bypassing AI recommendation. **Figures 9** and **10** present examples of such cases where evaluators adjusted decisions based on AI-recommendation and also based on RT outcome estimation by-passing

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

AI-recommendation. Note that here we make a distinction between belief as overall trust and trust level as a trust on individual recommendation, because we observed that, although, both evaluators #7 and #8 didn't change any decisions, they reported nonzero AI trust level for all recommendations. The minimum, median, and maximum AI trust level reported by evaluator #7 were 1, 3, and 5 respectively, whereas evaluator #8 reported an AI trust level of 5 for all patients.

Analysis of decision confidence level further showed differences in collaborative decision-making behavior between disease types and treatment regimens. In both NSCLC and HCC, we observed a significant positive correlation between AI-assisted decision confidence and AI trust level indicating clinical experts were more confident in their decision when they trusted a particular AI-recommendation. However, only the HCC cohort exhibited a significant positive correlation between change in confidence level and AI-trust level indicating that only HCC expert's confidence grew with AI-assistance. The insignificant correlation for NSCLC is consistent with the fact that correlation between unassisted and AI-assisted confidence were much stronger for NSCLC as shown in **supplementary Figure S6**.

We found a positive correlation between decision confidence level and closeness of decisions with the standard of care for both cases which indicates that evaluators were generally more confident in prescribing dose fractionation closer to the clinical practice. While this is a reasonable behavior, the correlation value and significance level for NSCLC were much lower than that of HCC. This may be due to the large difference in treatment options—2 Gy/fx for NSCLC administered in 30 fractions vs 10 Gy/fx for HCC administered in 5 fractions—which can result in a large difference in normal tissue complication probability; and due to a difference in organ, cancer type, and treatment protocols. Moreover, we observed that the correlation between decision confidence level and closeness was higher for AI-assisted decisions which may be due to evaluators feeling validated by the AI. When AI recommends doses closer to SOC, we would expect an amplification of confidence level due to the compounding effect of the evaluators being comfortable in repeating day-to-day practice and AI's confirmation. We note that our definition of SOC oversimplifies clinical practice, however, we find that this analysis provides valuable insights into collaborative decision making with respect to SOC.

From the analysis of adjusted decisions, we witnessed model-specific decision-making behavior. Evaluators mainly adjusted decisions focusing on TCP for NSCLC and NTCP for HCC. Upon further investigation, we found that for most HCC patients, TCP changed only slightly for the full range of dose decision values. Such AI behavior can be attributed to the training data distribution where the outcome class imbalance was very extreme, i.e. 95 out of 99 patients showed local control (**Table S10** in Niraula et al.[22]). Whereas higher local control is a clinically desirable endpoint and reflects the fact that SBRT is highly successful in treating HCC, it affects the model sensitivity of outcome prediction. As such, TCP was less sensitive to changes in dose than NTCP for HCC. Note that such behavior is similar to a type of AI bias shown by language models where AI learns slightly differently for underrepresented samples.[13]

The user-reported remarks provided miscellaneous feedback providing us with additional insights into the collaborative decision-making process. We found that the organs at risk and the patient's treatment response were the top priorities, consistent with the clinical practice where toxicity is the dose-limiting factor. For NSCLC, evaluators were concerned about esophagus and lung toxicity, and for HCC, liver function, and heart toxicities. We found that evaluators carefully inspected AI's outcome prediction for not only the AI-recommendation but also for the whole range of decisions. In particular, evaluators closely monitored the NTCP and sometimes bypassed AI's recommendation to lower the complication probability. This is a clinically desired behavior, as over-reliance on AI recommendations could lead to erroneous decisions. However, we note that since we received an unequal number of remarks from a handful of evaluators, the summary will be swayed toward the evaluators with the higher number of remarks.

## 4 Limitations

Adaptive treatment strategies are yet to be widely incorporated in standard clinical practice. Besides KBR-ART, there exist several alternative and complementary adaptive strategies, for instance, adaptation of the number of dose fractionation with fixed dose per fraction,[22,44] during treatment anatomical adaptation via CT-guided[45] or MRI-guided strategies,[46] and gene expression-based adaptation where the full course of radiotherapy is personalized.[47] Thus, evaluators' level of trust and confidence on AI recommendation would have been higher and more representative if the AI had been designed for standard clinical practice. However, since adaptive strategies are yet to be standardized, our study provides a direction for future collaborative decision-making study on AI for standard adaptive strategies. Our current study design directly asked evaluators to input their decision confidence instead of asking a series of questions

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

to assess AI trust level. As a result, we found that the analysis of decision confidence resulted in fuzzy trends and non-significant statistics. This suggests injecting objectivity and structure during the design of the evaluation questionnaire. Tschandl et al.36, in their human-computer collaboration study, used the time needed to reach a diagnosis as a surrogate marker for confidence. We could define confidence in a similar fashion, or by dividing total confidence into multiple contributions.

Although this study generated 144 evaluations, because the data is two dimensional and from two studies, the data size for NSCLC is 8 by 9 and for HCC is 9 by 8. As such, the result from our analysis will have low inferential capability. An obvious way is to increase the sample size. However, because of the two-dimensional nature of the data, increasing sample size increases the effort quadratically ($\sim n^2$). To complicate the matter, we found that the collaborative decision-making process depends on many factors and is sensitive to situational change. A similar behavior was reported in a Human-AI interaction study by McIntosh et al.,[48] where they found that medical professionals behaved more conservatively during prospective study than in retrospective settings. As such, to infer clinical collaborative decision-making behavior, conducting a prospective study should take priority followed by ensuring a sufficient sample size.

We did not analyze decision-making behaviors of sub-groups of evaluators stratified by experience, specialty, and affiliation, because further stratifying data would push to an extreme the sample size-related limitation. However, we want to note results from previous studies: Tschandl et al.,[36] in their human-computer collaboration for skin cancer recognition study, found that that the least experienced clinicians gained the most from diagnostic AI support, however, faulty AI could mislead all clinicians irrespective of experience level. Reverberi et al.,[15] reported similar behavior, where non-expert endoscopist (less than 500 colonoscopies performed) gained more from AI-assistance (improved accuracy) however, they found that experts were less able to discriminate between good and bad AI advice, and expert's average confidence was lower toward both their own judgments and AI recommendation. Similarly, Sun et al.[38] reported that AI-assistance helped inexperienced (resident and fellows) evaluators to increase their accuracy of assessing the bladder cancer treatment response. They found that with AI-assistance, inexperienced and experienced evaluators attained a similar accuracy. In the other hand, they found different performance among evaluators stratified by specialty: oncologists' accuracy improved more than that of radiologists. In contrast, Lee et al.[37] reported that the evaluator's characteristics, including experience level, were not associated with accurate AI-assisted readings of chest radiographs.

## 5 Conclusions

Human-AI interaction in dynamic decision-making has high variability as it depends on complex interrelationship between the expert's prior knowledge and preferences, the patient's state, disease site, treatment modality, and AI behavior. The collaborative decision-making for treating advanced diseases can be summarized as follows: (i) clinicians may not believe in an AI system, completely disregarding AI's recommendation (ii) clinicians may believe in the AI system but will critically analyze AI recommendations on a case-by-case basis, (iii) When clinicians find AI recommendations beneficial to patients they will adjust their decision as necessary, (iv) When clinicians do not find AI recommendations beneficial they will either stick to their own decision or, if an outcome prediction model is available, will search for an optimal decision on their own bypassing AI-recommendation. AI-assistance can reduce inter-physician variability and improving model transparency and explainability helps in balancing the reliance on AI. Clinicians are generally more comfortable making decisions that align with the standard clinical practice but will prescribe non-standard treatment if deemed optimal and necessary especially to lower treatment side effects.

## Funding Statement

This work was partly supported by the National Institute of Health (NIH) grant R01-CA233487 and its supplement.

## Declaration of Interest

D Niraula, KC Cuneo, ID Dinov, JB Jamaluddin, J Jin, Y Luo, RK Ten Haken, AK Bryant, MP Dykstra, JM Frakes, CL Liveringhouse, SR Miller, MN Mills, RF Palm, SN Regan, and A Rishi have no conflicting interests. BD Gonzalez reports fees unrelated to this work from Sure Med Compliance and Elly Health. MM Matuszak reports research funding from Varian, a licensing agreement with Fuse Oncology, and serves in the AAPM Board of Directors and is the Co-Director of

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

MROQC, funded by BCBSM. TJ Dilling is a member of the National Comprehensive Cancer Network (NCCN) NSCLC panel. JF Torres-Roca reports stock ownership and leadership in Cvergenx, Inc. He reports IP and royalty rights in RSI, GARD, RxRSI. HHM Yu reports funding or fees unrelated to this work from the National Institute of Health, UpToDate, Novocure and Bristol-Myers Squib. I El Naqa is on the scientific advisory of Endectra, LLC., co-founder of iRAI LLC, deputy editor for the journal of Medical Physics, co-Chief editor of British Journal of Radiology (BJR)-AI and receives funding from the National Institute of Health (NIH), foundations, and Department of Defense (DoD).

A PCT patent application for ARCliDS has been filed. Patent Applicant: H Lee Moffitt Cancer Center IP office in conjunction with University of Michigan IP office. Inventors: Dipesh Niraula, Issam El Naqa, Randall K. Ten Haken, Wenbo Sun, Judy Jin, Ivo Dinov, Kyle Cuneo, Martha M Matuszak, and Jamalina Jamaluddin. Application Number: US2023/075004. Status of Application: Pending. Specific aspect of manuscript covered in patent application: The patent covers the underlying model-based decision-making framework of ARCliDS.

## Contributors Role

I El Naqa conceived the study. KC Cuneo, ID Dinov, J Jin, MM Matuszak, RK Ten Haken, and I El Naqa supervised the project. D Niraula., ID Dinov, BD Gonzalez, J Jin, and I El Naqa designed the study. D Niraula designed, developed, and deployed the ARCliDS software and the evaluation modules. D Niraula and JB Jamaluddin collected and curated patient data. KC Cuneo, JB Jamaluddin, ID Dinov, J Jin, MM Matuszak, RK Ten Haken, and I El Naqa evaluated the module design. D. Niraula developed the tutorial videos. D Niraula, KC Cuneo, TJ Dilling, and I El Naqa conducted evaluator search. D Niraula, KC Cuneo, and I El Naqa conducted initial pre-evaluation information sessions with potential evaluators. AK Bryant, TJ Dilling, KC Cuneo, MB Dykstra, JM Frakes, CL Liveringhouse, SR Miller, MN Mills, RF Palm, SN Regan, A Rishi, JF Torres-Roca, and HHM Yu participated in the study and critically evaluated the module. D Niraula administered the evaluation study and provided tech support as needed. D Niraula analyzed the results. ID Dinov, BD Gonzalez, J Jin, and Y Luo provided feedback on the quantitative analysis. All authors carefully analyzed the methods and results. D Niraula drafted the manuscript and all authors critically reviewed and contributed to the final version.

## References

[1] Scott EC, Baines AC, Gong Y, et al. Trends in the approval of cancer therapies by the FDA in the twenty-first century. Nat Rev Drug Discov 2023;22(8):625–40.

[2] Debela DT, Muzazu SG, Heraro KD, et al. New approaches and procedures for cancer treatment: Current perspectives. SAGE Open Med 2021;9:205031212110343.

[3] Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. Front Oncol 2020;10.

[4] Wei L, Niraula D, Gates EDH, et al. Artificial intelligence (AI) and machine learning (ML) in precision oncology: a review on enhancing discoverability through multiomics integration. Br J Radiol 2023;96(1150).

[5] Kosorok MR, Moodie EEM, editors. Adaptive Treatment Strategies in Practice. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2015.

[6] Chakraborty B, Murphy SA. Dynamic Treatment Regimes. Annu Rev Stat Appl 2014;1(1):447–64.

[7] Engelhardt D, Michor F. A Quantitative Paradigm for Decision-Making in Precision Oncology. Trends Cancer 2021;7(4):293–300.

[8] The Lancet Oncology. Can artificial intelligence improve cancer care? Lancet Oncol 2023;24(6):577.

[9] El Naqa I, Karolak A, Luo Y, et al. Translation of AI into oncology clinical practice. Oncogene 2023;42(42):3089–97.

[10] Perez-Lopez R, Reis-Filho JS, Kather JN. A framework for artificial intelligence in cancer research and precision oncology. NPJ Precis Oncol 2023;7(1):43.

[11] Luchini C, Pea A, Scarpa A. Artificial intelligence in oncology: current applications and future perspectives. Br J Cancer 2022;126(1):4–9.

[12] Senthil Kumar K, Miskovic V, Blasiak A, et al. Artificial Intelligence in Clinical Oncology: From Data to Digital Pathology and Treatment. American Society of Clinical Oncology Educational Book 2023;(43).

[13] Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P. Towards a standard for identifying and managing bias in artificial intelligence. 2022.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[14] Weng W-H, Sellergen A, Kiraly AP, et al. An intentional approach to managing bias in general purpose embedding models. Lancet Digit Health 2024;6(2):e126–30.

[15] Reverberi C, Rigon T, Solari A, et al. Experimental evidence of effective human–AI collaboration in medical decision-making. Sci Rep 2022;12(1):14952.

[16] Knop M, Weber S, Mueller M, Niehaves B. Human Factors and Technological Characteristics Influencing the Interaction of Medical Professionals With Artificial Intelligence-Enabled Clinical Decision Support Systems: Literature Review. JMIR Hum Factors 2022;9(1):e28639.

[17] Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. J Med Internet Res 2020;22(6):e15154.

[18] USFDA. Proposed Regulatory Framework for Modifications to Artificial Intelligience/Machine Learning (AI/ML)-Based Software as a Medical Device(SaMD)) [Internet]. 2019 [cited 2024 Apr 24]. Available from: https://www.fda.gov/media/122535/download

[19] USFDA. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan [Internet]. 2021 [cited 2024 Apr 24]. Available from: https://www.fda.gov/media/145022/download

[20] Giddings R, Joseph A, Callender T, et al. Factors influencing clinician and patient interaction with machine learning-based risk prediction models: a systematic review. Lancet Digit Health 2024;6(2):e131–44.

[21] Niraula D, Jamaluddin J, Matuszak MM, Haken RK Ten, Naqa I El. Quantum deep reinforcement learning for clinical decision support in oncology: application to adaptive radiotherapy. Sci Rep 2021;11(1):23545.

[22] Niraula D, Sun W, Jin J, et al. A clinical decision support system for AI-assisted decision-making in response-adaptive radiotherapy (ARCliDS). Sci Rep 2023;13(1):5279.

[23] Tseng H-H, Luo Y, Ten Haken RK, El Naqa I. The Role of Machine Learning in Knowledge-Based Response-Adapted Radiotherapy. Front Oncol 2018;8.

[24] Sun W, Niraula D, El Naqa I, et al. Precision radiotherapy via information integration of expert human knowledge and AI recommendation to optimize clinical decision making. Comput Methods Programs Biomed 2022;221:106927.

[25] Niraula D, El Naqa I, Ten Haken R, et al. Adaptive Radiotherapy Clinical Decision Support Tool and Related Methods. PCT US2023/075004 filed September 25, 2023, Provisional patent U.S. Patent 63272888 filed October 28, 2021, (pending).

[26] Niraula D, Sun W, Jin J, et al. A Decision Support Software for AI-Assisted Decision Making in Response-Adaptive Radiotherapy — An Evaluation Study. International Journal of Radiation Oncology*Biology*Physics 2022;114(3):e101–2.

[27] Kong F-M, Ten Haken RK, Schipper M, et al. Effect of Midtreatment PET/CT-Adapted Radiation Therapy With Concurrent Chemotherapy in Patients With Locally Advanced Non–Small-Cell Lung Cancer. JAMA Oncol 2017;3(10):1358.

[28] Jackson WC, Suresh K, Maurino C, et al. A mid-treatment break and reassessment maintains tumor control and reduces toxicity in patients with hepatocellular carcinoma treated with stereotactic body radiation therapy. Radiotherapy and Oncology 2019;141:101–7.

[29] Cui S, Traverso A, Niraula D, et al. Interpretable artificial intelligence in radiology and radiation oncology. Br J Radiol 2023;96(1150).

[30] Issam El Naqa, editor. A Guide to Outcome Modeling In Radiotherapy and Oncology: Listening to the Data. 1st ed. Boca Raton : CRC Press; 2018.

[31] Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. 2017 [cited 2024 Apr 24];Available from: https://doi.org/10.48550/arXiv.1702.08608

[32] Pataranutaporn P, Liu R, Finn E, Maes P. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. Nat Mach Intell 2023;5(10):1076–86.

[33] Dipesh Niraula. Tutorial on Human-AI interaction type clinical evaluation scheme for ARCliDS – HCC [Internet]. YouTube; 2022 [cited 2024 Apr 24]. Available from: https://www.youtube.com/watch?v=WYS1WoAwhr0

[34] Dipesh Niraula. Tutorial on Human-AI interaction type clinical evaluation scheme for ARCliDS - NSCLC [Internet]. YouTube; 2022 [cited 2024 Apr 24]. Available from: https://www.youtube.com/watch?v=R_1D3ZvUt1E

[35] Niraula D, Sun W, Jin J (Judy), et al. ARCliDS: A Clinical Decision Support System for AI-assisted Decision-Making in Response-Adaptive Radiotherapy. medRxiv 2022;

[36] Tschandl P, Rinner C, Apalla Z, et al. Human–computer collaboration for skin cancer recognition. Nat Med 2020;26(8):1229–34.

[37] Lee JH, Hong H, Nam G, Hwang EJ, Park CM. Effect of Human-AI Interaction on Detection of Malignant Lung Nodules on Chest Radiographs. Radiology 2023;307(5).

[38] Sun D, Hadjiiski L, Alva A, et al. Computerized Decision Support for Bladder Cancer Treatment Response Assessment in CT Urography: Effect on Diagnostic Accuracy in Multi-Institution Multi-Specialty Study. Tomography 2022;8(2):644–56.

[39] Edgington E, Onghena P. Randomization Tests. 4th ed. New York: Chapman and Hall/CRC; 2007.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
**(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[40] Howell DC. Randomization Test on Means of Matched Pairs. https://www.uvm.edu/~statdhtx/StatPages/Randomization%20Tests/RandomMatchedSample/RandomMatchedSample.html. 2015;

[41] Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 1979;86(2):420–8.

[42] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1(1):30–46.

[43] Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation – A discussion and demonstration of basic features. PLoS One 2019;14(7):e0219854.

[44] Glide-Hurst CK, Lee P, Yock AD, et al. Adaptive Radiation Therapy (ART) Strategies and Technical Considerations: A State of the ART Review From NRG Oncology. Int J Radiat Oncol Biol Phys 2021;109(4):1054–75.

[45] Lavrova E, Garrett MD, Wang Y-F, et al. Adaptive Radiation Therapy: A Review of CT-based Techniques. Radiol Imaging Cancer 2023;5(4).

[46] Keall PJ, Brighi C, Glide-Hurst C, et al. Integrated MRI-guided radiotherapy — opportunities and challenges. Nat Rev Clin Oncol 2022;19(7):458–70.

[47] Scott JG, Berglund A, Schell MJ, et al. A genome-based model for adjusting radiotherapy dose (GARD): a retrospective, cohort-based study. Lancet Oncol 2017;18(2):202–11.

[48] McIntosh C, Conroy L, Tjong MC, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. Nat Med 2021;27(6):999–1005.

[49] Luo Y, McShan DL, Matuszak MM, et al. A multiobjective Bayesian networks approach for joint prediction of tumor local control and radiation pneumonitis in nonsmall-cell lung cancer (NSCLC) for response-adapted radiotherapy. Med Phys 2018;45(8):3980–95.

[50] Luo Y, Cuneo KC, Lawrence TS, et al. A human-in-the-loop based Bayesian network approach to improve imbalanced radiation outcomes prediction for hepatocellular cancer patients with stereotactic body radiotherapy. Front Oncol 2022;12.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

# Supplementary Materials

## S1 Background

### S1.1 KBR-ART

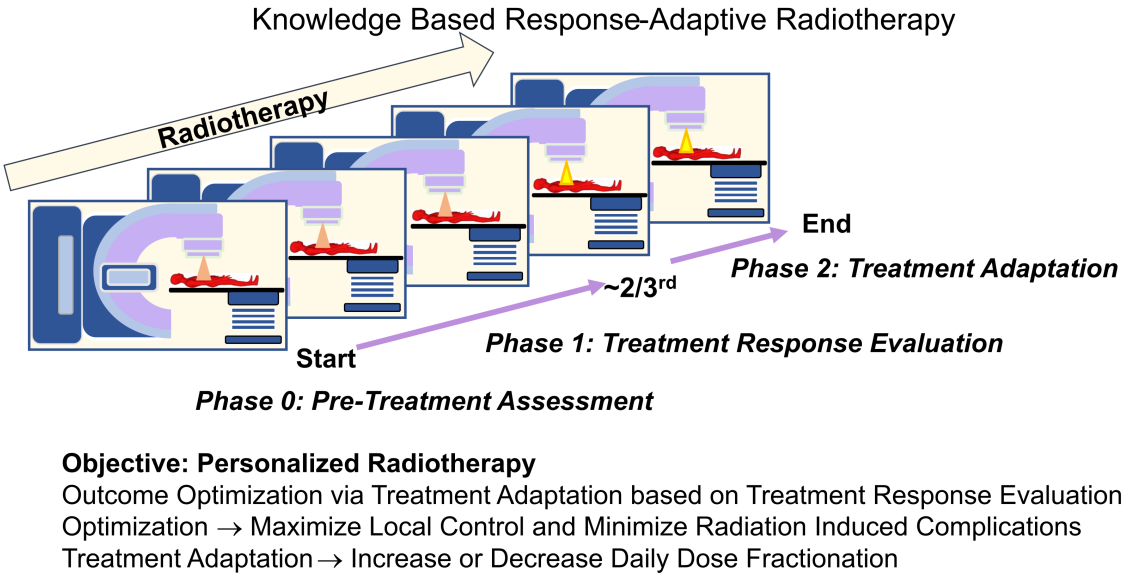



**Figure S1: Schematic of Knowledge-Based Response Adaptive Radiotherapy**. In KBR-ART, a pre-treatment assessment is conducted, and optimal treatment plan is selected. After 2/3rd of treatment, patients' treatment response is evaluated, and an optimal treatment adaptation is planned and executed. Figure adapted from Niraula D, Sun W, Jin J, et al. Sci Rep 2023;13(1): 527919.[22]

Knowledge Based response-adaptive Radiotherapy (KBR-ART[21–24]) is a dynamic interventive treatment strategy, collectively known as dynamic treatment regimen[5,6] that consists of at least three phases: Pre-Treatment Assessment, Treatment Response Evaluation (evaluation phase) and Treatment Adaptation (adaptation phase). The evaluation phase beings with the start of treatment and lasts up to the intervention; the adaptation phase follows and lasts for the remaining treatment period. In the pre-treatment phase, a patient's disease and condition is assessed and a treatment plan is tailored. In the evaluation phase, a patient's treatment response is evaluated by comparing pre and mid treatment multi-omics information changes. Based on the treatment responses, the patient's associated outcome probabilities are estimated. In the adaptation phase, treatment planning is adapted for a personalized and an optimal outcome. A schematic of KBR-ART is presented in **Figure S1**.

### S1.2 ARCliDS

ARCliDS is a web-based clinical decision support software for AI-assisted optimal decision-making in KBR-ART.[22] Given a patient's pre and during treatment multi-omics information, ARCliDS can estimate treatment response and recommend an optimal intervention for the remaining treatment period. The treatment response is estimated in terms of TCP and NTCP, and the intervention is recommended in terms of daily dose fractionation. The optimal intervention corresponds to maximum TCP and minimum NTCP estimate. Additionally, ARCliDS provides uncertainty estimates via statistical ensemble for both outcome estimates and dose recommendation.

ARCliDS can be divided into two main components: artificial radiotherapy environment (ARTE) and optimal decision-maker (ODM). ARTE is composed of radio-biologically constrained transition functions for capturing patient's state dynamics, RT outcome estimator for predicting patient outcomes, and reward function for representing clinical goal. ODM is an artificially intelligent agent trained using model-based deep reinforcement learning approach. ODM applies patient's pre and mid treatment information into the ARTE, and based of the reward signals, can learn optimal decision-making process.

For improved explainability and transparency of AI's decision-making process, along with optimal recommendation,

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

ARCliDS includes outcome space plot, reward signals, features distribution, and uncertainty estimates. ARCliDS presents its recommendations in the Outcome Space plot, spanned by TCP and NTCP, and contoured and colored with reward signal. Given a patient's information, the outcome space shows treatment outcomes and uncertainty estimate for a range of daily dose fractions [NSCLC: 1.5-4 Gy/fx; HCC: 1-15 Gy/fx]. The optimal dose recommendation and model uncertainty is marked with green diamonds. Additionally, since some of the multi-omics information is not used in the current clinical practice, ARCliDS provides patient-specific feature information in Population Distribution plots. Knowing the patient's feature value and its relative position to the population helps end-users to visualize patient's "whereabouts".

## S1.3 ARCliDS-NSCLC Module

ARCliDS' NSCLC module was trained on dataset obtained from UMCC 2007-123 phase II dose escalation clinical trial NCT01190527, where inoperable or unresectable NSCLC patients were administered with 30 daily dose fractions.[27] The patients received roughly 50 Gy [Gray $=$ J/Kg] equivalent dose in 2 Gy fractions (EQD2) in the evaluation phase and up to a total dose of 92 Gy EQD2 in the adaptation phase. The evaluation phase lasted for roughly two-thirds of the 6-week treatment period. For simplicity, the training dataset was divided into the evaluation phase of 20 fractions and the adaptation phase of 10 fractions. Two binary endpoints were considered for training: local control (LC) and radiation-induced pneumonitis of grade 2 or higher (RP2).

NSCLC patient's state was defined as the multi-omics features resulting from a multi-objective Markov Blanket feature selection process,[49] which found that the features were important predictors for both LC and RP2. The selected features were cytokines: pretreatment interleukin 4 (pre-IL4), pre-IL15 and slope of Interferon gamma-induced protein 10 (slope-IP10); Tumor PET imaging features/Radiomics: pretreatment Metabolic Tumor Volume (pre-MTV), relative difference (RD) of Gray-level size zone matrices (GLSZM)-large zone low gray-level (LZLGE) and RD-GLSZM-zone size variance (RD-GLSZM-ZSV); Dosimetry: Tumor gEUD and Lung gEUD; Genetics (single nucleotide polymorphism [SNP]): Cxcr1- Rs2234671, Ercc2-Rs238406, and Ercc5-Rs1047768; and MicroRNA: miR-191-5p and miR-20a-5p.

## S1.4 ARCliDS-HCC Module

ARCliDS' HCC module was trained on a dataset obtained from the adaptive arm of the clinical trials NCT01519219, NCT01522937, and NCT0246083514, where HCC patients received adaptive SBRT in a 3-2 split.[28] In the evaluation phase, patients received 3 high daily dose fractions followed by 1 month break, and in the adaptation phase, a suitable sub-population of the patients received 2 additional daily doses. Two binary endpoints were considered for training: local control (LC) and liver toxicity (LT) ($\geq$ 2 points increase in Child-Pugh score during any point in the treatment.)

Similarly, HCC patient's state was defined from multi-omics feature resulting from a human-in-the-loop based multi-objective Bayesian Network study.[50] The selected features were clinical: sex, age, pretreatment cirrhosis status (pre-cirrhosis), pretreatment Eastern Cooperative Oncology Group Performance Status (pre-ECOG-PS), number of active liver lesions (active lesions), pretreatment albumin level (pre-albumin); Tumor PET Imaging: gross tumor volume (GTV) and liver volume minus GTV (Liver-GTV); Dosimetry: GTV gEUD and Liver-GTV volume; and cytokines/signaling molecule: relative difference of Transforming growth factor beta (RD-TGF-$\beta$), Cluster of Differentiation 40 receptor's Ligand (RD-CD40L), and Hepatocyte growth factor (RD-HGF).

More details can be found in Niraula et al.'s work and its supplementary material.[22]

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

# S2 Evaluation Study Design Principle and Implementation

We designed two-phase evaluation modules, divided into Unassisted Phase and AI-assisted Phase, for ARCliDS-NSCLC and ARCliDS-HCC.[22] We chose a sequential design, where a unit of evaluation consisted of Unassisted Phase followed by AI-assisted phase for each patient. Doing so enabled us to perform a matched pair type of analysis in contrast to comparing whole distribution. In addition, we added text boxes for remarks and encouraged evaluators to provide comments which would provide valuable insight to their decision-making in the form of unstructured data.

To simplify the evaluation process, we designed the modules to be a stand-alone, interactive, and auto-saving web application; limited the patient count so that an evaluation could be completed in under an hour; developed tutorial videos; and conducted an initial pre-evaluation information session with the evaluators. To make it a stand-alone application that could operate without a standard treatment planner, we wrote python scripts to preprocess treatment planning DICOM files into standard 3D NumPy arrays, then built treatment plan viewer using Plotly library and incorporated it into the modules using the reticulate library. For interactive functions, we built the modules using R Shiny user-interface and Plotly graphing library. For auto-saving the evaluation, we linked the modules to google sheets using googesheet4 library. For web accessibility, we hosted the modules in R shinyapps.io server. In addition to limiting the evaluation time, we included an online account system so that the evaluation could be completed in multiple sessions if needed. We developed tutorial videos explaining the motivation, functionality, and a brief overview of ARCliDS and evaluation modules, which was played to all evaluators during the initial pre-evaluation information session in addition to a demonstration of evaluation on a sample case.

To minimize biases, external influence, and systematic errors, we incorporated elements from randomization and isolation at different levels in this study. First, evaluation modules randomly initialized ordering of patients, so that each evaluator would interact with the same group of patients in a different ordering. Second, modules isolated the decision-making process by restricting evaluators to revisit the Unassisted Phase once they have seen the AI-recommendation and by deliberately excluding Unassisted Phase decisions from the AI-assisted Phase page. Third, we enrolled diverse evaluators from multiple institutions, different career levels and different specializations.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

# S3 Evaluation Modules

## S3.1 Workflow

**Figure 1** summarizes the workflow between the main 3 components of the evaluation modules (EMs): User account and Data Frames, Unassisted page, and AI-Assisted page. The modules start at the Welcome Page which leads to either Log In page for returning evaluators or Create Account page for new evaluators. All information is auto saved into Data Frames and depending on the status of the evaluator, they are prompted to either Unassisted page to start or the latest phase they left at.

In the Unassisted Phase, we present necessary clinical information, treatment plan, and images from the KBR-ART Evaluation Phase. Based on the provided information, the evaluators are asked to input their decision for the following KBR-ART Adaptation Phase, their decision confidence level, and any remarks they may have. Once the evaluators are satisfied with their input, they can submit, which leads them to the AI-assisted Phase. In the AI-assisted Phase, we provide full access to ARCliDS. After seeing the ARCliDS' recommendation and outcome estimates, the evaluators are asked to re-enter their decision, decision confidence level, and their trust level on the AI Recommendation. Again, once the evaluators are satisfied with their input, they can submit, which completes the evaluation process for that patient. This process is repeated until the patient list is exhausted, which then leads them to the Exit Page which presents the input summary and a thank you message. In this study, we included 8 NSCLC patients and 9 HCC patients, which were selected based on availability of imaging information and also to limit the evaluation time to under one hour.

## S3.2 User Account, Data Frames, and Tutorial Video

We built a rudimentary account system consisting of a Welcome page, Create Account page, Log In page, and Data Frame to allow users to finish the evaluation in multiple sessions if needed. The Welcome page contains web links[33,34] to a tutorial video and ARCliDS manuscript.[15] The 10-minute-long tutorial video—created using PowerPoint and hosted in YouTube—contains description of KBR-ART; clinical trials and training dataset on which ARCliDS was trained;[27,28] ARCliDS architecture and its graphical user interface; and evaluation questionnaire. Two separate training videos were created for each of the two diseases: NSCLC[33] and HCC.[34]

The welcome page contains two push buttons for new and returning evaluators. The new evaluator button is linked to the Create Account page, where the new evaluators must input their name, level (physicians or resident), affiliation, specialization, experience (number of years), unique user ID, and 8-digit PIN, to create a new account. The user ID is actively validated against the existing accounts and the 8-digit PIN is actively validated for the length. Once all the information is provided, clicking on the Start Evaluation button will save the information in the cloud Data Frame and take the evaluators to the Unassisted Phase page. The returning evaluator button is linked to the Log In page, where the returning evaluators must input their existing user ID and 8-digit PIN to continue their evaluation.

Data Frames are stored as google sheets. We use two data frames: the first to save all the evaluation material and the second to store the account information, validate the account information, randomly initialized patient ordering, and to keep track of the latest evaluated patient and the phase for each evaluator. The latter record is used as a guide for the returning users.

## S3.3 Unassisted Page

In the Unassisted page, we recreated clinical workflow by presenting patient's information, Evaluation Phase treatment plan including PET/MRI images, and questionnaires as detailed in subsequent subsections.

### S3.3.1 Patient Information

A summary of the Patient information is listed in **Table S1**. In NSCLC module, we presented patient's sex, age, cancer stage, smoking history (binary), chronic obstructive pulmonary disease status (COPD, binary), cardiovascular disease status (CVD, binary), hypertension status (binary), histology (categorical), chemo status (binary), Karnofsky performance status (KPS), and gross tumor volume (GTV, in cc). In HCC module, we presented patient's sex, age, cirrhosis status (binary), number of active lesion, portal vein thrombosis (PVT) status (binary), number of pre-SBRT line of systemic therapies, number of pre-SBRT liver-directed therapy, presence of extrahepatic disease status (binary),

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

number of prior liver occurrence, previous treatment status (binary), eastern cooperative oncology performance status (ECOG-PS), gross tumor volume (GTV, in cc), and Liver minus GTV (in cc). In addition, we presented pre-treatment and mid-treatment liver functions in terms of Albumin level (g/DL), Bilirubin level (g/DL), ALBI score, and Child-Pugh (CP) score.

**Table S1: Patient's Clinical Information**

**NSCLC**

| Pat No | Sex | Age | Cancer Stage | Smoking History | COPD | CVD | Hypertension | Histology | Chemo | KPS | GTV [cc] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Female | 56-60 | 1 | Yes | No | No | No | Poorly Differentiated | Yes | 90 | 19 |
| 2 | Male | 56-60 | 3 | Yes | No | No | Yes | Adenocarcinoma | Yes | 100 | 207 |
| 3 | Female | 81-85 | 3 | No | No | No | No | Adenocarcinoma | Yes | 80 | 19 |
| 4 | Female | 56-60 | 3 | Yes | No | No | Yes | Adenocarcinoma | Yes | 90 | 58 |
| 5 | Male | 56-60 | 3 | Yes | No | No | Yes | Squamous Cell Carcinoma | Yes | 80 | 359 |
| 6 | Female | 56-60 | 3 | Yes | No | No | No | Adenocarcinoma | Yes | 90 | 102 |
| 7 | Male | 56-60 | 3 | Yes | No | No | No | Squamous Cell Carcinoma | Yes | 90 | 180 |
| 8 | Female | 61-65 | 3 | Yes | Yes | No | No | Squamous Cell Carcinoma | Yes | 70 | 113 |

**HCC**

| Pat No | Sex | Age | Cirrhosis | No of Active lesion | PVT | No of Pre-SBRT Systemic Therapies | No of Pre-SBRT Liver-Directed Therapies | Presence of Extra-hepatic Disease | No of Prior Liver Occurrences | Previously Treated | ECOG-PS | GTV [cc] | Liver minus GTV [cc] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Female | 76-80 | Yes | 1 | No | 0 | 2 | No | 1 | Yes | 0 | 17.02 | 1641 |
| 2 | Female | 76-80 | Yes | 2 | No | 0 | 4 | No | 3 | Yes | 0 | 0.75, 0.92 | 2546 |
| 3 | Female | 71-75 | Yes | 4 | No | 0 | 2 | No | 2 | Yes | 1 | 4.32, 1.92, 6.32, 0.56 | 1346 |
| 4 | Female | 76-80 | Yes | 1 | No | 0 | 2 | No | 2 | No | 1 | 2.2 | 1416 |
| 5 | Male | 56-60 | Yes | 1 | No | 0 | 2 | No | 3 | No | 0 | 9.9 | 1714 |
| 6 | Female | 56-60 | Yes | 1 | No | 0 | 3 | No | 1 | Yes | 1 | 4.3 | 2287 |
| 7 | Female | 56-60 | Yes | 1 | Yes | 0 | 0 | No | 0 | No | 0 | 469 | 1604 |
| 8 | Female | 76-80 | Yes | 1 | No | 0 | 5 | No | 4 | Yes | 0 | 5.51 | 1281 |
| 9 | Female | 61-65 | Yes | 1 | No | 0 | 1 | No | 0 | No | 0 | 2.04 | 1533 |

**Table S1: Patient's clinical information provided to the evaluators**. For NSCLC, the columns names are patient's sex, age (exact age was provided during evaluation), cancer stage, smoking history (binary), chronic obstructive pulmonary disease status (COPD, binary), cardiovascular disease status (CVD, binary), hypertension status (binary), histology (categorical), chemo status (binary), Karnofsky performance status (KPS), and gross tumor volume (GTV, in cc). For HCC, the column names are patient's sex, age, cirrhosis status (binary), number of active lesion, portal vein thrombosis (PVT) status (binary), number of pre-SBRT line of systemic therapies, number of pre-SBRT liver-directed therapy, presence of extrahepatic disease status (binary), number of prior liver occurrence, previous treatment status (binary), eastern cooperative oncology performance status (ECOG-PS), gross tumor volume (GTV, in cc), and Liver minus GTV (in cc).

### S3.3.2 Treatment Plan, PET, and MRI Viewers

We designed evaluation modules as stand-alone applications for which we wrote python scripts using pydicom and dicompylercore library to convert and co-register various DICOM files into NumPy arrays and then developed treatment plan viewers in R using Plotly library to independently incorporate them into the modules as shown in **Figures S2** and **S3**. For each treatment plan we took patient's CT slices, RT structures, and 3D Dose Distribution DICOM files,

1. performed pixel to co-ordinate transformation,
2. selected overlapping region between CT and 3D Dose array,
3. enlarged and interpolated the 3D dose images to match CT resolution via scipy.ndimage.zoom function,
4. combined all RT structures into one 3D grid by,
   (a) assigning unique integer value to the pixels of different structures and
   (b) adding the structures together, and saved all the transformed images into compressed 3D NumPy arrays.

In parallel, we extracted cumulative dose volume histogram (DVH) from 3D dose distribution DICOM files. In addition, to replicate the decision-making process in the adaptive RT clinical trials, we presented pre- and mid-treatment PET images in the NSCLC module and pre-treatment MRI in HCC module. We carried out the same procedure for registering PET and MRI images onto CT and RT Structure.

To reduce rendering time, we chose to show 2D slices of the compressed 3D NumPy arrays. To maintain the ease of

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

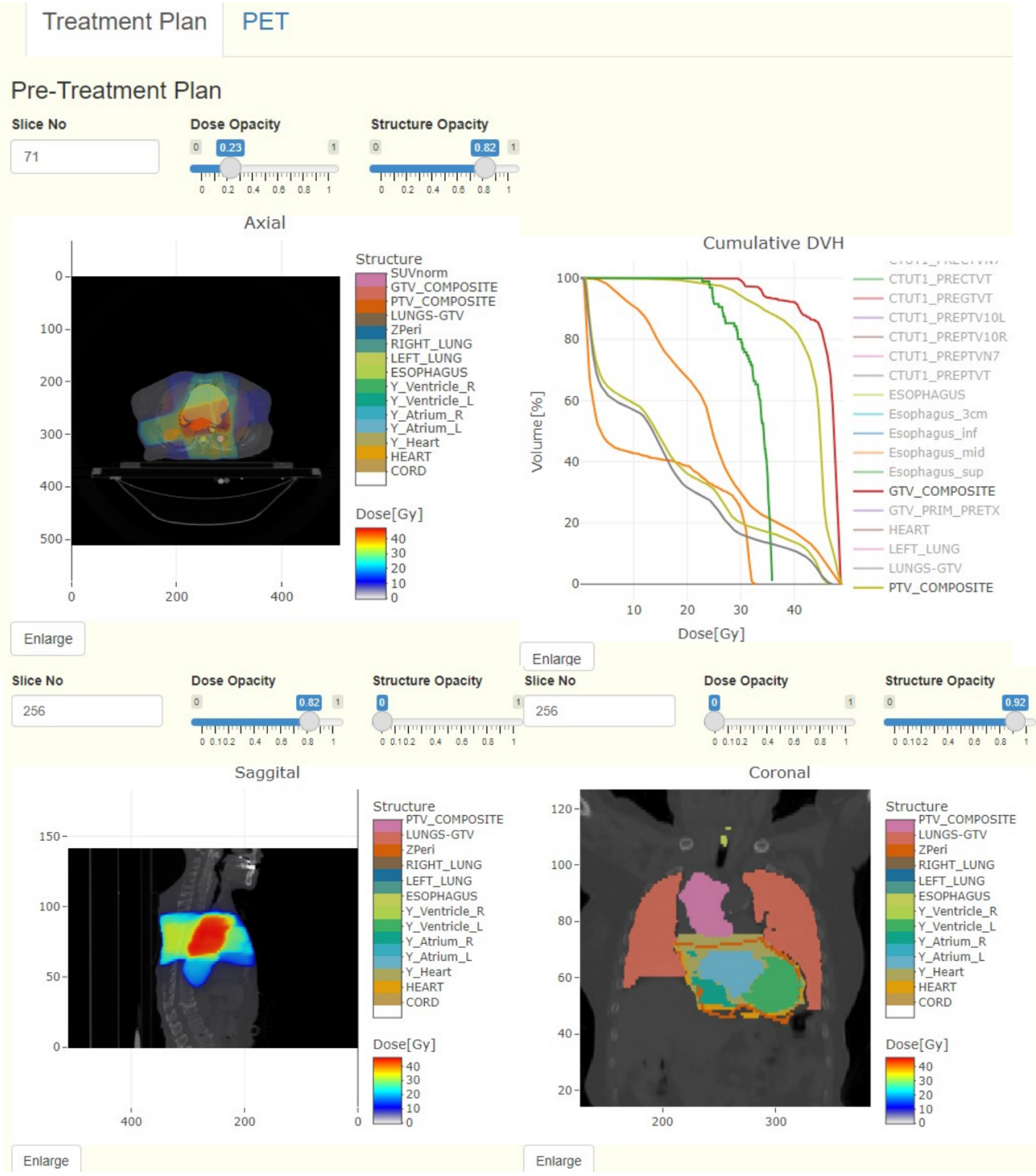



**Figure S2: Snippets of Unassisted page treatment plan viewer**. The viewer, built on Plotly library, includes Axial, Sagittal, and Coronal view of patients' superimposed CT scan, 3D dose distribution, and Structures, along with their cumulative DVH. Structure and dose values are color coded with discrete and continuous color map, respectively. In addition to the native Plotly controls, the viewer includes three extra controls: slice number input box, dose opacity slider, and structure opacity slider. To demonstrate the controls' functionality, the Axial view has been adjusted to show both dose and structure, the Sagittal to show only dose, and Coronal to show only structure. To demonstrate Plotly's native interactive control, the Coronal View has been zoomed with Plotly's zoom option which has been tweaked to preserve the physical aspect ratio. The native Plotly's option in cumulative DVH viewer also lets user select wanted or deselect unwanted histograms. Both NSCLC and HCC Evaluation module follow the same design.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

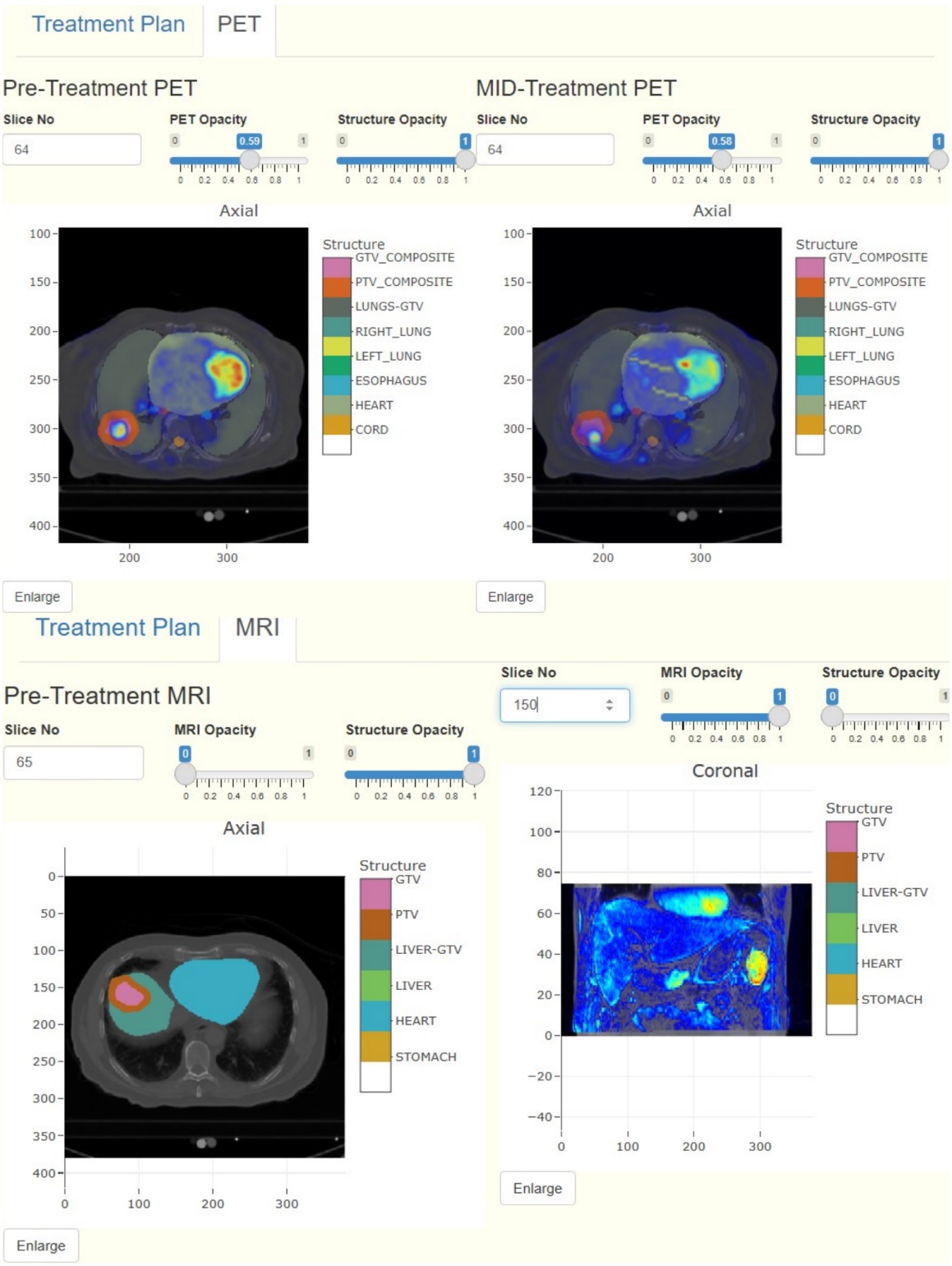



**Figure S3: Snippets of Unassisted page PET/MRI viewer**. Top row shows pre- and mid-treatment PET scans of NSCLC patient in Axial viewer placed in a side-by-side fashion for easy visual comparison. Not shown here are the Sagittal and Coronal Viewers. Bottom row shows pre-treatment MRI scans of HCC patients in Axial and Coronal viewer. Not shown here is the Sagittal Viewer. The viewers superimpose CT scan, PET/MRI image, and structure and includes three controls: slice number input box, PET/MRI opacity slider, and structure opacity slider. To demonstrate the controls' functionality, the top row viewers are adjusted to show both PET image and Structure, the Axial MRI viewer is adjusted to show only Structure, and the Coronal MRI viewer to show only MRI scans. The top row viewers have been zoomed in using Plotly's native interactive control which has been tweaked to preserve physical aspect ratio.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

viewing, we followed the standard procedure of showing images in Axial, Coronal, and Sagittal axis. In each viewer, in addition to native Plotly controls, we added three more controls:

1. slice number input box for selecting the slice number,
2. Dose/PET/MRI opacity slider for changing the intensity of the images, and
3. Structure opacity slider for changing the intensity of the structure.

The intensity of the CT background is left fixed while discrete color bars were included for Structure names and continuous color bar for Dose value. We separately presented the treatment plan and PET/MRI images in two tabs. We added a separate viewer for DVH in the treatment plan, in which wanted/unwanted histogram can be selected/unselected using the native Plotly controls. In NSCLC module's PET tab, for an easy comparison, we added three axial viewers for pre-treatment PET and another three for mid-treatment PET in a side-be-side fashion. In HCC module's MRI tab, we included three pre-treatment MRI images similar to the dose plan. We modified the native Plotly zoom controls to maintain the original aspect ratio and also included an Enlarge button for every viewer for viewing the images in a large pop-up Modal Dialog Box.

### S3.3.3 Questionnaire

We asked the following two questions:

i. Unassisted Dose Decision: Recommend a dose adaptation value between 1.5 to 4.0 Gy/frac (NSCLC) | 1.0 to 15.0 Gy/Frac (HCC) that best fits the current patient.
ii. Decision Confidence Level: On a level of 0 (lowest) to 5 (highest), how confident are you in your decision?

In addition, we included a textbox for remarks. Once satisfied with the inputs, evaluator could then click on the next button to submit the evaluation and go to the AI-assisted page.

## S3.4 AI-Assisted Page

**Figure S4** summarizes the AI-assisted page which consists of AI recommendation, outcome space, feature distribution, and questionnaires. We included a few help-push buttons through the page containing the following description.

### S3.4.1 AI Recommendation

We presented recommendations from an ensemble of 5 AI models. The AI-recommendation is provided as mean $\pm$ sem daily dose fractionation for the KBR-ART adaptive phase. In addition, we provided corresponding total equivalent dose in 2 Gy fractions (EQD2)—a more clinically meaningful metric.

### S3.4.2 Outcome Space

We provided outcome estimation from an ensemble of 5 AI models in the outcome space spanned by TCP and NTCP corresponding to adaptive daily dose fractionation value ranging from 1.5 to 4.0 Gy/frac for NSCLC and from 1.0 to 15.0 Gy/frac for HCC. The outcome estimates are color-coded continuously from yellow (lowest dose) to red (highest dose) color. The outcome corresponding to the AI-recommendation are marked with the green diamond markers and the model uncertainty (sem) is given as error bars in both TCP and NTCP direction. The background of the outcome space is colored according to the AI reward function, $r = tcp(1 - ntcp)$, which is highest at clinically desired outcome, $(tcp, ntcp) = (1, 0)$.

### S3.4.3 Feature Distribution

For improving interpretability, we presented distribution plots for all patient's features, where feature values are presented with population density in the background to provide "whereabouts" about the patient. Each feature plot could be enlarged by clicking on the background. Feature descriptions table from the work of Niraula et al.[22] including weblink to relevant literature were provided in the help box.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

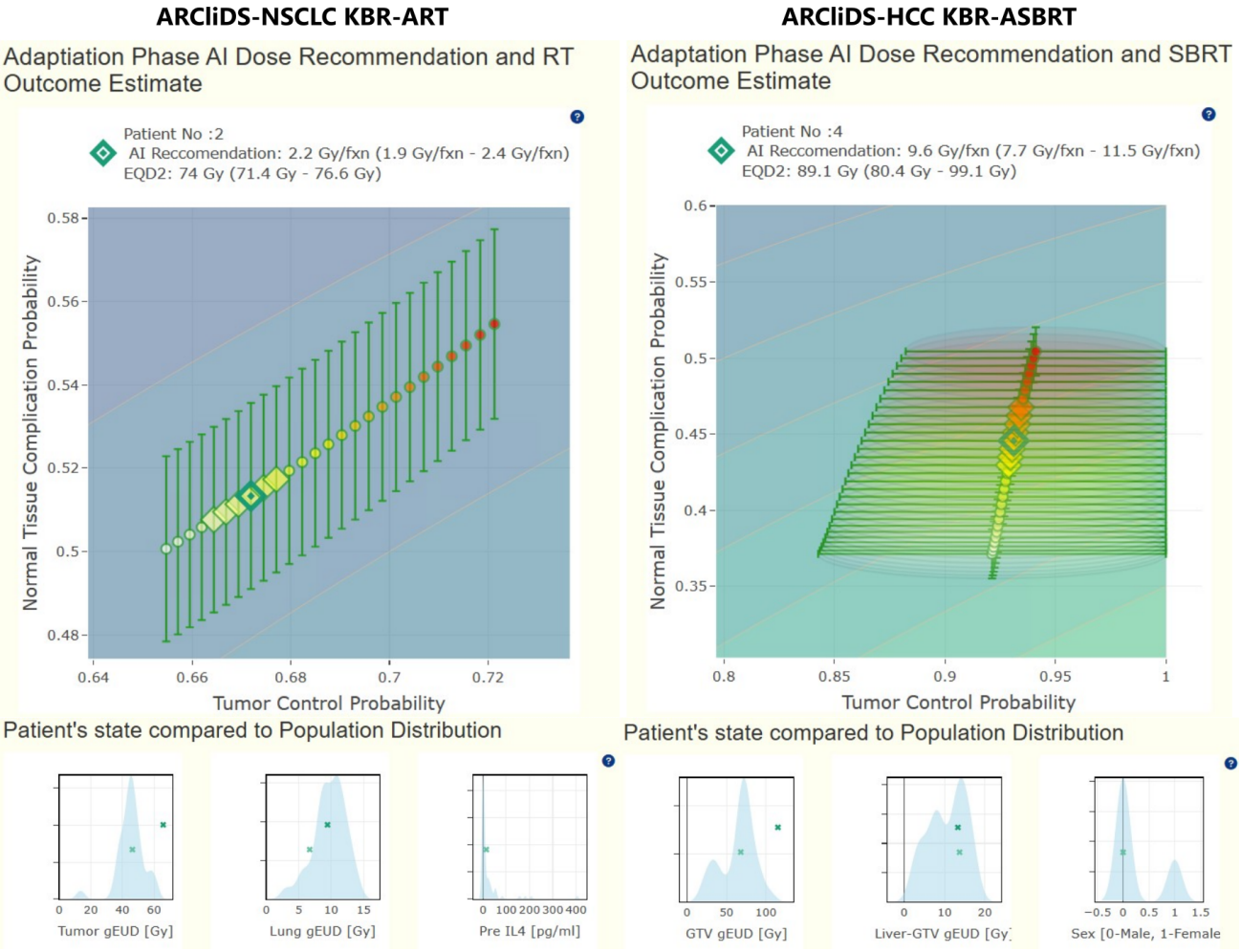



**Figure S4: Snippets of AI-assisted Pages showing ARCliDS recommendation, estimated outcome, and feature distribution panels**. An aggregation of 5 AI-models is included in the Evaluation Module. The AI-recommendation is provided as mean $\pm$ sem daily dose fractionation for the adaptive phase of KBR-ART/KBR-ASBRT treatment paradigm. In addition, corresponding total equivalent dose in 2 Gy fractions (EQD2) is provided. Estimated outcome is provided in the outcome space spanned by tumor control probability (TCP) and normal tissue complication probability (NTCP) for adaptive daily dose fractionation value ranging from 1.5 to 4.0 Gy/frac for NSCLC KBR-ART and from 1.0-15.0 Gy/frac for HCC KBR-ASBRT and color-coded continuously from yellow (lowest dose) to red (highest dose). The outcome corresponding to the AI-recommendation are marked with the green diamond markers and the model uncertainty (sem) is given as error bars in both TCP and NTCP direction. Note that data label class-imbalance can give rise to asymmetricity in uncertainty estimate. The background of the outcome space is colored according to the AI reward function, $r = tcp(1 - ntcp)$, which is highest at $(tcp, ntcp) = (1, 0)$. Patient's feature value is presented with population density in the background to provide "whereabouts" about the patient. Not shown here are the feature distribution plots for remaining 10 features. Help push buttons are included for additional information including feature descriptions. All the viewers are built using Plotly and comes with interactive native controls, such as zooming, selection, etc.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

### S3.4.4 Questionnaire

First, we re-asked the same two questions as the Unassisted phase as following:

i. AI-assisted Dose Decision: Having seen AI's prediction and recommendation, re-recommend a dose adaptation value between 1.5 to 4.0 Gy/frac (NSCLC EM) | 1.0 to 15.0 Gy/Frac (HCC EM) that best fits the current patient.
ii. Decision Confidence Level: On a level of 0 (lowest) to 5 (highest), how confident are you in your decision?

Second, to objectively quantify evaluators' trust level on ARCliDS recommendation, we asked following four multiple choice questions related to estimated outcome estimation and its associated uncertainty level and AI recommendation and its associated uncertainty level.

1. RT Outcome Estimation:
    i. How does the model's estimation for outcomes for the dosage range seem to you? [1pt]
        a. Reasonable
        b. Unrealistic
    ii. How does the range of uncertainty for outcomes seem to you? [1pt]
        a. Reasonable
        b. Unreasonable
2. AI Recommendation:
    i. What is your view on the AI Recommendation Dose Range? [2 pts]
        a. Agree
        b. Disagree-Go Higher
        c. Disagree-Go Lower
    ii. How does the range of uncertainty for AI dosage recommendation seem to you? [1pt]
        a. Reasonable
        b. Unreasonable: Too Small
        c. Unacceptable: Too Large

We chose the 4 multiple choice questions to total to 5 points, as same as the confidence level. We assigned 2 points to questions relating to evaluators view of the AI recommendation and 1 point to the rest. Only the first option, either Reasonable or Agree, carried the points and remaining options carried 0 points. Note that both the null options for AI-Recommendations, i.e. Disagree-Go Higher, Disagree-Go Lower, Unreasonable: Too Small, and Unacceptable: Too Large received 0 points.

We provided the following instruction/description in the help box.

AI Recommendation Trust Level (0-5): Calculated based on the following 4 questions.

1. RT Outcome Estimation
    i. Reasonableness of Estimation [1 pt]: Whether the outcome estimation is reasonable: if the shape of the path made by the tuple $(TCP, NTCP)$ makes sense, is physical, or is unrealistic or conflicting. For instance, both TCP and NTCP curves are assumed to monotonically increase with increasing dosage.
    ii. Uncertainty level [1pt]: Whether the Uncertainty envelope is reasonable or unacceptable: too small, too big, etc.
2. AI Recommendation
    i. Agreement with AI Recommended Dose Range [2 pts]: whether the evaluator agrees with the Recommended dose range i.e. mean $\pm$ sem or finds if the mean dose should be higher or lower.
    Help: Divide the TCP-NTCP outcome space into four quadrants which corresponds to four clinical outcome events. top-left: $(TC = 0, NTC = 1)$, top-right: $(TC = 1, NTC = 1)$, bottom-left: $(TC = 0, NTC = 0)$, and bottom-right: $(TC = 1, NTC = 0)$. Only the bottom-right: $(TC = 1, NTC = 0)$ is clinically desirable.
    ii. Uncertainty level [1pt]: See if the Uncertainty level (sem) in AI Recommended dosage is acceptable or unacceptable. It could be clinically unacceptable if it's too large or unreasonable if it's too small.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

# S4 Module Deployment

We deployed both NSCLC and HCC evaluation modules on shinyapps.io servers. Prior to deployment, we obtained institutional cyber security clearance for hosting the modules externally, which took about 2 weeks from filling out an application to getting clearance but took was little over 4 months in total to navigate to the application in the first place.

We purchased the standard shinyapps.io subscription capable of hosting a maximum of 5 instances (virtualized server, docker containers) at a time and at most 8 GB per instance, meaning that we had to take measures to limit module size to well below 8GB. We preprocessed DICOM images and saved 3D NumPy arrays in float16 data format, then we opted to show 2D image slices instead of 3D volumes, we precalculated and saved AI-recommendation and outcome estimation for all the patients, and we used Google Sheets to auto-save all the evaluation data instead of putting load to the server memory. Besides R, we had to set up Python interpreter during deployment.

# S5 List of Software and Library

| RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Python (v3.9.12) | Python Software Foundation | RRID:SCR_008394;https://www.python.org/ |
| NumPy (v1.21.5) | Python package | RRID:SCR_008633; http://www.numpy.org |
| SciPy (v1.7.3) | Python package | RRID:SCR_008058; http://www.scipy.org/ |
| Pandas (v1.4.2) | Python package | RRID:SCR_018214; https://pandas.pydata.org |
| PyTorch (v1.11.0) | Python package | RRID:SCR_018536; https://pytorch.org/ |
| PyTorch Geometric (v2.0.4) | Python package | https://pypi.org/project/torch-geometric/ |
| PyDicom (v2.3.1) | Python package | RRID:SCR_002573; https://pydicom.github.io/ |
| Dicompyler-Core (v0.5.5) | Python package | https://github.com/dicompyler/dicompyler-core |
| Scikit-Image (v0.19.3) | Python package | RRID:SCR_021142; https://scikit-image.org/ |
| R (v4.2.1) | R Foundation | RRID:SCR_001905; http://www.r-project.org/ |
| Shiny (v1.8.0) | R package | RRID:SCR_001626; http://www.rstudio.com/shiny/ |
| Reticulate (v1.34.0) | R package | https://rstudio.github.io/reticulate/ |
| Plotly (v4.10.3) | R package | RRID:SCR_013991; https://plotly.com/r/ |
| Googlesheets4 (v1.1.1) | R package | https://github.com/tidyverse/googlesheets4 |
| GoolgeDrive (v2.1.1) | R package | https://github.com/tidyverse/googledrive |
| shinyapps.io | Posit | https://www.shinyapps.io/ |
| Power Point (v.2401) | Microsoft Crop. | RRID:SCR_023631 |

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

# S6 Enrollment, Training, and Evaluation

We advertised the study via a combination of group emails and in-person invitations in two institutions: Moffitt Cancer center and Michigan Medicine. Initially, a total of 20 volunteers showed interest and initial pre-evaluation information sessions were conducted, mostly in a one-on-one virtual meeting, two in-person meetings, and one group virtual meeting with three volunteers. Initially, we found a bug in the software, which was corrected, and the five evaluators who had taken the evaluation were requested to retake the evaluation to which only two out of five agreed, and thus the remaining 3 evaluations were discarded. Similarly, three evaluators didn't follow up and one evaluator only completed evaluation for one patient. In total, 13 evaluators completed the evaluation, and 4 evaluators volunteered to take both NSCLC and HCC evaluation, with a total of 17 completed evaluations (9 NSCLC, and 8 HCC).

As shown in **Table S2**, the evaluators consisted of both physicians and residents from a variety of specializations and a range of experiences, out of which two evaluators were residents during the initial training/meeting and physicians during the evaluation and were accordingly grouped in a special classification in **Table S2**. We have de-identified the name and institution and assigned numbers for those who completed the evaluation and alphabets for the rest.

For consistency, during the initial pre-evaluation information sessions, all evaluators were shown tutorial video in addition to a demonstration. The evaluators then completed the evaluation in their own time and technical support was provided whenever needed, either via zoom, email, or phone call. The evaluation study took a little over five months from the first advertisement to the completion of the last evaluation. Altogether, 144 set of decision samples were collected from 17 evaluations on 17 patients: for NSCLC, we collected 72 datapoints (8 patients $\times$ 9 evaluators) and for HCC, we collected 72 datapoints (9 patients $\times$ 8 evaluators). Data analysis was conducted after the end of the data collection period.

**Table S2: Evaluators Summary**

| NSCLC | | | | | | HCC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **SN** | **Eval ID** | **Exp (yrs)** | **Spec** | **Inst** | **Status** | **SN** | **Eval ID** | **Exp (yrs)** | **Spec** | **Inst** | **Status** |
| **Residents** | | | | | | | | | | | |
| 1 | 1 | 4 | Breast | 1 | Complete | 13 | 3 | 4 | General | 1 | Complete |
| 2 | 3 | 4 | General | 1 | Complete | 14 | 6 | 2 | General | 2 | Complete |
| 3 | 4 | 2 | General | 2 | Complete | 15 | 7 | 3 | General | 2 | Complete |
| 4 | 6 | 2 | General | 2 | Complete | 16 | 13 | 4 | General | 1 | Complete |
| 5 | 7 | 3 | General | 2 | Complete | 17 | D | 3 | General | 2 | Didn't Retake |
| **Resident during training/Physician during Evaluation** | | | | | | | | | | | |
| 6 | 2 | 5 | Prostate | 1 | Complete | 18 | 12 | 5 | GI | 1 | Complete |
| **Physicians** | | | | | | | | | | | |
| 7 | 5 | 16 | CNS | 1 | Complete | 19 | 5 | 16 | CNS | 1 | Complete |
| 8 | 8 | 21 | Prostate | 1 | Complete | 20 | 10 | 14 | Liver | 2 | Complete |
| 9 | 9 | 17 | Lung | 1 | Complete | 21 | 11 | 7 | GI | 1 | Complete |
| 10 | A | 30 | GU | 1 | Didn't retake | 22 | E | 24 | GI | 2 | Didn't retake |
| 11 | B | 8 | Thoracic | 1 | Didn't Complete | 23 | F | - | Liver | 1 | Didn't Take |
| 12 | C | - | Lung | 1 | Didn't Take | 24 | G | - | GI | 1 | Didn't Take |

**Table S2: Summary of volunteer evaluators** and their relevant information including those who completed the evaluation, didn't take evaluation after initial training/meeting, and those who took the evaluation initially but declined to retake after discovery of a bug and debugging. The name of evaluators and institution have been de-identified. Four out of 12 evaluators volunteered to take both evaluations hence 17 evaluations by 12 evaluators. Evaluators who completed the evaluation has been de-identified with numbers and rest with alphabets. Note the special classification for the two evaluators who were resident during the initial training/meeting and went on the pass their board certification and formally accepting physician position before completion of evaluation. **Abbreviation**: **Eval ID**: Evaluator Identification number, **Exp**: Experience, **Spec**: Specialization, **Inst**: Institution.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

# S7 Statistics

## S7.1 Matched Pair Randomization T-test

Since both patients' and evaluators' sample size were small, we used randomization test to investigate the level of AI-influence in decision-making. Randomization test14 is a non-parametric test that does not assume random sampling, normality of the population, or estimates population parameters such as mean and variance. In randomization, a test statistic is calculated from the observed data then compared with the distribution of the test statistics obtained from resampling the data as opposed to comparing with the standard distribution. Exchangeability under null is the central idea of randomization test, which roughly states that exchanging data under null should not alter the sample statistics significantly. We performed a two-tailed matched pair randomization t-test on: $H_0 : \Delta = 0$, $H_\alpha : \Delta \neq 0$; $\Delta \equiv un - aia$, where the null hypothesis states that AI has no significant influence on the clinical decisions, or alternatively if AI fails to influence the decision, there should be no difference between unassisted decision and AI-assisted decision. In our case, the exchangeability under null translates to if AI had no influence, then an unassisted decision could have equally likely come from the group of ai-assisted decisions. The test was carried out following the codes from David C Howell.[40]

## S7.2 Correlation Analysis

Besides hypothesis testing, we conducted correlation analysis between several observed and derived variables. We primarily used Spearman rank correlation ($\rho$) which quantifies the monotonic relationship between two variables. Unlike Pearson correlation ($r$), which quantifies the linear relationship, spearman correlation is less sensitive to extreme values (outliers). In addition, we report p-value for correlation coefficient, which corresponds to hypothesis test, $H_0 : \rho = 0$, $H_\alpha : \rho \neq 0$; where the null hypothesis states that the correlation does not significantly differs from zero.

In contrast, we use Pearson correlation and scatter plot to investigate the hypothesis $H_0 : un = aia$, $H_\alpha : un \neq aia$ and $H_0 : un\,conf = aia\,conf$, $H_\alpha : un\,conf \neq aia\,conf$. We know that, under null, $un$ and $aia$ should show perfect (or near perfect) correlation and the tuple $(un, aia)$ should distribute close to the $un = aia$ line. Additionally, under null, the best linear fit line to the data should coincide with the null hypothesis line, $un = aia$. Same applies to the tuple $(un\,conf, aia\,conf)$.

## S7.3 Derived Quantities

In addition to the observed variables such as un,aia,un conf,aia conf, and AI trust, we derived number of quantities to investigate collaborative decision-making process.

1. *Decision adjustment frequency* is the number of cases where the decisions were adjusted after AI access, i.e. $\sum_{i=1}^{n} \delta_{un_i\,aia_i}$ , where $\delta_{un_i\,aia_i} = 0$ if $un_i = aia_i$ else 1. Decision adjustment frequency is zero if all unassisted decisions are pairwise equal to AI-assisted decision.

2. *Decision adjustment level* is the difference between AI-assisted and Unassisted decision, measured in Gy/fx, i.e. $(aia - un)$.

3. *Dissimilarity in decision-making with AI* is the difference between AI recommendation and Unassisted decision, measured in Gy/fx, i.e. $(ai - un)$.

4. *Agreement with AI* is the additive inverse of the absolute difference between AI-assisted Decision and AI recommendation, i.e. $-|aia - ai| \in (-\infty, 0]$. The level of agreement peaks at 0 which corresponds to the absence of difference between $aia$ and $ai$, and decreases when the difference between $aia$ and $ai$ increases in either direction.

5. *Closeness to Standard of Care* is the additive inverse of the absolute difference between decision and $SOC$, i.e. $-|d - SOC| \in (-\infty, 0]$, where $d \in \{un, aia\}$. Closeness peaks at 0 when $d = SOC$ and decreases when the difference between decision and $SOC$ increases in either direction.

## S7.4 Intraclass correlation coefficient

We performed a concordance analysis on the decisions by comparing the Intraclass correlation coefficient (ICC)[41–43] of $un$ and $aia$. We chose McGraw and Wong's formulation of ICC[42] for this study and assumed a two-way random

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

effect linear model following the fact that both patients and evaluators were chosen at random from a larger pool, i.e. $d_{ij} = \mu + a_i + b_j + ab_{ij} + e_{ij}$, $d \in \{un, aia\}$, $i \in \{1, ..., n\}$, $j \in \{1, ..., k\}$, where $n$ in the number of patients, $k$ is the number of evaluators, $\mu$ is the population mean, $a_i$ is the inter-patient heterogeneity, $k$ is inter-evaluator (physician) variability, $ab_{ij}$ is the patient-evaluator interaction variability and $e_{ij}$ is the random error. For completeness we calculated both ICC types: Consistency ($C$) and Absolute Agreement ($A$); for both units: Single rater (1) and Average rater ($k$), resulting in four combinations: $ICC(C, 1)$, $ICC(A, 1)$, $ICC(C, k)$, and $ICC(A, k)$. ICC type Consistency measures the symmetric differences between the decisions of the evaluators, and Absolute Agreement measures the absolute differences, making the latter a stricter form of coefficient. Similarly, ICC unit Single rater corresponds to using the decision from a single evaluator as the basis for measurement and $ICC$ unit Average corresponds to using the average decision from k evaluators. The absolute value of ICC is study-dependent and hard to interpret on its own. However, comparing ICC of pairwise variables from the same study under identical conditions is meaningful. Thus, regardless of the strength of ICC, we can draw a conclusion on the inter-evaluator agreement by comparing ICC between decisions made by the same group of evaluators without and with AI-assistance. We used irr R-package to compute the ICCs.

## S7.5 Toxicity Free Local Control Scoring Schema

In the absence of ground truth, we analyzed the adjusted decision ($un \neq aia$) based on a scoring schema $TCP(1 - NTCP) \in [0, 1]$, which reflects the clinical goal of achieving toxicity free local control. Mathematically, the scoring schema is the likelihood of achieving tumor control without a normal tissue complication ($1-NTCP$). It has a maximum value of 1 for ideal outcome of $(tcp, ntcp) = (1, 0)$ and minimum value of 0 for dose limiting factor, $ntcp = 1$. In addition, the scoring function is a quadratic function having a higher sensitivity than a probability metric, for instance, the scoring function for TCP and NTCP in percentages would be $TCP(100 - NTCP)$ and would range from 0 to 10,000, thus we use four decimal places for analysis.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

# S8 Decision Adjustment Level

The magnitude of decision adjustment level ($aia - un$) in Gy/fx for NSCLC and HCC grouped by individual evaluator and patient is presented in **Figure S5**. (Supporting data for **section 2.1**)

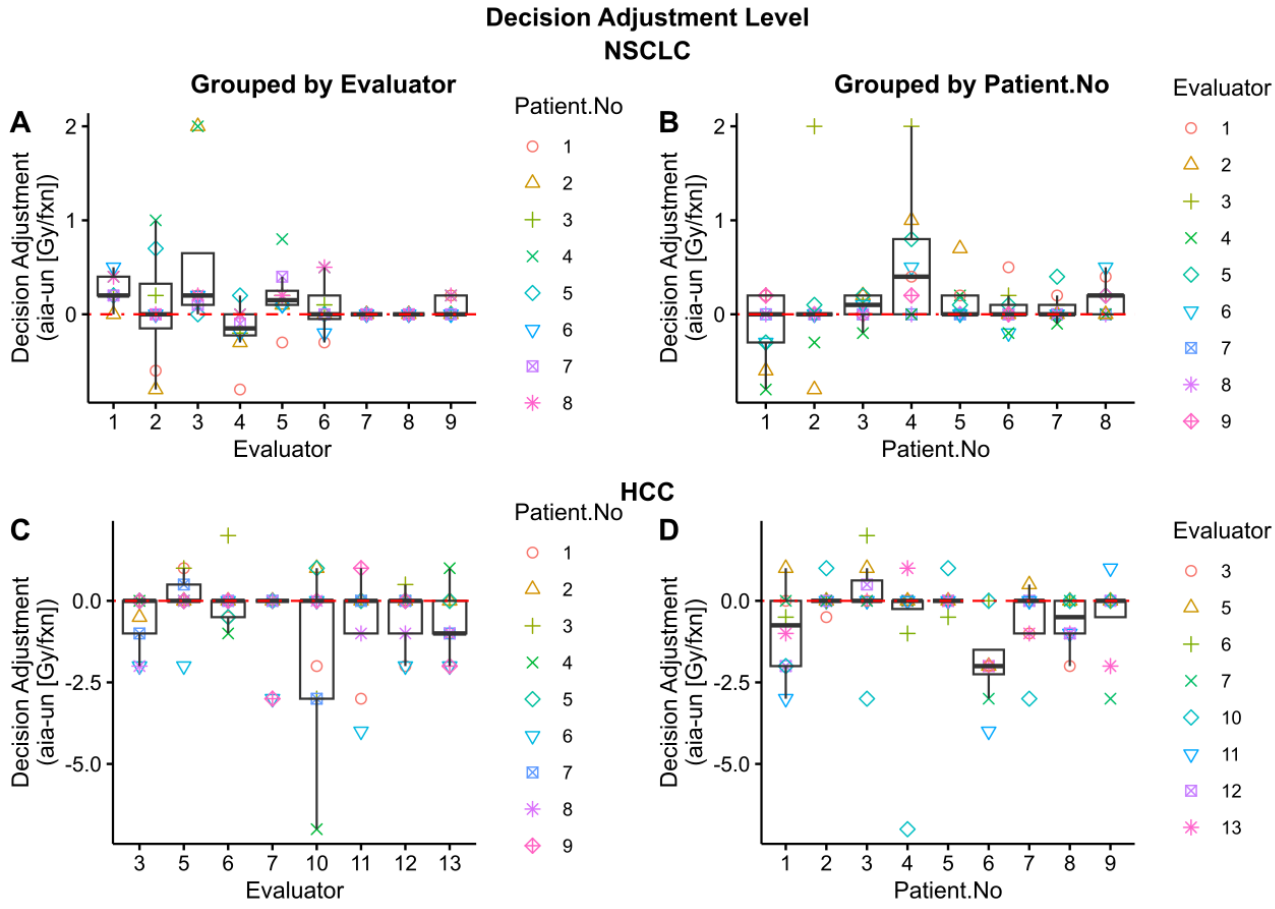



**Figure S5: Decision Adjustment Level grouped by Evaluators and Patient Number**. Box plots **A** and **B** summarizes decision adjustment for NSCLC and box plots **C** and **D** for HCC. The plots grouped by evaluators are marked and color coded by patients and vice versa.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

# S9 Correlation between Unassisted and AI-assisted Decision and Decision Confidence.

We investigated the correlation between $un$ and $aia$. Under null, $un$ and $aia$ should show perfect (or near perfect) correlation and the tuple $(un, aia)$ should distribute close to the $un = aia$ line; however, as shown in **Figures S5A** and **S5C**, for NSCLC, we found Pearson correlation coefficient $(r)$ of 0.39 $(p < 0.001)$ and spearman correlation coefficient $(\rho)$ of 0.61 $(p < 0.001)$, and for HCC, $r = 0.74\,(p < 0.001)$ and $\rho = 0.77\,(p < 0.001)$ which are not near perfect. Pearson coefficient represents level of linear relationship between $un$ and $aia$ however, unlike spearman rank correlation, it is very sensitive to extreme value which is seen from the drastic difference between NSCLC $r$ and $\rho$. In addition, we investigated evaluators who adjusted at least one of their decisions. As expected, we found a decrease in the correlation value as shown in **Figure S5B**, i.e. $r = 0.35\,(p = 0.0083)$ and $\rho = 0.57\,(p < 0.001)$. The best fit line for both cases did not align with null hypothesis line $un = aia$. The combination of results from hypothesis testing, frequency analysis, and correlation analysis indicated that AI influences decisions on a case-by-case basis.

Then, we investigated the relationship between evaluators self-reported Unassisted decision confidence and AI-assisted decision confidence. Under null (no AI-influence), we would expect $aia\,conf$ and $un\,conf$ to show a perfect (or near perfect) correlation and the tuple $(un\,conf, aia\,conf)$ to distribute close to the $un\,conf = aia\,conf$ line. As shown in **Figures S5D**, and **S5F**, we found a positive correlation between $aia\,conf$ and $un\,conf$ decision: for NSCLC, $r = 0.81\,(p < 0.001)$ and $\rho = 0.76\,(p < 0.001)$ and for HCC, $r = 0.31\,(p = 0.0073)$ and $\rho = 0.34\,(p = 0.0037)$. The correlation coefficient is not as strong as we would expect for the null case. Moreover, for NSCLC excluding evaluators with zero decision adjustment, we obtained $r = 0.5\,(p < 0.001)$ and $\rho = 0.53\,(p < 0.001)$, as shown in **Figure S5E**, which is considerably less than that of the overall NSCLC, consistent with the alternative hypothesis.

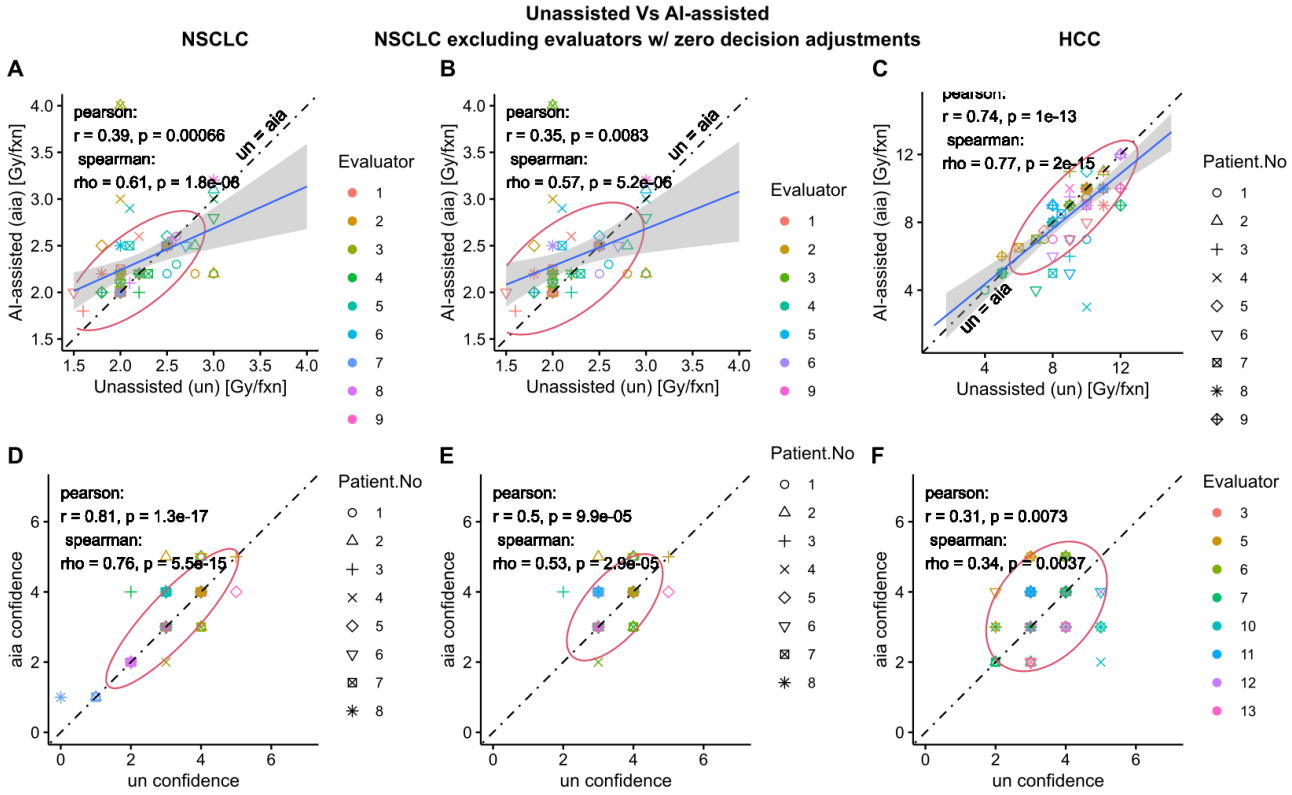



**Figure S6: Analysis of decision confidence and confidence level**. The first, second, and third column of figures correspond to NSCLC, NSCLC excluding Evaluators with zero decision adjustment, and HCC, respectively. All evaluators in HCC adjusted at least one of their decisions. Scatter plots **A**, **B**, and **C** show the relationship between unassisted $(un)$ and AI-assisted decision $(aia)$ and present the Pearson and Spearman correlation coefficient with p-value, covariance ellipse (95% confidence), a solid linear fit line and uncertainty, and the dot-dashed null-hypothesis line $un = aia$, which represents absence of AI-influence. The $ai - un = 0$ line corresponds to complete agreement between unassisted decision and AI-recommendation, whereas $aia - un = 0$ line corresponds to absence of decision change. 2D Scatter plots **D**, **E**, and **F** show the relationship between evaluators' unassisted decision $(un)$ confidence and AI-assisted decision $(aia)$ confidence (scaled 0-5, 5 being highest).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

# S10 Individual AI-Trust Level Contribution vs Agreement with AI-recommendation

The self-reported AI trust level is composed of 4 components as shown in **Figure S7**.

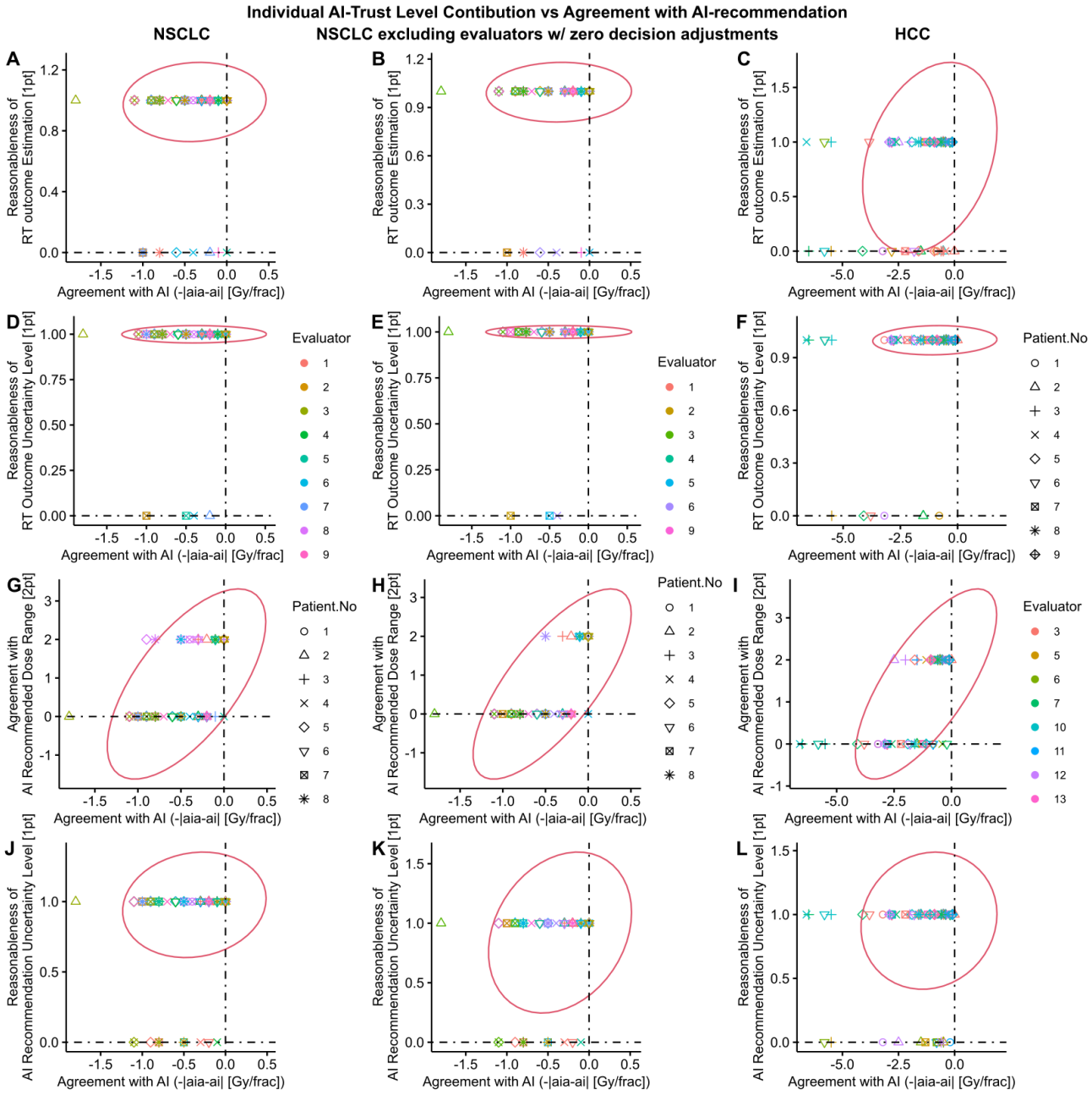



**Figure S7: Analysis of individual contribution to AI Trust level with respect to agreement between AI-assisted Decision and AI-recommendation**. The first, second, and third column of figures correspond to NSCLC, NSCLC excluding Evaluators with zero decision adjustment, and HCC, respectively. All evaluators in HCC adjusted at least one of their decisions. Evaluator's level of trust on AI (scaled 0-5, 5 being the highest) was determined based off of their self-reported agreement with AI's RT outcome estimate (first two rows) and AI-recommendation (last two rows) which was further divided into average trend (first and third row) and uncertainty estimate (second and last rows). The individual contributions were Reasonableness of RT outcome Estimation (1 pt), Reasonableness of RT outcome Uncertainty Level (1pt); Agreement with AI recommended Dose range (2pts), and Reasonableness of AI Recommendation Uncertainty Level (1pt). For visual insight, covariance ellipses are included.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint
**(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

# S11 Evaluators' Remark and Authors' Summary

**Table S3** presents evaluators' remarks for NSCLC and **Table S4** presents evaluators' remark for HCC. Note that only fraction of evaluators provided remarks, thus the table doesn't contain the entire data.

## Table S3: Evaluators' Remark for NSCLC - I

| Pat | Eval | un Gy/fx | ai Gy/fx | ai TCP | ai NTCP | aia Gy/fx | un conf | aia conf | ai trust | RTOE: avg | RTOE: un-cert | ai: avg | ai: un-cert | Unassisted Remark | Unassisted Remark Authors Summary | AI-assisted Remark | AI-assisted Remark Authors Summary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 9 | 2 | 2.2 | 0.685 | 0.631 | 2.2 | 4 | 3 | 5 | 1 | 1 | 2 | 1 | concerned about hypofractionation due to the fact that the mid-tx PET shows disease right at the esophagus. | concerned about OAR (esophagus) | - | - |
| 2 | 9 | 2 | 2.2 | 0.669 | 0.511 | 2 | 4 | 4 | 3 | 1 | 1 | 0 | 1 | disease remains peri-esophageal at the mid-tx PET | concerned about OAR (esophagus) | Can achieve same TCP/NTCP at 2 Gy/fx, so why suggest something higher? I'm worried about the accuracy of the NTCP calculation. By escalating to recommended dose, the TCP is still the same, so why risk it? | relation between RTOE and AI-recommendation seems inconsistent, concerned about RTOE-NTCP accuracy |
| 3 | 2 | 1.8 | 2.1 | 0.604 | 0.489 | 2 | 5 | 5 | 5 | 1 | 1 | 2 | 1 | High V20 | concerned about OAR (lung) | - | - |
| | 4 | 2.2 | | | | 2 | 2 | 4 | 5 | 1 | 1 | 2 | 1 | The plan appears to be inadequately covering the primary, so more than adapting the dose of this plan up or down, I would advocate for the plan to be modified to improve coverage of the primary. | Anatomical Adaptation instead of dose adaptation | - | - |
| | 9 | 2 | | | | 2 | 4 | 4 | 4 | 0 | 1 | 2 | 1 | tumor is peri-esophageal (station 7 on pre-tx PET. On post-tx PET, station 7 continues to enhance. Pushing the dose would be very unwise! | concerned about OAR (esophagus). Disagree with dose escalation. | Esophagus is VERY sensitive to RT and is typically dose-limiting structure. I think the NTCP is too low here, if anything, looking at the estimates at the higher dose per fraction. | concerned about RTOE-NTCP accuracy w.r.t. OAR (esophagus) |
| 4 | 3 | 2 | 2.9 | 0.56 | 0.51 | 4 | 3 | 2 | 2 | 1 | 1 | 0 | 0 | Unclear why 2.5 Gy/fx is recommended as the base evaluation dose in this scenario, this is not standard for practice anywhere to my knowledge. Standard dose/fractionation is 60-66 Gy in 2 Gy daily fractions. Additionally, the DVH does not make sense to me, if 2.5 Gy/fx is recommended, the PTV should be getting 50 Gy however the DVH shows that almost none of the PTV is receiving 50 Gy, and only 95% is getting 36 Gy. | Disagree with Evaluation Phase dose fractionation and potential inconsistency with DVH | Again, reason for AI recommendation appears unclear. Absolute difference between TCP and NCTP continues to increase with increasing dose/fx according to this model - why not go higher then? | relation between RTOE and AI-recommendation seems inconsistent; AI-recommendation must have been higher according to the RTOE |

**Abbreviation- Pat**: Patient No; **Eval**: Evaluator; **un**: Unassisted Decision; **ai**: AI recommendation; **ai TCP**: Tumor Control Probability for AI recommended dose; **ai NTCP**: Normal Tissue Complication Probability for AI recommended dose; **aia**: AI-assisted Decision; **un conf**: Unassisted Decision Confidence; **aia conf**: AI-assisted Decision Confidence; **ai trust**: AI recommendation trust level; **RTOE: avg**: Reasonableness of RT outcome estimation; **RTOE: uncert**: Reasonableness of RT Outcome Uncertainty Level; **ai:avg**: Agreement with AI Recommended Dose Range; **ai: uncert**: Reasonableness of AI Recommendation Uncertainty Level; **OAR**: Organs at Risk from RT; **DVH**: Cumulative Dose Volume Histogram



medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

**Table S3: Evaluators' Remark for NSCLC - II**

| Pat | Eval | un Gy/fx | ai Gy/fx | ai TCP | ai NTCP | aia Gy/fx | un conf | aia conf | ai trust | RTOE: avg | RTOE: un-cert | ai: avg | ai: un-cert | Unassisted Remark | Unassisted Remark Authors Summary | AI-assisted Remark | AI-assisted Remark Authors Summary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 4 | 3 | 2.9 | 0.56 | 0.51 | 3 | 4 | 4 | 4 | 1 | 1 | 2 | 0 | I'm somewhat confused by the DVH since I was expecting to see coverage with 50Gy after 2.5Gy over 20 fractions in the evaluation phase. But since it looks like the tumor only got ¡40 Gy over the first 20 fractions, has persistent uptake on PET, and there is some room on OARs (esp lungs-GTV), I am comfortable escalating dose. | potential inconsistency with DVH, potential dose escalation | A 4% increase in NTCP for an additional 10Gy seems low to me, especially with such small error bars. | concerned about RTOE-NTCP accuracy: low risk (NTCP) over risk (dose) |
| | 6 | 2 | | | | 2.5 | 3 | 3 | 1 | 0 | 0 | 0 | 1 | - | - | Seems to have a very tight range for NTCP despite a wide dose range | concerned about RTOE-NTCP accuracy: low risk (NTCP) over risk (dose) |
| | 9 | 2 | | | | 2.2 | 3 | 3 | 3 | 1 | 1 | 0 | 1 | Had trouble reading the mid-PET (pixelated). Mid-esoph dose very high on the DVH curve, so this is a case to "slow down" (but I don't have data to say you can go below 2 Gy/fx without adversely affecting clinical control) | Trouble with PET treatment Planner, OAR dose high, potential dose de-escalation | - | - |
| 5 | 4 | 1.8 | 3.1 | 0.676 | 0.647 | 2 | 3 | 3 | 3 | 1 | 1 | 0 | 1 | Patient appears to have had a good response to therapy at the 20 fraction mark. Given that many critical OARs have received substantial dose from the first 20 fractions (Lung V20, esophagus, etc), I would prioritize protecting OARs over escalating dose to tumor. In a clinical setting, I would also consider the patient's clinical status (overall tolerance of treatment, degree of dysphagia, etc) | Positive response by patient during KBRART evaluation phase. OAR. Consider other clinical factors. | Having this calculation is helpful to me in weighing the risks and benefits of dose adaptation, but I would like to know additional details. For example, if this algorithm was trained on patients treated after durva was incorporated into the standard of care, after which the pneumonitis risk has been higher. I am also not certain which organs are accounted for in the NTCP- to what degree is this accounted for by lung, esophagus, bronchus, etc. | uncertain about RTOE-NTCP modeling wrt to OAR lung, esophagus, bronchus, and incorporation of chemo information |
| | 7 | 2.2 | | | | 2.2 | 3 | 3 | 3 | 1 | 1 | 0 | 1 | reasonable amount of residual disease | Positive response by patient KBR-ART evaluation phase | - | - |
| | 9 | 2 | | | | 2 | 5 | 4 | 3 | 1 | 1 | 0 | 1 | - | - | I worry about esophageal toxicity with the proposed dose! | concerned about OAR (esophagus) toxicity for the recommendation |

**Abbreviation- Pat**: Patient No; **Eval**: Evaluator; **un**: Unassisted Decision; **ai**: AI recommendation; **ai TCP**: Tumor Control Probability for AI recommended dose; **ai NTCP**: Normal Tissue Complication Probability for AI recommended dose; **aia**: AI-assisted Decision; **un conf**: Unassisted Decision Confidence; **aia conf**: AI-assisted Decision Confidence; **ai trust**: AI recommendation trust level; **RTOE: avg**: Reasonableness of RT outcome estimation; **RTOE: uncert**: Reasonableness of RT Outcome Uncertainty Level; **ai:avg**: Agreement with AI Recommended Dose Range; **ai: uncert**: Reasonableness of AI Recommendation Uncertainty Level; **OAR**: Organs at Risk from RT; **DVH**: Cumulative Dose Volume Histogram



medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

**Table S3: Evaluators' Remark for NSCLC - III**

| Pat | Eval | un Gy/fx | ai Gy/fx | ai TCP | ai NTCP | aia Gy/fx | un conf | aia conf | ai trust | RTOE: avg | RTOE: un-cert | ai: avg | ai: un-cert | Unassisted Remark | Unassisted Remark Authors Summary | AI-assisted Remark | AI-assisted Remark Authors Summary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 4 | 3 | 2.2 | 0.458 | 0.424 | 2.8 | 3 | 3 | 3 | 1 | 1 | 0 | 1 | - | - | I think there should be error bars going in both directions. Some graphs have had error bars regarding NTCP, others have had error bars for TCP. | concerned about RTOE error bars |
| | 9 | 2.5 | | | | 2.5 | 3 | 3 | 3 | 1 | 1 | 0 | 1 | There is enhancement where the esophagus lies (perhaps evidence of esoph irritation/edema and not presumptively not tumor since not positive ot pre-tx pet. Esoph is on the edge of the tx field. I presume it might be ok to continue at same dose or perhaps dose escalate a little, if esophagus spared, but don't have much clinical experience. | potential dose escalation if OAR spared | Based on this graph I probably would push to 2.5 Gy/fx, not 2.25 (1% increase in NTCP and TCP) by doing so – trade-off seems reasonable (if those estimates are accurate, of course) | relation between RTOE and AI-recommendation seems inconsistent |
| 7 | 2 | 2 | 3 | 0.708 | 0.653 | 2 | 4 | 4 | 1 | 0 | 0 | 0 | 1 | - | - | Although constraints are met, target volumes are big for hypofractionation to 3Gy or higher. | target volumes are large for hypofractionation |
| | 3 | 2 | | | | 2.1 | 4 | 3 | 3 | 1 | 1 | 0 | 1 | - | - | Again, don't understand why recommendation is not for largest absolute difference between TCP and NTCP | relation between RTOE and AI-recommendation seems inconsistent |
| | 9 | 2 | | | | 2 | 4 | 4 | 2 | 0 | 1 | 0 | 1 | I'm really worried about esoph toxicity, looking at the treatment plan. I have no data re: reducing dose per fraction below 2 Gy, so will say 2 Gy, but this patient is going to clinically suffer if dose escalated! Mid-tx PET shows peri-esophageal PET positivity! | concerned about OAR (esophagus). Potentially dose deescalate | - | - |

**Abbreviation- Pat**: Patient No; **Eval**: Evaluator; **un**: Unassisted Decision; **ai**: AI recommendation; **ai TCP**: Tumor Control Probability for AI recommended dose; **ai NTCP**: Normal Tissue Complication Probability for AI recommended dose; **aia**: AI-assisted Decision; **un conf**: Unassisted Decision Confidence; **aia conf**: AI-assisted Decision Confidence; **ai trust**: AI recommendation trust level; **RTOE: avg**: Reasonableness of RT outcome estimation; **RTOE: uncert**: Reasonableness of RT Outcome Uncertainty Level; **ai:avg**: Agreement with AI Recommended Dose Range; **ai: uncert**: Reasonableness of AI Recommendation Uncertainty Level; **OAR**: Organs at Risk from RT; **DVH**: Cumulative Dose Volume Histogram



medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**Table S3: Evaluators' Remark for NSCLC - IV**

| Pat | Eval | un Gy/fx | ai Gy/fx | ai TCP | ai NTCP | aia Gy/fx | un conf | aia conf | ai trust | RTOE: avg | RTOE: un-cert | ai: avg | ai: un-cert | Unassisted Remark | Unassisted Remark Authors Summary | AI-assisted Remark | AI-assisted Remark Authors Summary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8 | 3 | 2 | 3 | 0.669 | 0.634 | 2.2 | 4 | 4 | 2 | 1 | 1 | 0 | 0 | - | - | I chose 2.2 Gy/fx as it provided the largest absolute difference between TCP and NCTP (65% vs 61%) of the information provided. Increasing to the recommended dose level of 3 Gy/fx only increases the TCP by 2% while increasing the NTCP by 3%, so it does not seem worthwhile to escalate beyond 2.2 Gy based on this model | decision adjustment made based on RTOE, bypassing AI-recommendation |
| | 7 | 2 | | | | 2 | 0 | 1 | 2 | 0 | 1 | 0 | 1 | apparent dose response, may not need 66Gy | Positive response by patient during KBR-ART evaluation phase. Potential Dose de-escalation. | with multiple level 1 trials showing no benefit of dose escalation, it's hard to have any confidence in these recommendations | Confirmation bias on prior clinical trials showing no benefits of dose escalation |
| | 9 | 3 | | | | 3.2 | 3 | 3 | 3 | 1 | 1 | 0 | 1 | It appears teh esophagus is not involved in this case (no station 7/8/9 LN) so I think it would be ok to escalate, though I have no personal experience doing it, so picked 3 Gy. | OAR seems safe. Potential dose escalation. Not enough knowledge. | The esophagus wasn't treated to a very high dose here, so would have perhaps pushed the dose escalation a little bit harder. (?) | OAR (esophagus) seems safe. Potential dose escalation. |

**Abbreviation**- **Pat**: Patient No; **Eval**: Evaluator; **un**: Unassisted Decision; **ai**: AI recommendation; **ai TCP**: Tumor Control Probability for AI recommended dose; **ai NTCP**: Normal Tissue Complication Probability for AI recommended dose; **aia**: AI-assisted Decision; **un conf**: Unassisted Decision Confidence; **aia conf**: AI-assisted Decision Confidence; **ai trust**: AI recommendation trust level; **RTOE: avg**: Reasonableness of RT outcome estimation; **RTOE: uncert**: Reasonableness of RT Outcome Uncertainty Level; **ai:avg**: Agreement with AI Recommended Dose Range; **ai: uncert**: Reasonableness of AI Recommendation Uncertainty Level; **OAR**: Organs at Risk from RT; **DVH**: Cumulative Dose Volume Histogram



medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

## Table S4: Evaluators' Remark for HCC - I

| Pat | Eval | un Gy/fx | ai Gy/fx | ai TCP | ai NTCP | aia Gy/fx | un conf | aia conf | ai trust | RTOE: avg | RTOE: un-cert | ai: avg | ai: un-cert | Unassisted Remark | Unassisted Remark Authors Summary | AI-assisted Remark | AI-assisted Remark Authors Summary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 11 | 10 | 6.8 | 0.925 | 0.578 | 7 | 4 | 4 | 4 | 1 | 1 | 2 | 0 | Change in ALBI mid treatment so would pause but not sure if that's representative of less liver reserve.... but would stick with BED of 100 for CP A disease. | negative Evaluation Phase response (LF), aim for 100 BED, CP A disease | I would back off on the dose with assistance and would ensure I keep the BED above 72 like I would with CP B disease. | Agreement with AI, lower dose to BED¿72 |
| | 13 | 8 | | | | 7 | 3 | 3 | 5 | 1 | 1 | 2 | 1 | I don't believe that HCC requires a full 50Gy/5fx for local control – my script would probably be lower – so hard to know what the right adaptation dose would be. | potential dose de escalation | - | - |
| 2 | 10 | 8 | 9.5 | 0.416 | 0.344 | 9 | 3 | 2 | 4 | 0 | 1 | 2 | 1 | - | - | This curve show improved control with less toxicity with higher dose. | Decision Adjustment with RTOE |
| 3 | 11 | 10 | 11.5 | 0.933 | 0.543 | 10 | 3 | 4 | 5 | 1 | 1 | 2 | 1 | - | - | I would consider going higher per AI recommendation if liver constraints and OARs were ok; otherwise would stick with BED of 100 (10 Gy x5) | Disagreement with AI, concerned and uncertain if OAR safe |
| 4 | 10 | 10 | 9.6 | 0.93 | 0.44 | 3 | 5 | 2 | 3 | 1 | 1 | 0 | 1 | - | - | There seems to be a big increase in NTCP for a small change in TCP for this case. Surprised the AI wants to go so high. | relation between RTOE and AI-recommendation seems inconsistent; |
| | 11 | 9 | | | | 9 | 4 | 5 | 5 | 1 | 1 | 2 | 1 | Great liver function (CPA, ALBI 1), low volume disease, and low normal liver dose: would try for BED of 100; therefore 9 Gy x2 for the last 2 fractions would reach that threshold. | positive Evaluation Phase response (LF), aim for 100 BED | This seems very reasonable since within standard of EQD2 or BED for HCC with CP A disease | Agreement with AI |
| 5 | 10 | 10 | 9.1 | 0.922 | 0.313 | 11 | 4 | 3 | 3 | 1 | 1 | 0 | 1 | - | - | Here, there seems to be an advantage for more dose. | decision adjustment made based on RTOE, bypassing AI-recommendation |
| | 11 | 9 | | | | 9 | 3 | 4 | 5 | 1 | 1 | 2 | 1 | Close to heart so would lower dose to reach BED of 100 for last 2 txts | concern with OAR (heart), dose deescalate, aim for 100 BED | - | - |

**Abbreviation**- **Pat**: Patient No; **Eval**: Evaluator; **un**: Unassisted Decision; **ai**: AI recommendation; **ai TCP**: Tumor Control Probability for AI recommended dose; **ai NTCP**: Normal Tissue Complication Probability for AI recommended dose; **aia**: AI-assisted Decision; **un conf**: Unassisted Decision Confidence; **aia conf**: AI-assisted Decision Confidence; **ai trust**: AI recommendation trust level; **RTOE: avg**: Reasonableness of RT outcome estimation; **RTOE: uncert**: Reasonableness of RT Outcome Uncertainty Level; **ai:avg**: Agreement with AI Recommended Dose Range; **ai: uncert**: Reasonableness of AI Recommendation Uncertainty Level; **OAR**: Organs at Risk from RT; **DVH**: Cumulative Dose Volume Histogram; **ALBI**: Albumin-Bilirubin Score; **LF**: Liver Function; **CP**: Child Pugh Score; **BED**: Biologically Effective Dose



medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

## Table S4: Evaluators' Remark for HCC - II

| Pat | Eval | un Gy/fx | ai Gy/fx | ai TCP | ai NTCP | aia Gy/fx | un conf | aia conf | ai trust | RTOE: avg | RTOE: un-cert | ai: avg | ai: un-cert | Unassisted Remark | Unassisted Remark Authors Summary | AI-assisted Remark | AI-assisted Remark Authors Summary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 7 | 7 | 4.2 | 0.925 | 0.29 | 4 | 3 | 2 | 3 | 1 | 1 | 0 | 1 | - | - | with TCP near identical per dose, it seems like omitting last two fractions should be an option | decision adjustment made based on RTOE, while agreeing with AI-recommendation, Steep Rise (NTCP) over run (dose) seems low |
| | 10 | 10 | | | | 10 | 5 | 4 | 2 | 0 | 1 | 0 | 1 | - | - | There seems to be more improved in TCP with higher dose relative to increase in NTCP. | decision adjustment made based on RTOE, bypassing AI-recommendation, Steep Rise (NTCP) over run (dose) seems low |
| | 11 | 9 | | | | 5 | 3 | 3 | 2 | 1 | 1 | 0 | 0 | Keeping BED of 100 since CP A disease like other cases. | aim for 100 BED | Keeping BED above 72 and within the range but on the higher end | decision adjustment made based on RTOE, while agreeing with AI-recommendation |
| 7 | 10 | 8 | 7.8 | 0.916 | 0.521 | 5 | 4 | 3 | 3 | 1 | 1 | 0 | 1 | - | - | Seems to be a small increase in TCP for higher increase in NTCP | decision adjustment made based on RTOE, bypassing AI-recommendation, Steep Rise (NTCP) over run (dose) seems low |
| | 11 | 8.5 | | | | 8.5 | 3 | 4 | 5 | 1 | 1 | 2 | 1 | Large tumor; preserved liver function based on ALBI and CP after 3 txts so would try for BED ¿72 like CP B disease phase 1/2 trial | Large tumor, positive Evaluation Phase response (LF), aim for >72 BED, CP B disease | - | - |
| 8 | 11 | 11 | 9.6 | 0.93 | 0.368 | 10 | 3 | 4 | 5 | 1 | 1 | 2 | 1 | Liver dose very low and tolerated 11 Gy × 3 well | positive Evaluation Phase response | - | - |
| 9 | 10 | 8 | 9.1 | 0.94 | 0.344 | 8 | 4 | 4 | 3 | 1 | 1 | 0 | 1 | Not the best liver function, would plan for 8 Gy × 5 up front. | Bad LF | - | - |
| | 11 | 8 | | | | 9 | 3 | 4 | 5 | 1 | 1 | 2 | 1 | Initial CP B disease so backing off of the dose but keep BED still high (>100). | CP B Disease, aim for BED >100 | 8-9 Gy seems reasonable | Agreement with AI |

**Abbreviation**- **Pat**: Patient No; **Eval**: Evaluator; **un**: Unassisted Decision; **ai**: AI recommendation; **ai TCP**: Tumor Control Probability for AI recommended dose; **ai NTCP**: Normal Tissue Complication Probability for AI recommended dose; **aia**: AI-assisted Decision; **un conf**: Unassisted Decision Confidence; **aia conf**: AI-assisted Decision Confidence; **ai trust**: AI recommendation trust level; **RTOE: avg**: Reasonableness of RT outcome estimation; **RTOE: uncert**: Reasonableness of RT Outcome Uncertainty Level; **ai:avg**: Agreement with AI Recommended Dose Range; **ai: uncert**: Reasonableness of AI Recommendation Uncertainty Level; **OAR**: Organs at Risk from RT; **DVH**: Cumulative Dose Volume Histogram; **ALBI**: Albumin-Bilirubin Score; **LF**: Liver Function; **CP**: Child Pugh Score; **BED**: Biologically Effective Dose



medRxiv preprint doi: https://doi.org/10.1101/2024.04.27.24306434; this version posted April 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .