Original Article

Assessing the robustness of artificial intelligence powered planning tools in radiotherapy clinical settings—a phantom simulation approach

Martin Hito[1], Wentao Wang[2], Hunter Stephens[2], Yibo Xie[2], Ruilin Li[2], Fang-Fang Yin[2], Yaorong Ge[3], Q. Jackie Wu[2], Qiuwen Wu[2], Yang Sheng[2]

[1]Department of Computer Science, Princeton University, Princeton, NJ, USA; [2]Department of Radiation Oncology, Duke University Medical Center, Durham, NC, USA; [3]Department of Software and Information Systems, University of North Carolina at Charlotte, Charlotte, NC, USA

Contributions: (I) Conception and design: M Hito, FF Yin, Y Ge, QJ Wu, Q Wu, Y Sheng; (II) Administrative support: QJ Wu; (III) Provision of study materials or patients: QJ Wu; (IV) Collection and assembly of data: M Hito, W Wang, H Stephens, Y Xie, R Li, Y Sheng; (V) Data analysis and interpretation: M Hito, W Wang, H Stephens, Y Xie, R Li, Y Sheng; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Yang Sheng, PhD. Department of Radiation Oncology, Physics Division, Duke University Medical Center, Durham, NC 27710, USA. Email: Yang.Sheng@duke.edu.

Background: Artificial intelligence (AI) based radiotherapy treatment planning tools have gained interest in automating the treatment planning process. It is essential to understand their overall robustness in various clinical scenarios. This is an existing gap between many AI based tools and their actual clinical deployment. This study works to fill the gap for AI based treatment planning by investigating a clinical robustness assessment (CRA) tool for the AI based planning methods using a phantom simulation approach.

Methods: A cylindrical phantom was created in the treatment planning system (TPS) with the axial dimension of 30 cm by 18 cm. Key structures involved in pancreas stereotactic body radiation therapy (SBRT) including PTV25, PTV33, C-Loop, stomach, bowel and liver were created within the phantom. Several simulation scenarios were created to mimic multiple scenarios of anatomical changes, including displacement, expansion, rotation and combination of three. The goal of treatment planning was to deliver 25 Gy to PTV25 and 33 Gy to PTV33 in 5 fractions in simultaneous integral boost (SIB) manner while limiting luminal organ-at-risk (OAR) max dose to be under 29 Gy. A previously developed deep learning based AI treatment planning tool for pancreas SBRT was identified as the validation object. For each scenario, the anatomy information was fed into the AI tool and the final fluence map associated to the plan was generated, which was subsequently sent to TPS for leaf sequencing and dose calculation. The final auto plan's quality was analyzed against the treatment planning constraint. The final plans' quality was further analyzed to evaluate potential correlation with anatomical changes using the Manhattan plot.

Results: A total of 32 scenarios were simulated in this study. For all scenarios, the mean PTV25 V25Gy of the AI based auto plans was 96.7% while mean PTV33 V33Gy was 82.2%. Large variation (16.3%) in PTV33 V33Gy was observed due to anatomical variations, a.k.a. proximity of luminal structure to PTV33. Mean max dose was 28.55, 27.68 and 24.63 Gy for C-Loop, bowel and stomach, respectively. Using $D_{0.03cc}$ as max dose surrogate, the value was 28.03, 27.12 and 23.84 Gy for C-Loop, bowel and stomach, respectively. Max dose constraint of 29 Gy was achieved for 81.3% cases for C-Loop and stomach, and 78.1% for bowel. Using $D_{0.03cc}$ as max dose surrogate, the passing rate was 90.6% for C-Loop, and 81.3% for bowel and stomach. Manhattan plot revealed high correlation between the OAR over dose and the minimal distance between the PTV33 and OAR.

Conclusions: The results showed promising robustness of the pancreas SBRT AI tool, providing important evidence of its readiness for clinical implementation. The established workflow could guide the process of assuring clinical readiness of future AI based treatment planning tools.

Keywords: Radiation therapy; artificial intelligence (AI); machine learning; treatment planning; knowledge-based planning

Submitted Feb 01, 2021. Accepted for publication Jun 07, 2021.
doi: 10.21037/qims-21-51
View this article at: https://dx.doi.org/10.21037/qims-21-51

Introduction

Artificial intelligence (AI) has been widely investigated and implemented in various fields and showing promising results. Radiation therapy has traditionally relied on experienced physicians, physicists and dosimetrists performing multiple tasks during the radiation therapy treatment workflow, starting with contouring, treatment planning, and quality assurance (QA). Recent years saw substantial amount of efforts in improving efficiency of the radiation therapy workflow while maintaining quality (1). These efforts include auto contouring (2), treatment planning automation (3-16), and QA automation (17-21).

With the tremendous success of AI applications in many domains, it is expected that more and more clinical deployments are forthcoming. Translating an AI tools to clinical use is a non-trivial task; as has been reported previously, many clinical uses have seen unexpected pitfalls despite promising initial development and validation results (22). It is of great importance to establish a proper process to ensure that the AI tools are safely and effectively deployed for clinical use. Unfortunately, there have been very few efforts reported that address this critical issue.

Like other clinical tools routinely used in a modern radiation oncology department, a proper QA program is needed to establish understanding, confidence, and the proper evaluation of the device/procedure's rigor in delivering the service as initially described. A traditional QA program is composed of initial acceptance, commissioning, and periodic QA (23,24). A successful QA program relies on proper documentation of the device functionality, the QA procedure, recommended frequency and tolerance, and the timely installation of recommended QA items. The AI based tools in the clinical setting should follow a similar assurance/assessment procedure as other techniques especially before delivering to clinical use, with the inclusion of additional items as needed by the unique features of the specific tool. In contrast to routine clinical QA, the assessment program for AI based tools should focus on volume testing and validation before finalizing the product and only needs to be done once. Due to the black box nature of many AI techniques, it is of extreme importance to understand the strength and weakness of the technique to avoid unexpected and unnoticed deviation in performance.

AI based tools for radiation treatment planning have seen increased interest in clinics to improve workflow efficiency while still maintaining high plan quality. An example of such tool is RapidPlan® from Varian Medical Systems (Palo Alto, CA), which is capable of generating dose-volume histogram (DVH) predictions to guide inverse optimization (25). There have been many publications which validate the performance of RapidPlan® in multiple treatment sites (26-28). RapidPlan® has demonstrated that the key to successful clinical implementation is composed of three pillars: (I) volume testing, (II) performance robustness against patient anatomy variation, (III) user friendliness. Before delivering the product to the end user, the developers often share the responsibility of testing and validating the performance of the tool in various clinical scenarios. It is perhaps the most important to evaluate and understand the product's performance under conditions where outlier data is presented. Unlike the development phase of the AI tool which focuses highly on optimizing performance, clinical deployment necessitates and prioritizes overall robustness.

Deep learning approaches have taken a step forward in automating treatment planning in recent years (9-11,29). Unlike traditional machine learning techniques such as the one used in RapidPlan®, the complexity of deep learning models increases exponentially with the depth and width of the network, often making it difficult to understand its limitations. Therefore, it is even more important to validate a deep learning based AI tool's overall robustness against various clinical scenarios. In previous studies, we have developed AI based fluence map prediction tools using deep convolutional neural networks (NN) (10). We used 85 cases to train the NN and additional 15 cases to test it. In this project, we propose a simulation based clinical robustness

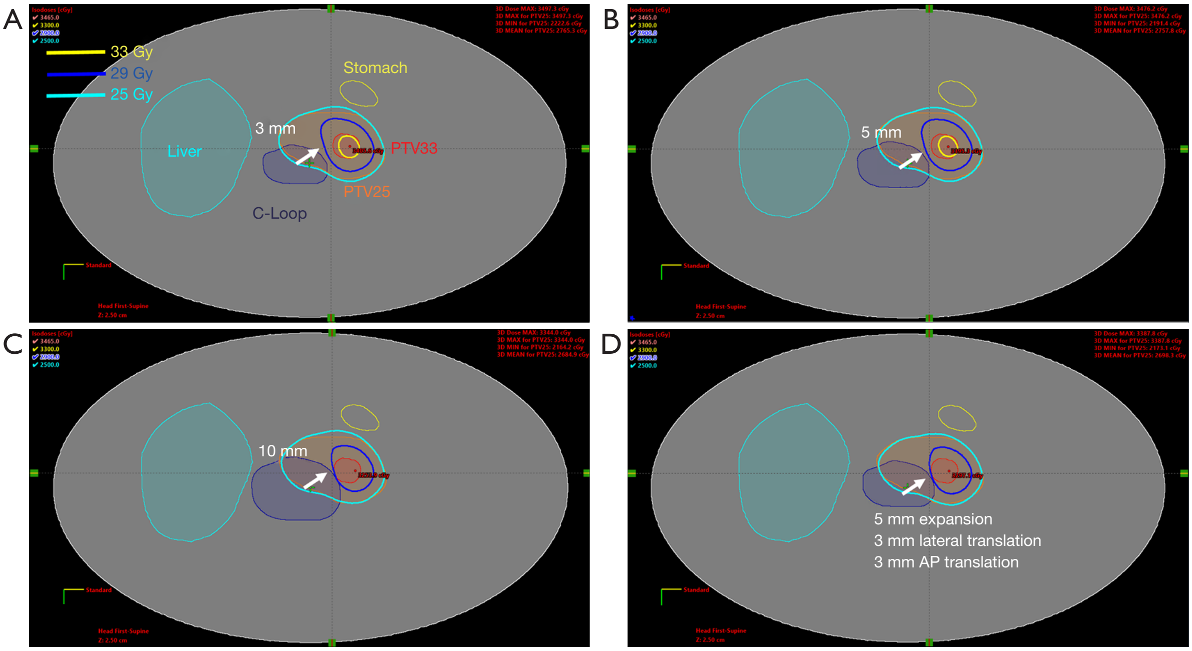

Figure 1 Example anatomy variation of the C-Loop under scenario (A) 3 mm, (B) 5 mm, (C) 10 mm dilation, and (D) 5 mm compounded with 3 mm translation in either of axial directions. Result isodose line is shown in yellow for 33 Gy, dark blue for 29 Gy and cyan for 25 Gy.

assessment (CRA) approach for assessing AI based planning tools and we apply this approach to assessing the pancreas fluence map prediction tool in terms of its robustness in various simulated clinical scenarios. Unlike traditional periodic QA program for routine clinical tasks, the CRA approach focuses more on the pre-clinical testing and validation for robustness. Our proposed CRA approach focuses on the first two pillars of the key items to ensure safe delivery: high volume testing and overall robustness against anatomical variation. We simulated various anatomical variations using real patient organ anatomy. We evaluated the performance of the AI tool against the proposed anatomy variation while expanding the size of the independent test cases.

Methods

Dataset

In this study, the simulation was carried out for pancreas stereotactic body radiation therapy (SBRT). The prescription was 25 Gy in 5 fractions with simultaneous integral boost (SIB) to 33 Gy, in accordance with our clinical practice. The luminal organ-at-risk (OAR) max dose was constrained to a maximum dose of 29 Gy, which included the stomach, C-Loop of the duodenum and bowel. All simulation plans were evaluated in the Eclipse® treatment planning system (TPS) version 15.6 (Varian Medical Systems, Palo Alto, CA) for a TrueBeam® Linear Accelerator (Linac) with Millennium 120 multi-leaf collimator (MLC). The leaf sequence was calculated with Smart LMC version 13.7.14, and the dose was calculated with the AAA dose calculation algorithm version 15.6.06.

In this study, one cylindrical phantom was generated for simulating the abdominal trunk of a patient. The body structure was generated as an ellipse with long axis of 30 cm (left-right) and short axis of 18 cm (anterior-posterior), with the density of water. Anatomical structures from a patient underwent pancreas SBRT were extracted and transferred to the phantom, including two planning target volume, PTV25 (prescribed to 25 Gy), PTV33 (prescribed to 33 Gy), and four OARs (liver, stomach, C-Loop and bowel). An axial view of the structure geometry is shown in *Figure 1*.

Study design and workflow

The AI tool tested in this study is capable of predicting the fluence map of pancreas SBRT as reported by Wang *et al.* (10). The algorithm, in short, is composed of two sequential deep NN, structure contours are fed in and the algorithm makes a prediction of the fluence map for a nine-field intensity modulated radiation therapy (IMRT) plan. The first NN predicts the optimal dose distribution and projects it to each field's beam's eye view (BEV). The second NN subsequently predicts the optimal fluence map for each field. The entire workflow includes data preparation, a beam dose prediction NN, dose projection, and a fluence map prediction NN. All the processes in the workflow were fully automated. The first beam dose prediction NN used a customized encoder-decoder design, which combines a downsampling block, upsampling blocks, and a convolutional block. For the beam dose NN, the input images were the axial slices including the PTV contours and OAR contours with 6 adjacent superior and inferior slices. All contour masks were assigned a value of the ratio of the max dose constraint (25 Gy) over the PTV33 prescription dose (33 Gy) for OARs, and the ratio of the PTV25 prescription dose (29 Gy) over the PTV33 prescription dose (33 Gy). The output for the beam dose NN was the beam dose distribution for the current slice. All predicted beam dose slices were stacked to create a 3D beam dose which was subsequently projected onto the BEV to feed into the second fluence map NN. The fluence map NN took beam dose projections as input and output fluence map for the specific beam. The predicted fluence maps were then imported into the TPS. The plan is finalized after leaf sequencing and dose calculation with no additional manual editing.

This AI tool has been validated with 15 clinical patient cases with promising results as reported by Wang *et al*. (10). The mean target dose difference was 0.1% between the auto-plan and the manually generated benchmark plan. For OARs, the mean and max dose difference (0.1 cc) was 0.2% and 4.4%, respectively. The predicted fluence map was deemed deliverable based on the gamma index comparing the optimal fluence map (before leaf sequencing) and actual fluence map (after leaf sequencing). The mean gamma index was 97.69% for the auto-plan group as compared to 98.14% for the benchmark plan group.

Due to a limited number of clinical pancreas SBRT cases, the overall robustness and sensitivity against various clinical scenarios is preferred to be tested using augmented clinical dataset and simulated on the phantom. It is due to the fact that the clinically available dataset for validation and testing is too scarce to cover a wide range of anatomical variations in pancreas SBRT context. In this study, we simulated anatomical variations using real patient anatomy. We evaluated the performance of the AI tool against the proposed anatomy variation while expanding the size of the independent test cases.

The test case augmentation was broken down into two categories: (I) single structure anatomy variation, and (II) dual structure anatomy variation. Under the first category, we simulated four scenarios, which included five individual structures involved in pancreas SBRT treatment planning: GTV (PTV33), PTV (PTV25), C-Loop, stomach and bowel. Anatomical variation includes organ translation, dilation, rotation and/or combination of two. Under the second category, we focus on the dual-structure anatomy variation, including GTV (PTV33)/C-Loop, GTV (PTV33)/stomach, and GTV (PTV33)/Bowel. The details of the investigated scenarios are outlined in *Table 1*. For each scenario, the anatomical variation was executed in Matlab® R2019b (Natick, MA) and subsequently exported to Python for fluence map prediction using the AI tool described above. The automatically generated plan (auto-plan) was assessed to the desired prescription constraints.

Evaluation criteria

All auto-plans for the simulation test cases were evaluated against the standard dose constraints used in the study. To re-iterate, PTV coverage was assessed for PTV25 using V25Gy (%) and for PTV33 using V33Gy (%). Volume coverage at prescription levels over 95% is ideal, with 90–95% to be acceptable. For all OARs including C-Loop, bowel and stomach, the percentage of cases passing max dose constraint of 29 Gy was reported. In addition, we recorded the percentage of cases passing an alternative max dose constraint of $D_{0.03cc}$ <29 Gy. For cases exceeding the constraint, the volume receiving more than 29 Gy was recorded. We define the cases having more than 1 cc of OAR volume exceeding 29 Gy as unacceptable and recorded the number of such cases.

In order to assess the correlation of the tool performance with changing geometry, we utilized a Manhattan plot (30,31) to visualize the DVH coupled with the geometry information, which is the minimum distance between the OAR and the PTV33. Manhattan plot is 4D visualization, where in this study the base two dimensions are dose and volume in 10% bins, the height of the bar is the number of cases falling into the specific dose-volume bin and the color map is the percentage of cases with close proximity to the higher prescribed PTV volume PTV33. The threshold distance was chosen as 5 mm which is the average distance

Table 1 Description of all scenarios simulated

Category	Scenario	Simulation detail (shift/margin/rotation)
Category 1 (single structure variation)	Scenario 1 (PTV33 displacement)	Scenario 1-1: PTV33 with X =5 mm shift
		Scenario 1-2: PTV33 with X =–5 mm shift
		Scenario 1-3: PTV33 with Y =5 mm shift
		Scenario 1-4: PTV33 with Y =–5 mm shift
	Scenario 2 (PTV25 displacement)	Scenario 2-1: PTV25 with X =7 mm shift
		Scenario 2-2: PTV25 with X =–7 mm shift
		Scenario 2-3: PTV25 with Y =7 mm shift
		Scenario 2-4: PTV25 with Y =–7 mm shift
	Scenario 3 (C-Loop expansion)	Scenario 3-1: C-Loop with 3 mm margin
		Scenario 3-2: C-Loop with 5 mm margin
		Scenario 3-3: C-Loop with 10 mm margin
		Scenario 3-4: C-Loop with 5 mm margin and X=3 mm, Y=–3 mm shift
	Scenario 4 (Stomach rotation and shift)	Scenario 4-1: stomach with 15 degrees rotation and X=–10 mm, Y=10 mm shift
		Scenario 4-2: stomach with 30 degrees rotation and X=–15 mm, Y=15 mm shift
		Scenario 4-3: stomach with –15 degrees rotation and X=–10 mm, Y=10 mm shift
		Scenario 4-4: stomach with –30 degrees rotation and X=–15 mm, Y=15 mm shift
	Scenario 5 (Bowel expansion and shift)	Scenario 5-1: bowel with 5 mm margin and Y=0 mm shift
		Scenario 5-2: bowel with 5 mm margin and Y=5 mm shift
		Scenario 5-3: bowel with 10 mm margin and Y=0 mm shift
		Scenario 5-4: bowel with 10 mm margin and Y=5 mm shift
Category 2 (dual structure variation)	Scenario 1 (PTV33 displacement and C-Loop shift)	Scenario 1-1: PTV33 with X=5 mm shift and C-Loop with 3 mm margin
		Scenario 1-2: PTV33 with X=–5 mm shift and C-Loop with 3 mm margin
		Scenario 1-3: PTV33 with Y=5 mm shift and C-Loop with X=3 mm, Y=–3 mm shift
		Scenario 1-4: PTV33 with Y=–5 mm shift and C-Loop with X=3 mm, Y=–3 mm shift
	Scenario 2 (PTV33 displacement and stomach shift)	Scenario 2-1: PTV33 with X=5 mm shift and stomach with 15 degrees rotation
		Scenario 2-2: PTV33 with X=–5 mm shift and stomach with –15 degrees rotation
		Scenario 2-3: PTV33 with Y=5 mm shift and stomach with 15 degrees rotation and X=–10 mm, Y=10 mm shift
		Scenario 2-4: PTV33 with Y=–5 mm shift and stomach with -15 degree rotation and X=–10 mm, Y=10 mm shift
	Scenario 3 (PTV33 displacement and bowel shift)	Scenario 3-1: PTV33 with X=5 mm shift and bowel with 5 mm margin and Y=0 mm shift
		Scenario 3-2: PTV33 with X=–5 mm shift and bowel with 10 mm margin and Y=0 mm shift
		Scenario 3-3: PTV33 with Y=5 mm shift and bowel with 10 mm margin and Y=5 mm shift
		Scenario 3-4: PTV33 with Y=–5 mm shift and bowel with 5 mm margin and Y=5 mm shift

X, Y and Z are left-right (LR), anterior-posterior (AP) and superior-inferior (SI) direction, respectively. Margin is isotropic in X, Y and Z direction. Positive value indicates left, posterior and superior direction.

Table 2 Geometric and dosimetric data summary for all five structures simulated in the study: PTV25, PTV33, C-Loop, bowel and stomach

Structure	Metrics	Mean	SD
PTV25	Volume (cc)	48.58	0.00
	V25Gy (%)	96.7	1.4
PTV33	Volume (cc)	1.45	0.02
	V33Gy (%)	82.2	16.3
C-Loop	Volume (cc)	26.27	10.66
	Max dose (Gy)	28.55	0.82
	D0.03cc (Gy)	28.03	0.78
	V29Gy (cc)	0.02	0.06
Bowel	Volume (cc)	346.34	61.60
	Max dose (Gy)	27.68	2.54
	D0.03cc (Gy)	27.12	2.52
	V29Gy (cc)	0.22	0.52
Stomach	Volume (cc)	82.57	0.00
	Max dose (Gy)	24.63	3.89
	D0.03cc (Gy)	23.84	3.81
	V29Gy (cc)	0.14	0.38

Mean and standard deviation were reported for all 32 simulated scenarios. SD, standard deviation.

for dose fall off to 88% (29 Gy/33 Gy).

Results

An example dose distribution of the auto-plan is shown in *Figure 1* for category 1 scenario 3. *Figure 1* shows an example scenario which simulates the C-Loop variation (dark blue). *Figure 1A,B,C* simulates the dilation of the C-Loop of 3, 5 and 10 mm respectively. And *Figure 1D* simulates 5mm dilation compounded with 3 mm translation in either of axial direction. The data suggested that the proposed AI tool is robust against the anatomical variation by providing conformal dose distribution for both 25 Gy (cyan) and 33 Gy (yellow). The critical OAR sparing goal 29 Gy isodose line has been tailored to not only conform to the PTV but also respect the OAR structure (C-Loop in dark blue) as shown in the white arrow, indicating overall robustness of the proposed AI tool. It is capable of recognizing the PTV (PTV25), GTV (PTV33) and OAR anatomy changes while still respecting the dose prescription, OAR sparing goal, and more importantly the choice of priority between two.

Structure volume and key dosimetric endpoints for both PTVs and three luminal OARs are listed in *Table 2*. The auto-plan provided satisfactory coverage for the PTV25 with a mean V25Gy of 96.7%. Mean PTV33 V33Gy has reached 82.2%, which is highly dependent on the proximity of the OAR to the PTV33 for each scenario as indicated by the standard deviation of 16.25%. Mean max dose was 28.55, 27.68 and 24.63 Gy for C-Loop, bowel and stomach, respectively. Mean $D_{0.03cc}$ was 28.03, 27.12 and 23.84 Gy for C-Loop, bowel and stomach, respectively. Corresponding mean volume over the dose limit of 29 Gy was 0.02, 0.22 and 0.14 cc for C-Loop, bowel and stomach.

Key dosimetric endpoints were further partitioned and evaluated for different ranges for the PTV and an alternative max dose endpoint for the OAR in *Table 3*. All PTV25 received satisfactory coverage as indicated in *Table 3*. For PTV33, 68.75% cases received less than optimal coverage (<90%). Among these cases, the mean minimum distance to the PTV33 for OARs was 2.7 mm, while the corresponding value was 6.7 mm for the group with over 90% PTV33 coverage (P<0.001, Wilcoxon Rank-Sum test). A boxplot of PTV25 V25Gy, PTV33 V33Gy and their

Table 3 Number of cases and percentage of cases that fall into specific range of for various dosimetric endpoints

Structure	Count/percentage	V >95%	90%< V ≤95%	V ≤90%	D_{max} <29 Gy	$D_{0.03cc}$ <29 Gy	V29Gy >1 cc
PTV25 V25Gy(%)	Count	31	1	0	–	–	–
	Percentage	96.90%	3.10%	0.00%	–	–	–
PTV33 V33Gy(%)	Count	4	6	22	–	–	–
	Percentage	12.50%	18.80%	68.80%	–	–	–
C-Loop	Count	–	–	–	26	29	0
	Percentage	–	–	–	81.30%	90.60%	0.00%
Bowel	Count	–	–	–	25	26	3
	Percentage	–	–	–	78.10%	81.30%	9.40%
Stomach	Count	–	–	–	26	26	1
	Percentage	–	–	–	81.30%	81.30%	3.10%

PTV25 V25Gy(%) and PTV33 V33Gy(%) were reported for range (0, 90%), (90%, 95%) and (95%, 100%). For OARs, Dmax and D0.03cc less than 29 Gy, V29Gy over 1 cc were reported, respectively.

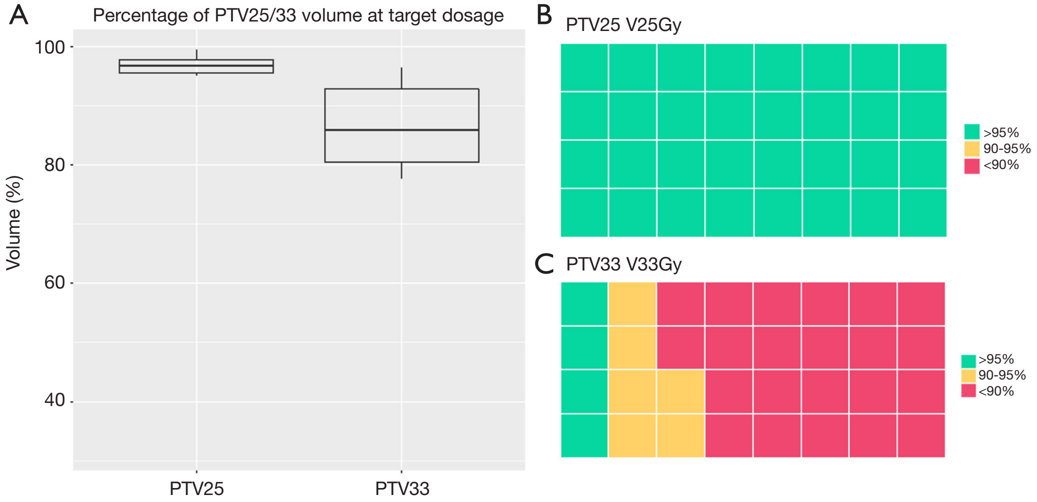

Figure 2 Dosimetric statistics of PTV25/33 volumetric coverage at prescription. (A) Boxplot of the prescription dose coverage for PTV25 (left) and PTV33 (right); (B) color-coded PTV25 V25Gy distribution in range over 95% (green), 90–95% (yellow) and below 90% (red); (C) color-coded PTV33 V33Gy distribution in range over 95% (green), 90–95% (yellow) and below 90% (red).

concentration for various ranges is shown in *Figure 2*. The max dose constraint of 29 Gy was met for 81.25% cases for the C-Loop and stomach, and 78.13% for the bowel. An alternative max dose constraint of $D_{0.03cc}$ indicated a 90.63% passing rate for the C-Loop, and 81.25% for the bowel and stomach. Significant overdosed (V29Gy >1 cc) were not observed for the C-Loop, while 3 and 1 cases were observed for bowel and stomach, respectively. A boxplot of the OAR max dose and pie charts showing OAR constraint passing rate are shown in *Figure 3*.

To visualize the correlation between the DVH and the geometry, a.k.a. the minimum distance between the OAR and PTV33, Manhattan plot was generated as shown in *Figure 4*. It shows C-Loop, bowel and stomach in *Figure 4A*, *B* and *C*, respectively. The base 2D is percentage dose and percentage volume both in 10% bin. The height is the number of cases falling into the bin. The color map is the percentage of cases that have minimum distance between

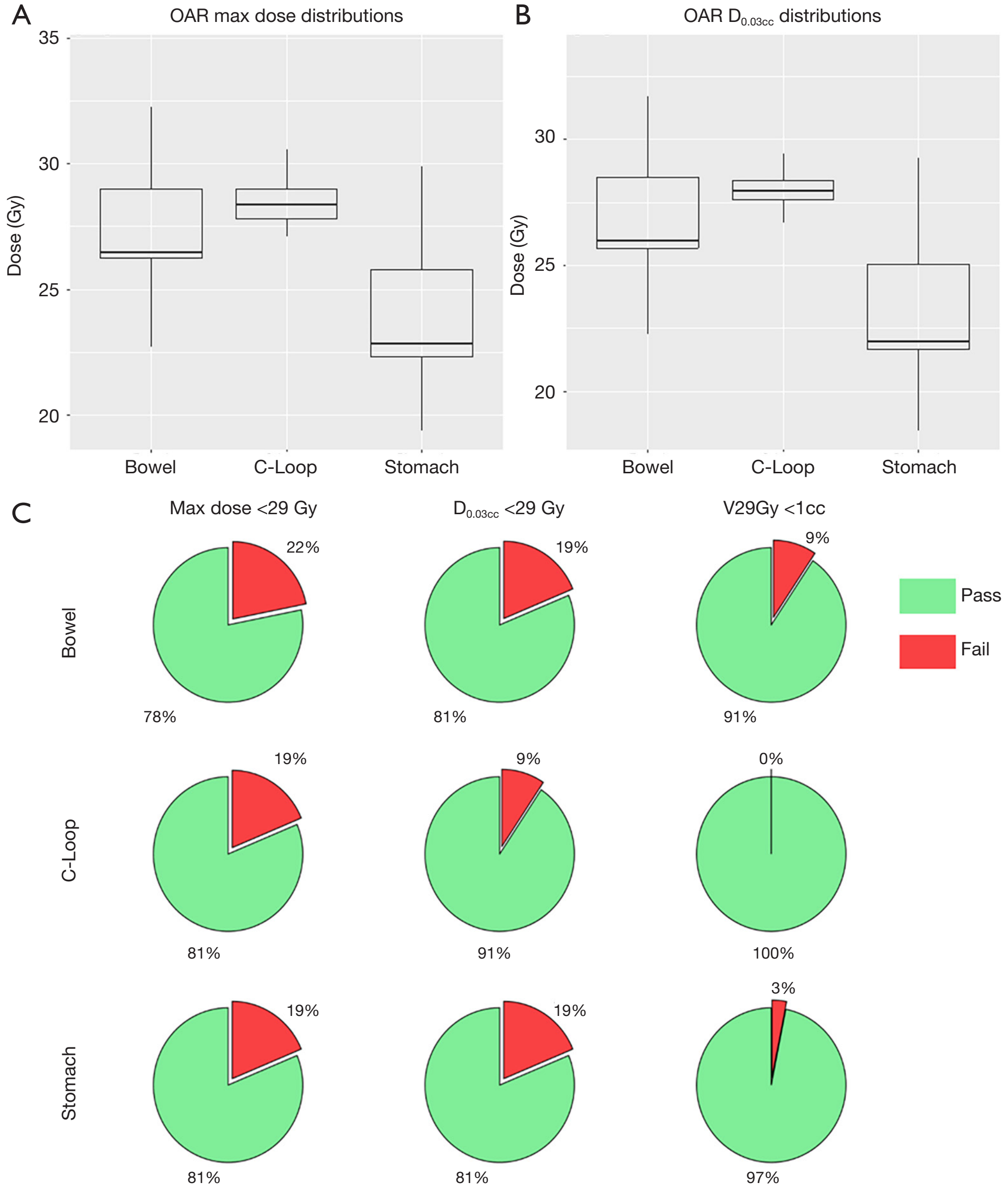

Figure 3 Boxplot for OAR (A) max dose and (B) $D_{0.03cc}$ distribution for bowel, C-Loop and stomach; (C) pie chart for pass fail percentage of max dose <29 Gy (left column), $D_{0.03cc}$ <29 Gy (middle column) and V29Gy <1 cc (right column) for bowel (first row), C-Loop (second row) and stomach (bottom row), respectively. OAR, organ-at-risk.

OAR and PTV less than 5 mm. It is noticeable that for both the C-Loop in (*Figure 4A*) and bowel in (*Figure 4B*), the excessive max dose is highly correlated with the minimum distance (dark red). For stomach as indicated in (*Figure 4C*), the overdose (higher percentage dose for each volume row) is strongly associated with the minimum distance in all volume rows. The results indicated that the auto-plan's quality in OAR dose is reflective of anatomical variation.

Discussion

In this study, we designed a CRA approach for systematically validating and testing the overall robustness of a pancreas SBRT automatic planning AI tool against various geometries in a simulated phantom. This tool can be used to systematically validate the robustness and sensitivity of an AI tool based on the available anatomy variation

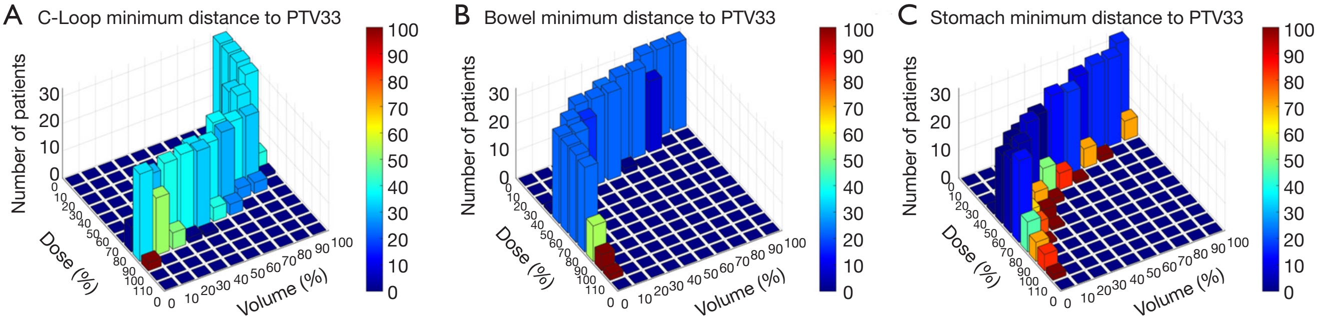

Figure 4 Manhattan plot of three OARs: (A) C-Loop, (B) bowel and (C) stomach, with respect to the endpoint of minimum distance to PTV33 with the threshold of 5 mm. The base two dimensions are percentage dose and percentage volume with 10% bin. The height of the bar indicates the number of cases falling into each bin. The color map is indicator of the percentage of cases that have minimum distance to PTV33 less than 5 mm. OAR, organ-at-risk.

before releasing it for clinical use. We believe this is the first study ever to investigate the CRA approach for an AI based planning tool. The simulated scenario can mimic multiple possible scenarios related to geometry change that are commonly seen in gastro-intestinal (GI) treatment planning. The results have shown promising robustness of the AI tool, boosting the confidence of it providing solid performance once used in the clinic. In addition, the design of the CRA approach augmented the number of cases and scenarios that the AI tool can be tested on, which has long been the limitation for developing AI tools development due to the scarcity of data.

In this study, we also noticed that the AI tool struggles with scenarios where the constraints were hard to meet, e.g., the minimum distance between OAR and PTV is small and therefore the PTV coverage would be compromised to prioritize OAR sparing. Such scenario represents an "extrapolation" scenario which was not fully accounted for during the training phase due to the data scarcity. It is more prominent when the overlap volume of the luminal structure increases with the PTVs. It is worth mentioning that such scenarios with conflicting constraints would also substantially increase the complexity of manual planning. The proposed AI tool provides valuable assistance in generating the plan in a significantly reduced amount of time. We believe that, based on its current performance, the AI tool will bring tremendous value to treatment planning teams once it is rolled out to the clinic.

Designing CRA approaches for AI based treatment planning tools is a new area which has been rarely investigated. Ensuring safe delivery and quality control is extremely important to maintain and guarantee high quality care provided to the patients. This study is the first attempt in addressing this topic by systematically analyzing and evaluating the performance of the AI tool. A complete CRA approach for AI based automation in radiation oncology departments is worth the effort to ensure quality. Once the algorithm is thoroughly tested and validated for stability and robustness, proper documentation of the functionality and version control is necessary. Designated personnel responsible for development, CRA, documentation should be appointed. Recommended check items for acceptance, commissioning and periodic CRA and its frequency shall be documented and properly followed.

The proposed CRA approach could also assist in determining the subpar performance of an AI planning tool. Augmenting the simulation dataset and classifying them into subclinical groups could help identify the area where the AI planning tool performs less than optimal, while also ensuring such performance is clinically acceptable. Simulating multiple scenarios is also beneficial in detecting geometric outliers that could deviate the AI tool's performance, which could help alert the clinical user to monitor the performance.

This study showed an example application of CRA approach for a pancreas SBRT AI based planning tool for our clinic. It worth mentioning that this approach allows users from different centers to customize the approach for their own local data sample. It is also important that such approach is adopted when the AI tool is updated or re-calibrated, based on model retraining due to data evolution.

We demonstrated the CRA approach for pancreas SBRT treatment planning in this study, which can be further expanded to other treatment sites, such as head-and-neck

(HN) and lung. Attention should be made to address the fact that the anatomy variation pattern could be different across different treatment sites. Therefore, simulating realistic anatomy variation distribution is key to successful validation. In addition, site specific consideration should be fully considered in the simulation process, such as using bolus in HN treatment planning, etc.

As more and more AI application are delivered to the clinical side, tremendous effort is needed to establish the pipeline of ensuring continuous safety practice regarding AI technology. A few studies have summarized the current and future practice of AI tools in the clinic (32,33). We believe the traditional acceptance-commissioning-QA pipeline would be a straightforward pathway of implementing the AI tool in the clinic. However, understanding the black box nature of the AI tool is the key component of ensure overall safety, which is not often seen in the software/hardware QA program in the modern radiation oncology department. This study focuses on addressing this issue by implementing the CRA pipeline, which facilitates the understanding of the candidate tool and benefits human-AI integration once deployed in the clinic. There are still many areas that need to be investigated regarding ensuring safe deployment and assisting clinical team understand and accept the AI tool in the future.

The results in this study show overall promising results. However, some limitations exist. For example, as mentioned before, the AI tool is not strictly following the luminal max dose constraint when it is challenging to meet both PTV and OAR constraints due to the anatomy. In future studies, it would be beneficial for the AI tool to appraise the performance on its own and provide a confidence level for the human operator. It is highly important that the planner is alerted when potential deviation could occur. Another limitation is that this simulation is based on a homogenous phantom. It worth mentioning that the AI tool tested in this study accounts for the heterogeneity in the CT images by incorporating the static dose in the deep NN design. It would be of clinical interest to use real patient's CT images to validate the tool's overall robustness in this regard. Using patient CT images to carry out simulation would be the next step in future studies. Finally, as the final step of clinical readiness evaluation, we would like to involve the clinician in appraising the plan acceptability from their perspective in the future study. The clinician will grade the plan as "acceptable", "acceptable with minor improvement" or "not acceptable" as reported by Cox *et al*. (34) and Wang *et al*. (35).

Conclusions

In this study, we developed a CRA approach to validate the performance and robustness of the pancreas SBRT AI based treatment planning tool. The results showed overall robustness which supports subsequent clinical implementation. This is the first attempt to systematically evaluate and test an AI based tool using simulated data for data augmentation. This approach can be used to establish a testing workflow for future endeavors in AI based treatment planning. Complete and systematic CRA workflow would ensure safe delivery and quality control of AI applications in real clinical settings.

Acknowledgments

Funding: This work is partially supported by NIH R01CA201212 research grant and Varian master research grant.

Footnote

Provenance and Peer Review: With the arrangement by the Guest Editors and the editorial office, this article has been reviewed by external peers.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://dx.doi.org/10.21037/qims-21-51). The special issue "Artificial Intelligence for Image-guided Radiation Therapy" was commissioned by the editorial office without any funding or sponsorship. WW reports funding support from NIH R01CA201212 research grant and Varian master research grant. FFY reports patent titled "Systems and methods for automatic, customized radiation treatment plan generation for cancer" was filed. YG reports funding support from NIH R01CA201212 research grant. Patent titled "Systems and methods for automatic, customized radiation treatment plan generation for cancer" was filed. QJW and YS report funding support from NIH R01CA201212 research grant and Varian master research grant. Patent titled "Systems and methods for automatic, customized radiation treatment plan generation for cancer" was filed. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by DUHS Institutional Review Board (Protocol ID: Pro00046706) and informed consent was taken from all individual participants.

References

1. Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, van Elmpt W. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol 2020;153:55-66.
2. Teguh DN, Levendag PC, Voet PW, Al-Mamgani A, Han X, Wolf TK, Hibbard LS, Nowak P, Akhiat H, Dirkx ML, Heijmen BJ, Hoogeman MS. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. Int J Radiat Oncol Biol Phys 2011;81:950-7.
3. Sheng Y, Li T, Yoo S, Yin FF, Blitzblau R, Horton JK, Ge Y, Wu QJ. Automatic Planning of Whole Breast Radiation Therapy Using Machine Learning Models. Front Oncol 2019;9:750.
4. Sheng Y, Li T, Zhang Y, Lee WR, Yin FF, Ge Y, Wu QJ. Atlas-guided prostate intensity modulated radiation therapy (IMRT) planning. Phys Med Biol 2015;60:7277-91.
5. Yuan L, Ge Y, Lee W, Yin FF, Kirkpatrick J, Wu Q. Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans. Med Phys 2012;39:6868-78.
6. Appenzoller LM, Michalski JM, Thorstad WL, Mutic S, Moore KL. Predicting dose-volume histograms for organs-at-risk in IMRT planning. Med Phys 2012;39:7446-61.
7. Wu B, Ricchetti F, Sanguineti G, Kazhdan M, Simari P, Jacques R, Taylor R, McNutt T. Data-driven approach to generating achievable dose-volume histogram objectives in intensity-modulated radiotherapy planning. Int J Radiat Oncol Biol Phys 2011;79:1241-7.
8. Babier A, Boutilier JJ, McNiven AL, Chan TCY. Knowledge-based automated planning for oropharyngeal cancer. Med Phys 2018;45:2875-83.
9. Nguyen D, Long T, Jia X, Lu W, Gu X, Iqbal Z, Jiang S. A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. Sci Rep 2019;9:1076.
10. Wang W, Sheng Y, Wang C, Zhang J, Li X, Palta M, Czito B, Willett CG, Wu Q, Ge Y, Yin FF, Wu QJ. Fluence Map Prediction Using Deep Learning Models - Direct Plan Generation for Pancreas Stereotactic Body Radiation Therapy. Front Artif Intell 2020;3:68.
11. Li X, Zhang J, Sheng Y, Chang Y, Yin FF, Ge Y, Wu QJ, Wang C. Automatic IMRT planning via static field fluence prediction (AIP-SFFP): a deep learning algorithm for real-time prostate treatment planning. Phys Med Biol 2020;65:175014.
12. Wang W, Sheng Y, Yoo S, Blitzblau RC, Yin FF, Wu QJ. Goal-Driven Beam Setting Optimization for Whole-Breast Radiation Therapy. Technol Cancer Res Treat 2019;18:1533033819858661.
13. Zhang J, Wang C, Sheng Y, Palta M, Czito B, Willett C, Zhang J, Jensen PJ, Yin FF, Wu Q, Ge Y, Wu QJ. An Interpretable Planning Bot for Pancreas Stereotactic Body Radiation Therapy. Int J Radiat Oncol Biol Phys 2021;109:1076-85.
14. Zhang J, Ge Y, Sheng Y, Wang C, Zhang J, Wu Y, Wu Q, Yin FF, Wu QJ. Knowledge-Based Tradeoff Hyperplanes for Head and Neck Treatment Planning. Int J Radiat Oncol Biol Phys 2020;106:1095-103.
15. Zhang J, Ge Y, Sheng Y, Yin FF, Wu QJ. Modeling of multiple planning target volumes for head and neck treatments in knowledge-based treatment planning. Med Phys 2019;46:3812-22.
16. Zhang J, Wu QJ, Xie T, Sheng Y, Yin FF, Ge Y. An Ensemble Approach to Knowledge-Based Intensity-Modulated Radiation Therapy Planning. Front Oncol 2018;8:57.
17. El Naqa I, Irrer J, Ritter TA, DeMarco J, Al-Hallaq H, Booth J, Kim G, Alkhatib A, Popple R, Perez M, Farrey K, Moran JM. Machine learning for automated quality assurance in radiotherapy: A proof of principle using EPID data description. Med Phys 2019;46:1914-21.
18. Guterres Marmitt G, Pin A, Ng Wei Siang K, Janssens G, Souris K, Cohilis M, Langendijk JA, Both S, Knopf A, Meijers A. Platform for automatic patient quality assurance via Monte Carlo simulations in proton therapy. Phys Med 2020;70:49-57.

19. Jenkins CH, Naczynski DJ, Yu SJS, Yang Y, Xing L. Automating quality assurance of digital linear accelerators using a radioluminescent phosphor coated phantom and optical imaging. Phys Med Biol 2016;61:L29-L37.
20. Li J, Wang L, Zhang X, Liu L, Li J, Chan MF, Sui J, Yang R. Machine Learning for Patient-Specific Quality Assurance of VMAT: Prediction and Classification Accuracy. Int J Radiat Oncol Biol Phys 2019;105:893-902.
21. Valdes G, Scheuermann R, Hung CY, Olszanski A, Bellerive M, Solberg TD. A mathematical framework for virtual IMRT QA using machine learning. Med Phys 2016;43:4323.
22. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020;2:e489-92.
23. Kutcher GJ, Coia L, Gillin M, Hanson WF, Leibel S, Morton RJ, Palta JR, Purdy JA, Reinstein LE, Svensson GK, Weller M, Wingfield L. Comprehensive QA for radiation oncology: report of AAPM Radiation Therapy Committee Task Group 40. Med Phys 1994;21:581-618.
24. Klein EE, Hanley J, Bayouth J, Yin F-F, Simon W, Dresser S, Serago C, Aguirre F, Ma L, Arjomandy B, Liu C, Sandin C, Holmes T. Task Group 142 report: Quality assurance of medical accelerators. Med Phys 2009;36:4197.
25. Varian Medical Systems. RapidPlan Knowledge-based Planning-Global Knowledge Sharing. 2015. Available online: https://varian.widen.net/view/pdf/ckzqo9ivqs/RapidPlan_ProductBrief_RAD10306C_May2015.pdf?u=wefire
26. Fogliata A, Belosi F, Clivio A, Navarria P, Nicolini G, Scorsetti M, Vanetti E, Cozzi L. On the pre-clinical validation of a commercial model-based optimisation engine: application to volumetric modulated arc therapy for patients with lung or prostate cancer. Radiother Oncol 2014;113:385-91.
27. Tol JP, Delaney AR, Dahele M, Slotman BJ, Verbakel WF. Evaluation of a knowledge-based planning solution for head and neck cancer. Int J Radiat Oncol Biol Phys 2015;91:612-20.
28. Delaney AR, Tol JP, Dahele M, Cuijpers J, Slotman BJ, Verbakel WF. Effect of Dosimetric Outliers on the Performance of a Commercial Knowledge-Based Planning Solution. Int J Radiat Oncol Biol Phys 2016;94:469-77.
29. Nguyen D, Jia X, Sher D, Lin MH, Iqbal Z, Liu H, Jiang S. 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture. Phys Med Biol 2019;64:065020.
30. Otter S, Schick U, Gulliford S, Lal P, Franceschini D, Newbold K, Nutting C, Harrington K, Bhide S. Evaluation of the Risk of Grade 3 Oral and Pharyngeal Dysphagia Using Atlas-Based Method and Multivariate Analyses of Individual Patient Dose Distributions. Int J Radiat Oncol Biol Phys 2015;93:507-15.
31. Lu L, Sheng Y, Zhang G, Li Y, OuYang PY, Ge Y, Xie T, Chang H, Deng X, Wu JQ. Temporal lobe injury patterns following intensity modulated radiotherapy in a large cohort of nasopharyngeal carcinoma patients. Oral Oncol 2018;85:8-14.
32. Wang C, Zhu X, Hong JC, Zheng D. Artificial Intelligence in Radiotherapy Treatment Planning: Present and Future. Technol Cancer Res Treat 2019;18:1533033819873922.
33. Wang M, Zhang Q, Lam S, Cai J, Yang R. A Review on Application of Deep Learning Algorithms in External Beam Radiotherapy Automated Treatment Planning. Front Oncol 2020;10:580919.
34. Cox BW, Teckie S, Kapur A, Chou H, Potters L. Prospective Peer Review in Radiation Therapy Treatment Planning: Long-Term Results From a Longitudinal Study. Pract Radiat Oncol 2020;10:e199-e206.
35. Wang W, Sheng Y, Palta M, Czito B, Willett C, Hito M, Yin FF, Wu Q, Ge Y, Wu QJ. Deep Learning-Based Fluence Map Prediction for Pancreas Stereotactic Body Radiation Therapy With Simultaneous Integrated Boost. Adv Radiat Oncol 2021;6:100672.

Cite this article as: Hito M, Wang W, Stephens H, Xie Y, Li R, Yin FF, Ge Y, Wu QJ, Wu Q, Sheng Y. Assessing the robustness of artificial intelligence powered planning tools in radiotherapy clinical settings—a phantom simulation approach. Quant Imaging Med Surg 2021;11(12):4835-4846. doi: 10.21037/qims-21-51