5-2023

The Safe and Effective Clinical Deployment of Artificial Intelligence Tools

Kelly Nealon

The Safe and Effective Clinical Deployment of Artificial Intelligence Tools

By

Kelly Nealon, M.S.

APPROVED:

Laurence Court

Laurence E. Court, Ph.D.

Advisory Professor

Eunyounghan

Eun Young Han, Ph.D.

Stephen Kry, Ph.D.

V. Reed MD

Valerie Reed, MD

Samantha Simiele

Samantha Simiele. Ph.D.

APPROVED:

Dean, The University of Texas

MD Anderson Cancer Center UTHealth Graduate School of Biomedical Science

The Safe and Effective Clinical Deployment of Artificial Intelligence Tools

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Kelly Nealon, M.S.
Houston, Texas

May, 2023

Acknowledgments

I am grateful to have so many people to acknowledge who have contributed to this project and to my development as a researcher and as an individual. First, I would like to thank my advisor Dr. Laurence Court for his constant support and guidance. Every day I was reminded how lucky I am to have a mentor who can both motivate great science and foster a kind, collaborative and exciting environment to work in. His passion for improving global health is truly what inspired me to pursue my Ph.D., and I wouldn't have wanted to work with anyone else.

I would also like to thank my advisory committee (Drs. Stephen Kry, Eun Young Han, Samantha Simiele, and Valerie Reed) whose guidance and support have been a critical part of my learning and success in graduate school. I would like to especially acknowledge Dr. Eun Young Han who has been an amazing mentor to me and assisted with the development and execution of several of the experiments in this study. She taught me that an important part of being a great scientist is being persistent.

A key to my success in this project is of course all of the wonderful members of the Court Yard, whom I was so lucky to have the opportunity to work with. I've been fortunate to be part of a community of people who provide unconditional support and enthusiasm to each other. I wouldn't have survived my Ph.D. without the amazing friends I have made at MD Anderson, and I'm so grateful for all of you.

I owe a special thank you to my wonderful family, who have always believed in me. My parents, Cathy and Shawn, have always encouraged me to go after my dreams, and never once doubted that I could achieve whatever I set my mind to. I

couldn’t have done this without their constant love and support. And to my siblings, Kate, Tim, and Danny, who have always inspired me to question everything. The love of learning and the curious nature that is essential to great science is something I learned from them.

To my best friend, and the best medical physicist I know, Eric Welch. We met when I was struggling to find my footing at Vanderbilt, and it’s because of you that I learned to love medical physics. You have helped me through hard times and taught me how to be a better student, better clinician, and better person. I have continued to be inspired by you, and I’m so grateful that I have a friend like you in the field.

I couldn’t have done any of this without my amazing partner Andrew Wallace. When I decided to move halfway across the country to pursue my Ph.D., I was lucky enough to have a partner willing to go on this adventure with me. He kept me steady and optimistic through many months of working from home during COVID and has continued to believe in me even on days when I didn’t believe in myself. Having you by my side has helped me through some of the hardest days, and I’m so excited to begin the next stage of our adventure together in Boston.

And of course, I have to thank my dog Boson for the positive impact he has had on my mental health throughout this process. Boson has given me reasons to smile even on days when the world of research and academia seemed too daunting. After sitting by my side through 80% of my Ph.D., I think he has earned the title of Dr. Boson from this point forward.

THE SAFE AND EFFECTIVE CLINICAL DEPLOYMENT OF ARTIFICIAL INTELLIGENCE TOOLS

Kelly Nealon, M.S.

Advisory Professor: Laurence Court, Ph.D.

18 million new cancer cases are diagnosed each year. Roughly half of these patients will be treated with radiation therapy, a complex technique that requires an interdisciplinary team of clinical staff and expensive equipment to be delivered safely. Cancer centers in Low- and Middle-Income Countries (LMIC) have an especially difficult time meeting the demands of radiation therapy as the complexity of treatment techniques increase, with only 37% of patients in these regions having access to the care they need. Artificial Intelligence (AI) based tools are being developed to simplify the treatment planning and quality assurance processes to increase the number of patients who can be treated, as well as improving the quality of their treatment plans. While AI techniques have shown great promise, with any new technology it is important to not only assess the potential benefits, but also the associated risk. To this end, we have performed a risk assessment of our in-house automated treatment planning system, the Radiation Planning Assistant, to identify points of risk and subsequently develop appropriate quality assurance and training resources to minimize patient risk.

To identify points of risk, a failure mode and effects analysis was performed by a multidisciplinary team of clinicians and software developers. Changes were then made to limit the risk of 76% of high-risk failures. These risk points were then incorporated into hazard testing, and we found that 62% of errors could be detected

before a plan was created in the RPA. The user interface was then modified to limit the number of errors that will be propagated into the automatic planning process.

Following the changes made to optimize the safety of the user interface, the efficacy of error detection during the plan review process was assessed. A custom checklist was developed to guide the review of automatically generated treatment plans, based on the results of our FMEA and AAPM TG-275. During final physics plan checks, when utilizing the customized checklist, we found an increase in the rate of error detection by 20% for physicists and 17% for medical physics residents.

An end-to-end test was then performed to evaluate the entirety of the RPA training and deployment procedure for new users. Users were asked to review training materials and generate 10 treatment plans, including all treatment sites available in the RPA. Following training, 100% of the errors present in these plans were detected and users reported that the developed training materials provided them with all information needed to generate safe, high-quality, treatment plans.

Finally, a real-time contour monitoring system was developed to limit the risk of systematic errors and detect abnormalities in the contouring process that could be attributed to software error, off-label use, or automation bias.

In conclusion, we have optimized the safety and efficacy of the RPA training, quality assurance, and deployment processes. This evaluation has allowed us to not only maximize the impact of our automated treatment planning tool, the RPA, but has also generated results that should be used to inform the development of safe AI software and clinical deployment procedures, in future clinical environments.

Table of Contents

List of Illustrations

List of Tables

Chapter 1: Introduction

Roughly half of the 18 million new cancer patients diagnosed each year are treated with radiation therapy, a complex technique that requires an interdisciplinary team of clinical staff and expensive equipment to be delivered safely.[1] This complexity has led to a deficit in the number of patients who have access to the care they need, as well as large variability in planning outcomes.[2] In order to address this deficiency, Artificial Intelligence (AI) based tools are being introduced into radiotherapy workflows to assist with organ contouring and treatment planning. These tools use complex algorithms to perform similar functions as human planners while lowering staffing demands and creating more consistent and often higher-quality treatment plans.[3]

Manual treatment planning has been the standard of care for decades, however human factors, such as planner experience, fatigue, and training can lead to a large variation in the final plan output. Inter-observer variability in target volume contouring has been found to introduce the largest uncertainty in the treatment planning process for many tumor sites, an error that could potentially result in geographic miss in dose delivery, and ultimately lower the probability of tumor control.[4] The treatment planning process is also susceptible to inter-observer variability, with clear variations in planning techniques, and output being evident even within a single institution.[5] The planning process is prone to errors with as much as 33% of near-miss errors identified in patients' treatments originating in the treatment planning step of the workflow.[6] To address this problem, AI-based tools

are being introduced to clinical workflows to standardize the contouring and planning processes.[7]

Automated contouring and treatment planning tools are now being used clinically, with both commercial and institutionally developed options available.[8–10] The Radiation Planning Assistant (RPA) is an automated contouring and treatment planning tool under development which will lower the workload for low-resource clinics, potentially increasing the number of patients who can receive radiation treatments.[11] Thus far, we have developed auto-planning tools for cervix, chest wall, and head and neck cancers, with other anatomical sites in progress.[12–18] Using the RPA, the time and resources needed will decrease, allowing for consistent, high-quality treatment plans for all patients, regardless of their resources or location.

Automated tools have the potential to greatly simplify workflows, including in radiotherapy, however, they will also introduce new complexities and uncertainties which have yet to be adequately evaluated.[19] In 2005, a radiation accident occurred in New York, due to errors in an automated process. A patient with oropharyngeal cancer was treated with a higher dose than planned due to a failure in automation and the subsequent treatment plan exporting process. A treatment plan was generated which was missing the intended multi-leaf collimator sequences and the patient received treatment with an unshielded treatment plan, rather than the intended plan which included shielding of critical organs. Due to inconsistency in procedure and quality assurance (QA) mechanisms, this patient was treated for three days with the corrupted plan, leading to overdose, unintended toxicities, and

ultimately patient death.[20] In order to prevent incidents like this from happening, new technologies must be thoroughly evaluated to mitigate unintended consequences.

To maximize patient safety when introducing AI tools, a risk assessment must be performed and recommendations for the standardization and regulation of AI must be developed to ensure that the radiotherapy workflow remains safe and efficient for all patients. Using the RPA as a case study, this work will focus on developing best practice guidelines for how to produce, perform quality assurance for and clinically deploy safe and useful automated tools.

Chapter 2: Purpose and Central Hypothesis

Central Hypothesis:

We hypothesized that 90% of clinically relevant errors introduced by automated treatment planning tools can be prevented or detected by establishing a robust risk evaluation process and developing a thorough quality assurance and deployment procedure.

Specific Aim 1:

Aim: A risk analysis of the impact of AI in clinical practice

Hypothesis: Risk assessments can be used to reduce the risk profile of utilizing AI in radiotherapy treatments.

By developing an in-depth understanding of the risk profile when utilizing automation in clinical practice, it is possible to maximize the safety of these processes and subsequently mitigate risk. To achieve this, we performed a failure mode and effects analysis (FMEA) to assess the risk associated with an artificial intelligence-based treatment planning system. Based on the results of this analysis, changes were made to improve the risk profile of our system. Simulated hazard scenarios were used to evaluate and quantify the final risk profile of our automated system.

Aim 1.1: An FMEA risk assessment of the Radiation Planning Assistant

Aim 1.2: Simulate hazard scenarios to identify weaknesses in the Radiation Planning Assistant Workflow

The work towards Specific Aim 1.1 is presented in Chapter 3: A Risk Analysis of the Impact of AI in Clinical Practice.

The work towards Aim 1.2 is presented in Chapter 4: Using Hazard Scenarios to Identify Points of Weakness.

Specific Aim 2:

Aim: Maximize the role of supported quality assurance to ensure the safe deployment of AI-based contouring and planning tools

Hypothesis: Developing quality assurance aids, such as checklists and clear plan documentation, for treatment plans created using Artificial Intelligence can help mitigate risk and improve overall safety.

To reduce the risk of errors reaching and negatively impacting patients, quality assurance mechanisms are put in place as part of the radiotherapy process. Due to the changes made to clinical practice when automated tools are introduced, quality assurance resources must be adapted to address and mitigate these previously unexplored points of error. In Aim 2, a plan reporting and quality assurance system was developed for use with AI-based tools, to prevent clinical errors from occurring. We examined and quantified the impact that the AI-supported workflow has on the efficiency and effectiveness of plan QA, using plans with simulated errors and surveys of clinical staff. We also evaluated the deployment of automated contouring

and planning tools for a variety of anatomical sites, both internally and at partner institutions by performing an end-to-end assessment of the training and deployment procedure. This allowed us to optimize the training process to ensure that all users are properly equipped to identify and address any errors which could arise when utilizing automated tools.

Aim 2.1: Evaluate the impact the AI-supported workflow has had on the efficiency and effectiveness of plan QA

Aim 2.2: Perform an End-to-End test of the RPA deployment and training strategy

The work towards Specific Aim 2.1 is presented in Chapter 5: Maximize the Role of Supported Quality Assurance to Ensure the Safe Deployment of AI-based Contouring and Planning Tools

The work towards Specific Aim 2.2 is presented in Chapter 6: Performing an End-to-End test of the RPA deployment and training strategy: A Pilot Study

Specific Aim 3:

Aim: Evaluating the clinical use and acceptability of automatically generated contours

Hypothesis: Monitoring of patient contour edits can lead to increased detection of systematic errors, such as those caused by software error, automation bias, or off-label use.

While careful software development, thorough quality assurance, and robust training processes can limit the amount of error that occurs when utilizing automated tools, unexpected errors may still arise during clinical use. These errors could be due to failures of the software, such as the generation of low-quality contours, automation bias, in which users do not carefully review the provided contours, and off-label use, in which contours are generated for a site outside of the intended scope of the automated system. To address these abnormalities in clinical use, a monitoring system has been developed to perform a real-time evaluation of the magnitude of edits made to automatically generated contours. Based on the results of this evaluation, action thresholds can be set which will flag contours that fall outside of the expected range of edits for further review.

> Aim 3.1: Perform real-time monitoring of automated contouring tools in the clinic, to evaluate the edits needed to achieve clinical acceptability.

The work towards Specific Aim 3 is presented in Chapter 7: Evaluating the Clinical Use and Acceptability of Automatically Generated Contours.

Chapter 3: A Risk Analysis of the Impact of AI in Clinical Practice

This chapter is based on the following article:

Nealon KA, Balter PA, Douglas RJ, Fullen DK, Nitsch PL, Olanrewaju AM, Soliman M, Court LE. Using Failure Mode and Effects Analysis to Evaluate Risk in the Clinical Adoption of Automated Contouring and Treatment Planning Tools. *Pract Radiat Oncol*. Published online 2022. doi:10.1016/J.PRRO.2022.01.003

3.1 – Introduction

Each year, the number of people diagnosed with cancer continues to increase, with more than 19 million new cases in 2020.[21] Of these new cancer cases, roughly half will be treated with radiation therapy, a technique that requires a team of interdisciplinary clinical staff and expensive equipment. Cancer centers in low and middle-income countries (LMIC) have an especially difficult time meeting the demands of radiation therapy as the complexity of treatment techniques increases, with only 37% of patients in these countries having access to the care they need.[22,23]

Artificial intelligence (AI)-based tools are being developed at an unprecedented rate to simplify treatment planning and quality assurance processes, thus increasing the number of patients who can be treated and improving the quality of their treatment plans. Current applications of AI in radiation oncology include automated contouring, dose optimization, and plan quality assurance, with both research and commercial solutions becoming available.[8,9,19,24]

One approach to applying AI in radiation therapy is that of the Radiation Planning Assistant (RPA), an automated contouring and treatment planning tool that is currently being developed. This software uses AI to simplify the time-consuming and user-intensive planning process, which compensates for staffing deficiencies and increases the availability of high-quality plans in low and middle-income countries[11]. The current iteration makes use of deep learning techniques to provide planning options for treating head and neck, cervix, and chest wall cancers.[12–18]

AI techniques have shown great promise, but as with any new technology, it is important to not only assess the potential benefits but also the associated risks[25]. Understanding risks is important, both in the design of the tool itself and in how it is integrated into the clinical workflow. A failure modes and effects analysis (FMEA) has been used to examine the weaknesses of new technologies, helping to predict and mitigate potential errors.[26–28]

In this paper, we applied the FMEA approach to the RPA to understand the risk of deploying this tool in clinics locally and in low and middle-income countries. Based on the results of our FMEA, changes were made to the RPA workflow to reduce the associated risk.

3.2 - Methods and Materials

For this study, we assembled a multidisciplinary team of RPA users consisting of one physician, three physicists, one dosimetrist, two members of the RPA development team, and a representative from our institutional radiation therapy quality improvement team. All members were from the same institution, and each clinical group member had at least 3 years of experience and understanding of the types of errors that occur in a typical radiation therapy workflow. Before beginning the FMEA risk assessment, each user was given an overview of the functionality of the RPA and was instructed to perform the contouring and treatment planning workflow for several test patients to familiarize themselves with the system.

We then developed a process map to identify each step in the RPA workflow that required user intervention. Potential failure modes and causes were identified for each step in the process. When considering potential failures, we began by focusing on operator error and software error, two commonly identified themes in the radiation therapy literature. We then evaluated additional potential problems that could arise from the introduction of new techniques and technologies into a previously established workflow. We closely examined errors that could occur because of an intentional or unintentional unwillingness to adjust the RPA. We also considered what errors could occur because of inadequate user training. Finally, we looked closely at any potential automation bias-based errors, which are caused by the tendency to rely too heavily on the outputs of automated systems. For example, an automatically generated plan may not be reviewed as thoroughly as a plan

created by a human planner because many people expect the output of automated systems to be consistent and of high quality.[29]

Each failure mode was discussed in the group to prevent bias from any individual user and scored by the entire FMEA team. Each criterion—severity, occurrence, and detectability—was scored on a scale from 1 to 10, as suggested by the American Association of Physicists in Medicine Task Group 100. For severity, a qualitative scale was used in which a score of 1 indicates that the error had no impact on patient care and a score of 10 indicates that the error could lead to patient fatality. Occurrence and detectability were also scored using the quantitative 10-point scale provided in TG-100 (Table 1).[30]

Table 2. Failure modes and effects analysis scoring criteria for occurrence, severity, and detectability from TG-100.

Rank	Occurrence (*O*)		Severity (*S*)		Detectability (*D*)
	Qualitative	Frequency in %	Qualitative	Categorization	Estimated Probability of failure going undetected in %
1	Failure unlikely	0.01	No effect		0.01
2		0.02	Inconvenience	Inconvenience	0.2
3	Relatively few failures	0.05			0.5
4		0.1	Minor dosimetric error	Suboptimal plan or treatment	1.0
5		<0.2	Limited toxicity or tumor underdose	Wrong dose, dose distribution, location, or volume	2.0
6	Occasional failures	<0.5			5.0
7		<1	Potentially serious toxicity or tumor underdose		10
8	Repeated failures	<2			15
9		<5	Possible very serious toxicity or tumor underdose	Very wrong dose, dose distribution, location, or volume	20
10	Failures inevitable	>5	Catastrophic		>20

Criteria were scored using a combination of RPA plan data and clinical experience from FMEA committee members. The risk priority number (RPN) for each failure mode was calculated by multiplying each failure's occurrence, severity,

and detectability scores together. A failure mode with a higher RPN indicates that more risk is associated with the error than with another error with a lower RPN.

After each failure mode was scored, the total list of failures was reviewed by the FMEA team to ensure that the scoring was consistent throughout. Failure modes with high RPNs were discussed to assess how the associated risk could be lowered. For this portion of the study, we focused on failure modes with an RPN greater than 125 or a severity score greater than 7. We discussed making changes to the system and ultimately implementing them into the RPA. Failure modes were then rescored to reflect the changes.

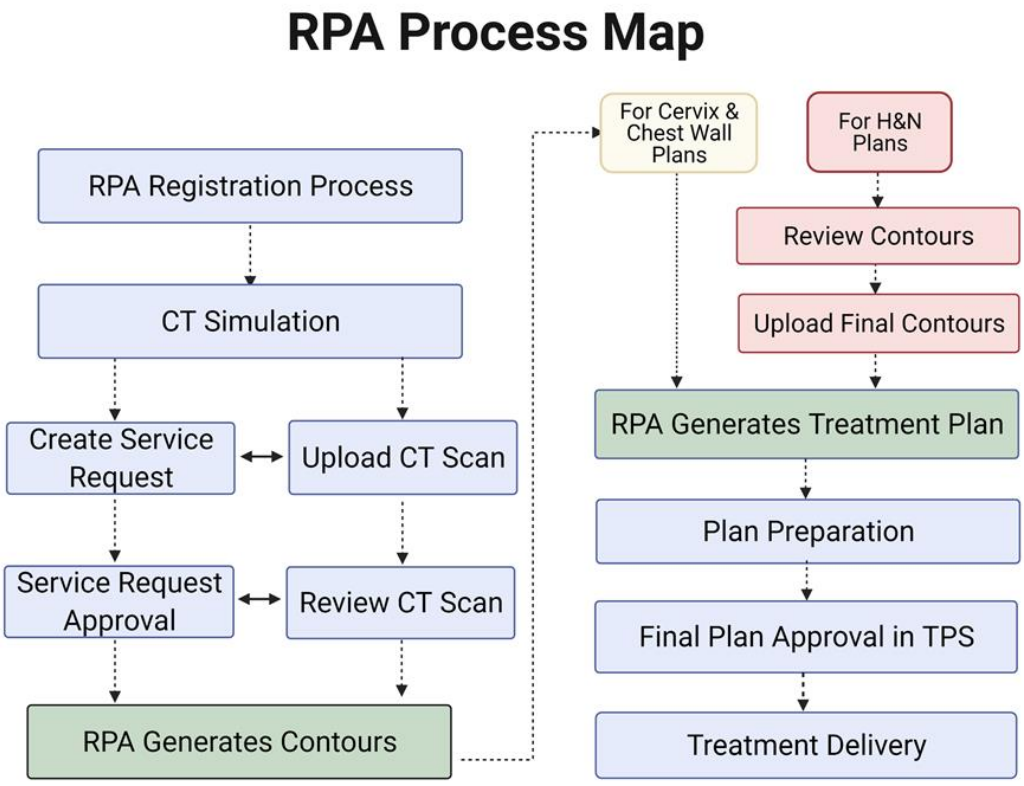


Fig. 7. Process map for the radiation planning assistant (RPA). CT, computed tomography; H&N, head and neck; TPS, treatment planning system.

3.3 - Results

3.3.1 - Process Map

Rather than examining the RPA's internal processes, we developed a process map to demonstrate the flow of the RPA system (Figure 1) from the user's perspective. Any step that required user intervention or action was isolated, and failure modes were assigned to each step individually (Figure 2). Figure 3 shows the distribution of identified high-risk (RPN >125) failure modes for each step in the workflow during which they occur.

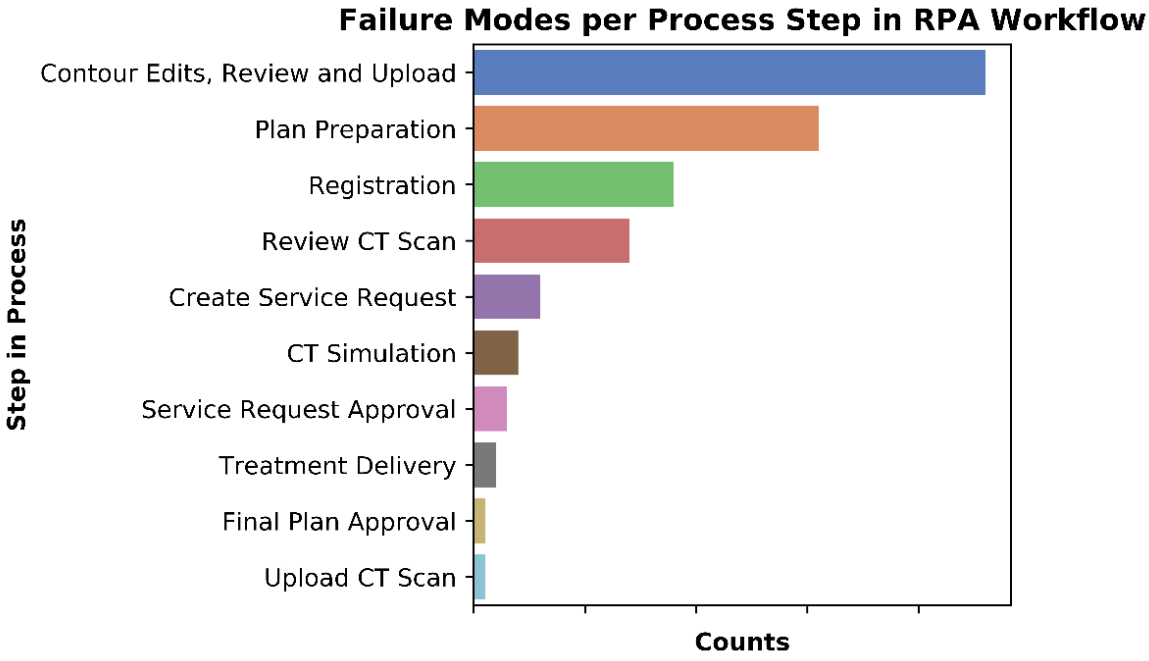


Fig. 2. All identified failure modes, sorted by the step in workflow during which they occurred. RPA, Radiation Planning Assistant; CT, computed tomography.

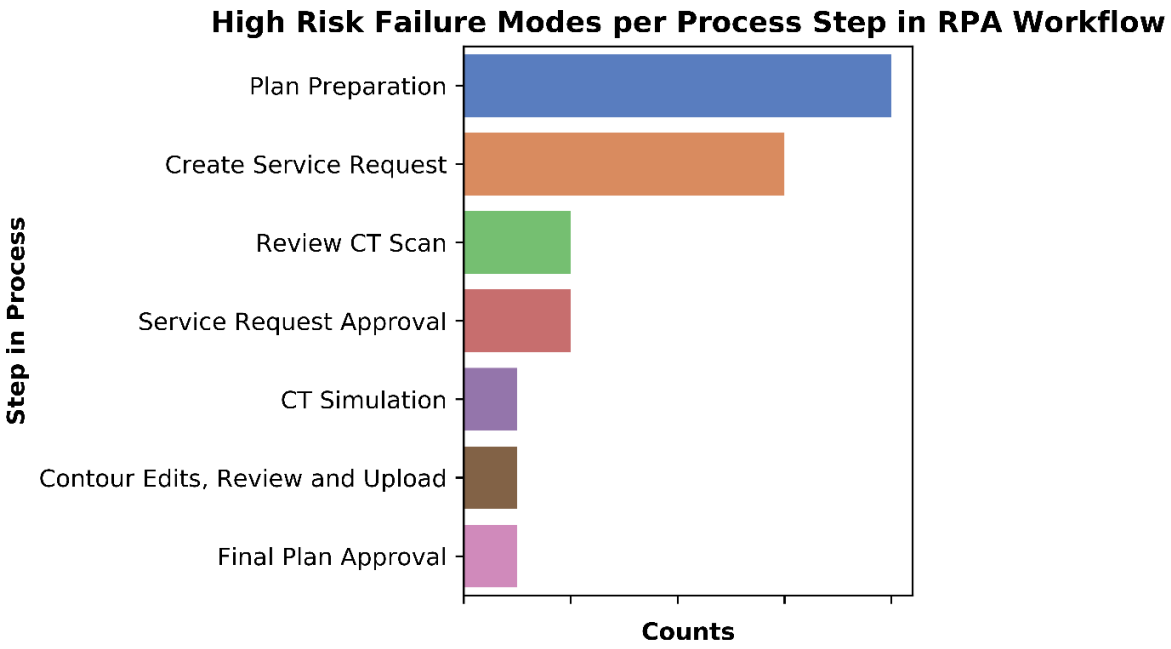


Fig. 8. All identified high-risk (>125 risk priority number) failure modes, sorted by the step in the workflow during which they occurred. RPA, Radiation Planning Assistant, CT, computed tomography.

3.3.2 - Cause of Error

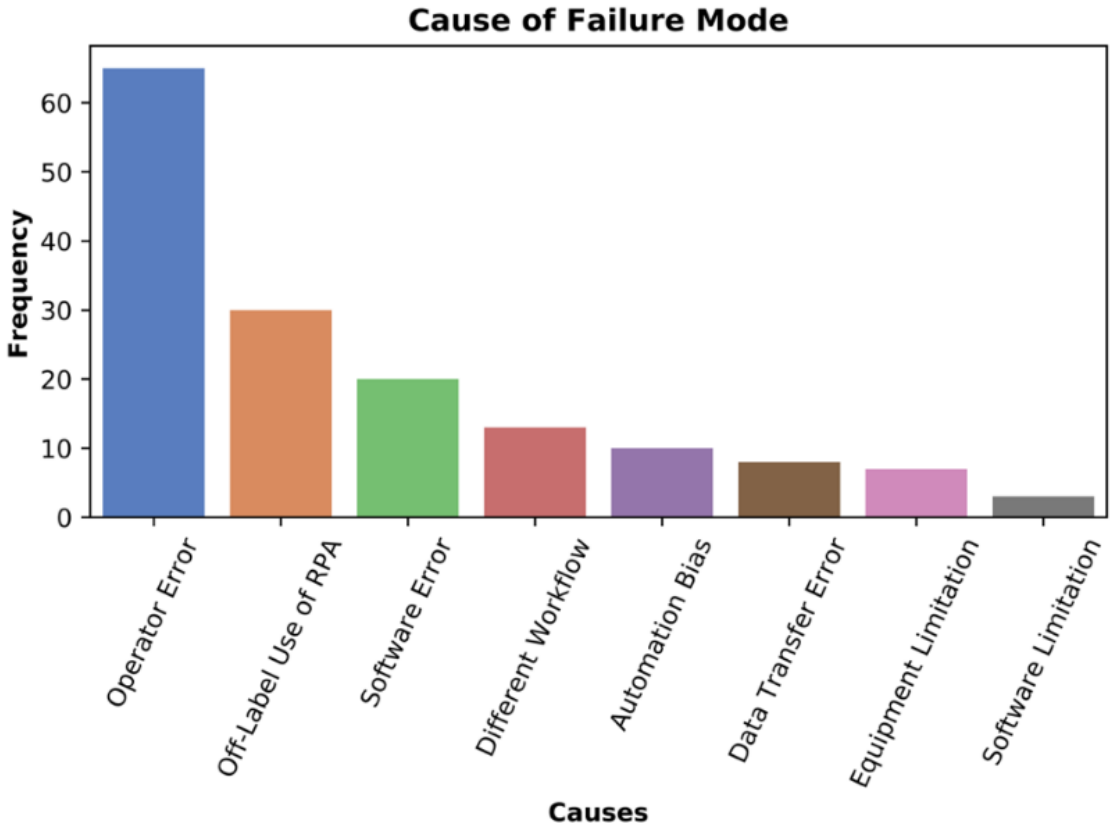


Fig. 9. Distribution of causes for all identified failure modes. RPA, Radiation Planning Assistant.

Each failure mode was also classified according to the cause of the error (Figures 4 and 5). By isolating the cause of each failure mode, we can more accurately tailor the corrective action to reduce the associated risk.

While assigning causes to each failure mode, we identified eight error categories: operator error, different workflow, software error, software limitation, data transfer error, equipment limitation, automation bias, and off-label use of the RPA. Each of these are defined below.

Operator error was defined as any accidental user input, such as clicking the wrong button or typing information incorrectly. *Different workflow* was defined as an error that occurred because clinicians were accustomed to treating patients according to a workflow that was incompatible with the use of the RPA. *Software error* was defined as an issue within the RPA software that either negatively impacted the plan's quality or limited the plan's creation. *Software limitation* was defined as an error that occurred if the user attempted to create a plan that was outside of the RPA's capabilities. *Data transfer error* was defined as the inability of plan data to be entirely or accurately transferred between the RPA and a user's local system. *Equipment limitation* was defined as errors caused by the inability of the user's local equipment or infrastructure to meet the RPA's requirements (e.g., a low-quality CT scanner or unstable internet connection). *Automation bias* was defined as the tendency of humans to depend too heavily on software tools, resulting in inadequate output examination. The final error category, *off-label use of the RPA*, was defined as unintended software use. The RPA has specific site and treatment guidelines in place to ensure that users receive a usable and safe treatment plan for every patient. Plans outside of the appropriate statements of use could lead to mistreatment. While it was important for us to acknowledge this possibility, these failure modes were removed from our final evaluation because off-label application use is not specific to the RPA. To mitigate off-label RPA use, the risk will be communicated to all users, plan auditing will be performed, and a variety of training materials will be available.

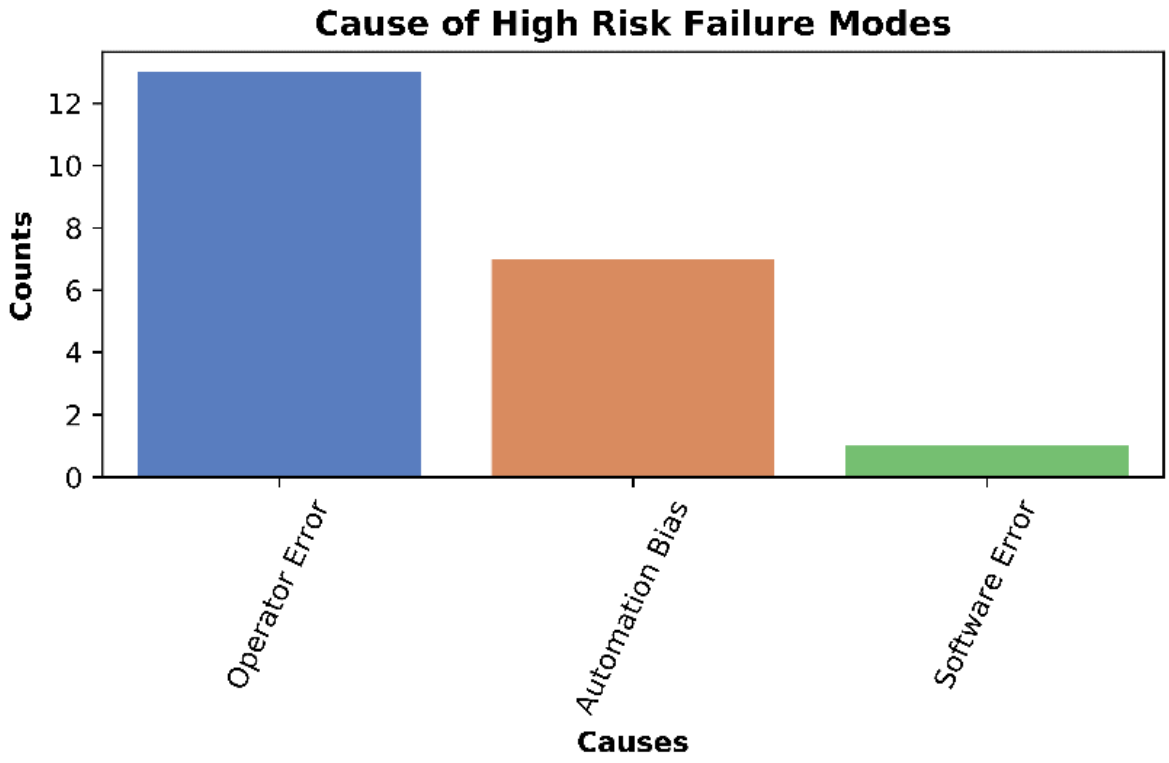


Fig. 10. Distribution of causes for high-risk (risk priority number > 125) failure modes

3.3.3 - Failure Modes

In the current RPA workflow, we identified 290 failure modes. Of these, 126 were specific to the RPA workflow and 164 could occur in a conventional (i.e., manual) radiation therapy treatment planning workflow. To focus on the RPA-attributed risks, we omitted these 164 general failure modes from the analysis. The remaining 126 failure modes were classified according to the step in the workflow and the cause of the error, as shown above. The full list of RPA-specific failure modes can be found in Appendix A. The mean RPN of these failure modes was 56.3; 105 errors had an RPN below 125, which is the TG-100-recommended threshold at which action should be taken to reduce the risk. The 10 highest-risk failure modes are included in Table 2. To provide an understanding of how failure

modes were identified and scored in this study, we explain how three high-scoring failure modes were scored below.

The “plan not reviewed carefully prior to approval” error, a result of automation bias, was defined as users relying too heavily on the RPA’s capabilities, resulting in the user trusting that the plan was clinically acceptable and not reviewing it diligently. While the RPA is a reliable tool, it still has the potential to generate plans that are not ideal for a given patient. For example, if contours need to be corrected, the plan must be recalculated in the local treatment planning system (TPS) for the linear accelerator being used. If the plan is not reviewed, the patient could be seriously injured or exposed to toxic levels of radiation. Thus, we assigned the “plan not reviewed carefully prior to approval” error a severity score of 9. Automation bias is a common phenomenon, so we assigned this error an occurrence score of 6, indicating a relatively high probability. Because the plan is reviewed at the end of the workflow, the likelihood of this error being detected is low; therefore, we assigned this error a detectability score of 9. Taken together, these scores result in an RPN of 486, indicating the necessity of a careful plan review by all clinicians.

The “RPA printout used as plan documentation” error arises from a situation in which the clinician decides to use the final plan report generated by the RPA as the primary record for a patient. The plan printout briefly summarizes treatment planning and provides results for the necessary internal quality assurance checks. Although this document contains useful information, it is only a brief snapshot of the treatment plan. The DICOM file must still be imported into the users’ local TPS and recalculated. The physician should review the plan's quality, and necessary quality

assurance should be performed. The final review of the plan should occur in the local TPS. Because the RPA plan printout summarizes an intermediate step in the overall treatment planning process, using the printout as documentation could result in low-quality treatment. Because of the potential for organ at risk (OAR) toxicity or tumor underdose, we assigned the “RPA printout used as plan documentation error” a severity score of 7, indicating a severe error. Because the printout is readily available, the occurrence of this error is moderate; thus, we assigned it an occurrence score of 6, signifying that automation bias is likely. Because this error would most likely not be detected, we assigned it a detectability score of 10. If the clinical procedure uses the plan PDF, then the potential flaws in the plan, which are only visible in the TPS review, would never be identified. Taken together, these scores resulted in a final RPN of 420.

The “contoured target incorrectly” error is applicable in head and neck cancer treatment planning, in which CTV2 and CTV3 are automatically generated by the RPA to include the lymph node regions specified by the user. A review of previously generated RPA plans found that minor edits were needed for the autogenerated CTV volumes in roughly 5% of cases. Thus, we assigned the “contoured target incorrectly” error an occurrence score of 9. The target would be missed or organs would be subjected to additional toxicity if the target were contoured inaccurately; thus, we assigned this error a severity score of 5. While the error could be detected when the physician reviews the plan, the differences could be fairly subtle or only problematic in small regions. We, therefore, determined that this error was only moderately detectable and assigned it a detectability score of 6. Together, the scores resulted in a total RPN of 270 for this software error.

Failure Mode	Step	Cause	S	O	D	RPN
Plan not reviewed carefully prior to approval	Final Plan Approval	Automation bias	9	6	9	486
RPA printout used as plan documentation	Plan Preparation	Automation bias	7	6	10	420
Necessary physics QA was not performed	Plan Preparation	Automation bias	10	4	9	360
Request approved without careful physician review	Approve Service Request	Automation bias	9	6	6	324
Contoured target incorrectly	Contouring	Software error	5	9	6	270
Incorrect identification of CTV2 and CTV3 for head and neck plans	Create Service Request	Automation bias	5	5	10	250
Incorrect identification of prior radiation	Create Service Request	Operator error	10	4	6	240
Incorrect answer on pregnancy questionnaire	Create Service Request	Operator error	10	4	5	200
Incorrect plan downloaded	Plan Preparation	Operator error	9	3	7	189
Incorrect plan imported into TPS	Plan Preparation	Operator error	9	3	7	189

Table 2. Ten highest-scoring failure modes identified in the RPA workflow. S, severity; O, occurrence; D, detectability; RPN, risk priority number; RPA, Radiation Planning Assistant; QA, quality assurance; TPS, treatment planning system.

3.3.4 - Corrective Actions

Corrective actions were applied to any failure mode with an RPN greater than 125. We identified 21 failure modes above this threshold, resulting in changes to the RPA workflow and infrastructure. Changes to the software included 1) the removal of patient information questions from the RPA service request, 2) verification checkboxes to ensure that correct target coverage was selected, and 3) the addition of redundancy checks to ensure that the correct laterality was selected. To reduce the occurrence of mistakes, we made guidance on reference point placement and

statements of appropriate use more easily accessible to users. Finally, RPA plan reports now include statements of use and reminders to the user that final checks must be made after the plan has been imported into their local TPS. More detailed training tools are also being developed to assist users with making informed and safe decisions about RPA usage.

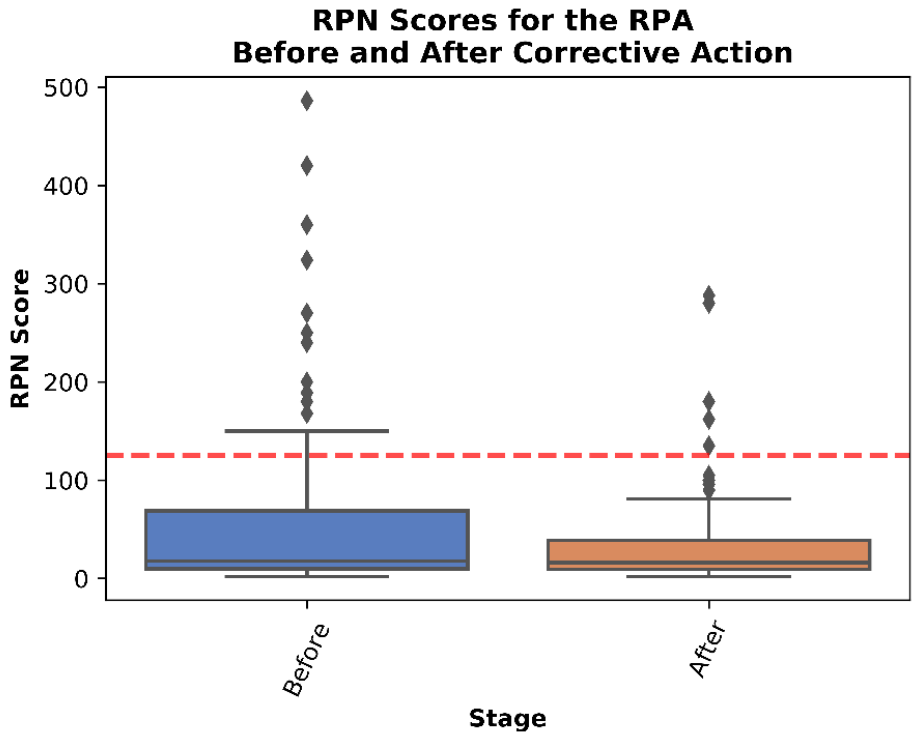


Fig. 11. RPNs for the RPA workflow before and after corrective actions were made to limit risk to patients. RPN, risk priority number; RPA, Radiation Planning Assistant.

Following the implementation of the aforementioned risk reduction techniques, failure modes were rescored to reflect the updated system, with a final mean RPN of 33.7 and a final maximum RPN of 288 (Figure 6). This decrease in RPN indicates that the changes were able to successfully reduce the risk associated with the use of the RPA system. The number of failure modes that exceeded the action threshold of RPN 125 decreased from 21 to 5, showing a 76% reduction in high-risk errors. We were also able to increase the detectability of 15 of the 21 errors (71%), ensuring that any errors that do occur can be more easily detected by users.

The 10 highest scoring modes after the application of risk reduction techniques can be seen in Table 3. The full list of rescored failure modes can be found in Appendix B.

Failure Mode	Step	Cause	S	O	D	RPN
Plan not reviewed carefully prior to approval	Final Plan Approval	Automation bias	9	4	8	288
RPA printout used as plan documentation	Plan Preparation	Automation bias	7	4	10	280
Necessary physics QA was not performed	Plan Preparation	Automation bias	7	2	9	180
Request approved without careful physician review	Approve Service Request	Automation bias	9	3	6	162
Contoured target incorrectly	Contouring	Software error	5	9	3	135
Incorrect identification of CTV2 and CTV3 for head and neck plans	Create Service Request	Automation bias	5	2	10	100
Incorrect BB placement	CT Simulation	Operator error	9	3	3	81
Marked isocenter not verified in TPS	Plan Preparation	Operator error	9	3	3	81
Shift not validated in TPS	Plan Preparation	Operator error	9	3	3	81
Request is edited by another user prior to approval	Approve Service Request	Operator error	9	4	2	72

Table 3. Final 10 highest scoring failure modes, rescored after risk mitigation adjustments were made to the Radiation Planning Assistant workflow. S, severity; O, occurrence; D, detectability; RPN, risk priority number; QA, quality assurance; BB, radiopaque markers; CT, computed tomography; TPS, treatment planning system.

3.4 – Discussion

After we rescored the high-risk failure modes to reflect the changes made, five still exceeded the TG-100-established threshold of RPN of 125. Of these, four can be attributed to automation bias or a user's overreliance on the RPA. To combat this potential problem, training will be provided that emphasizes the importance of performing thorough plan checks and quality assurance, including a discussion of the risks of overreliance on automated solutions. We are also developing checklists for physics and radiation oncology plan checks that are designed to support the plan

review process and emphasize its importance. The final high-scoring error was "contoured target incorrectly" as mentioned above; as a result of this study, an automated target contouring quality assurance tool, similar to one already employed for normal tissue, is currently being developed for the RPA to reduce the detectability score of this error[31]. This quality assurance would flag any contours without good agreement between a secondary auto contouring method, warning the user not to proceed with the plan without careful review.

3.4.1 - Integration with TPS

The current version of the RPA is a web-based tool. This design was chosen to maximize availability and minimize cost, thus allowing future use by clinics with limited budgets and resources. However, this approach introduces the need for additional data transfer between the user's systems and the RPA website, and 32 of the identified failure modes were related to data transfer. Although these errors were not high-risk (i.e., their RPNs were less than 125), further integration of the RPA tools with the local planning system would eliminate these modes, resulting in an inherently safer process. The gains from integrating the RPA with the local planning system are similar to those resulting from integrating other automatic tools with a planning system (such as in commercial solutions) and from improving integration with other software tools in radiation therapy (e.g., planning systems, auto contouring systems, and oncology information systems).

3.4.2 - Automation Bias

Automation bias is a well-established phenomenon in decision-making scenarios aided by automated tools; it was first recognized as a point of risk in the airline industry when pilots were provided with, and ultimately depended too heavily on, electronic flight planning tools in the 1990s[32]. Since then, a number of studies have assessed the frequency of and solution to automation bias in medicine[32–34]. While a clear consensus has yet to be reached, Goddard et al recommend that automation bias be mitigated in both the development and deployment steps. Presenting users with clear confidence levels for outputs, highlighting the importance of user responsibility, and substantial user training in the AI software can help to lower the likelihood of inappropriate automation bias[33]. Based on the results of these studies, the risk mitigation techniques used for automation bias in the RPA have included 1) verification boxes to ensure that the user has knowingly consented to treatment objectives, 2) clear internal quality assurance metrics for the final plan output, and 3) the development of a variety of training and tools for new and established users. Clear transparency regarding the risk associated with automation bias will be provided during user training.

3.4.3 - Operator Error

Operator error was found to be the most frequent cause of failure in the risk assessment of the RPA system, accounting for more than half of the identified failure modes. While the impact of operator error is potentially large, a literature review indicates that operator error is also a prevalent concern in manual processes.

Operator error has been identified as a cause of failure in FMEAs in many aspects of the radiation therapy workflow, showing that this problem is not unique to the RPA[30, 35–37]. This raises the question: Is operator error more common in automated workflows than in manual workflows?

Wexler et al performed an FMEA for the commissioning of TPSs and found that the number of high-risk failures due to human error in automation-aided workflows decreased compared to that in manual workflows as did the average and maximum RPNs for the processes[38]. To limit the impact of operator error on the RPA workflow, several changes have been implemented to reduce the occurrence and increase the detectability of these errors. First, redundancy checks have been established to verify patient information, laterality, and prescription. This is accomplished by performing automatic checks of the information (laterality) or forcing the user to perform a secondary review of the input information prior to proceeding. Automated quality assurance checks have also been added, which will display a failed result to improve the user's detectability of plan errors. Additionally, an RPA-specific plan checklist is being developed that will focus on the high-risk operator-based errors that were identified in this study.

3.4.4 - Impact on Deployment and Staff Training

The RPA is being developed to provide automated contouring and treatment planning tools to clinics with limited resources, potentially saving the clinical teams hours of preparation time for each patient. Our study has shown that it is vital that

the teams review the plan before it is implemented; such a review may play an even more important role in the RPA workflow than in manual processes. A plan review by clinicians is particularly important as training programs do not always sufficiently emphasize these reviews during clinical training[39,40]. Additionally, the time taken to perform these checks must be considered when evaluating any potential workflow benefits of the RPA or other automated processes.

3.5 – Conclusions

An FMEA was performed on the current version of the RPA, an automated contouring and treatment planning program. As a result of this analysis, changes were made to the RPA interface, training tools, and workflow to limit risk. Following these improvements, the number of high-risk failure modes decreased by 76%. The detectability of these high-risk failure modes also improved by 71%, ensuring that any errors that do occur can be more easily detected by users. The vast majority of identified high-risk failure modes were related to automation bias or operator error, especially actions related to plan quality review and quality assurance. Thus, when automation is added to the radiation therapy process, plan review by physicians, physicists, and other clinical staff is important, and staff must be thoroughly trained in this process.

Chapter 4: Using Hazard Scenarios to Identify Points of Weakness

This chapter is based on the article “Hazard testing to reduce risk in new clinical workflows“ currently under review with the Journal of Applied Clinical Medical Physics.

4.1 – Introduction

To address the increasing complexities in radiation therapy, various commercial and in-house automated solutions are being introduced into treatment planning workflows to assist with contouring, planning, and quality assurance.[8–10,41–46] These tools can improve both the consistency and the quality of patients' final treatment plans; however, they also introduce new steps into the clinical workflow that must be evaluated for safety, reliability, and usability. Traditionally, when new technologies are introduced into the radiotherapy workflow, they undergo commissioning to ensure that they can be used safely and accurately. The American Association of Physicists in Medicine has released task group reports that provide recommendations on how to commission treatment planning systems, linear accelerators, intensity-modulated radiation therapy systems, and other technologies.[47–49] Many of these recommendations focus on preventing errors with the software, equipment, or calculations that could impact patient safety. While these factors are important and must be considered, the reports have often omitted the value of risk assessments for identifying additional points of weakness. In one prospective risk assessment—a failure mode and effects analysis—to assess the

clinical implementation of automated tools, the most common errors were caused not by issues with software or equipment, but rather by mistakes made by human users.[50] Similarly, a study of reported safety incidents by Weintraub et al.[51] found that while the use of automation can contribute to improvements in clinical workflows, it also creates an increased need for mindfulness from users to ensure patient safety.

One way to assess and optimize safety when introducing automated tools into the clinic is by performing a hazard analysis of the workflow, as recommended by IEC 62366: Application of usability engineering to medical devices.[52] A hazard scenario is a problematic or dangerous situation that could arise when an error is introduced into a workflow. If such an error were to go undetected by members of a radiation therapy team, it would increase risk and compromise patient safety. By performing a hazard analysis, the cause of such hazard scenarios can be determined and subsequently mitigated by implementing additional safeguards. One example of hazard analysis is a study by Pawlicki et al.[53], who utilized a tool called system theoretic process analysis (STPA) to identify and eliminate potential hazards in clinical radiation therapy workflows. Another study showed the benefits of using hazard analysis to assess the clinical safety of using the Halcyon treatment system.[54]

In this study, we used hazard testing to evaluate the human component of radiation oncology workflows, which is often not addressed in commissioning processes, using the Radiation Planning Assistant (RPA), an automated contouring and treatment planning software tool, as a case study.[11]

4.2 – Methods

4.2.1 - The Radiation Planning Assistant

The RPA is an automated contouring and treatment planning tool currently under development to provide high-quality radiation therapy treatments to low-resource communities throughout the world.[11] The RPA is a web-based system that uses artificial intelligence to simplify the planning process. The current version of the RPA can create plans for treating cancers of the head and neck, chest wall, cervix, and whole brain, with more sites under development.[12,13,15–18,55–58]

For the RPA to generate treatment plans, the user must input a computed tomography (CT) scan and the prescription information for each patient. That information is then verified by the users to ensure that it matches the intended final treatment plan before the RPA begins the automated planning process. Figure 7 shows each step of the user-facing workflow.

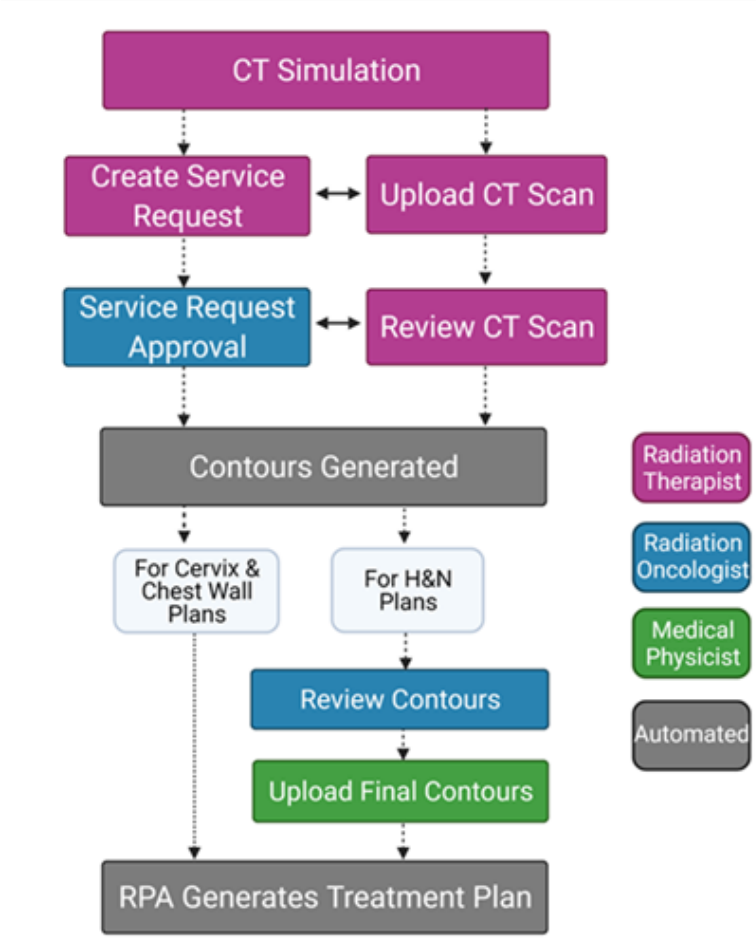


Fig. 7. Process map of the Radiation Planning Assistant workflow. H&N, head and neck.

To assess the potential for detection of hazards in the RPA, we examined three steps of the process: service request approval, review CT scan, and the review and upload of contours. These steps were selected because after each piece of data is approved, information is sent to the RPA to generate plans; therefore, the accuracy of the data at these points is imperative for the creation of a safe treatment plan.

A service request in the RPA is a document containing patient information, planning techniques, prescription, and dose constraints. A service request is created for every patient for whom the RPA will be used for contouring or planning. The information included must be correct, as it cannot be changed later in the planning process without the creation of an entirely new plan. While a service request can be created by any member of the clinical team, it must be approved by a radiation oncologist. Therefore, to assess the effectiveness of data review at this stage, radiation oncologists were recruited to review and approve service requests.

Each plan created using the RPA also requires the upload of the patient's CT scan. Following upload, the CT must be reviewed to ensure it is for the correct patient and the correct treatment site. Users must also ensure that the image quality and field of view are acceptable and that any artifacts are minimal. They must also verify that the isocenter was identified correctly by the RPA before proceeding. We anticipated that this step of the workflow would likely be performed by a radiation therapist immediately after performing the CT scan, so radiation therapists were recruited to review and approve CT scans.

Following the approval of the service request and the CT scan, the RPA generates contours for head and neck or cervix plans. These contours can then be manually edited by the treating physician or dosimetrist if corrections for organs at risk are needed or to create additional target volumes. They are then re-uploaded into the RPA for a final review before the final plan generation. Following upload, a PDF document is generated which requires users to review the contours on axial images slice by slice before approving for planning. The PDF also reports any edits

that users make to the autogenerated contours to highlight modifications that may require more careful review. As a preliminary contour review would have been completed by physicians before uploading the structure set into the RPA, errors that occur in this step in the workflow are likely to be due to failures in data transfer, rather than issues with contour quality. Therefore, the details of contour approval overlap with checks typically performed during physics plan review. As such, we anticipate that the contour review and approval step will be the responsibility of medical physicists so for this study, physicists were recruited to review and approve the final contours.

Hazards were selected from a failure mode and effects analysis of the RPA based on three criteria.[50] First, the error must have occurred previously, as this indicates it may happen again. Next, all hazards were scored with a severity greater than three (on a scale from one to ten, with one indicating a minor inconvenience and ten representing possible patient fatality), indicating that if they occurred, at a minimum the final plan would contain a dosimetric error. Finally, all hazards must be able to be simulated for testing. The final list of hazards selected for testing is in Table 4.

	Description	S	O	D	RPN	Relevant data input task	Hazard category
1	Isocenter position not identified correctly	9	3	3	81	CT upload/review	RPA error
2	Reference point at the wrong position	6	4	5	150	CT upload/review	Human error
3	Inappropriate CT field of view	6	4	8	192	CT upload/review	Human error
4	Error in data entry in service request – wrong CTV for H&N	5	4	2	40	Service request	Human error
5	Error in data entry for service request – wrong prescription	9	6	3	162	Service request	Human error
6	Contours approved despite non- contiguous slices	4	3	5	120	Contour approval	Human error
7	Failure in automated contouring – target (CTV2)	5	9	6	270	Contour approval	RPA error, Automation bias

Table 4. Hazards evaluated in this study. S, severity; O, occurrence; D, detectability; RPN, risk priority number; CT, computed tomography; CTV, clinical target volume; H&N, head and neck; RPA, Radiation Planning Assistant.

4.2.2 - Hazard Testing

When recruiting participants for this study, we aimed to include a heterogeneous population to ensure the scalability of these results to many types of institutions. To do this, we included reviewers from multiple countries and institutions, with varying experience levels.

The testing process was conducted over a virtual meeting between a participant and a member of the RPA team. Participants were told they would be evaluating the usability of the system and were not told that errors could be included in the data set. The session began with a training video for the task participants would be performing, which detailed how to navigate the interface and what features they should be paying attention to on the screen prior to approval.

The participant then shared their screen, logged into the RPA system, and began performing the prescribed task for a previously assembled set of patients. Participants were instructed to vocalize any concerns or questions they may have while reviewing patient data, and all feedback was recorded.

4.2.2.1- Service Request Approval

During the service request approval step, two errors were identified for testing, both of which focused on the correctness of patient information. First, we tested incorrect nodal level coverage for a patient with head and neck cancer. If the wrong nodes are selected for treatment, the target volumes will be generated incorrectly, and the treatment plan will not cover the desired regions. Next, we created service requests containing the wrong dose prescription for a head and neck cancer patient, which could lead to over- or undertreatment. Both errors in patient information should be detected when physicians compare the patient's in-house prescription document to the RPA service request.

For the service request approval, radiation oncology residents compared a PDF of each patient's prescription to the information included in the RPA service

request to ensure the correct transfer of information. If there was inconsistency between the prescription and the service request or if participants had concerns, they were instructed to reject the request to indicate that it needed to be corrected. If the prescription was accurate, the service request was approved. Each oncologist reviewed a set of ten patient prescriptions that included chest wall, cervix, and head and neck treatment plans. In each set of ten patients, two plans contained errors to be detected.

4.2.2.2 - CT Approval

For the CT approval step, three errors were identified for testing. Incorrect identification of the isocenter describes a scenario in which the RPA is unable to automatically identify the isocenter based on the position of three fiducial markers on the patient's skin. Rather than place the isocenter at the intersection of those points, the software can incorrectly place the isocenter in a different region of the body. This could lead to the creation of an inaccurate treatment plan and the irradiation of unintended tissues if undetected. When reviewing CT scans, users can also place reference points to be used to set boundaries for the treatment plan. For cervix plans, the reference point is used to set the superior border of the treatment field. For chest wall plans, the reference point sets the inferior border of the treatment field. If these points are placed incorrectly, the plan generated could over- or undertreat the patient, affecting tumor control. Finally, CT scans can be uploaded which, if used for planning, could lead to a less accurate treatment plan. One example of this is a CT image in which portions of the patient are cut off owing to an

inappropriate field of view. This could lead to errors in creating the plan and calculating dose.

For CT scan approval, radiation therapists were asked to review CT scans to verify acceptable image quality and scan location. They were provided with the axial, sagittal, and coronal views and were encouraged to navigate slice-by-slice through the images. Before approving the CT scan for planning, users responded to six yes-or-no questions: (1) correct patient, correct CT scan, and correct orientation; (2) correct number of CT slices; (3) acceptable image quality (no large artifacts, implants); (4) correct identification of marked isocenter (except chest wall cases); (5) sufficient axial field of view and craniocaudal extent; and (6) correct position of the reference point (essential for chest wall, optional for cervix 4-field box pelvis). If “no” is selected for any of these questions, the CT scan cannot be used for planning, and a more appropriate CT scan must be used for that patient.

4.2.2.3 - Contour Approval

For the contour matching and approval step of the workflow, two errors were introduced for testing. First, we deleted ten slices of an autogenerated clinical target volume (CTV) contour to be used for the treatment of cervical cancer. This left gaps within the contour that could be seen by scrolling through the axial slices of the CT scans. Next, we made edits to an autogenerated CTV contour to delete one side of the contour from all slices, creating asymmetry and inconsistency. While both of these edits were visually detectable, the RPA’s report also includes warnings when contours have been edited to help direct the users' focus onto those contours. Both

errors would lead to a warning about large edits, which we expected would increase the errors' detectability by our physicist reviewers.

For the contour approval portion of the testing, medical physicists were asked to upload the patient's final contours (DICOM structure file) into the RPA to be used for treatment planning. Once uploaded, the user was required to match each clinical structure's name to the name used for each organ contour in the RPA. Following the matching, a PDF containing each slice of the axial CT scan, with all contours present, was created. Users then reviewed the document to ensure that the contours look clinically appropriate before performing final approval of the contours for planning.

4.2.3 - Usability Testing

Following their completion of the review and approval of each step, users were informed of the hazards present and asked if they had any comments or concerns regarding the safety, effectiveness, ease of use, and user satisfaction of RPA. They were also asked to respond to the questions "How confident are you that you completed this task completely?" on a scale of 1-5, where 1 = not confident and 5= very confident, and "How easy was this task to complete?" where 1 = difficult and 5 = very easy. This feedback will be used to improve our training tools and the usability of the system and to address any safety concerns raised by the users.

4.3 – Results

4.3.1- Service Request

	Errors Detected	Confidence in Use	Ease of Use
Resident 1	50%	4	5
Resident 2	100%	4	4
Resident 3	50%	5	4
Resident 4	100%	4	4
Mean	75%	4.25	4.25
	System Update		
Resident 5	100%	3	3
Resident 6	100%	5	4
Resident 7	100%	5	5
Resident 8	100%	4	4
Resident 9	100%	4	4
Mean	100%	4.2	4

Table 5. Errors detected by radiation oncology residents at the service request approval portion of the RPA treatment planning workflow.

Four radiation oncology residents from four academic institutions in the US reviewed service requests. Of the two errors included in the tests, errors in nodal

coverage went undetected by 50% of the participants. Based on feedback from reviewers, testing was paused while updates were made to the organization of information in the service request document. Testing was then repeated with five new residents to validate the changes. Ultimately, 100% of errors were detected by residents following upgrades to the service request document (Table 5).

4.3.2 - CT Scan Approval

	Errors Detected	Confidence in Use	Ease of Use
Therapist 1	100%	4	5
Therapist 2	100%	5	5
Therapist 3	100%	5	5
Therapist 4	33%	5	5
Therapist 5	100%	5	5
Mean	87%	4.8	5

Table 6. Errors detected by radiation therapists at the CT approval step of the RPA treatment planning workflow.

All of the radiation therapist reviewers reported that the CT review task was clear and easy to complete. In addition, 80% of these reviewers were able to detect and appropriately respond to all hazard scenarios included in the provided set of CT scans (Table 6). Therapist 4 vocalized all errors and showed a clear understanding

of the clinical risk, but due to environmental distractions did not respond accordingly and approved all CTs for planning.

4.3.3 - Contour Approval

	Errors Detected	Confidence in Use	Ease of Use
Physicist 1	50%	5	5
	System Update		
Physicist 2	0%	4	4
Physicist 3	0%	4	3
Physicist 4	0%	4	4
Mean	0%	4	3.7

Table 7. Errors detected by medical physicists at the contour approval portion of the RPA treatment planning workflow.

Four physicists performed a review of the uploaded final contours. The first physicist reviewed 10 patients' plans and detected 50% of errors. During testing, a software bug that affected the display of contour edits was identified, and testing was stopped. Testing then started from scratch, this time using only five patient plans due to time limitations. One error, missing CT slices, was included in the remaining patient plans. This error went undetected by all physicists (Table 7).

4.3.4 - Usability Scores

Overall, at the conclusion of testing, users reported confidence in their understanding of how to use the RPA for treatment planning tasks based on their review of relevant training materials. Therapists felt especially confident, with a mean rating of 4.8/5 on their confidence in their ability to perform the task and a mean rating of 5/5 on the ease of using the RPA. Therapists did report that the task of placing a reference point could have been more clearly explained in the training video. This feedback will be used to clarify user training materials.

Usability scores were slightly lower for the radiation oncology residents, with a mean confidence rating of 4.2/5 and a mean ease-of-use score of 4/5. Multiple residents reported that they did not rate their confidence as 5/5 owing to their lack of experience with the RPA and indicated that their confidence would increase with continued use. The ease-of-use scores were likely lower than the radiation therapists' scores because residents felt that while navigating the interface was easy, approving the prescription information in both in their local treatment planning system and the RPA would add an additional task to their workload.

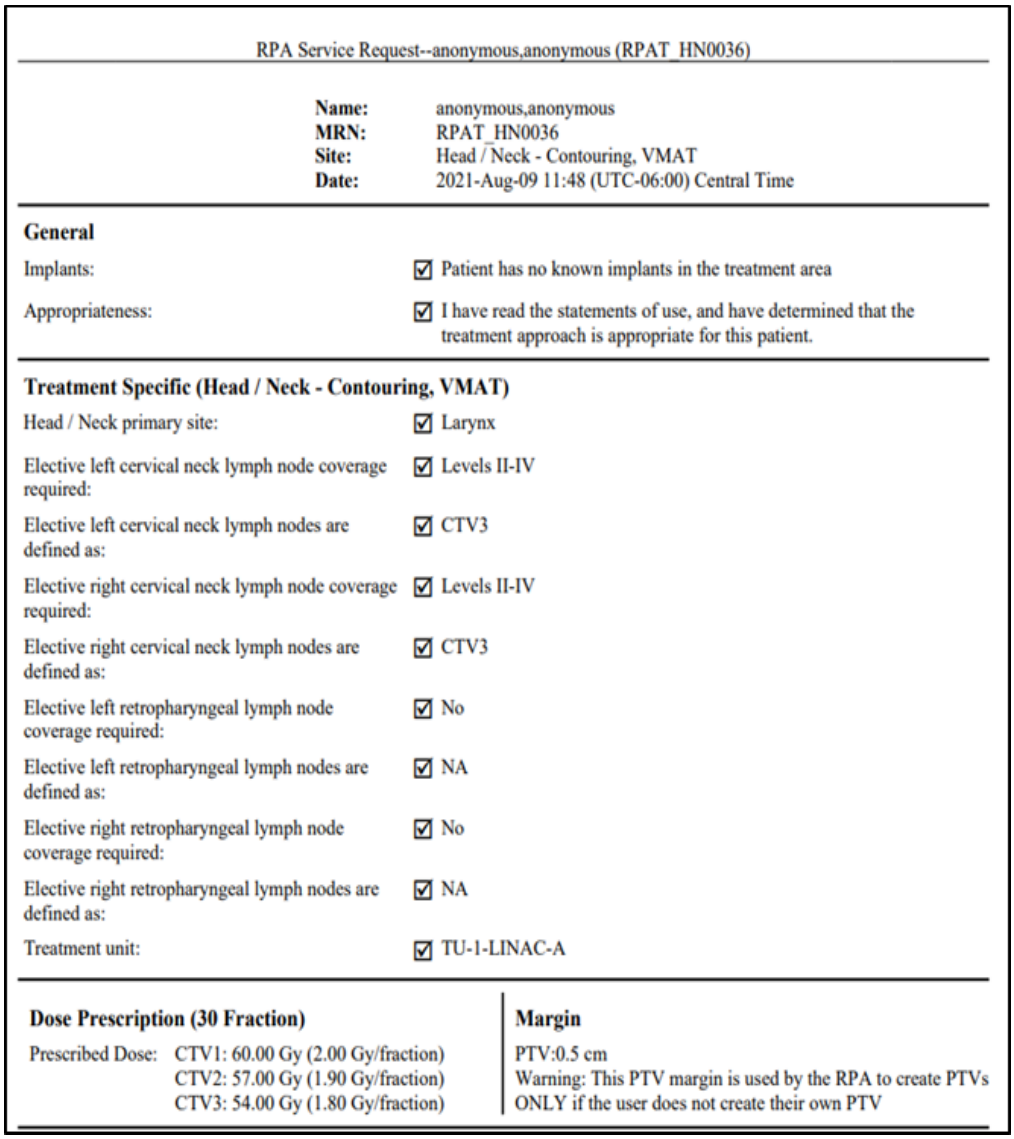
Finally, following the removal of inaccurate CTV contouring from the hazards under evaluation, physicists rated their confidence in the use of the RPA at 4/5 and the ease of use at 3.7/5. Physicists reported that their confidence was not higher because they were unaccustomed to being responsible for contour review. The lower ease-of-use score was attributed to confusing coloring of organs at risk and the overwhelming length of the PDF contour report. This feedback will be incorporated into the final iteration of the contour review process.

4.4 – Discussion

4.4.1 - System Updates

When the radiation oncology residents reviewed service requests, the error most frequently missed was the incorrect selection of nodal coverage, which went undetected by two of the four initial residents. All four of these residents reported concerns about the display of information on the service request, particularly for head and neck patients, for whom the selection of coverage for three separate CTVs is required. Initially, the coverage selections for each patient were presented in a list format (Figure 8).

Residents reported that it was very easy to overlook mistakes when the information was so condensed, and with further discussion, it was suggested that separating nodal selection by laterality would lead to a more intuitive review process. The service request document was updated accordingly (Figure 9). Following the updates, five new residents were asked to review and approve the same ten patient prescriptions, with the updated service request format, for treatment planning. The new cohort of residents all correctly detected both errors (incorrect nodal coverage and prescription dose), validating that this change improved the detectability of errors.



RPA Service Request--anonymous,anonymous (RPAT_HN0036)

Name: anonymous,anonymous
MRN: RPAT_HN0036
Site: Head / Neck - Contouring, VMAT
Date: 2021-Aug-09 11:48 (UTC-06:00) Central Time

General

Implants:	☑ Patient has no known implants in the treatment area
Appropriateness:	☑ I have read the statements of use, and have determined that the treatment approach is appropriate for this patient.

Treatment Specific (Head / Neck - Contouring, VMAT)

Head / Neck primary site:	☑ Larynx
Elective left cervical neck lymph node coverage required:	☑ Levels II-IV
Elective left cervical neck lymph nodes are defined as:	☑ CTV3
Elective right cervical neck lymph node coverage required:	☑ Levels II-IV
Elective right cervical neck lymph nodes are defined as:	☑ CTV3
Elective left retropharyngeal lymph node coverage required:	☑ No
Elective left retropharyngeal lymph nodes are defined as:	☑ NA
Elective right retropharyngeal lymph node coverage required:	☑ No
Elective right retropharyngeal lymph nodes are defined as:	☑ NA
Treatment unit:	☑ TU-1-LINAC-A

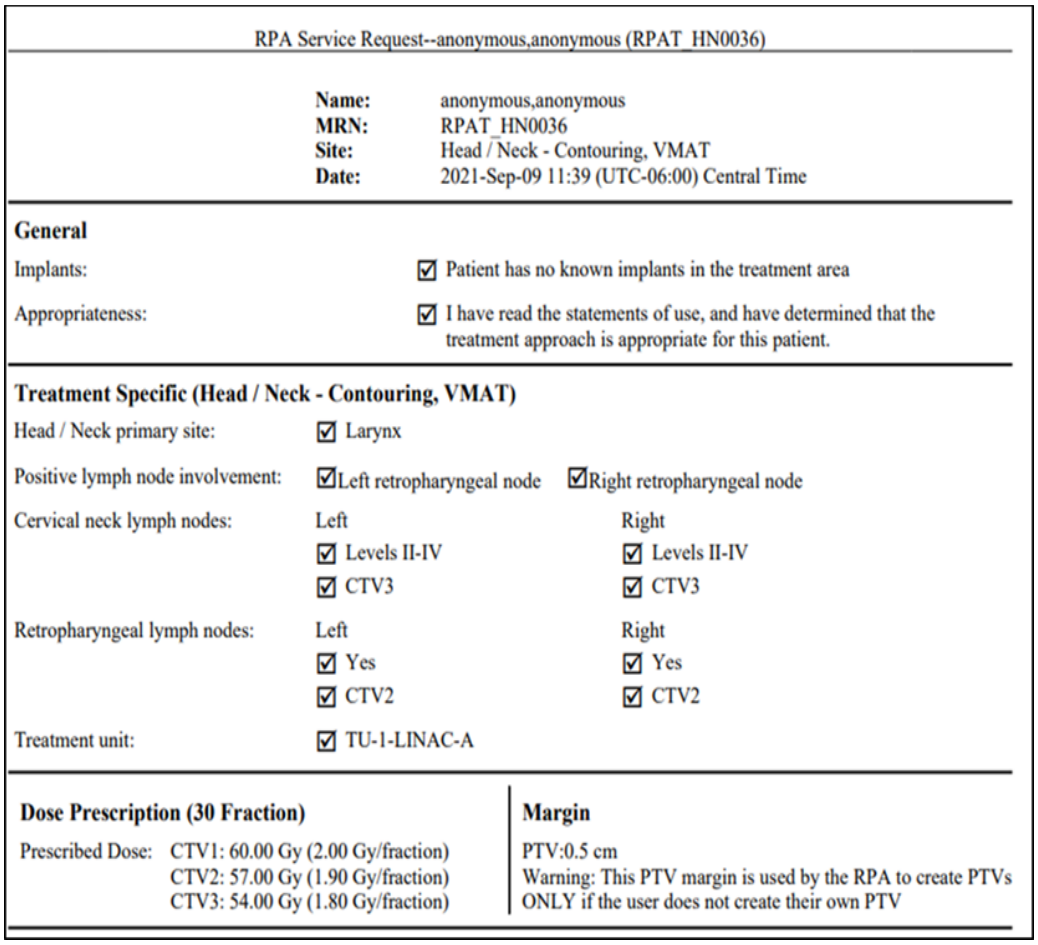
Dose Prescription (30 Fraction)

Prescribed Dose: CTV1: 60.00 Gy (2.00 Gy/fraction)
CTV2: 57.00 Gy (1.90 Gy/fraction)
CTV3: 54.00 Gy (1.80 Gy/fraction)

Margin

PTV:0.5 cm
Warning: This PTV margin is used by the RPA to create PTVs ONLY if the user does not create their own PTV

Fig. 8. Original format of head and neck service request document.



RPA Service Request--anonymous,anonymous (RPAT_HN0036)

Name: anonymous,anonymous
MRN: RPAT_HN0036
Site: Head / Neck - Contouring, VMAT
Date: 2021-Sep-09 11:39 (UTC-06:00) Central Time

General

Implants: ☑ Patient has no known implants in the treatment area

Appropriateness: ☑ I have read the statements of use, and have determined that the treatment approach is appropriate for this patient.

Treatment Specific (Head / Neck - Contouring, VMAT)

Head / Neck primary site: ☑ Larynx

Positive lymph node involvement: ☑Left retropharyngeal node ☑Right retropharyngeal node

Cervical neck lymph nodes:

Left	Right
☑ Levels II-IV	☑ Levels II-IV
☑ CTV3	☑ CTV3

Retropharyngeal lymph nodes:

Left	Right
☑ Yes	☑ Yes
☑ CTV2	☑ CTV2

Treatment unit: ☑ TU-1-LINAC-A

Dose Prescription (30 Fraction)

Prescribed Dose: CTV1: 60.00 Gy (2.00 Gy/fraction)
CTV2: 57.00 Gy (1.90 Gy/fraction)
CTV3: 54.00 Gy (1.80 Gy/fraction)

Margin

PTV:0.5 cm
Warning: This PTV margin is used by the RPA to create PTVs ONLY if the user does not create their own PTV

Fig. 9. Updated service request document, based on user feedback that organizing nodal coverage by laterality would simplify the review of patient information.

When the first physicist reviewed the assigned ten contour sets for approval, several issues were identified that needed to be corrected. First, the training and review of 10 contour sets took significantly longer than the hour that had been allotted for testing. To address this, the patient set was reduced to five to respect the time of our volunteers. Next, we received feedback that physicists would be unlikely to review CTV contours for accuracy at this stage in the workflow, as that task belongs to the physician. Instead, the physicist's task would be to review contours

for integrity to ensure no error in data transfer would occur during the upload process back into the RPA system. As a result of this feedback, we decided to remove the patient who contained inaccurate CTV contouring from the set. Therefore, for the reviewers moving forward, only one error (missing contour slices) was present. Finally, we identified a software bug that caused the system to show that no edits had been made to the contour set despite several slices having been deleted. Testing was paused and the system was updated before proceeding with the remaining physicist reviewers.

4.4.2 - Contour Approval Task

As shown in the results, the contour upload and approval task had an extremely low rate of hazard detection among all participants (0% for the final iteration of the study). Discussing this revealed two primary issues with the design of this step of the test. First, rather than evaluating a simple data quality assurance step as the other cohorts did, the physicists were required to perform several tasks for each patient: (1) find and upload the DICOM RTStruct file for each patient to the RPA; (2) match each final contour to the appropriate name in the RPA; and (3) verify that all contours appeared reasonable and approve for treatment planning. As the contour upload and approval step was primarily seen as a necessary task to move the planning workflow forward, users often did not consider it to be a quality assurance step. The need for this verification and approval step will be emphasized in training to ensure that users are encouraged to review all relevant planning data at each stage of the workflow to limit patient risk. Next, our reviewers identified that

assigning the review of targets to physicists rather than physicians was a weakness of this study. Several of our reviewers commented that contouring and treatment planning were not part of their clinical responsibilities and therefore they did not feel comfortable questioning the output of the RPA or the clinical judgment of the physicians. They also stated that reviewing contours on PDFs rather than in the treatment planning system made it easy to overlook errors unless there were contours explicitly flagged as needing review on the provided contour report.

These results show that unless contour review and approval are performed by a dosimetrist or physician, it is unrealistic to expect quality assurance of contours to occur at this step. To mitigate the potential risk from contouring errors, we will remind all users, especially physicians, to perform a thorough review of contour quality before final plan approval.

4.5 - Conclusion

Hazard testing was used to test an automated contouring and treatment planning process at points where human interaction is necessary. Several failure points were identified and resolved, resulting in a high error detection rate for key process steps. For the review of CT scans, we found an 87% rate of error detection. For the review of service requests, 100% of errors were detected following a system update. We found that one workflow step (contour upload and approval), however, was not performed by members of the radiation therapy team trained to perform contour quality assurance, and no errors (0%) were detected. Therefore, to catch errors, we must highlight the need for a radiation oncologist's review of the final contours and dose distribution to ensure safe operation.

Chapter 5: Development of a custom checklist for use with automatically generated radiotherapy treatment plans.

This chapter is based on the following article:

Nealon KA, Court LE, Douglas RJ, Zhang L, Han EY. Development and validation of a checklist for use with automatically generated radiotherapy plans. *J Appl Clin Med Phys*. 2022;1-7. doi:10.1002/acm2.13694

5.1 - Introduction

Radiotherapy is a complicated treatment technique that is used to treat approximately half of all cancer patients[21]. Radiotherapy requires several components: a CT image of the patient's anatomy, manually or automatically created contours to identify targets and organs at risk, and a treatment plan generated using complex algorithms to model the patient dose. Each step is susceptible to error; as such, a thorough review of the final treatment plan must be performed to limit patient risk. This includes a physics plan review of many aspects of the treatment plan, including patient information, plan dosimetry, and treatment parameters.[59,60]

According to a study by Ford et al, a physics pretreatment plan review is the step of the planning process that is most likely to detect errors before they impact patient treatment[61]. Recommendations have been made regarding the content, frequency, and methods of plan reviews to maximize effectiveness.[25] Checklists have been shown to improve the rate of error detection.[62–66]

While American Association of Physicists in Medicine (AAPM) task group 275 provides recommendations on how to perform a physics plan review, this report was written prior to the automation boom that is currently occurring in radiotherapy.[19] New treatment planning tools automate aspects of the planning process, including contouring, planning, and quality assurance.[8–14,17,18,67] Automation can streamline the process, limiting the need for human interaction and decreasing the planning time.[68–70] While this lack of human input could limit human error, it could also decrease the error detection rate because of the lack of human review. Because of the different workflows used in automated contouring and treatment planning tools, the effectiveness of manual checklists in the physics review process, specifically in automated plans, should be evaluated.

In this study, we developed a customized checklist to improve the rate of errors detected during the review of treatment plans that had been automatically generated by the Radiation Planning Assistant (RPA), an automated contouring and treatment planning tool that is currently under development.[11] Planning errors were simulated, and the physics plan review was performed both without and with the custom checklist. Based on feedback from reviewers, the checklist was modified to optimize the effectiveness for use with automatically generated plans. Although it

was tested with a specific automated process (the RPA), the study results will apply to the automated processes that are increasingly available in commercial treatment planning systems.

5.2 - Methods and Materials

5.2.1 - Checklist development

A customized plan review checklist was developed using guidance from AAPM task groups 275 and 315 (Medical Physics Practice Guideline 11.a).[25,60] Based on the results of a failure modes and effects analysis of the clinical integration of the RPA, the checklist was modified to address additional high-risk points of error[50]. This checklist directly addresses known, common and critical errors which could occur in the RPA planning process. The preliminary checklist (Figure 10) contained 90 items to be checked for each RPA-generated plan; these fell into the categories of general, demographic, prescription and plan directive, simulation, plan information, plan summary, dose calculation, beam's eye views, isodose images, dose verification, and task scheduling. This comprehensive checklist was reviewed by two physicists and several developers from the RPA team, to verify clarity before proceeding.

5.2.2. - Study 1

To evaluate the effectiveness of our plan review checklist we assembled a group of eight physicists from MD Anderson with at least 2 years of clinical

experience, including review of external beam radiotherapy treatment plans. These physicists were also provided with training on how to safely use the RPA as part of this study. The training included videos providing step-by-step instructions for how to generate treatment plans in the RPA, as well as how to review the final plan report. These videos discuss what errors could occur in the plan generation process, and how to detect them in the final plan and report. Users were also provided with all user documentation for the RPA planning system. Participants were instructed to review all training materials and to follow up if they had any questions.

We provided each physicist with 10 automatically generated treatment plans (four cervical cancer plans, three chest wall plans, and three head and neck plans) and imported them into RayStation, along with the corresponding RPA plan report as a PDF file. Details of the automatic algorithms used to generate the plans are described elsewhere.[12,14,56] Of the ten plans provided to each reviewer, five contained deliberate errors, all of which were identified as high risk in our *failure modes and effects analysis* study.[50] These errors included incorrect treatment laterality, unidentified isocenter, incorrect coverage of the target, inappropriate dose normalization, and incorrect placement of the reference point and were introduced in the automated planning process.

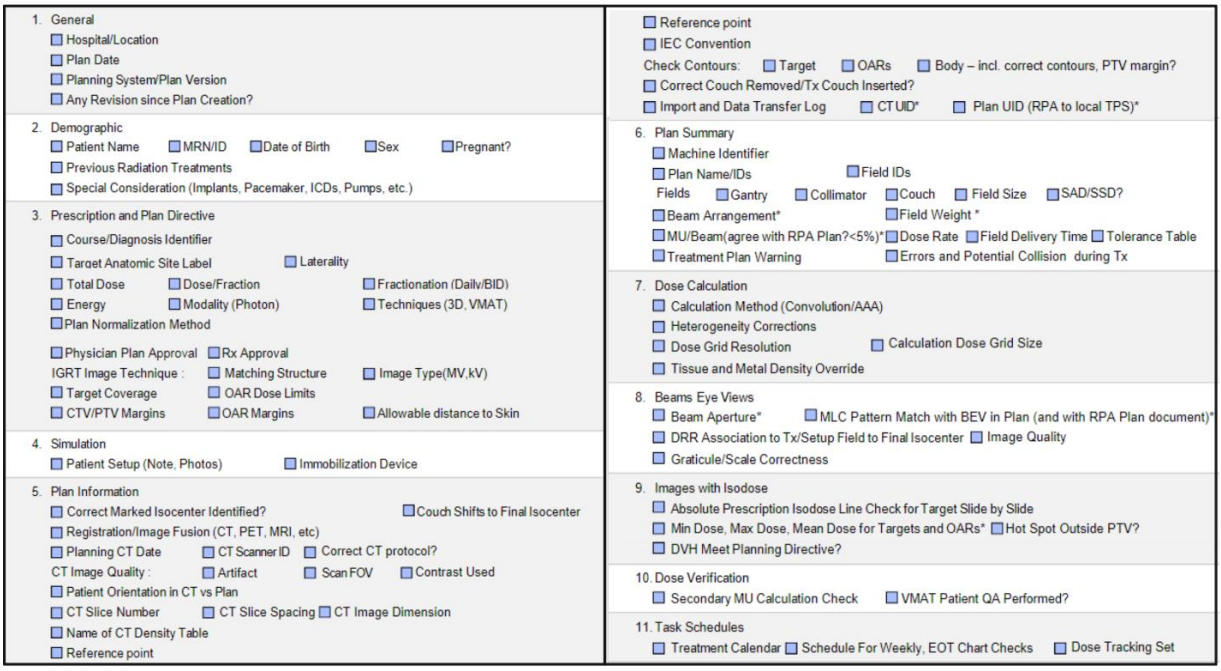


Fig. 10. Checklist (version 1). Items for review were included based on recommendations from AAPM TG-275 and TG-315 and the results of a failure mode and effects analysis of the Radiation Planning Assistant (RPA). 90 total items were included to be reviewed.

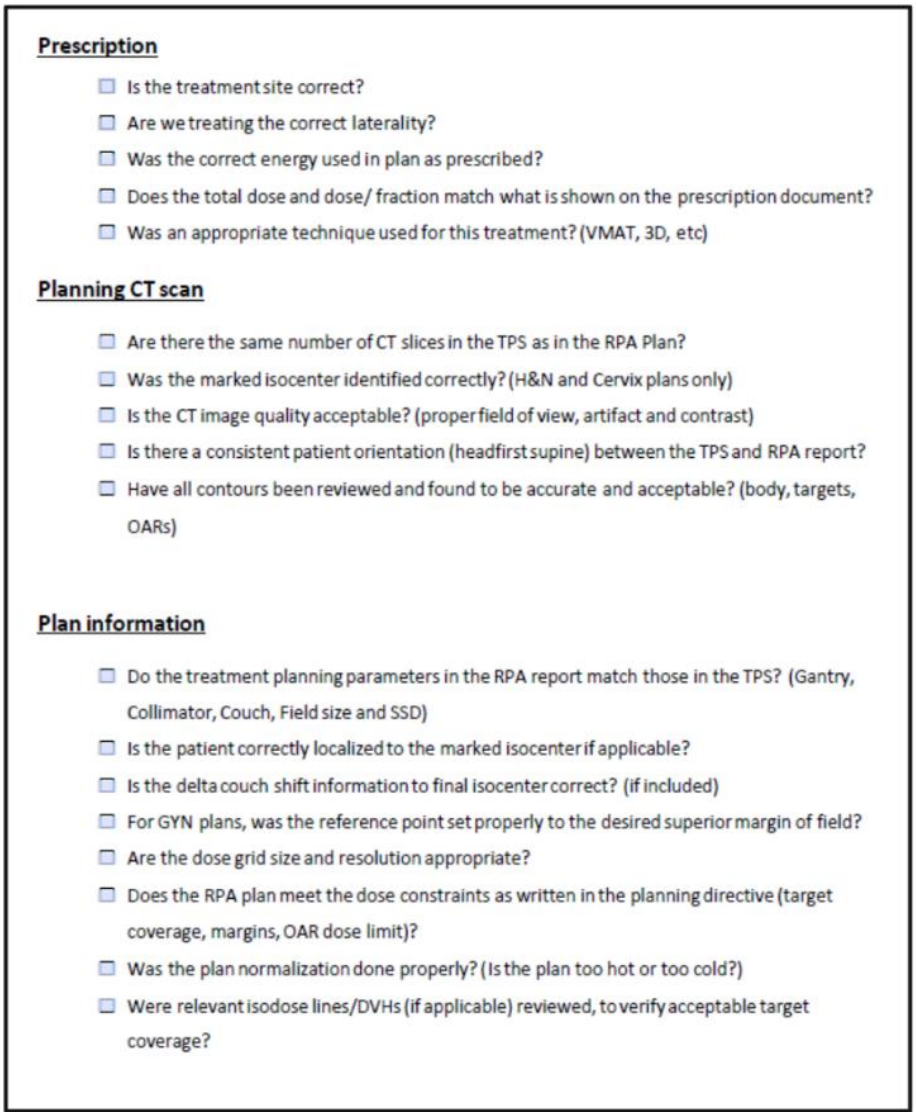
The 8 physicists performed plan quality and safety checks according to their normal process for the 10 automatically generated treatment plans without the checklist and recorded any errors that they found. They also rated the plans based on clinical acceptability and provided written feedback on the plan check process. We then created an additional 10 plans, featuring a similar distribution of treatment site and planning errors, which the participants reviewed following a 2-week break with our customized checklist (Figure 10). After all plan checks had been completed and each plan had been scored, participants were provided with a document summarizing the errors present in each plan and a final, anonymous survey to

evaluate the process. The survey collected information about the overall RPA plan quality, the time needed to check each plan, with and without the checklist, the clarity of the RPA plan report, the usefulness of the plan checklist, and any suggestions to improve the checklist or plan review process.

Modifications were made to the checklist to reflect the results of the survey. The checklist was reduced from 90 items to 18 based on feedback that there was substantial overlap with recommendations from AAPM task group 275, leading to redundancy in the plan review process. While the initial checklist contained items for the entire plan check process (RPA output and final treatment parameters), the revised version focused specifically on the review of the RPA output. The majority of the redundant items were removed, excluding basic planning parameters, and all checks related to identifying automatically generated plan failure modes were preserved.

5.2.3 - Study 2

A second study was then performed with 14 senior medical physics residents from a variety of CAMPEP-accredited residency programs within the United States. Participants were again provided with RPA training materials before proceeding with the plan review process. Six of the participants were chosen at random to be provided with the updated checklist (Figure 11) to assist with their review, and eight were given no checklist. This uneven split was caused by participants from the checklist cohort dropping out prior to completing the study and was not intentional.



Prescription

- ☐ Is the treatment site correct?
- ☐ Are we treating the correct laterality?
- ☐ Was the correct energy used in plan as prescribed?
- ☐ Does the total dose and dose/ fraction match what is shown on the prescription document?
- ☐ Was an appropriate technique used for this treatment? (VMAT, 3D, etc)

Planning CT scan

- ☐ Are there the same number of CT slices in the TPS as in the RPA Plan?
- ☐ Was the marked isocenter identified correctly? (H&N and Cervix plans only)
- ☐ Is the CT image quality acceptable? (proper field of view, artifact and contrast)
- ☐ Is there a consistent patient orientation (headfirst supine) between the TPS and RPA report?
- ☐ Have all contours been reviewed and found to be accurate and acceptable? (body, targets, OARs)

Plan information

- ☐ Do the treatment planning parameters in the RPA report match those in the TPS? (Gantry, Collimator, Couch, Field size and SSD)
- ☐ Is the patient correctly localized to the marked isocenter if applicable?
- ☐ Is the delta couch shift information to final isocenter correct? (if included)
- ☐ For GYN plans, was the reference point set properly to the desired superior margin of field?
- ☐ Are the dose grid size and resolution appropriate?
- ☐ Does the RPA plan meet the dose constraints as written in the planning directive (target coverage, margins, OAR dose limit)?
- ☐ Was the plan normalization done properly? (Is the plan too hot or too cold?)
- ☐ Were relevant isodose lines/DVHs (if applicable) reviewed, to verify acceptable target coverage?

Fig. 11. Checklist (version 2). The initial checklist was revised based on feedback from study participants that the checklist had too much overlap with prior clinical practice. The revised version focuses specifically on the errors which could be present during the review of RPA output, as identified in a prior failure mode and effects analysis.

Each resident reviewed 10 automatically generated plans, five of which contained errors. Residents were given 1 month to complete their review to prevent the study from interfering with their clinical training. The final survey was then

repeated after all plan reviews had been completed.

5.3 – Results

5.3.1 - Study 1

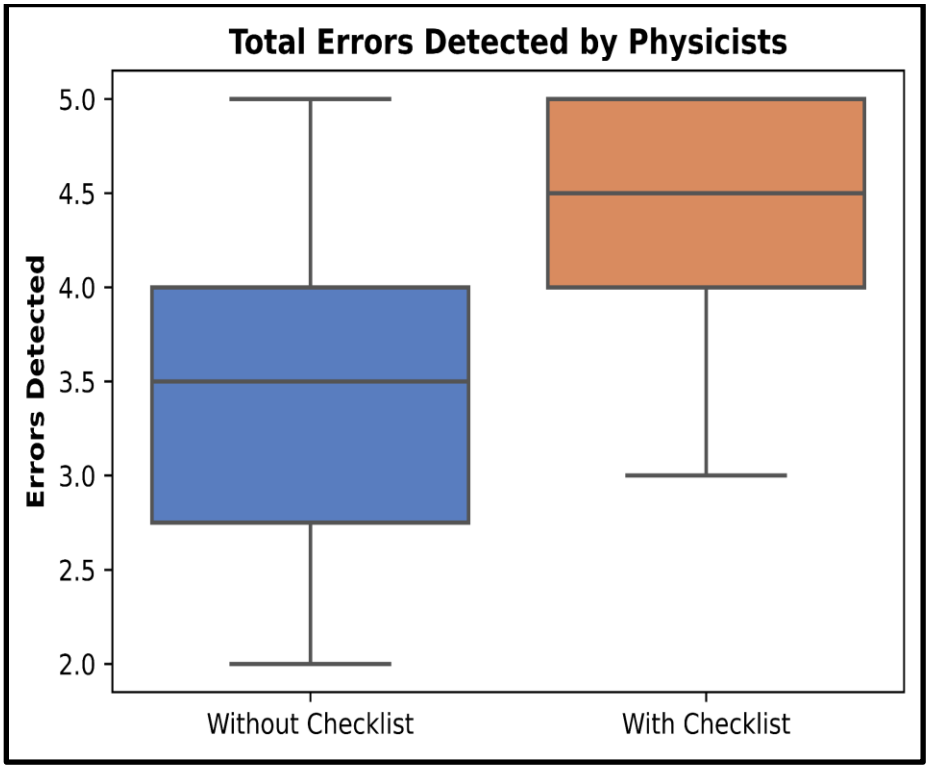


Fig. 12. Errors detected without and with the initial checklist in study 1 (physicists).

All eight physicists completed 20 plan checks each, separated into two phases, without and with the customized checklist. Each phase contained five errors to be detected per physicist and 40 errors in total. In phase 1, 27 errors (68%) were detected, and in phase 2, 35 errors (88%) were detected. Without and with the checklist, the mean and standard deviation of errors detected per participant was 3.4 ± 1.1 and 4.4 ± 0.74, respectively (Figure 12). A t-test indicated that the

improvement in error detection was statistically significant (p=0.02) for the physicist cohort.

5.3.2 - Study 2

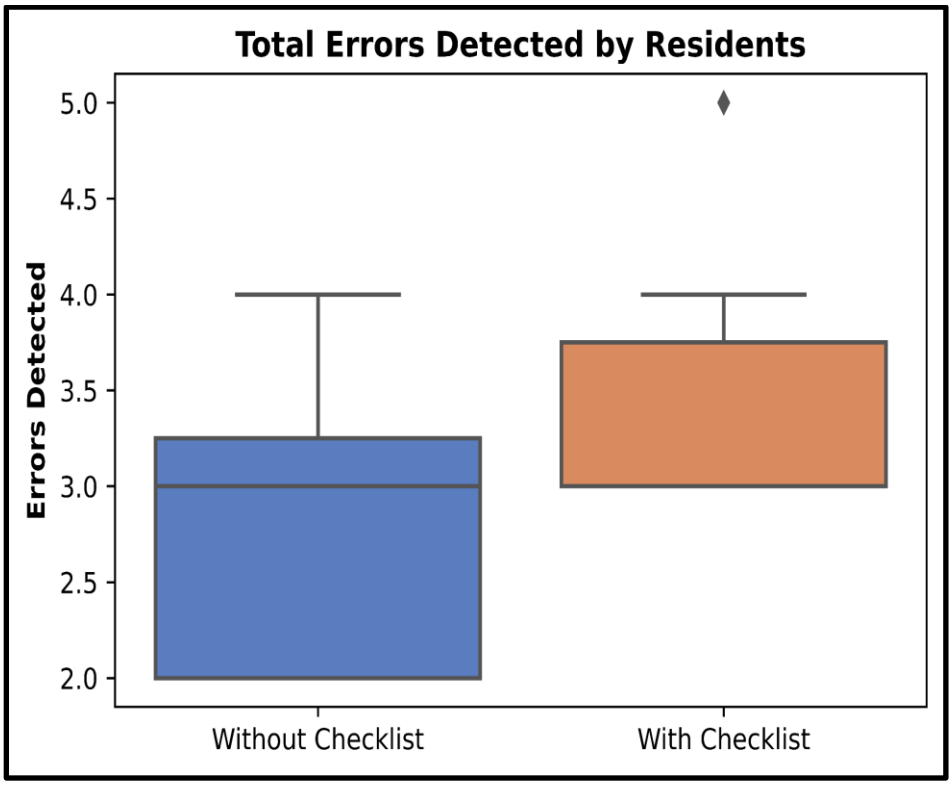


Fig. 13. Errors detected without and with the revised checklist in study 2 (residents).

The revised checklist was assessed by 14 physics residents who completed 10 plan checks each, five of which contained errors. Eight residents completed their reviews without the checklist, and the remaining six participants utilized the checklist. Without the checklist, 53% (21 out of 40) of errors were detected; with the checklist, 70% (21 out of 30) of errors were detected. Without and with the checklist, the mean and standard deviation of errors detected per participant was 2.9 ± 0.84

and 3.5 ± 0.84, respectively (Figure 13). A t-test indicated these results were not statistically significant ($p=0.08$) for the resident cohort, however, the increase in error detection when the checklist was utilized showed that there would be a clinical benefit when using the custom checklist to assist with plan reviews.

5.4 – Discussion

We developed a checklist specifically for use when performing physics reviews of automatically generated treatment plans. Two versions of the checklist were developed and tested with different groups of medical physicists and trainees. While the first checklist was found to be effective at increasing the detectability of errors in the treatment plan, the study participants were overwhelmingly dissatisfied with its length. The revised checklist was significantly shorter but still led to a similar improvement in error detection compared to no checklist.

5.4.1 - Error detection in physics plan review

Errors in treatment planning, both automated and manual, are inevitable. A study by Gopan et al. found that when physics plan checks were performed on a set of treatment plans containing 113 errors, only 67% of errors were detected at this step of the workflow[71]. Similarly, Ford et al. found that pretreatment plan review by physicists leads to the detection of 63% of errors[61]. While these numbers may seem discouraging, Ford et al. also identified that when physics plan review is used in

conjunction with other quality assurance checks, such as physician review, portal dosimetry, and therapist review, 97% of errors were detected before impacting the patient. This reinforces that every step of the treatment planning process should be used for quality assurance to increase redundant checks, and limit patient risk.

The rate of error detection for RPA plans when utilizing the custom checklist (88% for physicists and 70% for residents) is higher than in the prior studies, reinforcing that the final treatment plan and plan report from the RPA is clear and errors are evident. We recognize that the rate of error detection would have ideally been higher for both studies (physicists and residents) when utilizing the checklist. However, the improvement in error detection in both cohorts that used the checklist indicates the utility of this quality assurance aid. Physics plan review should never be used as the stand-alone quality assurance step, and we are confident that when evaluated in conjunction with other stages of the planning process, the rate of error detection will increase.

5.4.2 - Participant experience levels

As this checklist will be used by physicists or other clinicians with varying levels of experience, we evaluated its effectiveness in two separate populations of physicists: clinical faculty physicists with more than 2 years of experience in checking radiotherapy plans and therapeutic medical physics residents who were in their second year of a CAMPEP-accredited residency program.

While the rate of error detection was higher in both participant populations when the checklist was used, we identified a lower rate among the residents in study

2. While this could be a result of a lack of experience or the modifications that were made to the checklist, we also found that on average, the residents reported that they spent less time reviewing each plan than did the experienced physicists; therefore, the lower detection rate could be attributed to a less thorough review. Regardless, the rate was higher in both populations with the checklist (20% in the first study and 18% in the second), indicating that it is an impactful quality assurance aid.

5.4.3 - Trends in error detection

In study 1, when the custom checklist was not utilized for plan review, we found that physicists were least likely to detect an error in plan normalization, such as a dose that is too high or too low (25% detected). When the custom checklist was utilized, the rate of detection for improper plan normalization increased to 75%.

In study 2, we found that both cohorts (with and without the checklist) were unable to detect when the CTV coverage did not match the intended prescription. Without and with the checklist, 0% and 17% of participants detected this error, respectively. The low detection rate when reviewing CTVs can likely be attributed to the lower clinical experience level of the residents, as CTVs based on nodal regions can be difficult to visually delineate. This same error was detected by 50% and 100% of physicists, without and with the checklist respectively, showing the increase of detection with experience.

Incorrect reference point position and incorrect isocenter detection are two errors that are somewhat unique to the RPA workflow, however, we found that for

both studies these errors had the highest detection rates among both the checklist and no checklist cohorts. This highlights a strength of the RPA system – the clarity of the final plan report. When unique errors are easily detectable, this indicates that the presence of the error was effectively displayed on the plan documentation, simplifying the review process.

5.4.4 - Survey feedback

In the survey from study 1, we received feedback that the provided checklist was too long from 80% of the physicists. Respondents also indicated that the checklist presented limited utility due to redundancy with recommendations from TG-275, which inform the standard clinical review process. Thus, the checklist was revised to contain only critical errors that would be more likely to occur with automated planning systems. We expect this checklist to be used as an additional review step for automatically generated plans, in conjunction with the established plan review process Patient information checks in the record and verify system were removed.

Only 60% of participants in the first study reviewed all of the relevant RPA training videos and documentation that they had been provided with. In the second study, only 29% of physics residents reported that they had reviewed all provided training materials. We anticipate that had all training materials been used, the error detection rate would have increased, and the duration of plan review would have decreased, as the auto-generated plans would be more easily understood.

Most (83%) participants in both studies indicated that it took less than 30 minutes to review each plan, both without and with the provided custom checklist. Each participant surveyed reported that overall, the length of plan review was unchanged when using the checklist developed for use with automatically generated plans. We conclude that the use of a quality assurance checklist did not increase the time required to complete the plan review and ultimately increased the rate of error detection; thus, it will be an asset to the physics plan check process.

5.4.5 - Checklist development

This checklist was developed to assist with the physics plan review for treatment plans that are generated using automated tools. Rather than reiterating the recommendations that were made in AAPM task group 275, we generated a supplemental document that should be used in addition to the standard clinical procedure. The final checklist includes errors that were identified as more commonly occurring in plans generated using artificial intelligence-based tools and data from a failure modes and effects analysis study[50]. This decision led to more specific checks and a shorter checklist in the second study.

5.4.6 - Future Deployment

The final iteration of the custom checklist, included in the appendix, will be deployed to physicists for use with the RPA. Training will be provided to help guide the plan review process, with emphasis on possible high-risk failures. Users will then

be provided with test plans to review using the checklist, several of which will contain previously assessed errors. If all errors are detected, the user will be able to proceed with using the RPA for plan generation. If errors are not detected, additional training will be provided to the user and the checklist will be modified to add any missing items. The final iteration of our checklist will be evaluated as part of an end-to-end test of the RPA commissioning and training procedures.

5.4.7 – Limitations

This study included a limited number of participants because of the large time commitment required by each volunteer. In the first study, conducted with experienced clinical physicists, each volunteer participated in two rounds of plan checks, first without and then with the customized checklist. This format could introduce observer bias into the results: each participant was familiar with the RPA plan reports and performing plan reviews before the second phase of the study, which could have resulted in a higher number of errors detected with the checklist. To eliminate this factor from the second round of the study, each physics resident was randomly assigned to the checklist or no checklist cohort, and all plan checks were performed in one session.

This study, including the development of a checklist, was motivated by the results of a failure modes and effects analysis that focused on the RPA system. Therefore, the checklist will need to be adapted when applied to other systems. Our results confirm that checklists are useful with automated planning approaches, which

should apply to other systems, including those that we expect treatment planning system vendors to introduce in future versions.

5.5 – Conclusion

Our results indicate that the use of a customized checklist in the review of automated treatment plans will result in a higher error detection rate and, thus improved patient safety. When physicists completed their plan review utilizing the checklist, the error detection rate increased by 20%, to 88% of total errors being detected. When physics residents completed their plan review utilizing the checklist, the error detection rate increased by 17%, to 70% of total errors being detected. While this analysis was performed using the Radiation Planning Assistant as a case study, we anticipate the results will be scalable to other automated systems.

Chapter 6: Performing an End-to-End test of the RPA deployment and training strategy: A Pilot Study

6.1 – Introduction

Within the next few years, it is anticipated that approximately 20% of radiation oncology clinics will be making use of automated contouring and treatment planning tools[19]. Before utilizing automated tools, new users must have access to appropriate guidance and training to limit the risk that could be passed on to patients.[19,72]

Several failure modes and effect analyses have been performed to evaluate points of risk with the use of automated treatment planning tools and determined that automation introduces unique failure modes which may not be accounted for in standard quality management program.[27,50] While both manual and automatic quality assurance tools have been shown to detect errors introduced in automation-assisted workflows effectively, more thorough user testing must be performed on the entirety of the training and deployment process.[15,73,74]

When introducing new software into clinical practice, Carden et al. recommend that system testing be performed, in which the entirety of the process is evaluated.[75] This concept is often referred to as end-to-end testing and has become common practice in radiation therapy when evaluating new workflows.[76–81] While automation tools undergo piece-wise testing during the development stage, to properly assess their safety, they must be fully evaluated while mirroring the anticipated clinical workflow.[82] By creating test procedures that mimic the anticipated

training and subsequent use of automated tools, the occurrence of errors can be decreased while increasing the quality of the final plan output.[83]

In this work, an end-to-end test of the training and deployment workflow for a novel automated contouring and treatment planning system was performed to evaluate any weaknesses in software function, user training, quality assurance, or ease of use. Based on participant feedback, updates will be made to the system and the new-user onboarding process prior to the clinical release of the automated treatment planning tool.

6.2 - Methods and Materials

6.2.1 - The Radiation Planning Assistant

We performed an end-to-end assessment to evaluate the efficacy of the training and quality assurance resources for our in-house developed software, the Radiation Planning Assistant.

The Radiation Planning Assistant is an automated contouring and treatment planning tool developed to be used in conjunction with each institution's local treatment planning system.[11] The user uploads a patient's CT scan and inputs the desired prescription information, including treatment site, disease extent, fractionation, dose, and treatment machine. The RPA then generates relevant contours for organs-at-risk and target structures. The user reviews these contours, and when approved, a plan is automatically generated to meet the desired planning

objectives. The plan DICOM is then downloaded from the RPA, imported into the local TPS, and recalculated using the desired CT table, dose grid, and local linac parameters. Following the review of this plan, and the automated QA results provided in the RPA plan report, the plan can be modified as desired before clinical use. The process map for this workflow is shown in Figure 14. Treatment planning services are currently available for chest wall (3D), cervix (VMAT and 3D), head and neck VMAT, and whole brain treatments (3D), with other sites under development.[12–15,17,18,56,84]

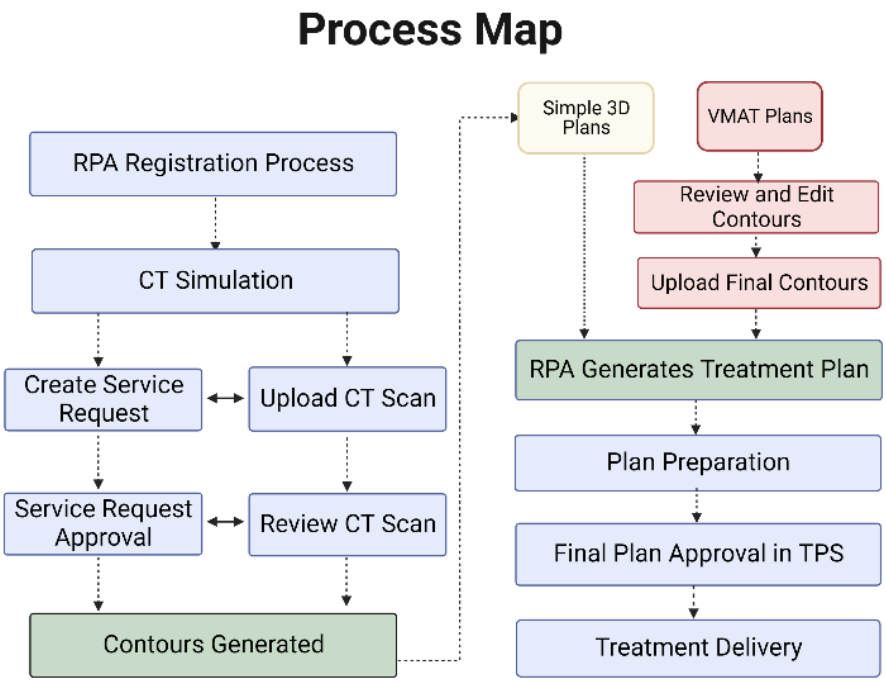


Fig. 124. Process map of the treatment planning workflow in the RPA

The goal is for the RPA to be available for low or no cost to low-resource sites, which could benefit from additional planning assistance. Prior to use, there will

be a required commissioning and training process all new users must complete to ensure that the tool is used correctly and safely.

6.2.2 - Proposed Training Procedure

First, users are asked to provide the RPA team with a list of all CT scanners, linear accelerators, and treatment planning systems available at their sites. This allows us to ensure that their machines are up to the minimum specifications needed to treat with the system. For example, if the available linear accelerators do not have certain energies available, or VMAT capabilities, they will not be given access to plans which utilize those functions.

Next, users who intend to interact with the RPA system must register for an account. As each clinical team member will play a different role in the planning process, different account types are assigned to physicists, therapists, dosimetrists, and radiation oncologists. Each account type has different approval privileges. For example, radiation oncologists have the ability to approve the prescription document, called the service request, whereas other users do not. This is to ensure that plans are not generated which have not received explicit approval from the oncologists to ensure proper planning objectives are being used.

Once accounts are created, each user must undergo a thorough training process to ensure that they have a clear understanding of the tool's capabilities, use cases, and points of risk. They are required to review all training videos which have been created to guide the user's understanding of the RPA. Topics covered in the videos include how to navigate the website, how to generate simple (3D) plans and

complex (VMAT) plans, how to review the output of the system, including the final plan summary, the contours, and the results of all internal quality assurance which was performed. Other videos teach users how to perform replans, delete patient data and register new users in the system. In total, there are nine videos, which take 60 minutes to review.

Users are also provided with the complete user guide for the RPA, which addresses any questions regarding the tool's functionality, and statements of appropriate use for each available treatment site. Statements of use describe the disease extent, patient setup, and treatment technique for which the RPA can be used for planning. If the patient scenario falls outside of these use cases, users are instructed to complete the planning manually without using our automated tools.

Once these materials have been well studied and users feel comfortable, users complete a set of mock treatment plans, consisting of both simple and complex techniques, in their entirety. After completing these test patients, the final plans are submitted to the RPA team to check for any issues that may lead to low-quality patient treatment, which must be corrected. Users then attend a live seminar with members of the RPA team to address any final questions or concerns. In this session, they are also reminded of the potential risk to patients if the tool is not used according to the planning guidelines and if plans are not reviewed diligently following generation.

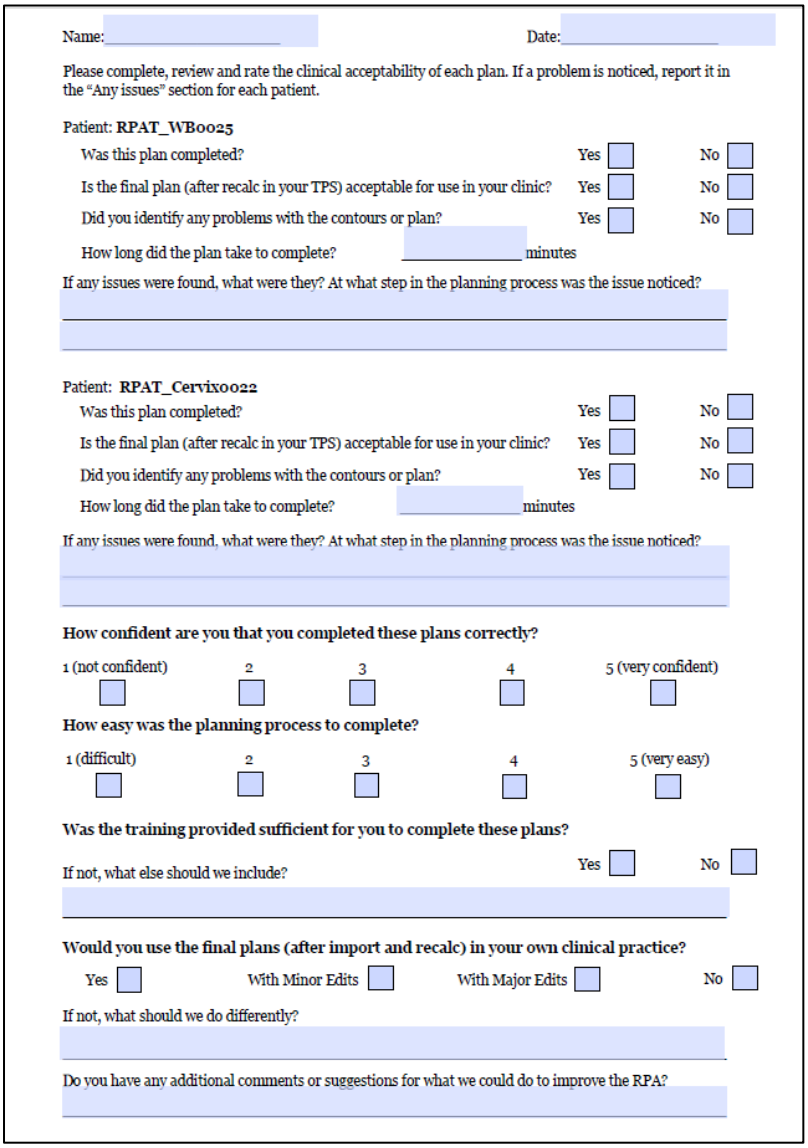
After completing the training, users are given access to the RPA for use in their clinical workflow.

6.2.3 - End-to-end testing

To assess the effectiveness of the proposed training process, a set of 10 test patients was assembled, which encompassed all of the planning capabilities of the Radiation Planning Assistant. The set consists of 2 chest wall, two head and neck VMAT, two whole brain, two cervix 4-field box, and two cervix VMAT patient plans. The patients were fully anonymized prior to proceeding.

Two practicing physicists were recruited, who had no prior experience with the automated contouring and planning tool being evaluated, to perform the end-to-end testing. These physicists were ABR-certified clinical physicists from two US academic hospitals. Participants were provided with the full suite of training materials that were developed for new users of the RPA, discussed above.

Following the review of this data, users were asked to complete two treatment plans: one whole brain patient and one cervix VMAT. They were given a scorecard (Figure 15) to provide feedback and address any issues they may have found. A custom plan review checklist was provided to participants in order to increase the efficacy of final plan checks[73]. Users were then asked about the clarity of training, the quality of plans generated, the ease of use of the tool, and finally, whether the RPA would be a positive addition to their clinical workflow.



Name: _____ Date: _____

Please complete, review and rate the clinical acceptability of each plan. If a problem is noticed, report it in the "Any issues" section for each patient.

Patient: **RPAT_WB0025**

Was this plan completed? Yes ☐ No ☐

Is the final plan (after recalc in your TPS) acceptable for use in your clinic? Yes ☐ No ☐

Did you identify any problems with the contours or plan? Yes ☐ No ☐

How long did the plan take to complete? _____ minutes

If any issues were found, what were they? At what step in the planning process was the issue noticed?

Patient: **RPAT_Cervix0022**

Was this plan completed? Yes ☐ No ☐

Is the final plan (after recalc in your TPS) acceptable for use in your clinic? Yes ☐ No ☐

Did you identify any problems with the contours or plan? Yes ☐ No ☐

How long did the plan take to complete? _____ minutes

If any issues were found, what were they? At what step in the planning process was the issue noticed?

How confident are you that you completed these plans correctly?

1 (not confident) ☐ 2 ☐ 3 ☐ 4 ☐ 5 (very confident) ☐

How easy was the planning process to complete?

1 (difficult) ☐ 2 ☐ 3 ☐ 4 ☐ 5 (very easy) ☐

Was the training provided sufficient for you to complete these plans?

Yes ☐ No ☐

If not, what else should we include?

Would you use the final plans (after import and recalc) in your own clinical practice?

Yes ☐ With Minor Edits ☐ With Major Edits ☐ No ☐

If not, what should we do differently?

Do you have any additional comments or suggestions for what we could do to improve the RPA?

Fig. 15. Scorecard used to collect participant feedback during end-to-end testing

Following the completion of these two plans, users attended a 30-minute seminar with a member of the team to address any concerns they identified and answer any final questions they may have about the planning process. Following this

seminar, users were provided with the remaining eight patients and asked to complete the planning workflow and again provide feedback.

While the process discussed thus far was designed to assess the usability of the RPA, we wanted to also use the opportunity to assess the detectability of errors in plans generated by the system. Prior to deployment, we must not determine only if the tool is effective but if it is safe to use and if issues that arise can be easily identified and mitigated.

To test this, two errors were inserted into our set of plans to evaluate how new users respond to their presence. First, a CT was provided of a head and neck patient whose scan had been inappropriately cropped. Due to the limited field of view, the RPA cannot provide a high-quality treatment plan. Throughout the training process, users were informed that if this were to happen, they should not proceed with planning.

Fig. 16. End-to-end testing workflow for the RPA

For the second error, a CT was provided for a patient with cervical cancer. When this CT scan is uploaded into the RPA, the system is unable to visually identify the marked isocenter for this patient, which is the standard workflow for RPA

planning. Throughout the training, users were instructed that if the isocenter is not identified, they should not proceed with planning with the automated tool. Treatment planning should instead be completed manually in their local treatment planning system. Each of these errors was identified as a potential risk point in an FMEA of the system[50].

After completing all of the treatment plans, users were asked to complete a final survey to provide any remaining feedback about improvements to the training materials and the system's overall usability.

6.3 - Results

6.3.1 - Round 1

In the first round of this study, each participant was asked to complete two treatment plans: one 3D whole-brain treatment and one cervix VMAT plan. Participant 1 successfully completed plans for both patients (100%). They reported that both plans (100%) were acceptable for use in their clinic following plan generation and minor edits. Each plan took 45 minutes to complete.

Participant 2 completed planning for one (50%) of the patients. The cervix VMAT plan could not be completed due to an incompatibility between machine treatment parameters for the RPA and their local TPS that prevented final recalculation. While the whole brain plan was finished, the final plan was not acceptable for clinical use. This plan took 10 minutes to complete.

Participants were asked how confident they were that the plans were completed properly on a scale of 1-5, where 1 = not confident and 5= very confident. Participant 1 scored their confidence at 3/5, and Participant 2 scored their confidence at 4/5 (Figure 17).

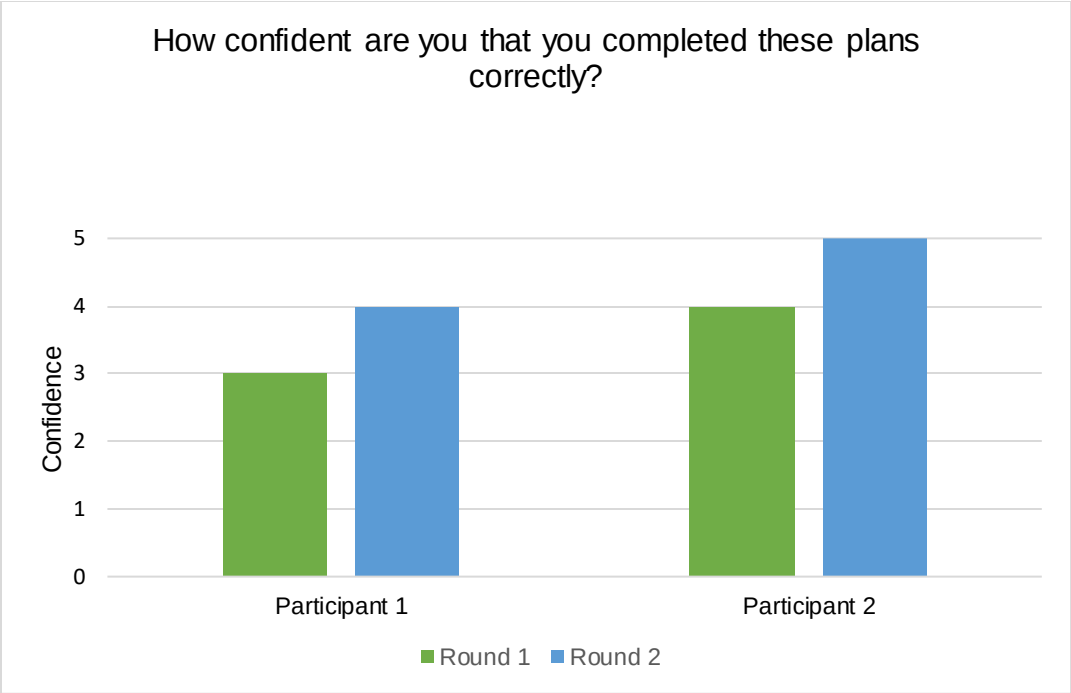


Fig. 17. Confidence scores for Participant 1 and Participant 2, in each round of the study, where 1=not confident and 5=very confident.

Participants were also asked how easy was the planning process to complete on a scale of 1-5, where 1 = difficult and 5 = very easy. Participant 1 scored the ease of use at 4/5, and Participant 2 scored their ease of use at 5/5. (Figure 17)

Both users (100%) reported that the training provided was sufficient and provided all information needed to complete the plans.

6.3.2 - Round 2

In the second round of this study, each participant was asked to complete eight treatment plans: two VMAT head and neck, one cervix VMAT, two cervix 4-field box, two 3D chest wall, and one whole brain treatment. The two errors introduced into this dataset had a 100% detection rate, with both participants identifying the errors before proceeding with planning.

Participant 1 completed plans for six of the eight (75%) patients, excluding only the patients with known errors. They reported that all six plans (100%) were acceptable for use in their clinic following plan generation and minor edits. On average, each plan took 60 minutes to complete.

Participant 2 completed plans for five of the eight (63%) patients, excluding the two patients with known errors and one patient plan that could not be recalculated in RayStation. Of these plans, two (40%) were rated as acceptable for use in their clinic. On average, each plan took 32 minutes to complete.

Participants were again asked how confident they were that the plans were created properly on a scale of 1-5, where 1 = not confident and 5= very confident. Participant 1 confidence score increased to 4/5, and Participant 2's score increased to 5/5. (Figure 18)

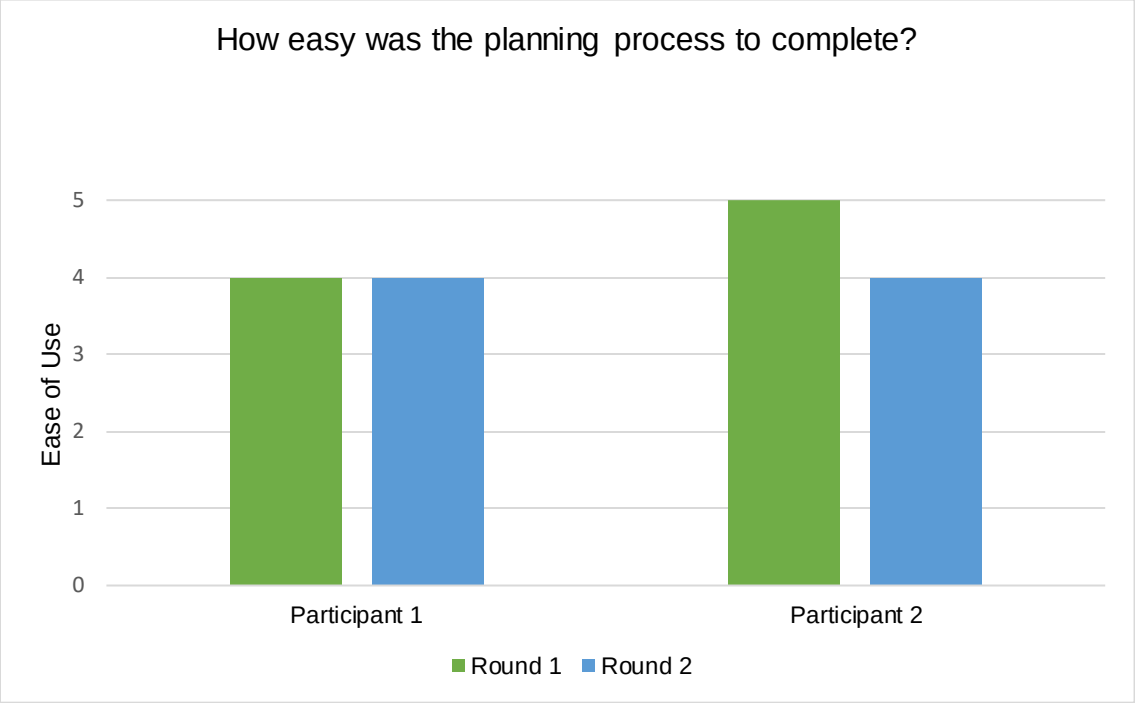


Fig. 18. Ease of use score for Participant 1 and Participant 2 in each round of the study, where 1 = difficult and 5 = very easy.

Participants were also asked how easy the planning process was to complete on a scale of 1 - 5, where 1 = difficult and 5 = very easy. Participant 1 again scored the ease of use at 4/5, and Participant 2's ease of use decreased to 4/5. (Figure 5)

Both users again (100%) reported that the training provided was sufficient and provided all information needed to complete the plans.

6.3.3 - Final Survey

After completing the end-to-end testing, both participants were sent a survey and asked to provide feedback on the training and planning process. The results from this survey are shown in Table 8.

Did the training materials provide you with all the necessary information to make safe, high-quality treatment plans? (1= no, 5= yes)	
Participant 1	4
Participant 2	5
Was the live seminar a useful addition to the training program? (1= no, 5= yes)	
Participant 1	4
Participant 2	2
You were also provided with a short, recorded video demonstration of a user generating an RPA plan. Do you feel that this video was as useful as the live seminar?	
Participant 1	Yes
Participant 2	Yes
What do you feel is an appropriate length of time to ask new users to commit to training for the RPA prior to introducing the tool into their clinical workflow?	
Participant 1	10 hours
Participant 2	1 hour
How much time did you commit to training and completing all 10 RPA Plans?	
Participant 1	8-10 hours
Participant 2	15-20 hours
Was the RPA workflow easy to understand and execute? (1= no, 5= yes)	
Participant 1	4
Participant 2	5
Was the RPA PDF Plan Report used as part of your review for all plans?	
Participant 1	Yes
Participant 2	Some of the Plans
Did you find the RPA Plan Report to be helpful for plan review? (1= no, 5= yes)	
Participant 1	3
Participant 2	2
Was the provided checklist used as part of your review for all plans?	
Participant 1	Yes
Participant 2	Yes
Did you find it to be helpful? (1= no, 5= yes)	
Participant 1	4
Participant 2	4
What TPS did you use for final plan recalculation and preparation?	
Participant 1	Eclipse
Participant 2	RayStation
Did you have any issues recalculating the RPA plan with your local machines?	
Participant 1	Yes
Participant 2	Yes

Table 8. Participant feedback from the final survey, administered upon completion of the end-to-end testing.

6.4 - Discussion

6.4.1 - The RPA Plan Report

In the current iteration of the RPA, a final Plan Report is generated once planning has been completed. This report contains copies of the patient's Service Request, the CT Approval checklist, the planning parameters, a slice-by-slice of the isodose distribution, the dose volume histogram (DVH), dose statistics, and the results of all automated QA tasks. It also contains a slice-by-slice of the patient's contours. While this document provides valuable information, which can provide planners with more guidance on reviewing the provided treatment plans, the document can exceed 50 pages for some planning techniques. Participants of the study reported that the exhaustive length of the plan report limited their ability to find the desired information and stated that they would not be inclined to review the report in its entirety. Both participants recommended removing the slice-by-slice of the contours and dose distribution to simplify the document and allow for a more streamlined review.

Based on this feedback, two versions of the plan report may be offered in future iterations of the RPA. First, a full version contains all of the currently available information, allowing users to examine each detail of the plan in the report before importing it into their own TPS. Alternatively, an abridged version will be available, allowing users to review the plan parameters, DVH, and results of the automated QA. Upon import, a more thorough plan review will occur in the user's local TPS.

6.4.2 - Dose Calculation Error

While generating the first set of treatment plans, Participant 2 identified an issue that prevented them from completing a Cervix VMAT patient plan. The RPA successfully created the contours. They were then edited in the user's local TPS (RayStation) and reimported into the RPA for final plan generation.

Once the plan was imported into RayStation for recalculation, an error occurred. The treatment planning parameters, including MLC positions and MU/gantry rotation, were incompatible with the thresholds set in RayStation. Due to this error, the user could not calculate the dose for the remaining VMAT treatment plans. Testing will be paused until this incompatibility in planning parameters is corrected in the RPA.

6.4.3 - Live Q&A Session vs. Videos Alone

Participants were surveyed to assess whether the live question and answer session, held between the two rounds of planning, was a helpful addition to the training program. Scores were requested between 1-5 when one indicates not beneficial, and five is very useful. Participant 1 scored this question a 4, while Participant 2 scored this question a 2.

An alternate approach was offered to participants, in which the seminar would be removed, and similar treatment planning guidance would be provided in a pre-recorded video. This video showed a live demonstration of the entirety of the plan generation process.

After reviewing the video, both participants felt that it was as effective of a resource as the live seminar. Participant 1 indicated that offering both options would allow us to cater better to users with different learning styles. In future iterations of end-to-end testing, participants will be provided with both options and asked for additional feedback.

6.4.4 – Training Time Commitment

Participants were asked to keep track of the time required for each treatment plan generated and the total time invested in the end-to-end workflow. Participant 1 reported that completing the required training and all treatment plans took between eight and ten hours, whereas Participant 2 reported that the process took between fifteen and twenty hours.

We expect some deviation in the time required for training, depending on each user's previous experience and planning preferences; however, we want to ensure that the time commitment is not a limiting factor in a user's ability to utilize the RPA in their clinic. To address this concern, participants were also asked what length of they would be willing to commit to the training process. Participant 1 reported that ten hours would be reasonable for training with a new treatment planning tool. Participant 1 indicated that training should not exceed one hour.

Due to the significant disagreement regarding the expected time commitment, additional feedback will be requested from the next round of participants to determine the best path forward.

If consensus is still not reached, we will consider offering two sets of participants different types of training: first, an abridged version that requires the review of training material and the generation of only two plans, one simple and one complex. The second group will be asked to follow the extended training discussed in this report.

Both cohorts would again provide feedback on the system's usability and their confidence in use. Suppose no difference is found between the scoring from each group. In that case, hazard testing will be performed to ensure that the condensed training program does not negatively impact the error detection rate.

6.4.5 – Updates to Testing

Participants of this study made several suggestions that will be implemented in future rounds of testing. First, users requested that we provide more explicit guidance regarding when to place the couch structure in their local TPS. Instructions were then updated to state that the appropriate couch structure for each patient can be placed at any time before calculating the treatment plan's final dose.

Both participants also requested that space be allocated on the provided physics plan checklist for recording notes about the plan. The checklist has been updated to include a small area for planner notes.

A bug in the RPA system was identified when participants reported that dose constraints were inconsistent for chest wall plans generated by the RPA. Upon investigation, it was determined that the lung constraints were not implemented as

instructed in the statements of use. The ipsilateral lung dose, specifically V17 Gy, should be used as a constraint for both the left and right chest wall. However, we determined that the left lung dose was reported and used as the constraint in the automatic QA check, regardless of the laterality of treatment. The dose was reported to the ipsilateral lung for left chest wall patients and the contralateral lung for right chest wall. This inconsistency will be corrected prior to additional rounds of testing.

6.4.5 - Future Work

Following needed updates to the RPA system, two additional phases of end-to-end testing will be performed. In the second phase, we will recruit five additional physicists to repeat the testing described in this report. These participants will be clinicians from the international partner institutions where the RPA will first be deployed. By focusing our testing on these participants, we can ensure that the training and planning procedures are optimized for our intended users. Based on feedback from this study, modifications will again be made to the RPA deployment strategy as needed.

Finally, a third testing phase will occur, for which 15 new physicists from various institutions will be recruited. Each participant will review the provided training information and will generate and review one treatment plan each. Feedback regarding the training resources, system usability, and final plan quality will be requested. By expanding our study to a broader variety of users, the scalability of the RPA system can be more clearly understood.

Following the completion of this study, the RPA will be ready to undergo phased clinical deployment to intended partner institutions.

6.5 – Conclusions

We performed a pilot study to assess the effectiveness of the RPA training and deployment workflow. We found that the training resources provided users with a clear understanding of how to generate treatment plans, with Participant 1 scoring the quality of training as 4 out of 5 and Participant 2 scoring it 5. Both participants reported high confidence in their planning capabilities and high scores for ease of use. Two errors were introduced into the planning process, which had a 100% detection rate, indicating that users can appropriately identify and respond to issues that may arise. This result supported our central hypothesis that 90% of errors in the automated treatment planning process can be prevented or detected with proper training and quality assurance resources. Updates were made to the system based on participant feedback, and additional testing will be performed with a new cohort of participants to further optimize the training and deployment procedures.

Chapter 7: Evaluating the clinical use and acceptability of automatically generated contours

7.1 – Introduction

Organ contouring is an essential part of the radiotherapy treatment planning process. Organ contouring has long been accepted as a variable process in which the volume of the final contour can depend on the contouring clinicians' individual stylistic choices. For example, Collier et al.[85] analyzed the clinical contours for the heart, esophagus, and spinal cord generated by 6 different dosimetrists of varying experience levels to determine the uncertainty in organ delineation. They observed substantial variations of up to several centimeters among the users and recommended mitigating this issue using automated methods. In another study, Jenkins et al.[86] investigated how this contour variability can impact the outcomes of prostate cancer radiotherapy. They compared manually drawn target contours with an automatically generated reference contour. As the size of the manual contour increased relative to the reference, they determined that the risk of biochemical recurrence increased by 8-24%/mm.

In recent years, the use of both commercially and in-house developed automated contouring tools has increased rapidly, with reported improvements in contouring consistency and efficiency.[87–89] Although these automated tools have many benefits, they also introduce risks into the clinical workflow that must be explored. A failure mode and effects analysis of an automated contouring and planning system demonstrated that the most frequent cause of high-risk failure was automation bias, in which users rely too heavily on the output of automated tools.[50]

Thus, formalized review of the output of autocontouring software is essential. Turchan et al.[90] surveyed 273 individuals at both community and academic centers regarding their departmental contour review procedures. Only 19% of these individuals reported having a formal process for physicians to review organ contours. Furthermore, 21% of them reported that the formal review process was rarely or never completed.

Although the quality of autocontouring solutions has improved significantly in recent years, these tools can still provide low-quality contours. Therefore, the lack of contour review can greatly impact patient safety and treatment outcomes. One technique that can be used to supplement clinical reviews and detect abnormalities in the use of automated contouring tools is statistical process control (SPC), in which statistical analysis of data acquired using the given workflow to be monitored is performed.[91,92] Groups have used SPC in many applications in medical physics, including quality assurance of couch positioning,[93] evaluating the acceptability of machine performance checks,[94] analyzing adaptive treatment plans,[95,96] deriving machine tolerance for proton quality assurance,[97] developing a predictive quality assurance system,[98] and exploring the dosimetric properties of automated planning tools.[99]

In this study, we applied SPC methods to identify abnormalities in the clinical use of automatically generated contours for 15 organs in the head and neck region in cancer patients. We developed a real-time automated contour monitoring system to notify users if abnormally small or large edits are made to the contours to help improve the clinical review process and the quality of provided automated contours.

7.2 - Methods and Materials

This study was performed to determine whether SPC can be used to monitor the magnitude of edits made to automatically generated contours. SPC uses the statistical properties of a set of data to identify systematic errors in a given workflow. By setting action and warning thresholds, abnormalities in data can be detected, and "out-of-control" processes can be investigated. In monitoring the magnitude of edits made to automatically generated contours, our hope was to improve our ability to detect 2 phenomena. First, abnormally large edits of deep-learning contours may indicate the failure of deep-learning models, which must be addressed. This could occur if the contouring tool was used on a different patient population than that which was used to train the model. One example would be using a contouring tool that was trained for adult patients, to generate OAR contours for a pediatric patient. Large edits could also indicate that automated contouring was performed for a different patient population than the one used for the training data. Off-label use of the contouring tool, in which the tool is used in a fashion unintended by developers, such as using a contouring model developed for the male pelvis to contour organs at risk (OARs) for a female patient, could also be occurring.

Second, if the number of contours that are not edited increases, or the magnitude of edits made decreases over time, automation bias may be occurring. Automation bias could also present as a user who consistently makes fewer edits from the start. Therefore, In addition to SPC, we investigated the initial magnitude of contour editing performed by each clinical user. For example, if some users

consistently make fewer edits than others, additional education about the risk of not thoroughly reviewing automatically generated contours may be needed.

7.2.1 - Monitoring for unusually large contour edits

Data were collected for a cohort of 500 head and neck cancer patients whose organs were contoured using in-house deep learning–based segmentation tools from October 2020 to December 2021[74]. The automated tool provided contours for several relevant OARs, consisting of the brain, brainstem, cochleae, esophagus, eyes, lens, mandible, optic nerves, parotid glands, and spinal cord. The automatically generated contours were saved into a research database prior to integration into the treatment workflow, to ensure that a copy was preserved for comparison. Each generated contour was then subjected to dosimetrist and physician review, and necessary edits were made until the contours were deemed clinically acceptable. The final set of contours was then approved for treatment planning.

The final approved clinical contours for each OAR were exported from our treatment planning system (RayStation) and two overlap metrics, Dice similarity coefficient[100] (DSC) and added path length[101] (APL) were calculated for comparing the automatically–generated contours with the final contours (after editing). The DSC was calculated due to its sensitivity to large volume changes, whereas APL was calculated because it was found to be the metric that most closely correlates with the time spent editing.[101]

These metrics were then used to calculate the appropriate warning and action thresholds for each OAR, following standard SPC methodology. To do this, a

random sample of 50 patients was selected from the initial cohort for use in defining our thresholds. For both DSC and APL, the mean (μ) and SD (σ) were calculated for every OAR in these 50 patients. The warning threshold was then set at $\mu \pm 2\sigma$, and the action threshold was set at $\mu \pm 3\sigma$. These limits are commonly accepted in SPC and have been applied in other areas of radiation therapy.[96,97,99] Control charts were then created for each OAR, in which the magnitude of edits made for each patient is plotted. Separate control charts were generated for DSC and APL.

In this portion of the study, we aim to detect contours that required exceptionally large edits. Therefore, patients will only be flagged who exceeded the large edit action threshold. Specifically, for the DSC plots, $\mu - 3\sigma$ was the large action threshold, as the DSC decreases with increased editing. For the APL plots, $\mu + 3\sigma$ was the action threshold because the APL increases with increased editing.

The overlap metrics for the remaining patients in the dataset were then plotted on the established control charts. Patients whose organs required edits that exceeded the warning or action threshold were then investigated to identify any clinical justification for these statistical abnormalities.

7.2.2 - Monitoring for automation bias

Automation bias occurs when users over-rely on automated tools and therefore do not review the output as thoroughly as they would for manually generated contours. This overreliance can lead to additional risk for patients and thus must be carefully monitored when deploying new automated tools in clinical

practice.[50]To detect automation bias, rather than looking at the magnitude of edits being made for each individual patient, trends in the data indicating that fewer edits are made over time were examined. To do this, moving mean control charts were used. These charts are generated by grouping a fixed number of data points and plotting their mean.[91] By averaging several consecutive data points, the data are smoothed, and rather than flagging for each individual patient with statistically abnormal edits, users will only be alerted if a change in editing practice is observed.

At our institution, dosimetrists are primarily responsible for running autocontouring tools and for the initial review and editing of the output. Therefore, the patients were according to the dosimetrist who was responsible for editing and approving the automatically generated contours. The data corresponding to the 5 dosimetrists who completed the most patient plans was then investigated further. A *t*-test ($P<.05$) was performed to determine whether editing preferences varied among the dosimetrists. The dosimetrists' practice patterns were inconsistent; therefore, automation bias was assessed for each individual who edited contours.

After grouping data according to the dosimetrist performing edits, contours were then sorted by the date of plan approval. For the moving mean control charts, the mean values were calculated for a series of 5 sequential patients. Every time a new patient was added to the system, a new mean was calculated using the 5 most recent data points. The first 10 mean values in each dosimetrist's data set were used to set the warning and action thresholds for the moving mean charts, again setting the warning threshold at $\mu \pm 2\sigma$ and the action threshold at $\mu \pm 3\sigma$. When using moving mean control charts, an important point is that these values (μ and σ)

are calculated not using each individual measurement of contour editing but rather using a subset of the mean values calculated[91]. In this study, we have chosen to use the first 10 mean values calculated to set our action and warning thresholds.

Because we looked for abnormally small edits in this portion of the study, the thresholds were set asymmetrically. For the DSC plot**s**, $\mu + 3\sigma$ was used as the action threshold because the DSC increases with decreased editing, with a DSC of 1 indicating no edits were made. For the APL plots, $\mu - 3\sigma$ was used as the action threshold because APL decreases with decreased editing and approaches 0 when no edits are made.

Using the moving range charts for each dosimetrist, changes in editing practice that could be indicative of automation bias events were flagged if 1) 1 point exceeded the action threshold, 2) 4 consecutive points fell between the warning and action thresholds, or 3) 12 consecutive points fell on the same side of the mean. This process was repeated for each OAR.

7.3 - Results

7.3.1 - SPC results for detection of abnormally large edits

We set the thresholds for each OAR by calculating the mean and SD of the magnitude of edits made for the first 50 patients whose OARs were contoured as described above. We then plotted the remaining data to facilitate the visual identification of patients whose edits exceeded these thresholds. Examples of SPC control plots for the brain and mandible are shown in Figure 19.

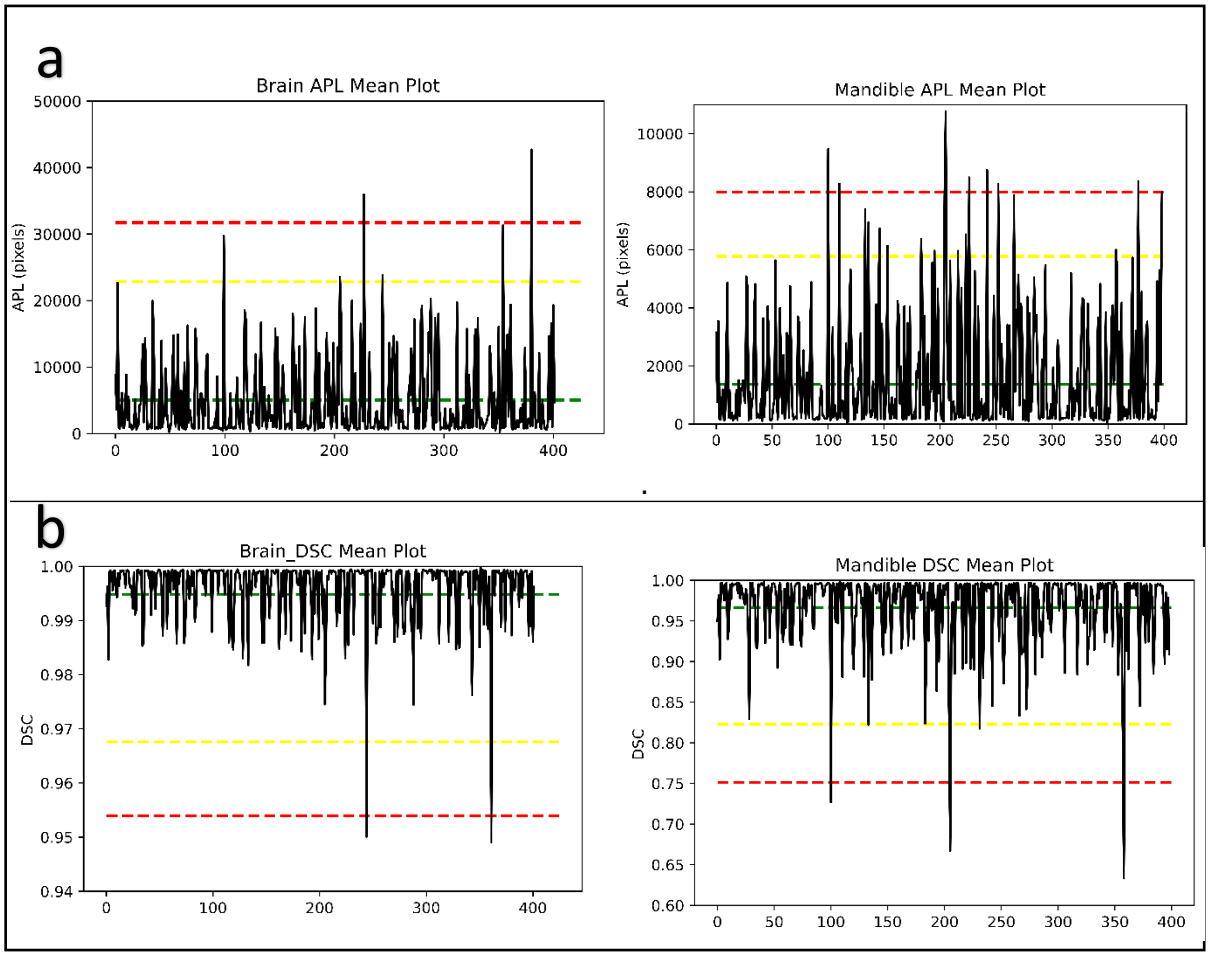


Fig.19. Mean control plots showing the magnitude of edits made to automatically generated brain (a) and mandible (b) contours.

The number of patients whose edits exceeded the warning and action thresholds when using both the DSC and APL, for each OAR is shown in Table 9. The percentage of automatically generated contours that required substantial edits and exceeded the action threshold for each OAR is as follows: 1.0% of brain contours, 3.1% of brainstem contours, 3.5% of left cochlea contours, 2.9% of right cochlea contours, 4.8% of esophagus contours, 4.1% of left eye contours, 4.0% of right eye contours, 2.2% of left lens contours, 4.9% of right lens contours, 2.5% of

mandible contours, 11% of left optic nerve contours, 6.1% of right optic nerve contours, 3.8% of left parotid gland contours, 5.9% of right parotid gland contours and 3.0% of spinal cord contours. For every patient whose edits exceeded the action threshold, we visually inspected both the automatically generated and final clinical contours to determine the cause of major edits.

		Above Action Threshold (*n*)				Above Warning Threshold (*n*)	
OAR	**No. of Patients**	**DSC**	**APL**	**No. of Patients Above Action Threshold**	**Percent of Patients Above Action Threshold**	**DSC**	**APL**
Brain	402	2	2	4	1.0%	0	4
Brainstem	418	7	8	13	3.1%	9	4
Left cochlea	342	1	12	12	3.5%	11	7
Right cochlea	342	0	10	10	2.9%	14	10
Esophagus	357	12	6	17	4.8%	11	11
Left eye	346	13	6	14	4.1%	16	4
Right eye	347	13	6	14	4.0%	19	4
Left lens	319	7	3	7	2.2%	4	2
Right lens	324	15	6	16	4.9%	6	4
Mandible	399	4	9	10	2.5%	4	11
Left optic nerve	326	21	27	35	11%	22	8
Right optic nerve	327	8	15	20	6.1%	12	9
Left parotid	391	3	14	15	3.8%	8	12
Right parotid	390	3	22	23	5.9%	4	10
Spinal cord	431	8	6	13	3.0%	10	6

Table 9. Percentage of patients who were flagged as having abnormally large edits made to the contours of each OAR in our dataset.

7.3.1.1 - Flagged Scenarios

Spinal Cord

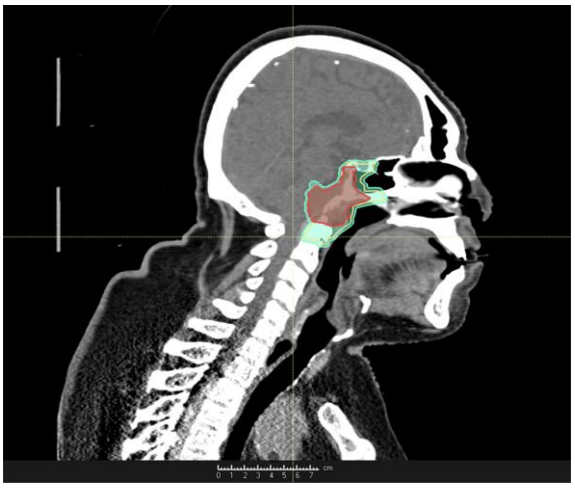


Fig. 20. CT scan showing the target volume for treatment of a skull base tumor.

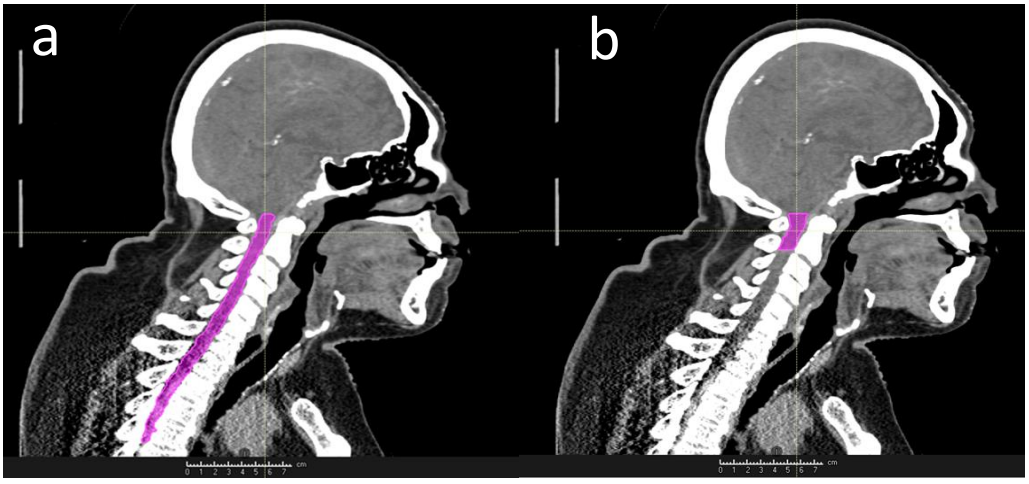


Fig. 21. CT scans showing (a) an automatically generated spinal cord contour and (b) the final, clinically approved spinal cord contour following edits made by a dosimetrist.

We found that 13 patients (3.0%) were flagged as having edits that exceeded the action threshold for the spinal cord contour. Eight of these patients exhibited a consistent deviation from standard clinical practice. One example is shown in Figure 20. For this base of skull treatment plan, the spinal cord edits exceeded our action threshold, with a DSC of 0.224. The figure shows the automatic segmentation of the spinal cord contour which was generated by our deep learning–based model. The automatic segmentation (Figure 21a) appeared to be reasonable with no evident failures. Figure 21b shows the final clinical spinal cord for this patient. The dosimetrist cropped the provided contour to contain only the small region of the total spinal cord volume that was proximal to the target volume and therefore was receiving the most radiation dose.

Upon further investigation, we determined that dosimetrists at our institution traditionally contoured only the portion of the spinal cord that falls within the treatment field, with a several-centimeter margin both superiorly and inferiorly. Although this is no longer the standard practice in our clinic, several dosimetrists in the present study did not update their practice when advanced tools became available. By identifying this deviation from standard practice, we were able to remind the dosimetrists in our clinic that the entire length of the spinal cord should be included in the structure set to maximize the out-of-field dose reported in the event of retreatment.

Cochlea Contour Preferences

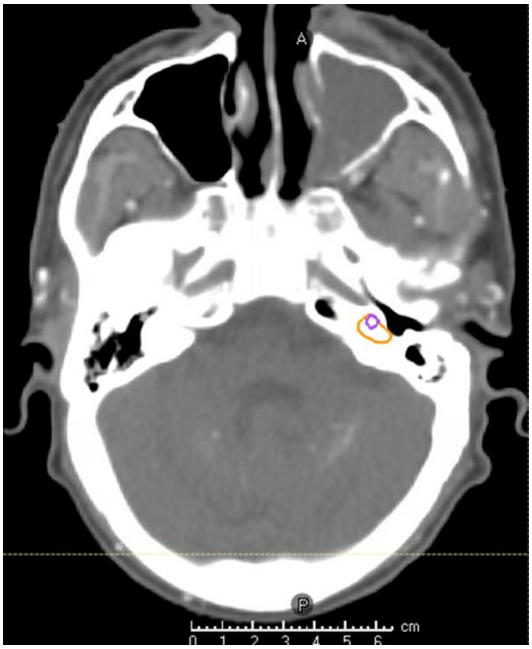


Fig. 22. A CT scan depicting the difference between the automatically generated left cochlea contour (in purple), and the final clinical contour (in orange).

When reviewing large contour edits made to the cochleae, we observed another trend. For the left cochlea, in 12 patients, the warning or action threshold was exceeded. For 10 of these patients, the automated contour was edited to increase both the diameter of the contour on each slice and the superior-inferior extent of the contour (Figure 22. Of these 10 patients, 5 of them were contoured by the same member of the dosimetry team performed contouring for 5 of these 10 patients, indicating a clear preference for larger cochlear volumes by some of the contouring staff than those provided by the automated contouring tool.

Using our real-time automated contour monitoring system, we identified contouring preferences for specific dosimetrists that do not align with the output of

our model. For all 10 patients whose left cochlea contours exceeded the action threshold, the final clinical contours were larger than those obtained using the deep learning model. This may indicate a need to create new autocontouring models for the cochleae, however, it also highlights the variability in cochlear contouring styles between dosimetrists in our clinic.

Inappropriate Use of Autocontouring Tool

Another aim of this contour monitoring system was to identify and subsequently address situations in which our automated contouring tools are used inappropriately. Each automated contouring model is trained to provide high-quality contours in certain situations, which are defined by the training and testing data sets. These parameters can include but are not limited to specific patient orientations, imaging field of view, immobilization devices used, and computed tomography (CT) quality. Although all users of automated systems are trained to only use these tools in appropriate scenarios, off-label use is a potential point of risk that has been identified.[50] When automated contouring models are used on patient images having features outside the acceptable parameters abnormal, unpredictable, or low-quality contours may be generated.

Fig. 23. A patient CT scan that was flagged by the monitoring system, due to large magnitude edits which were needed to 9 of the 15 provided contours. The failures occurred due to the atypical patient orientation during simulation.

While reviewing data for this study, we identified a patient whose edits exceeded the action threshold for 9 of the 15 contours, indicating that large edits were required prior to final clinical approval. Upon review of the patient CT and treatment plan, we determined that the cause of these model failures was atypical patient positioning (Figure 23). The alignment of this patient was not consistent with the patients used to train our automated contouring model, which led to the generation of suboptimal contours. This is an example of a case that falls outside the scope of the automated contouring tool, so the OARs should have been segmented manually to ensure accuracy. Because the automated contouring monitoring system flagged this patient for review, the clinical team can be reminded of the appropriate

situations in which to use automated contouring and the risk of not following these directions.

7.3.2 - Monitoring for Automation Bias

To determine if automation bias occurs in our clinic, we examined the magnitude of edits made by the 5 dosimetrists who contributed most frequently to this data set. To do this, we plotted the DSC and APL for patients edited by each of these 5 dosimetrists and compared them to the magnitude of edits made by all dosimetrists who contributed to this study (Figure 24). We repeated this process for each OAR. More information about the distribution of data for each dosimetrist is available in Appendix D.

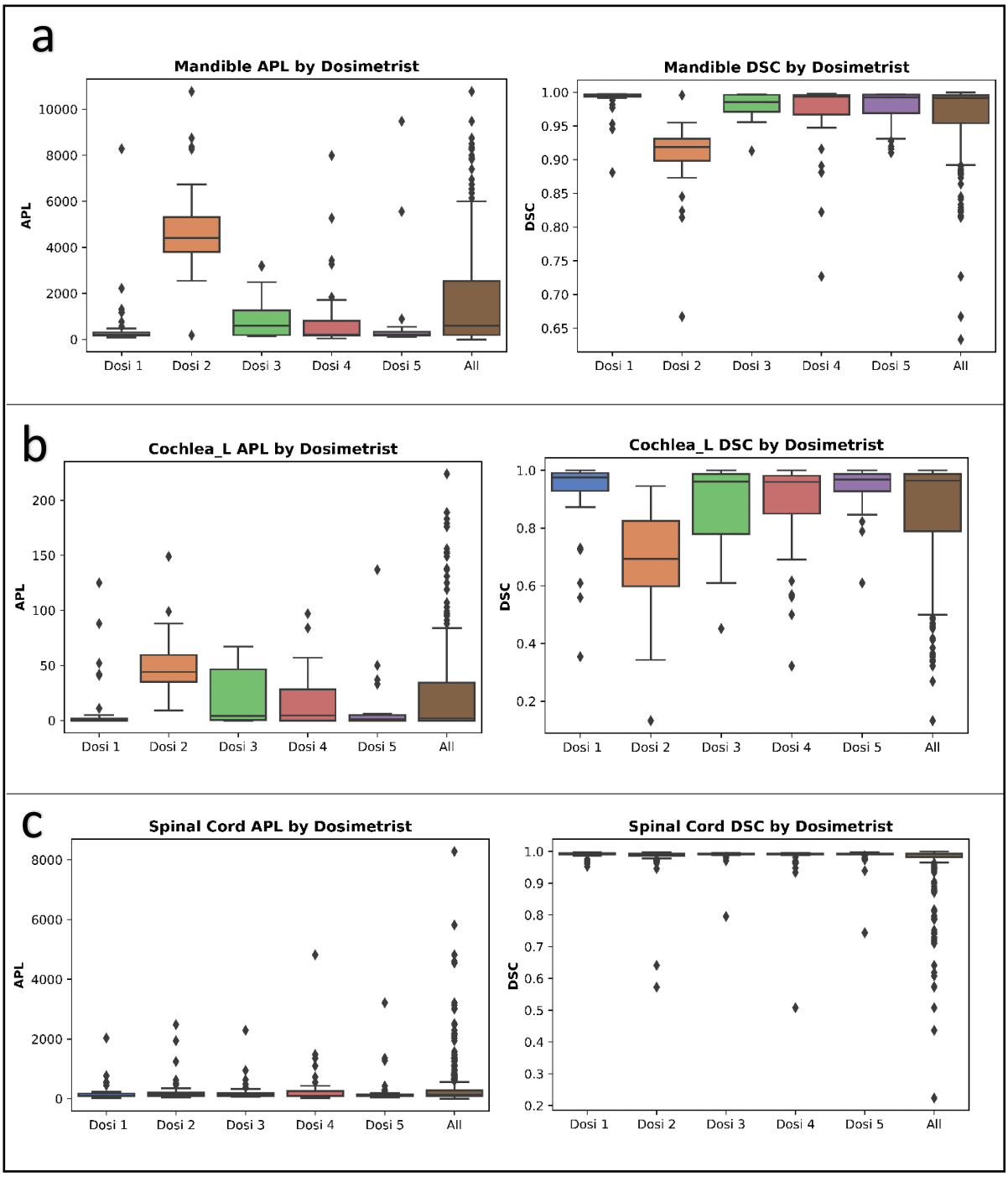


Fig. 24. Distribution of edits, by dosimetrist in this study for the (a) mandible, (b) left cochlea and (c) the spinal cord.

Trends in dosimetry practice

When examining the magnitude of edits made by the dosimetrists, we identified trends in user preference. In particular, we found that dosimetrist 1 consistently edited contours less than the mean DSC and APL for 14 of the 15 OARs evaluated in this study. Dosimetrist 2, however, made considerably more edits and exceeded the data set mean for 14 of the 15 OARs, indicating that Dosimetrist 2 found the automatically generated contours consistently required additional edits to achieve clinical acceptability.

We performed a *t*-test to assess the significance of the difference in edits made to each OAR based on the dosimetrist performing the task. We did so for both DSC and APL metrics and compared the data sets for dosimetrists 2 and 1. For most of the OARs, the results demonstrated a significant difference between the 2 dosimetrists' edits ($P < .05$). However, the *t*-test results demonstrated that for the optic nerves, parotid glands, and spinal cord, the difference was not significant ($P > .05$). Based on these results, the conservative approach was taken and for each OAR the presence of automation bias was assessed on an individual basis.

Moving Mean Control Charts

For the moving mean control chart, the mean values were calculated for a series of 5 sequential patients. Every time a new patient was added to the system, a new mean was calculated using the 5 most recent data points. The first 10 mean values in each dosimetrist's data set were used to set the warning and action thresholds for the moving mean charts. The remaining data points were then plotted

on the moving mean charts. Examples of these plots, which display the edits made to mandible contours, are shown in Figure 25.

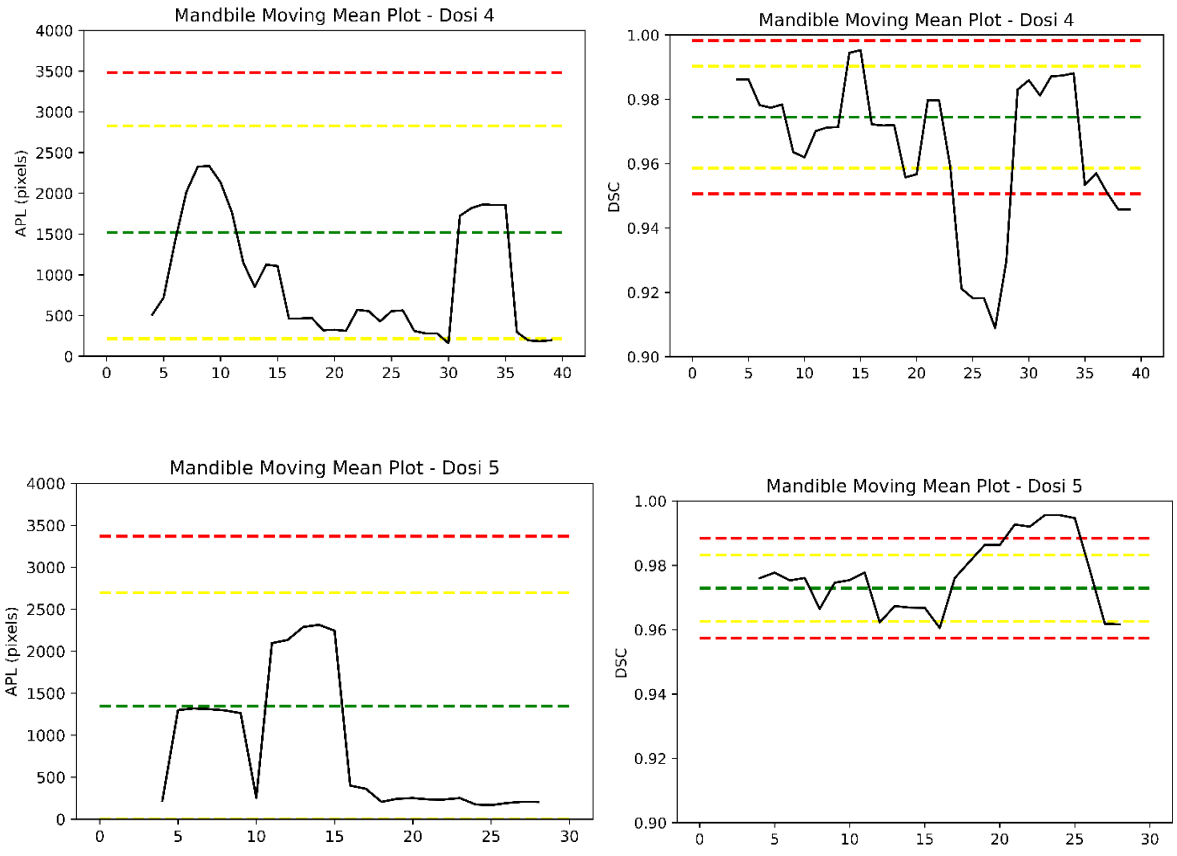


Fig. 25. Moving mean charts for edits made to the mandible, for each dosimetrist

10 moving mean charts were generated for each OAR in each dataset, including each of the 5 dosimetrists edits quantified with DSC and with APL. In total, 150 moving mean charts were created in this study to assess automation bias. Of these control plots, 27 were flagged due to the possibility of automation bias. Of these 23 charts, 8 were for OARs edited by dosimetrist 1, 9 were for OARs edited by dosimetrist 2, 4 were for OARs edited by dosimetrist 3, 2 were for OARs edited by dosimetrist 4, and 4 were for OARs edited by dosimetrist 5. Figure 25 shows examples of charts that were flagged for automation bias, in the APL moving mean charts for dosimetrists 3, 4, and 5. Table 10 summarizes the instances in which dosimetrists exceeded the moving mean thresholds for OARs.

	Dosimetrist				
	1	**2**	**3**	**4**	**5**
Mandible DSC	0	1	1	0	1
Mandible APL	0	1	1	1	1
Brain DSC	1	0	1	0	0
Brain APL	1	0	1	0	0
Brainstem DSC	1	0	0	0	0
Brainstem APL	1	0	0	0	0
Left cochlea DSC	0	0	0	0	0
Left cochlea APL	0	0	0	0	0
Right cochlea DSC	0	0	0	0	0
Right cochlea APL	0	0	0	0	0
Esophagus DSC	1	0	Trending	0	0
Esophagus APL	0	0	Trending	1	0
Left eye DSC	0	1	0	0	0
Left eye APL	0	1	0	0	0
Right eye DSC	0	0	0	0	0
Right eye APL	0	0	0	0	0
Left lens DSC	0	1	0	0	0
Left lens APL	0	1	0	0	0
Right lens DSC	1	0	0	0	0
Right lens APL	1	0	0	0	0
Left optic nerve DSC	0	0	0	0	0
Left optic nerve APL	0	0	0	0	0
Right optic nerve DSC	0	0	0	0	0
Right optic nerve APL	0	0	0	0	0
Left parotid gland DSC	0	0	0	0	Trending
Left parotid gland APL	1	1	0	0	Trending
Right parotid gland DSC	0	1	0	0	Trending
Right parotid gland APL	0	1	0	0	Trending
Spinal cord DSC	0	0	0	0	1
Spinal cord APL	0	0	0	0	1

Table 10. Number of flags for automation bias in all moving mean control plots, with 1 indicating that the dosimetrist was flagged for exceeding action thresholds corresponding to less edits over time, and 0 indicating that action thresholds were

not exceeded. “Trending” is used to indicate that the dosimetrist’s most recent patients were consistently receiving fewer edits.

Upon review, 6 additional control charts were noted as showing a trend in the data which indicated that fewer edits were being made over time, however, the action threshold had not yet been reached. 2 examples of this phenomenon can be shown in Figure 26. For the left and right parotid glands, dosimetrist 5’s APL moving mean control charts indicated fewer edits were being made over time, as shown in the plateau beginning at patient 19. However, the number of available data points was insufficient to flag the automated contour monitoring system and trigger an investigation. To mitigate the risk associated with potential trending automation bias, the dosimetrists exhibiting these trends will be contacted prematurely to discuss the motivation behind decreased contour editing.

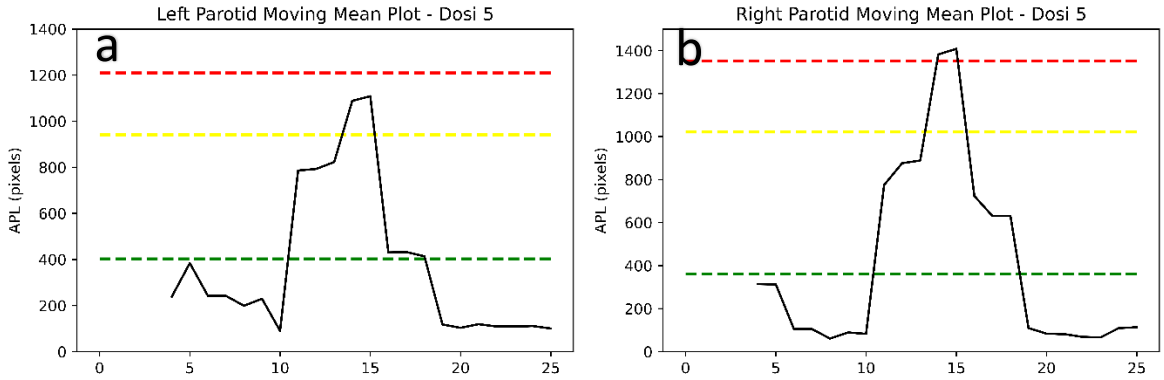


Fig. 26. Moving mean charts of the left (a) and right (b) parotid for edits made by dosimetrist 5.

7.4 - Discussion

In this study, we developed a monitoring system to detect when statistically abnormal edits are made to automatically generated contours. By investigating the causes of these large edits, we identified scenarios, such as contouring for the cochleae, in which the automatically generated contours were consistently modified to be more closely aligned with a dosimetrists preference. We also identified scenarios where large changes were made to the automatically generated contour which were inconsistent with standard clinical practice, including substantial cropping of the spinal cord contour. The spinal cord is a serial organ, and therefore our primary concern is the maximum dose received by the structure, which would still have been captured despite the cropping. Therefore, the issues identified in spinal cord editing were not dangerous to the patients under treatment, however other situations may arise in which patient safety could be compromised if a monitoring system to detect abnormal contour edits were not implemented.

Furthermore, we assessed trends in the use of autocontouring by dosimetrists in our clinic and found marked differences in the magnitude of the edits made between dosimetrists. By examining the editing data for each dosimetrist individually, we also identified situations indicative of automation bias, in which fewer edits were made over time.

7.4.1 - Deployment of the Automatic Contour Monitoring System

While we believe this method of monitoring contouring edits using SPC is scalable for use with other automated segmentation tools, the specific action and warning thresholds used may not be. Contouring styles and preferences vary greatly among both institutions and individual users. The tools in this paper can be used to identify differences in local clinical practice, which can inform where the appropriate action levels should be set depending on the cause of these differences.

When a new or updated automated contouring, model is introduced into practice, the baselines used for monitoring, including the means and warning and action thresholds, should be reset using data from the new patient cohort and the most up-to-date version of the model.

Over time, fewer edits may be made to contours as clinicians become more comfortable with the contours generated by the automatic contouring tools. The warning and action thresholds set in this process should be used as a starting point**,** but we recognize that this process is iterative, and thresholds may need to be adapted over time.

Furthermore, the goal of this automated contour monitoring system is not to regulate how autogenerated contours should be used but rather to gain information. Large or small edits are not inherently good or bad; however, examining the scenarios in which they occur could lead to a better understanding of clinical practice. Implementing a real-time automated contour monitoring system will enable the identification of weaknesses in our model and areas in our user training and

clinical deployment strategy that could be improved. While this study was performed using data obtained from the use of our in-house developed automatic contouring tool, we recommend monitoring systems should be built into other automatic contouring software programs, as well. By using a real-time automated contour monitoring system, we can ensure that automated segmentation models provide users with the most accurate, useful, and safe contours.

7.5 - Conclusions

By performing a quantitative assessment of the magnitude of edits required to achieve clinical acceptability of deep learning–generated contours, we have shown that it is possible to detect variations in the use of automatically generated contouring tools that may impact patient treatment.

Chapter 8 – Discussion and Conclusions

8.1 – Specific Aim One

In chapter three, we hypothesized that risk assessment techniques could be used to identify and decrease points of risk with automated treatment planning tools. A failure mode and effects analysis (FMEA) was performed to assess the risk that may be introduced into the clinical workflow when utilizing the RPA. A multidisciplinary team consisting of physicians, physicists, radiation therapists, dosimetrists, and members of the RPA development team generated a list of all possible failure modes in the current RPA workflow. Any error that was not specific to the auto-planning workflow and would also occur during manual planning was removed from this list. This allowed the team to focus on new failures or errors whose risk was made worse when utilizing the RPA.

In total, 126 failure modes were identified, which were unique to RPA workflow and may not be accounted for in established quality management programs. The mean RPN of these failure modes was 56.3; 21 errors had an RPN above 125, which is the TG-100-recommended threshold at which action should be taken to reduce the risk. In order to mitigate this risk, changes were made to the RPA workflow. These changes included updates to the user interface, implementing redundancy checks to limit the risk of possible human error, and adding user guidance documents to clearly communicate appropriate use scenarios and the importance of thorough quality assurance.

Following these system updates, the number of failure modes that exceeded the action threshold of RPN 125 decreased from 21 to 5, showing a 76% reduction in high-risk errors. We also increased the detectability of 15 of the 21 errors (71%), ensuring that users can more easily detect any errors that do occur. By prospectively identifying risk points in the RPA treatment planning workflow, we found that risk could be effectively reduced to increase the safety of patient treatment. The 76% reduction in high-risk failure modes, when added to other work in this thesis, indicates that we would reach our central hypothesis that 90% of errors can be prevented or detected prior to treatment.

In chapter four, we hypothesized that hazard scenarios could be used to identify and correct points of weakness in the RPA planning workflow. Errors were introduced into multiple stages of the RPA workflow, during which data needed to be reviewed and approved before submission for planning. Errors were selected that had been identified during the FMEA, which would negatively impact patient outcomes if they went undetected.

Radiation therapists reviewed provided CT scans and found 87% of all errors present. Radiation oncology residents reviewed the service request and detected 75% of errors. Feedback was requested about how the request could be improved to increase the detectability of errors, and the service request was updated accordingly. Following system updates, 100% of errors were detected by five new radiation oncologists, indicating that the change was effective.

Physicists were asked to review the final clinical contours, and 0% of errors were detected. Based on this result, we determined that physicists were not the

appropriate clinical team member to assess target contours for accuracy. Therefore, guidance will be provided to users reinforcing that contours should be thoroughly reviewed by physicians prior to patient treatment. In summary, 93.5% of errors were detected by radiation therapists and oncologists, supporting our central hypothesis. The low rate of error detection by physicists does not support our central hypothesis; therefore, the rate of error detection by physicists was further investigated in chapter five.

8.2 - Specific Aim Two

In chapter five, we hypothesized that we could increase the error detection rate by creating a custom checklist, which focuses the review on errors known to occur in the automated planning workflows. A customized plan review checklist was developed using guidance from AAPM task groups 275 and 315 [25,60] and modified to ensure all errors identified during our FMEA were represented.

To assess the effectiveness of this checklist, physicists were asked to review ten treatment plans created by the RPA, which contained five plans with known errors. These plan checks were first performed without the use of the checklist. Participants were then asked to review an additional ten treatment plans, utilizing the custom checklist to guide their review. The checklist was then modified based on physics feedback and updated prior to repeating the study with medical physics residents.

When physicists completed their plan review utilizing the checklist, the error detection rate increased by 20%, to 88% of total errors being detected. When

physics residents completed their plan review utilizing the checklist, the error detection rate increased by 17%, to 70% of total errors being detected. For both cohorts, the number of errors increased when the checklist was utilized, indicating that the safety of automated planning tools can be improved when checklists generated based on the results of an FMEA are used to guide plan review. The 20% increase in error detection for physicists and 17% increase in error detection by residents when the checklist was utilized supports our central hypothesis that 90% of errors can be prevented or detected prior to treatment by utilizing effective quality assurance resources.

In chapter six, we hypothesized that performing an end-to-end test of the RPA training and deployment procedure would identify any weaknesses that need correcting prior to full-scale deployment of the RPA. The new-user training process was simulated for two physicists who had no experience with the system. Each participant was asked to review all provided training materials and then use the RPA to generate ten treatment plans of varying sites and techniques. They were then instructed to thoroughly review the final plan and record any errors that would limit the plan's clinical acceptability. Two errors were included in these plans to assess the detectability of errors. Feedback was requested regarding improvements that could be made to the training process.

Both participants reported that the provided training documents were helpful and provided all information needed to generate safe, high-quality treatment plans. The errors included had a 100% detection rate, indicating that provided training and

quality assurance documents provided users with the guidance needed to respond appropriately to unsafe scenarios. Several weaknesses were identified, including planning incompatibility between the RPA and RayStation, and the limited functionality of the provided plan report. The 100% error detection rate supports our central hypothesis that 90% of errors can be prevented or detected prior to treatment.

Updates will be made based on user feedback, and additional rounds of testing will be performed with clinicians from international partner institutions who will be among the first to use the RPA in clinical practice.

8.3 – Specific Aim Three

In chapter seven, we hypothesized that performing monitoring of patient contour edits can lead to increased detection of systematic errors, such as those caused by software error, automation bias, or off-label use.

A monitoring system was developed that uses statistical process control (SPC) techniques to identify patients whose automatically generated contours required significant edits to achieve clinical acceptability. DSC and APL were calculated between the automatically generated contour and the final approved clinical contour for 15 OARs in the head and neck region. When the magnitude of edits exceeded the thresholds set using SPC, the clinical scenario was further investigated to determine the cause. Causes identified were contour style deviating from standard clinical practice, dosimetry contouring style not aligning with auto-

contouring guidelines, and inappropriate use case, in which automated contouring was performed on a patient who was simulated in an abnormal position.

Trends in the use of auto contouring by dosimetrists from our clinic were also assessed. We found a significant difference in the magnitude of edits made between dosimetrists. By examining the editing data for each dosimetrist individually, situations indicative of automation bias, in which fewer edits are made over time, were also identified. The detection of instances of off-label use and automation bias will be utilized to provide feedback to dosimetrists and reduce systematic errors in future clinical scenarios. This supports our central hypothesis that 90% of errors can be detected or prevented prior to patient treatment.

8.4 – General Discussion

The development of automated treatment planning tools requires consideration not only of the final output (i.e., contour or plan quality) but also of the interface's functionality from the perspective of the intended user. Autoplanning tools introduce new steps into the treatment planning workflow. This change in procedure can lead to additional points of risk, such as those caused by human error, automation bias, off-label use of the tool, and software error. Therefore, thorough human factors engineering (HFE) studies must be performed prior to introducing new systems into clinical practice. HFE studies do not examine how well tools can perform in optimal situations. Instead, they examine what occurs when humans, with all of their capabilities, limitations, and tendencies considered, interact with these systems[102]. In this study, we found that by using techniques consistent with HFE,

such as hazard testing, physics plan checks, and end-to-end testing, we could modify the user interface, training materials, and workflow to limit errors that could be passed on to patients.

To support the central hypothesis of this study, that 90% of clinically relevant errors could be prevented or detected prior to impacting patient care, an FMEA was first used to identify significant risk points in the RPA workflow from the perspective of the clinical user. Changes were then made to limit the risk of 76% of high-risk failures. These risk points were then incorporated into hazard testing, and we found that 62% of errors could be detected before a plan was created in the RPA. During final plan checks, when utilizing the customized checklist, we found a rate of error detection of 88% for physicists and 70% for medical physics residents. Following the optimization of provided training materials, 100% of the errors present were detected during an end-to-end test of the entirety of the RPA treatment planning workflow. A monitoring system was also developed to limit the risk of systematic errors and detect abnormalities in the contouring process that could be attributed to software error, off-label use, or automation bias.

Each of the studies in this report focused on optimizing various layers of defense in the quality management program for automated treatment planning tools. By evaluating the detectability of errors at several stages in the workflow, we are maximizing safety, not by focusing only on total detection of errors at the time of patient treatment, but by following the ‘swiss cheese’ approach to quality management. In the swiss cheese model, multiple layers of quality assurance are incorporated into the safety program which work together to minimize the risk which

could reach patients[103]. In this study we evaluated and mitigated risk at multiple stages: First, prospectively, when changes were made to the system based on a risk assessment. Next, hazard testing was performed, where error detection was evaluated at different steps in the plan generation process. Final physics plan checks were also assessed, during which each completed plan is evaluated for quality and safety prior to patient treatment. Finally, we developed a system to detect systematic errors through the monitoring of contour edits.

While not all studies in this report independently exceeded the stated goal of mitigating 90% of errors, when the developed resources are combined into a cohesive quality management program, the central hypothesis that 90% of clinically relevant errors can be detected or prevented prior to impacting patient safety was supported.

8.5 – Study Limitations

Although we have accomplished the goals we set out to achieve, the development of a robust clinical implementation strategy is an iterative process. One limitation of our study is that each experiment was conducted as a prospective simulation, and the results have yet to be validated in clinical practice. While this choice was made to ensure that risk has been mitigated prior to the global deployment of the RPA, it must be acknowledged that updates might be needed once clinical integration has occurred.

Additionally, the participants who assisted in this project were predominantly clinicians from US academic institutions. These users were selected due to their

availability and ability to commit multiple hours to research. We recognize that using participants from our target demographic, such as clinicians from international institutions who have expressed interest in utilizing the RPA in their practice, would have been preferred. However, due to the high clinical workload, and low staffing levels in these environments, participants from our partner institutions were often unavailable to participate. Therefore, adjustments may need to be made to the QA and training resources during the phased deployment of the RPA to reflect the clinical practice of our users.

8.6 – Future Direction

Additional end-to-end testing of the RPA training and deployment process will be performed to ensure that the provided resources allow users to create safe, high-quality treatment plans. In the next testing phase, additional physicists from the international partner institutions where the RPA will first be deployed will repeat the established end-to-end workflow. By performing testing with our international partners, we can ensure that the training and planning procedures are optimized for our intended users. Modifications will be made to the deployment strategy to increase user confidence, ease of use, and understanding of materials as needed. By participating in this testing, users will also become familiar with the RPA system before deployment into their clinic in late 2023.

A final round of testing will be performed in which 15 clinicians from various institutions will review all training materials, create one treatment plan each, and

provide feedback on the process. By expanding the study to a broader range of users, the scalability of the RPA system can be more clearly assessed.

Following the completion of these studies, the RPA will be ready to undergo phased clinical deployment.

8.7 – Conclusions

In this study, we presented the work done to optimize the safety and usability of an in-house automated treatment planning tool by focusing on three key categories: risk, quality assurance, and deployment. This work was encompassed in a final evaluation of the end-to-end RPA workflow, during which 100% of clinically relevant errors were detected, supporting our central hypothesis that 90% of clinically relevant errors introduced by automated treatment planning tools could be prevented or detected by establishing a robust risk evaluation process and developing a thorough quality assurance and deployment procedure. Furthermore, a monitoring system was developed to detect systematic errors that may occur when automatically generated contours are used clinically. This evaluation has not only maximized the impact of our own automated treatment planning tool but is also the first study into the potential risks associated with AI-based treatment planning tools, which also examines how best software and workflows can be modified to increase usability and safety. The recommendations presented in this report can be used to benefit AI software development further and inform the development of robust deployment procedures for other AI-based tools in future clinical environments.

Appendix A – Full Results of Failure Mode and Effects Analysis (Before system improvements)

Potential Failure Mode	Step	Cause	S	O	D	RPN
Plan not reviewed carefully prior to approval (plan that doesn't meet clinical standards approved)	Final Plan Approval	overreliance on system	9	6	9	486
Incorrect assignment of DVH tolerances - H&N	Create Service Request	off-label use of RPA	8	6	10	480
Request is approved by member of the team who is not physician (used others credentials)	Approve Service Request	off-label use of RPA	9	6	8	432
Incorrect treatment unit selected - Cervix	Create Service Request	off-label use of RPA	7	6	10	420
Request is approved for wrong linac	Approve Service Request	off-label use of RPA	7	6	10	420
RPA printout being used as plan documentation	Plan Preparation	off-label use of RPA	7	6	10	420
RPA printout being used as plan documentation	Plan Preparation	overreliance on system	7	6	10	420
Necessary physics QA was not performed	Plan Preparation	overreliance on system	10	4	9	360
Incorrect Physics QA procedures selected	Registration	off-label use of RPA	9	4	10	360
Selected a disease extent not suited for the patient - CW	Create Service Request	off-label use of RPA	7	5	10	350
Selected a disease extent not suited for the patient - Cervix	Create Service Request	off-label use of RPA	7	5	10	350
Assign User category incorrectly	Registration	off-label use of RPA	9	4	9	324
Physician changed prescription after plan generated	Final Plan Approval	off-label use of RPA	9	4	9	324

Request is approved without careful physician review	Approve Service Request	overreliance on system	9	6	6	324
Incorrect identification of positive nodes - H&N	Create Service Request	off-label use of RPA	6	5	10	300
Incorrect selection of CTV regions - H&N	Create Service Request	off-label use of RPA	6	5	10	300
Incorrect IMRT QA equipment selected	Registration	off-label use of RPA	7	4	10	280
Contoured target incorrectly	Contouring	software error	5	9	6	270
Incorrect identification of CTV 2 and CTV 3 - H&N	Create Service Request	off-label use of RPA	5	5	10	250
Incorrect identification of CTV 2 and CTV 3 - H&N	Create Service Request	overreliance on system	5	5	10	250
Incorrect prescription input for each CTV - H&N	Create Service Request	off-label use of RPA	9	3	9	243
Incorrect identification of prior radiation	Create Service Request	human error	10	4	6	240
Incorrect treatment unit selected - CW	Create Service Request	off-label use of RPA	4	6	10	240
Incorrect treatment unit selected - H&N	Create Service Request	off-label use of RPA	4	6	10	240
Request is approved for wrong site	Approve Service Request	off-label use of RPA	7	3	10	210
CT approved despite presence of artifact in treatment region	Review CT	off-label use of RPA	7	3	10	210
Incorrect answer on pregnancy questionnaire	Create Service Request	human error	10	4	5	200
Plan sent directly to console without importing to TPS	Treatment Delivery	off-label use of RPA	10	2	10	200
Plan not recalculated in TPS	Plan Preparation	off-label use of RPA	5	4	10	200
Incorrect prescription input - CW	Create Service Request	off-label use of RPA	4	7	7	196

Incorrect prescription input - Cervix	Create Service Request	off-label use of RPA	4	7	7	196
Request is approved with wrong prescription	Approve Service Request	off-label use of RPA	4	7	7	196
Incorrect plan downloaded from RPA (if multiple)	Plan Preparation	human error	9	3	7	189
Incorrect plan imported into the TPS	Plan Preparation	human error	9	3	7	189
Reference point added to incorrect location	Review CT	off-label use of RPA	9	3	7	189
Unexpected BB placement - wrong place	CT Simulation	human error	9	5	4	180
Marked isocenter not verified in TPS	Plan Preparation	human error	9	5	4	180
Incorrect shift not validated in TPS	Plan Preparation	human error	9	5	4	180
Not loading 3D because takes too long	Review CT	overreliance on system	6	7	4	168
Incorrect selection of laterality - CW	Create Service Request	human error	10	5	3	150
Reference point added to incorrect location	Review CT	human error	6	5	5	150
Request is edited by another user prior to approval	Approve Service Request	human error	9	4	4	144
CT approved despite poor image quality	Review CT	off-label use of RPA	6	6	4	144
Needed plan edits not made in TPS	Plan Preparation	human error	4	6	6	144
RPA printout being used as plan documentation	Plan Preparation	human error	7	2	10	140
Incorrect MRN (wrong patient)	Create Service Request	human error	9	5	3	135
Incorrect selection of CTV regions - H&N	Create Service Request	overreliance on system	6	3	7	126

Exported incorrect contours	Contouring	human error	7	5	3	105
Incorrect identification of department CT	Registration	human error	4	3	8	96
Forgets to review normal tissue contours	Contouring	overreliance on system	4	6	4	96
Imported contours conflict with existing contours	Plan Preparation	different workflow	9	5	2	90
Wires -> could be interpreted as BBs	CT Simulation	software limitation	9	5	2	90
Isocenter position not selected correctly	Review CT	software error	9	3	3	81
Exported contours from TPS, then made edits which were not imported to RPA	Contour Upload	human error	9	4	2	72
Exported incomplete contours from TPS	Contour Upload	human error	9	4	2	72
Imported contours conflict with existing contours	Plan Preparation	human error	9	4	2	72
Incorrect plan downloaded from RPA (if multiple)	Plan Preparation	software error	9	2	4	72
Contoured OAR that cannot be considered in RPA planning	Contouring	different workflow	8	3	3	72
Incorrect treatment machine parameters	Registration	off-label use of RPA	4	4	4	64
Incorrectly omitted presence of implants	Create Service Request	off-label use of RPA	6	5	2	60
Creating target as a new structure, with incorrect naming	Contouring	different workflow	5	2	6	60
Reference point not added when needed for planning (cervix)	Review CT	human error	5	4	3	60
Incorrect identification of calibration protocol	Registration	human error	9	3	2	54
Wrong calibration settings (SSD vs SAD, MU/cGy)	Registration	human error	9	3	2	54

Warning "not for clinical use" only in English (could lead to issues translating or understanding)	Plan Preparation	human error	7	3	2	42
Warning "not for clinical use" only in English (could lead to issues translating or understanding)	Plan Preparation	software limitation	7	3	2	42
Contours approved by a member of clinic unfamiliar with anatomy & targets	Contour Upload	off-label use of RPA	10	4	1	40
Deidentified plan matched to incorrect patient	Plan Preparation	software error	10	2	2	40
Added additional target volume that is not supported by the RPA	Contouring	different workflow	9	2	2	36
Inaccurate contour of OAR	Contouring	software error	9	2	2	36
Imported contours conflict with existing contours	Plan Preparation	software error	9	2	2	36
Contour report not reviewed prior to download	Contour Review & Download	overreliance on system	8	4	1	32
Did not review contours prior to approving	Contour Upload	overreliance on system	8	4	1	32
Deviation between RPA and recalculated plan are too large	Plan Preparation	equipment limitation	4	4	2	32
Incorrect selection of primary site - H&N	Create Service Request	off-label use of RPA	5	3	2	30
PDF from RPA does not match institutional standards (shifts in room space vs. shifts in patient space)	Plan Preparation	different workflow	2	5	3	30
Changed contour name to one not compatible with RPA	Contouring	different workflow	2	7	2	28
Incorrect patient orientation label (patient in correct position, label is incorrect)	CT Simulation	human error	2	7	2	28

Inaccurate image guidance techniques selected	Registration	human error	1	3	9	27
Deviation between RPA and recalculated plan are too great	Plan Preparation	software error	4	3	2	24
Inconsistency in plan and beam labels between RPA and clinical workflow	Plan Preparation	different workflow	2	6	2	24
RPA unable to properly load 3D CT (too slow, or crashes)	Review CT	equipment limitation	3	7	1	21
Did not match correct contours	Contour Upload	human error	10	2	1	20
Approve contours which were matched incorrectly	Contour Upload	human error	10	2	1	20
Contours approved by a member of clinic unfamiliar with anatomy & targets	Contour Upload	human error	10	2	1	20
Automatch could cause mismatch in contours	Contour Upload	software error	10	2	1	20
Multi isocenter scan with multiple bb's	CT Simulation	software limitation	2	5	2	20
Incorrect OIS selected	Registration	human error	1	2	10	20
Isocenter not found	Review CT	software error	6	3	1	18
Internet quality limits ability to download plans from RPA	Plan Preparation	equipment limitation	3	6	1	18
Incorrect treatment machine parameters	Registration	human error	3	3	2	18
Incorrect MLC type selected	Registration	human error	3	3	2	18
Misselection of treatment techniques to be used	Registration	human error	3	3	2	18
Inaccurate TPS Parameters shared	Registration	human error	2	3	3	18

Incorrect Physics QA procedures selected	Registration	human error	2	3	3	18
Incorrect IMRT QA equipment selected	Registration	human error	2	3	3	18
Error in CT data transmission into RPA (error in dicom tag)	Upload CT	data transfer error	8	2	1	16
Imported wrong contours to TPS	Contouring	human error	8	2	1	16
Selected wrong contours to upload - wrong patient	Contour Upload	human error	8	2	1	16
Did not review contours prior to approving	Contour Upload	human error	8	2	1	16
Mismatch of scale on report and TPS	Plan Preparation	software error	2	4	2	16
Institutional firewalls/ antivirus blocks ability download plan	Plan Preparation	equipment limitation	3	5	1	15
Downloaded contours for wrong patient	Contour Review & Download	human error	2	7	1	14
Didn't correctly correlate contour to patient once downloaded	Contour Review & Download	human error	2	7	1	14
Attempt to upload contours to incorrect patient	Contour Upload	human error	2	7	1	14
CT rejected even though OK	Review CT	different workflow	2	6	1	12
Downloaded contours to confusing location (not organized)	Contour Review & Download	different workflow	2	6	1	12
Incorrect institutional information	Registration	human error	2	3	2	12
Omitted team member from registration	Registration	human error	2	3	2	12
Assign User category incorrectly	Registration	human error	2	3	2	12
Wrong energies selected for machine	Registration	human error	2	3	2	12

CT not approved	Review CT	human error	2	6	1	12
Delete is selected instead of review	Review CT	human error	2	6	1	12
CT rejected even though OK	Review CT	human error	2	6	1	12
Did not download DICOM file for TPS import	Contour Review & Download	human error	2	6	1	12
Downloaded contours to confusing location (not organized)	Contour Review & Download	human error	2	6	1	12
Do not accept contours	Contour Upload	human error	2	6	1	12
Do not upload contours to RPA	Contour Upload	human error	2	6	1	12
Contours neither accepted nor rejected	Contour Upload	human error	2	6	1	12
Delete is selected instead of review	Contour Upload	human error	2	6	1	12
Plan not transferred to TPS	Plan Preparation	human error	2	6	1	12
Error in upload (some structures have errors)	Contour Upload	data transfer error	10	1	1	10
Plan not properly transferred to TPS	Plan Preparation	data transfer error	10	1	1	10
Plan not interpreted correctly by the TPS (coordinates flipped, etc)	Plan Preparation	data transfer error	10	1	1	10
Plan not interpreted correctly by the TPS (coordinates flipped, etc)	Plan Preparation	software error	10	1	1	10
Plan not recalculated in TPS	Plan Preparation	human error	5	2	1	10
Reference point not added when needed for planning (chest wall)	Review CT	human error	2	5	1	10
Incorrect contact information for triage team	Registration	human error	3	3	1	9

Do not upload contours to RPA	Contour Upload	different workflow	1	9	1	9
contour associates with incorrect CT	Contouring	software error	8	1	1	8
Does not import contours to TPS	Contouring	different workflow	2	4	1	8
Request is accidentally denied	Approve Service Request	human error	2	4	1	8
Does not import contours to TPS	Contouring	human error	2	4	1	8
Plan not able to be downloaded from the RPA (browser issues)	Plan Preparation	software error	2	4	1	8
Error in upload (catastrophic)	Contour Upload	data transfer error	3	2	1	6
Plan is not deliverable	Treatment Delivery	equipment limitation	3	2	1	6
Failure to submit a registration request	Registration	human error	3	2	1	6
The plan is not deliverable	Treatment Delivery	software error	3	2	1	6
CTV1 was not contoured but was matched appropriately	Contour Upload	different workflow	2	3	1	6
H&N OAR's not included in the scan range	Review CT	human error	2	3	1	6
CTV1 was not contoured but was matched appropriately	Contour Upload	human error	2	3	1	6
Forgot to match contoured structures	Contour Upload	human error	2	3	1	6
Necessary OARs for RPA were not contoured	Contour Upload	human error	2	3	1	6
Contours not downloaded	Contour Review & Download	data transfer error	2	2	1	4
Plan not transferred to TPS	Plan Preparation	data transfer error	2	2	1	4

File corrupted while downloading from RPA	Plan Preparation	data transfer error	2	2	1	4
H&N OAR's not included in the scan range	Review CT	different workflow	2	2	1	4
Does not import contours to TPS	Contouring	equipment limitation	2	2	1	4
TPS will not import the treatment plan	Plan Preparation	equipment limitation	2	2	1	4
RPA incorrectly connects contours to the wrong patient	Contour Review & Download	software error	2	2	1	4
Compatibility issue preventing import	Contouring	software error	2	2	1	4
Processing error - results in no report	Contour Upload	software error	2	2	1	4
Necessary OARs for RPA were not contoured	Contour Upload	software error	2	2	1	4
Incorrect selection of TPS	Registration	human error	1	2	2	4
RPA unable to properly load 3D CT (too slow, or crashes)	Review CT	software error	3	1	1	3
Does not import contours to TPS	Contouring	software error	2	1	1	2

Appendix B – High-Risk Failure Mode and Effects Analysis Results Following System Updates

Failure Mode	Step	Cause	S	O	D	RPN
Plan not reviewed carefully prior to approval	Final Plan Approval	overreliance on system	9	4	8	288
RPA printout being used as plan documentation	Plan Preparation	overreliance on system	7	4	10	280
Necessary physics QA was not performed	Plan Preparation	overreliance on system	10	2	9	180
Request is approved without careful physician review	Approve Service Request	overreliance on system	9	3	6	162
Contoured target incorrectly	Contouring	software error	5	9	3	135
Incorrect identification of CTV 2 and CTV 3 - H&N	Create Service Request	overreliance on system	5	2	10	100
Unexpected BB placement - wrong place	CT Simulation	human error	9	3	3	81
Marked isocenter not verified in TPS	Plan Preparation	human error	9	3	3	81
Incorrect shift not validated in TPS	Plan Preparation	human error	9	3	3	81
Request is edited by another user prior to approval	Approve Service Request	human error	9	4	2	72

Incorrect selection of CTV regions - H&N	Create Service Request	overreliance on system	6	2	6	72
RPA printout being used as plan documentation	Plan Preparation	human error	7	1	8	56
Incorrect MRN (wrong patient)	Create Service Request	human error	9	3	2	54
Needed plan edits not made in TPS	Plan Preparation	human error	4	2	6	48
Not loading 3D because takes too long	Review CT	overreliance on system	6	1	4	24
Incorrect selection of laterality - CW	Create Service Request	human error	10	2	1	20
Incorrect identification of prior radiation	Create Service Request	human error	10	1	1	10
Incorrect answer on pregnancy questionnaire	Create Service Request	human error	10	1	1	10
Was the correct plan downloaded? (if multiple)	Plan Preparation	human error	9	1	1	9
Was the correct plan imported into the TPS?	Plan Preparation	human error	9	1	1	9
Reference point added to incorrect location	Review CT	human error	6	1	1	6

Appendix C – Physics Checklist Created for Use for Radiation Planning Assistant-Generated Treatment Plans

Checklist for Review of RPA Output Plans

Prescription / Service Request

- ☐ Service Date
- ☐ Patient Name
- ☐ MRN
- ☐ Patient De-identified ID
- ☐ Treatment Site
- ☐ Disease Extent
- ☐ Laterality
- ☐ Treatment Machine
- ☐ Technique (VMAT, 3D)
- ☐ Appropriate Energy
- ☐ Total Dose
- ☐ Fractionation
- ☐ Target coverage

CT Scan

- ☐ Marked Isocenter correctly identified (**Head and Neck** or **Cervix** plans)
- ☐ Correct CT
- ☐ Consistent # of CT slices in RPA & TPS
- ☐ Patient Orientation (HFS)
- ☐ Acceptable CT image quality (artifacts, FOV, etc)
- ☐ Target names
- ☐ Target contours
- ☐ OAR names
- ☐ OAR contours
- ☐ Target margins
- ☐ Body/External Contour

Final RPA Plan

- ☐ Gantry angle
- ☐ Collimator angle
- ☐ Couch angle
- ☐ Planning constraints met
- ☐ Hot Spot Dose
- ☐ Hot Spot in Target
- ☐ RPA QA Checks Reviewed
- ☐ Shift from marked isocenter to final isocenter verified
- ☐ **If Cervix**: Reference point set to desired superior field margin (optional)
- ☐ **If Chest Wall**: Reference point set at desired inferior tangent field margin (required)

RPA and TPS Plans

- ☐ Plan/CT/Contour DICOM imported to TPS without error
- ☐ Beam parameters match between TPS and RPA (Gantry angle, collimator angle, table angle, jaw size, etc.)
- ☐ MLC pattern matches between TPS and RPA **(Cervix & Chest Wall)**
- ☐ MU/beam matches between TPS and RPA prior to recalculation.

Recalculate with preset values of MU from RPA

Notes: _____

Final Plan Review **MUST be performed in local TPS following plan recalculation**

Appendix D – Magnitude of contouring edits made by each of the five primary dosimetrists in our dataset compared to the mean of the entire dataset, using DSC and APL.

	Dosi 1	**Dosi 2**	**Dosi 3**	**Dosi 4**	**Dosi 5**	**All Patients**
Mandible - Patients	37	54	26	40	29	399
Mandible - DSC	0.989±0.021	0.910±0.044	0.980±0.019	0.967±0.056	0.976±0.028	0.967±0.047
Mandible – APL	591±1370	4780±1660	931±938	881±1570	7556±1920	1610±2050
Brain - Patients	36	52	25	38	27	402
Brain - DSC	0.998±0.003	0.986±0.008	0.999±0.001	0.998±0.003	0.998±0.004	0.996±0.006
Brain – APL	2330±3460	15300±5000	1710±1540	2340±3540	2830±5930	4840±6250
Brainstem- Patients	36	53	26	40	28	418
Brainstem - DSC	0.985±0.044	0.941±0.056	0.983±0.029	0.981±0.040	0.985±0.033	0.975±0.047
Brainstem - APL	222±462	1090±681	302±463	268±402	318±728	502±959

Cochlea_L- Patients	35	39	15	32	22	342
Cochlea_L - DSC	0.911±0.149	0.684±0.184	0.851±0.167	0.867±0.173	0.932±0.094	0.869±0.173
Cochlea_L – APL	11.2±27.4	50.3±26.4	22.8±25.5	17.9±25.9	13.1±31.0	22.6±37.6
Cochlea_R- Patients	35	39	15	32	22	342
Cochlea_R - DSC	0.904±0.157	0.748±0.144	0.895±0.113	0.842±0.161	0.974±0.087	0.877±0.148
Cochlea_R - APL	12.8±29.9	38.2±20.5	14.2±21.2	21.1±25.4	4.00±19.3	20.9±33.9
Esophagus- Patients	35	41	15	37	21	357
Esophagus - DSC	0.951±0.128	0.908±0.137	0.957±0.070	0.956±0.106	0.992±0.009	0.926±0.154
Esophagus - APL	229±351	620±475	247±197	336±480	102±101	351±507
Eye_L - Patients	34	39	15	28	23	346
Eye_L - DSC	0.989±0.016	0.939±0.036	0.962±0.041	0.988±0.021	0.962±0.059	0.971±0.070
Eye_L - APL	67.4±102	388±217	203±201	63.9±92.43	272±460	165±295

Eye_R - Patients	34	39	15	28	23	347
Eye_R - DSC	0.989±0.014	0.945±0.029	0.975±0.035	0.989±0.023	0.979±0.029	0.978±0.039
Eye_R - APL	66.0±85.4	345±164	163±228	64.1±120	162±277	151±273
Lens_L - Patients	29	39	13	25	21	319
Lens_L - DSC	0.948±0.075	0.816±0.140	0.897±0.100	0.921±0.147	0.897±0.158	0.865±0.130
Lens_L - APL	2.00±3.06	24.0±18.9	12.2±13.7	6.92±13.9	13.7±24.5	11.7±16.4
Lens_R - Patients	29	39	13	27	21	324
Lens_R - DSC	0.957±0.069	0.824±0.111	0.913±0.062	0.950±0.062	0.918±0.107	0.904±0.149
Lens_R – APL	2.79±6.50	22.2±14.6	9.15±9.31	5.39±9.29	13.9±27.9	11.7±19.7
OpticNrv_L- Patients	27	39	14	26	21	326
OpticNrv_L - DSC	0.915±0.153	0.828±0.116	0.801±0.236	0.940±0.098	0.893±0.130	0.888±0.147
OpticNrv_L - APL	44.3±105	39.6±36.9	63.9±105	18.0±37.4	50.9±120	60.3±138

OpticNrv_R- Patients	27	39	14	26	21	327
OpticNrv_R- DSC	0.886±0.186	0.811±0.133	0.878±0.122	0.942±0.120	0.871±0.157	0.888±0.143
OpticNrv_R - APL	54.4±109	49.6±48.7	35.6±72.2	20.4±46.5	61.1±131	61.3±127
Parotid_L - Patients	36	51	20	39	26	391
Parotid_L - DSC	0.977±0.045	0.959±0.097	0.961±0.045	0.979±0.037	0.985±0.029	0.972±0.056
Parotid_L - APL	486±971	558±953	594±950	373±670	356±717	489±844
Parotid_R - Patients	36	52	20	38	26	390
Parotid_R - DSC	0.986±0.020	0.968±0.067	0.965±0.048	0.981±0.031	0.981±0.042	0.974±0.050
Parotid_R – APL	356±608	482±793	608±1060	362±663	388±816	471±800
SpinalCord - Patients	41	57	27	42	36	431
SpinalCord - DSC	0.989±0.010	0.974±0.071	0.983±0.037	0.975±0.074	0.976±.057	0.968±0.080
SpinalCord- APL	212±339	257±416	269±445	355±772	280±570	374±757

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424. doi:10.3322/caac.21492

2. Datta NR, Samiei M, Bodis S. Radiation therapy infrastructure and human resources in low- and middle-income countries: Present status and projections for 2020. *Int J Radiat Oncol Biol Phys*. 2014;89(3):448-457. doi:10.1016/j.ijrobp.2014.03.002

3. Huynh E, Hosny A, Guthier C, Bitterman DS, Petit SF, Haas-Kogan DA, Kann B, Aerts HJWL, Mak RH. Artificial intelligence in radiation oncology. *Nat Rev Clin Oncol*. 2020;17(12):771-781. doi:10.1038/s41571-020-0417-8

4. Segedin B, Petric P. Uncertainties in target volume delineation in radiotherapy - Are they relevant and what can we do about them? *Radiol Oncol*. 2016;50(3):254-262. doi:10.1515/raon-2016-0023

5. Berry SL, Boczkowski A, Ma R, Mechalakos J, Hunt M. Interobserver variability in radiation therapy plan output: Results of a single-institution study. *Pract Radiat Oncol*. 2016;6(6):442-449. doi:10.1016/j.prro.2016.04.005

6. Novak A, Nyflot MJ, Ermoian RP, Jordan LE, Sponseller PA, Kane GM, Ford EC, Zeng J. Targeting safety improvements through identification of incident origination and detection in a near-miss incident learning system. *Med Phys*. 2016;43(5):2053-2062. doi:10.1118/1.4944739

7. Scaggion A, Fusella M, Roggio A, Bacco S, Pivato N, Rossato MA, Mariel L, Peña A, Paiusco M. Reducing inter- and intra-planner variability in radiotherapy plan output with a commercial knowledge-based planning solution. *Physica Medica*. 2018;53:86-93. doi:10.1016/j.ejmp.2018.08.016

8. *MACHINE LEARNING AUTOMATED TREATMENT PLANNING MACHINE LEARNING PLANNING IN RAYSTATION*.

9. AutoContouring, Adaptive Therapy, Deep Learning Contouring. Accessed February 25, 2021. https://mirada-medical.com/radiation-oncology/

10. Limbus AI - Automatic Contouring for Radiation Therapy. Accessed February 25, 2021. https://limbus.ai/

11. Radiation planning assistant - A streamlined, fully automated radiotherapy treatment planning system. *Journal of Visualized Experiments*. 2018;2018(134):57411. doi:10.3791/57411

12. Kisling K, Zhang L, Simonds H, Fakie N, Yang J, McCarroll R, Balter P, Burger H, Bogler O, Howell R, Schmeler K, Mejia M, Jhingran A, Court L, Beadle BM. Fully automatic treatment planning for external-beam radiation therapy of locally advanced cervical cancer: A tool for low-resource clinics. *J Glob Oncol*. 2019;2019(5):1-8. doi:10.1200/JGO.18.00107

13. Rhee DJ, Jhingran A, Kisling K, Cardenas C, Simonds H, Court L. Automated Radiation Treatment Planning for Cervical Cancer. *Semin Radiat Oncol*. 2020;30(4):340-347. doi:10.1016/j.semradonc.2020.05.006

14. Kisling K, Zhang L, Shaitelman SF, Anderson D, Thebe T, Yang J, Balter PA, Howell RM, Jhingran A, Schmeler K, Simonds H, du Toit M, Trauernicht C, Burger H, Botha K, Joubert N, Beadle BM, Court L. Automated treatment planning of postmastectomy radiotherapy. *Med Phys*. 2019;46(9):3767-3775. doi:10.1002/mp.13586

15. Kisling K, Cardenas C, Anderson BM, Zhang L, Jhingran A, Simonds H, Balter P, Howell RM, Schmeler K, Beadle BM, Court L. Automatic Verification of Beam Apertures for Cervical Cancer Radiation Therapy. *Pract Radiat Oncol*. 2020;10(5):e415-e424. doi:10.1016/j.prro.2020.05.001

16. Rhee DJ, Jhingran A, Rigaud B, Netherton T, Cardenas CE, Zhang L, Vedam S, Kry S, Brock KK, Shaw W, O'Reilly F, Parkes J, Burger H, Fakie N, Trauernicht C, Simonds H, Court LE. Automatic contouring system for cervical cancer using convolutional neural networks. *Med Phys*. 2020;47(11):5648-5658. doi:10.1002/mp.14467

17. Cardenas CE, Beadle BM, Garden AS, Skinner HD, Yang J, Rhee DJ, McCarroll RE, Netherton TJ, Gay SS, Zhang L, Court LE. Generating High-Quality Lymph Node Clinical Target Volumes for Head and Neck Cancer Radiation Therapy Using a Fully Automated Deep Learning-Based Approach. In: *International Journal of Radiation Oncology Biology Physics*. Vol 109. Elsevier Inc.; 2021:801-812. doi:10.1016/j.ijrobp.2020.10.005

18. Cardenas CE, Anderson BM, Aristophanous M, Yang J, Rhee DJ, McCarroll RE, Mohamed ASR, Kamal M, Elgohari BA, Elhalawani HM, Fuller CD, Rao A, Garden AS, Court LE. Auto-delineation of oropharyngeal clinical target

volumes using 3D convolutional neural networks. *Phys Med Biol*. 2018;63(21). doi:10.1088/1361-6560/aae8a9

19. Netherton TJ, Cardenas CE, Rhee DJ, Court LE, Beadle BM. The Emergence of Artificial Intelligence within Radiation Oncology Treatment Planning. *Oncology*. 2021;99(2):124-134. doi:10.1159/000512172

20. [Treatment Facility] Incident Evaluation Summary, CP-2005-049 VMS. . Published online 2005:1-12.

21. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi:10.3322/caac.21660

22. DAC list of ODA recipients by region. In: ; 2015:453-454. doi:10.1787/aid_glance-2015-82-en

23. Datta NR, Samiei M, Bodis S. Radiation therapy infrastructure and human resources in low- and middle-income countries: Present status and projections for 2020. *Int J Radiat Oncol Biol Phys*. 2014;89(3):448-457. doi:10.1016/j.ijrobp.2014.03.002

24. Limbus AI - Automatic Contouring for Radiation Therapy. Accessed August 19, 2022. https://limbus.ai/

25. Ford E, Conroy L, Dong L, de Los Santos LF, Greener A, Gwe-Ya Kim G, Johnson J, Johnson P, Mechalakos JG, Napolitano B, Parker S, Schofield D, Smith K, Yorke E, Wells M. Strategies for effective physics plan and chart

review in radiation therapy: Report of AAPM Task Group 275. *Med Phys*. 2020;47(6):e236-e272. doi:10.1002/mp.14030

26. Ford EC, Evans SB. Incident learning in radiation oncology: A review. *Med Phys*. 2018;45(5):e100-e119. doi:10.1002/mp.12800

27. Kisling K, Johnson JL, Simonds H, Zhang L, Jhingran A, Beadle BM, Burger H, du Toit M, Joubert N, Makufa R, Shaw W, Trauernicht C, Balter P, Howell RM, Schmeler K, Court L. A risk assessment of automated treatment planning and recommendations for clinical deployment. *Med Phys*. 2019;46(6):2567-2574. doi:10.1002/mp.13552

28. Rassiah P, Su FF, Huang YJ, Spitznagel D, Sarkar V, Szegedi MW, Zhao H, Paxton AB, Nelson G, Salter BJ. Using failure mode and effects analysis (FMEA) to generate an initial plan check checklist for improved safety in radiation treatment. *J Appl Clin Med Phys*. 2020;21(8):83-91. doi:10.1002/acm2.12918

29. Batumalai V, Jameson MG, King O, Walker R, Slater C, Dundas K, Dinsdale G, Wallis A, Ochoa C, Gray R, Vial P, Vinod SK. Cautiously optimistic: A survey of radiation oncology professionals' perceptions of automation in radiotherapy planning. *Tech Innov Patient Support Radiat Oncol*. 2020;16:58-64. doi:10.1016/J.TIPSRO.2020.10.003

30. Huq MS, Fraass BA, Dunscombe PB, Gibbons JP, Ibbott GS, Mundt AJ, Mutic S, Palta JR, Rath F, Thomadsen BR, Williamson JF, Yorke ED. The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation

therapy quality management. *Med Phys*. 2016;43(7):4209-4262. doi:10.1118/1.4947547

31. Rhee DJ, Cardenas CE, Elhalawani H, McCarroll R, Zhang L, Yang J, Garden AS, Peterson CB, Beadle BM, Court LE. Automatic detection of contouring errors using convolutional neural networks. *Med Phys*. 2019;46(11):5086-5097. doi:10.1002/mp.13814

32. Parasuraman R, Manzey D. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Hum Factors*. 2010;52(3):381-410. doi:10.1177/0018720810376055

33. Goddard K, Roudsari A, Wyatt JC. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*. 2012;19(1):121-127. doi:10.1136/amiajnl-2011-000089

34. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28(3):231-237. doi:10.1136/bmjqs-2018-008370

35. Teo PT, Hwang MS, Shields W (Gary), Kosterin P, Jang SY, Heron DE, Lalonde RJ, Huq MS. Application of TG-100 risk analysis methods to the acceptance testing and commissioning process of a Halcyon linear accelerator. *Med Phys*. 2019;46(3):1341-1354. doi:10.1002/mp.13378

36. Gilmore MDF, Rowbottom CG. Evaluation of failure modes and effect analysis for routine risk assessment of lung radiotherapy at a UK center. *J Appl Clin Med Phys*. Published online April 9, 2021. doi:10.1002/acm2.13238

37. Xu AY, Bhatnagar J, Bednarz G, Flickinger J, Arai Y, Vacsulka J, Feng W, Monaco E, Niranjan A, Lunsford LD, Huq MS. Failure modes and effects analysis (FMEA) for Gamma Knife radiosurgery. *J Appl Clin Med Phys*. 2017;18(6):152-168. doi:10.1002/acm2.12205

38. Wexler A, Gu B, Goddu S, Mutic M, Yaddanapudi S, Olsen L, Harry T, Noel C, Pawlicki T, Mutic S, Cai B. FMEA of manual and automated methods for commissioning a radiotherapy treatment planning system. *Med Phys*. 2017;44(9):4415-4425. doi:10.1002/mp.12278

39. Schubert LK, Hendrickson K, Miften M, McNulty M, Vinogradskiy YY, Thomas DH, Westerly D, Stuhr K, Corral J, Royer D. The Current State of Physics Plan Review Training in Medical Physics Residency Programs in North America. *Pract Radiat Oncol*. 2020;10(3):e166-e172. doi:10.1016/j.prro.2019.09.006

40. Wu SY, Sath C, Schuster JM, Dominello MM, Burmeister JW, Golden DW, Braunstein SE. Targeted Needs Assessment of Treatment Planning Education for United States Radiation Oncology Residents. *Int J Radiat Oncol Biol Phys*. 2020;106(4):677-682. doi:10.1016/j.ijrobp.2019.11.023

41. Expect More from Auto-Contouring | Radiation Oncology Automation. Accessed August 19, 2022. https://www.mimsoftware.com/radiation-oncology/contour-

protegeai?utm_source=ads&utm_medium=ppc&utm_term=auto contouring radiation therapy&utm_campaign=Contour+ProtégéAI+-+Atlas-Based+vs.+AI+Auto+Contouring&hsa_src=g&hsa_acc=2475176161&hsa_ver=3&hsa_ad=592927998898&hsa_cam=14795459221&hsa_grp=127208431669&hsa_net=adwords&hsa_mt=b&hsa_kw=auto contouring radiation therapy&hsa_tgt=kwd-874304132355&gclid=CjwKCAjw6fyXBhBgEiwAhhiZsrjoNC_Wr2w7G5QRtnlFO_YFtJofXS3WOYgmPV0rVm97m_QnONE1ohoCbS8QAvD_BwE

42. Automated Treatment Planning | RaySearch Laboratories. Accessed August 19, 2022. https://www.raysearchlabs.com/automated-treatment-planning/

43. Eckhause T, Al-Hallaq H, Ritter T, Demarco J, Farrey K, Pawlicki T, Kim GY, Popple R, Sharma V, Perez M, Park S, Booth JT, Thorwarth R, Moran JM. Automating linear accelerator quality assurance. *Med Phys*. 2015;42(10):6074-6083. doi:10.1118/1.4931415

44. Babier A, Mahmood R, McNiven AL, Diamant A, Chan TCY. Knowledge-based automated planning with three-dimensional generative adversarial networks. *Med Phys*. 2020;47(2):297-306. doi:10.1002/MP.13896

45. Purdie TG, Dinniwell RE, Fyles A, Sharpe MB. Automation and intensity modulated radiation therapy for individualized high-quality tangent breast treatment plans. *Int J Radiat Oncol Biol Phys*. 2014;90(3):688-695. doi:10.1016/J.IJROBP.2014.06.056

46. Schipaanboord BWK, Giżyńska MK, Rossi L, de Vries KC, Heijmen BJM, Breedveld S. Fully automated treatment planning for MLC-based robotic radiotherapy. *Med Phys*. 2021;48(8):4139-4147. doi:10.1002/MP.14993

47. Das IJ, Cheng CW, Watts RJ, Ahnesjö A, Gibbons J, Li XA, Lowenstein J, Mitra RK, Simon WE, Zhu TC. Accelerator beam data commissioning equipment and procedures: Report of the TG-106 of the Therapy Physics Committee of the AAPM. *Med Phys*. 2008;35(9):4186-4215. doi:10.1118/1.2969070

48. Ezzell GA, Burmeister JW, Dogan N, Losasso TJ, Mechalakos JG, Mihailidis D, Molineu A, Palta JR, Ramsey CR, Salter BJ, Shi J, Xia P, Yue NJ, Xiao Y. IMRT commissioning: Multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119. *Med Phys*. 2009;36(11):5359-5373. doi:10.1118/1.3238104

49. Fraass B, Doppke K, Hunt M, Kutcher G, Starkschall G, Stern R, Dyke J Van. *American Association of Physicists in Medicine Radiation Therapy Committee Task Group 53: Quality Assurance for Clinical Radiotherapy Treatment Planning*.; 1998. doi:10.1118/1.598373

50. Nealon KA, Balter PA, Douglas RJ, Fullen DK, Nitsch PL, Olanrewaju AM, Soliman M, Court LE. Using Failure Mode and Effects Analysis to Evaluate Risk in the Clinical Adoption of Automated Contouring and Treatment Planning Tools. *Pract Radiat Oncol*. Published online 2022. doi:10.1016/J.PRRO.2022.01.003

51. Weintraub SM, Salter BJ, Chevalier CL, Ransdell S. Human factor associations with safety events in radiation therapy. *J Appl Clin Med Phys*. 2021;22(10):288-294. doi:10.1002/ACM2.13420

52. International Electrotechnical Commission, International Electrotechnical Commission. Technical Committtee 62, International Electrotechnical Commission. Subcommittee 62A, International Organization for Standardization. Technical Committee 210. *Medical Devices. Part 2, Guidance on the Application of Usability Engineering to Medical Devices*.

53. Pawlicki T, Samost A, Brown DW, Manger RP, Kim GY, Leveson NG. Application of systems and control theory-based hazard analysis to radiation oncology. *Med Phys*. 2016;43(3):1514-1530. doi:10.1118/1.4942384

54. Pawlicki T, Atwood T, McConnell K, Kim GY. Clinical safety assessment of the Halcyon system. *Med Phys*. 2019;46(10):4340-4345. doi:10.1002/MP.13736

55. Kisling K, Zhang L, Shaitelman SF, Anderson D, Thebe T, Yang J, Balter PA, Howell RM, Jhingran A, Schmeler K, Simonds H, Toit M, Trauernicht C, Burger H, Botha K, Joubert N, Beadle BM, Court L. Automated treatment planning of postmastectomy radiotherapy. *Med Phys*. 2019;46(9):3767-3775. doi:10.1002/mp.13586

56. Olanrewaju A, Court LE, Zhang L, Naidoo K, Burger H, Dalvie S, Wetter J, Parkes J, Trauernicht CJ, McCarroll RE, Cardenas C, Peterson CB, Benson KRK, du Toit M, van Reenen R, Beadle BM. Clinical Acceptability of Automated Radiation Treatment Planning for Head and Neck Cancer Using

the Radiation Planning Assistant. *Pract Radiat Oncol*. 2021;11(3):177-184. doi:10.1016/J.PRRO.2020.12.003

57. McCarroll RE, Beadle BM, Balter PA, Burger H, Cardenas CE, Dalvie S, Followill DS, Kisling KD, Mejia M, Naidoo K, Nelson CL, Peterson CB, Vorster K, Wetter J, Zhang L, Court LE, Yang J. Retrospective validation and clinical implementation of automated contouring of organs at risk in the head and neck: A step toward automated radiation treatment planning for low- And middle-income countries. *J Glob Oncol*. 2018;2018(4):1-11. doi:10.1200/JGO.18.00055

58. Clinical acceptability of fully automated external beam radiotherapy for cervical cancer with three different beam delivery techniques. *Med Phys*. Published online July 26, 2022. doi:10.1002/MP.15868

59. AAPM medical physics practice guideline 10.a.: Scope of practice for clinical medical physics. *J Appl Clin Med Phys*. 2018;19(6):11-25. doi:10.1002/acm2.12469

60. Xia P, Sintay BJ, Colussi VC, Chuang C, Lo YC, Schofield D, Wells M, Zhou S. Medical Physics Practice Guideline (MPPG) 11.a: Plan and chart review in external beam radiotherapy and brachytherapy. *J Appl Clin Med Phys*. 2021;22(9):4-19. doi:10.1002/ACM2.13366/FORMAT/PDF

61. Ford EC, Terezakis S, Souranis A, Harris K, Gay H, Mutic S. Quality control quantification (QCQ): A tool to measure the value of quality control checks in

radiation oncology. *Int J Radiat Oncol Biol Phys*. 2012;84(3). doi:10.1016/j.ijrobp.2012.04.036

62. Medical Physics Practice Guideline 4.a: Development, implementation, use and maintenance of safety checklists. *J Appl Clin Med Phys*. 2015;16(3):37-59. doi:10.1120/jacmp.v16i3.5431

63. Chan AJ, Islam MK, Rosewall T, Jaffray DA, Easty AC, Cafazzo JA. The use of human factors methods to identify and mitigate safety issues in radiation therapy. *Radiotherapy and Oncology*. 2010;97(3):596-600. doi:10.1016/J.RADONC.2010.09.026

64. Xia P, Lahurd D, Qi P, Mastroianni A, Magnelli A, Murray E, Kolar M, Guo B, Meier T, Chao ST, Suh JH, Yu N. Combining automatic plan integrity check (APIC) with standard plan document and checklist method to reduce errors in treatment planning. *J Appl Clin Med Phys*. 2020;21(9):124-133. doi:10.1002/acm2.12981

65. Zeitman A, Palta J S. M. Safety is No Accident: A Framework for Quality Radiation Oncology and Care. 2019. *Astro*. Published online 2019.

66. Covington EL, Chen X, Younge KC, Lee C, Matuszak MM, Kessler ML, Keranen W, Acosta E, Dougherty AM, Filpansick SE, Moran JM. Improving treatment plan evaluation with automation. *J Appl Clin Med Phys*. 2016;17(6):16-31. doi:10.1120/jacmp.v17i6.6322

67. Rhee DJ, Jhingran A, Rigaud B, Netherton T, Cardenas CE, Zhang L, Vedam S, Kry S, Brock KK, Shaw W, O'Reilly F, Parkes J, Burger H, Fakie N,

Trauernicht C, Simonds H, Court LE. Automatic contouring system for cervical cancer using convolutional neural networks. *Med Phys*. 2020;47(11):5648-5658. doi:10.1002/mp.14467

68. Cilla S, Ianiro A, Romano C, Deodato F, Macchia G, Buwenge M, Dinapoli N, Boldrini L, Morganti AG, Valentini V. Template-based automation of treatment planning in advanced radiotherapy: a comprehensive dosimetric and clinical evaluation. *Scientific Reports 2020 10:1*. 2020;10(1):1-13. doi:10.1038/s41598-019-56966-y

69. Craft DL, Hong TS, Shih HA, Bortfeld TR. Improved planning time and plan quality through multicriteria optimization for intensity-modulated radiotherapy. *Int J Radiat Oncol Biol Phys*. 2012;82(1). doi:10.1016/J.IJROBP.2010.12.007

70. Cokelek M, Holt E, Kelly F, Rolfo A, Ng M, Foley BM, Ryan S, Ho H, Brown A, McAlpine J, Chao M. Automation: The Future of Radiotherapy. *Int J Radiat Oncol Biol Phys*. 2020;108(3):e314. doi:10.1016/J.IJROBP.2020.07.750

71. Gopan O, Smith WP, Chvetsov A, Hendrickson K, Kalet A, Kim M, Nyflot M, Phillips M, Young L, Novak A, Zeng J, Ford E. Utilizing simulated errors in radiotherapy plans to quantify the effectiveness of the physics plan review. *Med Phys*. 2018;45(12):5359-5365. doi:10.1002/mp.13242

72. Moore KL, Kagadis GC, McNutt TR, Moiseenko V, Mutic S. Vision 20/20: Automation and advanced computing in clinical radiation oncology. *Med Phys*. 2014;41(1):010901. doi:10.1118/1.4842515

73. Nealon KA, Court LE, Douglas RJ, Zhang L, Han EY. Development and validation of a checklist for use with automatically generated radiotherapy plans. *J Appl Clin Med Phys*. 2022;(March):1-7. doi:10.1002/acm2.13694

74. Rhee DJ, Cardenas CE, Elhalawani H, McCarroll R, Zhang L, Yang J, Garden AS, Peterson CB, Beadle BM, Court LE. Automatic detection of contouring errors using convolutional neural networks. *Med Phys*. 2019;46(11):5086-5097. doi:10.1002/mp.13814

75. Cardan RA, Covington EL, Popple RA. Code Wisely: Risk assessment and mitigation for custom clinical software. *J Appl Clin Med Phys*. 2021;22(8):273-279. doi:10.1002/ACM2.13348

76. Hickling S V., Veres AJ, Moseley DJ, Grams MP. Implementation of free breathing respiratory amplitude-gated treatments. *J Appl Clin Med Phys*. 2021;22(6):119-129. doi:10.1002/ACM2.13253

77. Brezovich IA, Wu X, Duan J, Popple RA, Shen S, Benhabib S, Huang M, Christian Dobelbower M, Fisher WS. End-to-end test of spatial accuracy in Gamma Knife treatments for trigeminal neuralgia a). *Med Phys*. 2014;41(11):111703. doi:10.1118/1.4896819

78. Zakjevskii V V., Knill CS, Rakowski JT, Snyder MG. Development and evaluation of an end-to-end test for head and neck IMRT with a novel multiple-dosimetric modality phantom. *J Appl Clin Med Phys*. 2016;17(2):497-510. doi:10.1120/JACMP.V17I2.5705

79. Kim Y, Modrick JM, Pennington EC, Kim Y. Commissioning of a 3D image-based treatment planning system for high-dose-rate brachytherapy of cervical cancer. *J Appl Clin Med Phys*. 2016;17(2):405-426. doi:10.1120/JACMP.V17I2.5818

80. Gallo JJ, Kaufman I, Powell R, Pandya S, Somnay A, Bossenberger T, Ramirez E, Reynolds R, Solberg T, Burmeister J. Single-fraction spine SBRT end-to-end testing on TomoTherapy, Vero, TrueBeam, and CyberKnife treatment platforms using a novel anthropomorphic phantom. *J Appl Clin Med Phys*. 2015;16(1):170-182. doi:10.1120/JACMP.V16I1.5120

81. Han EY, Wang H, Briere TM, Yeboa DN, Boursianis T, Kalaitzakis G, Pappas E, Castillo P, Yang J. Brain stereotactic radiosurgery using MR-guided online adaptive planning for daily setup variation: An end-to-end test. *J Appl Clin Med Phys*. 2022;23(3):e13518. doi:10.1002/ACM2.13518

82. Leveson NG, Weiss KA. Software System Safety. *Safety Design for Space Systems*. Published online January 1, 2009:475-505. doi:10.1016/B978-0-7506-8580-1.00015-4

83. Rosen II. Writing software for the clinic. *Med Phys*. 1998;25(3):301-309. doi:10.1118/1.598201

84. McCarroll RE, Beadle BM, Balter PA, Burger H, Cardenas CE, Dalvie S, Followill DS, Kisling KD, Mejia M, Naidoo K, Nelson CL, Peterson CB, Vorster K, Wetter J, Zhang L, Court LE, Yang J. Retrospective validation and clinical implementation of automated contouring of organs at risk in the head and

neck: A step toward automated radiation treatment planning for low- And middle-income countries. *J Glob Oncol*. 2018;2018(4). doi:10.1200/JGO.18.00055

85. Collier DC, Burnett SSC, Amin M, Bilton S, Brooks C, Ryan A, Roniger D, Tran D, Starkschall G. Assessment of consistency in contouring of normal-tissue anatomic structures. *J Appl Clin Med Phys*. 2003;4(1):17-24. doi:10.1120/JACMP.V4I1.2538

86. Jenkins A, Mullen TS, Johnson-Hart C, Green A, McWilliam A, Aznar M, van Herk M, Vasquez Osorio E. Novel methodology to assess the effect of contouring variation on treatment outcome. *Med Phys*. 2021;48(6):3234-3242. doi:10.1002/MP.14865

87. Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, van Elmpt W, Dekker A. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiotherapy and Oncology*. 2018;126(2):312-317. doi:10.1016/j.radonc.2017.11.012

88. Cardenas CE, McCarroll RE, Court LE, Elgohari BA, Elhalawani H, Fuller CD, Kamal MJ, Meheissen MAM, Mohamed ASR, Rao A, Williams B, Wong A, Yang J, Aristophanous M. Deep Learning Algorithm for Auto-Delineation of High-Risk Oropharyngeal Clinical Target Volumes With Built-In Dice Similarity Coefficient Parameter Optimization Function. *Int J Radiat Oncol Biol Phys*. 2018;101(2):468-478. doi:10.1016/j.ijrobp.2018.01.114

89. Schreier J, Genghi A, Laaksonen H, Morgas T, Haas B. Clinical evaluation of a full-image deep segmentation algorithm for the male pelvis on cone-beam CT and CT. *Radiotherapy and Oncology*. 2020;145:1-6. doi:10.1016/j.radonc.2019.11.021

90. Turchan WT, Arya R, Hight R, Al-Hallaq HA, Dominello MM, Joyce D, McCabe B, McCall ARR, Perevalova E, Stepaniak CJ, Yenice KM, Burmeister JW, Golden DW. Physician Review of Organ-At-Risk Contours and Image Fusion Accuracy During the Radiotherapy Treatment Planning Process. *International Journal of Radiation Oncology*Biology*Physics*. 2019;105(1):E622. doi:10.1016/j.ijrobp.2019.06.1156

91. Oakland JS, Amsterdam O, London B, York N, San P, San D, Singapore F, Tokyo S. *Statistical Process Control*.

92. Nguyen CI. *Advanced Statistical Process Control Techniques for Analysis of Medical Linear Accelerator Performance*.

93. Binny D, Lancaster CM, Trapp J V., Crowe SB. Statistical process control and verifying positional accuracy of a cobra motion couch using step-wedge quality assurance tool. *J Appl Clin Med Phys*. 2017;18(5):70-79. doi:10.1002/ACM2.12136

94. Binny D, Aland T, Archibald-Heeren BR, Trapp J V., Kairn T, Crowe SB. A multi-institutional evaluation of machine performance check system on treatment beam output and symmetry using statistical process control. *J Appl Clin Med Phys*. 2019;20(3):71-80. doi:10.1002/ACM2.12547

95. Wang H, Xue J, Chen T, Qu T, Barbee D, Tam M, Hu K. Adaptive radiotherapy based on statistical process control for oropharyngeal cancer. *J Appl Clin Med Phys*. 2020;21(9):171-177. doi:10.1002/ACM2.12993

96. Strand S, Boczkowski A, Smith B, Snyder JE, Hyer DE, Yaddanapudi S, Dunkerley DAP, St-Aubin J. Analysis of patient-specific quality assurance for Elekta Unity adaptive plans using statistical process control methodology. *J Appl Clin Med Phys*. 2021;22(4):99-107. doi:10.1002/ACM2.13219

97. Rana S, Eckert C, Singh H, Zheng Y, Chacko M, Storey M, Chang J. Determination of machine-specific tolerances using statistical process control analysis of long-term uniform scanning proton machine QA results. *J Appl Clin Med Phys*. 2020;21(9):163-170. doi:10.1002/ACM2.12990

98. Puyati W, Khawne A, Barnes M, Zwan B, Greer P, Fuangrod T. Predictive quality assurance of a linear accelerator based on the machine performance check application using statistical process control and ARIMA forecast modeling. *J Appl Clin Med Phys*. 2020;21(8):73-82. doi:10.1002/ACM2.12917

99. Mehrens H, Douglas R, Gronberg M, Nealon K, Zhang J, Court L. Statistical process control to monitor use of a web-based autoplanning tool. *J Appl Clin Med Phys*. 2022;23(12):e13803. doi:10.1002/ACM2.13803

100. Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology*. 1945;26(3):297-302. doi:10.2307/1932409

101. Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, van Elmpt W. Evaluation of measures for assessing time-saving of automatic

organ-at-risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol*. 2020;13:1-6. doi:10.1016/J.PHRO.2019.12.001

102. Gosbee J. Human factors engineering and patient safety. *Quality and Safety in Health Care*. 2002;11(4):352-354. doi:10.1136/QHC.11.4.352

103. Reason J. Human error: models and management. *BMJ : British Medical Journal*. 2000;320(7237):768. doi:10.1136/BMJ.320.7237.768

Vita

Kelly Anne Nealon was born in New York, the daughter of Catherine Mary Nealon and Shawn Timothy Nealon. After completing her work at Tamarac High School in 2011, she enrolled at Siena College in Loudonville, New York. She received the degree of Bachelor in Science with a major in Physics in May 2015. After matriculating into Vanderbilt University's graduate medical physics program, she earned her Masters of Science with a major in Medical Physics in May 2017. She then completed the Ohio State University's Residency in Therapeutic Medical Physics in June 2019. In August of 2019, she entered The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences.

Permanent address:

123 N Shore Road

Petersburg, NY 12138