Advances in Radiation Oncology

# Development and Clinical Implementation of an Automated Virtual Integrative Planner for Radiation Therapy of Head and Neck Cancer

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | ADVANCESRADONC-D-21-00367R1 |
| **Article Type:** | Scientific Article |
| **Section/Category:** | Physics Contribution |
| **Corresponding Author:** | Charles Mayo, Ph.D.<br>University of Michigan<br>Ann Arbor, MI UNITED STATES |
| **First Author:** | Elizabeth M Jaworski, MD, MS |
| **Order of Authors:** | Elizabeth M Jaworski, MD, MS |
| | Michelle L Mierzwa, MD |
| | Karen A Vineberg, MS |
| | John Yao, PhD |
| | Jennifer L Shah, MD |
| | Caitlin A Schonewolf, MD, MS |
| | Dale Litzenberg, PhD |
| | Laila A Gharzai, MD, LLM |
| | Martha M Matuszak, PhD |
| | Kelly C Paradis, PhD |
| | Ashley Dougherty, CMD |
| | Pamela Burger, CMD |
| | Daniel Tatro, CMD |
| | George Spencer Arnould, CMD |
| | Jean M Moran, PhD |
| | Choonik Lee, PhD |
| | Avraham Eisbruch, MD |
| | Charles Mayo, PhD |
| **Abstract:** | Purpose : Head and neck (HN) radiation (RT) treatment planning is complex and resource intensive. Deviations and inconsistent plan quality significantly impact clinical outcomes. We sought to develop a novel automated virtual integrative (AVI) knowledge-based planning application to reduce planning time, increase consistency, and improve baseline quality.<br>Materials and Methods: An in-house write-enabled script was developed from a library of 668 previously treated HN RT plans. Prospective hazard analysis was performed, and mitigation strategies were implemented before clinical release. The AVI-planner software was retrospectively validated in a cohort of 52 recent HN cases. A physician panel evaluated planning limitations during initial deployment, and feedback was enacted via software refinements. A final second set of plans was generated and evaluated. Kolmogorov-Smirnov (KS) test in addition to Generalized Evaluation Metric (GEM) and Weighted Experience Score (WES) were used to compare normal tissue sparing between final AVI-planner versus respective clinically treated and historically accepted plans. One-tailed T-test was used to compare the interactive time required for AVI-planner versus manual optimization.<br>Results: Initially, 86% of plans were acceptable to treat with 10% minor and 4% major revisions or rejection recommended. Variability was noted in plan quality among HN |

| | |
|---|---|
| | subsites, with high initial quality for oropharynx and oral cavity plans. Plans needing revisions were comprised of sinonasal, nasopharynx, p-16 negative SCC Unknown Primary or cutaneous primary sites. Normal tissue sparing varied within subsites, but AVI-planner significantly lowered mean larynx dose (median 18.5 Gy vs 19.7 Gy, $p<0.01$) compared to clinical plans. AVI-planner significantly reduced interactive optimization time (mean 2 vs 85 minutes, $p<0.01$). |

Cover Letter

MICHIGAN MEDICINE
UNIVERSITY OF MICHIGAN

Department of Radiation Oncology
University of Michigan Medical Center
1500 East Medical Center Drive
Ann Arbor, MI 48109

Advances in Radiation Oncology
July 10, 2021

Dear Editorial Team,

We are pleased to submit our work entitled, "**Development and Clinical Implementation of an Automated Virtual Integrative Planner for Radiation Therapy of Head and Neck Cancer**" for consideration of publication as a Scientific Article (Full-Length Article) in Advances in Radiation Oncology. This work has not been published in part or full with another journal as we feel it is best suited for your esteemed publication. Considering the scope of this project, the contributions of all 18 co-authors were critical in the development, testing, and clinical implementation of this work.

Radiation (RT) planning for head and neck (HN) cancer remains complex and resource-intensive, with deviations in plan quality directly impacting clinical outcomes. Automated planning rapidly creates acceptable HN RT plans; however, the quality of automated plans has never been formally compared among different HN subsites. Herein, we describe the development and implementation of a novel write-enabled autoplanning script based upon a large foundational plan library. We identified profound variability in plan quality among HN subsites, which improved with iterative testing and physician feedback. Our work cautions against interpreting that automated planning achievements are universal among all HN subsites.

Given the successes achieved with automated RT planning and interest in clinical adoption, we feel this is highly relevant to your readership and is well suited for publication in ARO. We hope you share our enthusiasm for the study, and look forward to any concerns or questions you have about it.

Sincerely,
Elizabeth Jaworski, MD, MS

Advances in Radiation Oncology
December 18, 2021

Dear Editor,
Thank you and the reviewers for your careful consideration of our manuscript entitled, "Development and Clinical Implementation of an Automated Virtual Integrative Planner for Radiation Therapy of Head and Neck Cancer" and your offer to consider a revised version of the manuscript. We have the following responses to the reviewers' suggestions.

Reviewer 1:
**While the manuscript under review appears to address an important issue in modern radiotherapy- that is, the development and implementation of an automated planner for H&N cases- the readers of this journal will benefit if, in according to the view of this referee, address some important issues.**
We thank the reviewer for these comments.

In a non exhaustive enumeration these issues include:
1) **In Methods, pg 6, "write-enabled Script Refinement". This part of the manuscript will benefit from re-evaluation of its presentation. In echoes the work of Prof Sokal ( Social Text #46/47, pp. 217-252 (spring/summer 1996), or the central point in Sidney Harris' work ( http://www.sciencecartoonsplus.com/images/home-page-miracle.gif).**
We take your point and have added the content suggested. Our focus for the manuscript is on a more detailed process, including diagnosis sub-sites, for quantifying statistical comparisons of plans created with automated planning, scripts with both historic and literature benchmark values. We believe these comparisons are important for any specific instance of an automated planning script.

Note this is not a machine learning algorithm. The reviewers comments and reference to https://mlbazaar.github.io is suggestive that they are operating under that premise. The algorithm uses statistical analysis of prior plans to set constraints for future plans. We've modified the language in the paper to help clarify.

However, as you point out readers are going to want more detail about the specific instance of the auto-planning algorithm and ability to use the code. We have added edits to provide additional detail and access, while still keeping the focus on the process for evaluation.

2) **The Script central in this manuscript accepts input from five other distinct modules ( Hazarda Analysis, Physics & Clinical Evaluation Testing and Clinical Deployment). Consider showing explicitly how this input is handled by the script and how it may modify the script itself.**
A paragraph on clinical deployment was added in the discussion section, since it pertains to actions after the results of the current study.

3) **Consider offering explicitly the "coupling" of this script to the calculation engine of Eclipse.**
We agree that programmatically integrating our script into the optimization engine in the manufacturer's code is highly desirable. However, it is not possible in the FDA cleared clinical version of the Eclipse Scripting Application Programming Interface. As now mentioned in the manuscript, the clinical version does not even allow interaction with the optimizer. This is only possible in the non-FDA cleared

research version. We strongly agree with you that this is highly desirable. We've added a call for it in our discussion section.

4) **Further, it would be beneficial both to the reviewers of this manuscript and to the subsequent readers of the publication if the complete work of the development and implementation become available, for instance, via GitHub or a similar publicly accessible platform as shown here ( MLBazaar <https://mlbazaar.github.io/>) Understandably, if protection of IP resulting from this manuscript has not yet been obtained, the author present elliptic information which obviously detracts from the usefulness of this work in its current phase.**
We appreciate your suggestion and have had several discussions about it. This is a significant ask, requesting that the source code be made open source and relinquishing licensing rights. We strongly believe in increasing use of standardization and automation in clinical processes to improve efficiency, quality and ability to aggregate and learn from data. We agree with the reviewer that making the code open source, could substantially aid of automated planning growth in the community. The work was done as part of a grant from Varian medical systems. They typically do not prevent publication, if so, they may only do so for a few months. We plan publish the code with the manuscript when it reaches a stage where neither requires anonymization of the authors or institution.

Reviewer 2
**This paper describes the development of an automated planning script for head and neck radiotherapy. The authors trained their script using a large cohort of cases, although the number of cases of some sub-classifications were a lot smaller. Two testing rounds were performed, although the first round was used as more of a validation to tweak the script based on failures. The second testing round involved repeating planning for the same cohort used in the first testing round using the revised script, which would not be considered best practice but I think is acceptable given the results of the first testing round. The work is generally well performed and well described. I have several comments below, including some regarding what I consider a lack of sufficient detail to replicate the authors' work.**
We thank the reviewer for these comments.

1) **The methods section does not give sufficient details to replicate the work. Specifically, more detail is required on the initial script development, the training phase, and how the developed script produces the automated plans. Elsewhere, the script is described as knowledge-based but the reader needs to understand what specific knowledge is used and how it is used in the initial script. For example, is there case matching based on target and OAR sizes/proximity, are overlap metrics used, is an expected DVH generated, is that converted into optimization objectives, are standard OAR constraints from the literature added, are optimization relative priorities decided by the script or is a standard wishlist used, etc. There is slightly more detail about the refinements to the script prior to 'Round 2', however the reader needs a clear understanding of how the script works during 'Round 1'.**
Thank you this is a valuable suggestion. We have added detail about how the algorithm works and about refinements made in response to Round 1 evaluation.

Our focus for the manuscript is on a more detailed process, including diagnosis sub-sites, for quantifying statistical comparisons of plans created with automated planning, scripts with both historic and literature benchmark values. We believe these comparisons are important for any specific instance of an

automated planning script. Adding the suggested detail helps to illustrate the importance of this iterative evaluation and value in quantifying results for several HN sub sites.

2) **I was unclear on the level of clinical review performed at 'Round 2'. On P10 line 33 the authors state that only the 7 failed 'Round 1' plans were clinically re-evaluated. However, earlier it states that all 52 plans were replanned in 'Round 2'. How was the continued clinical acceptability of the other 45 plans confirmed after optimization using the new version of the script?**
We appreciate this comment, and we agree with the need to clarify the level of clinical review in 'Round 2'. Using the same criteria, the continued clinical acceptability of the remaining 45 plans was evaluated by the same physician panel. All 45 remaining plans were considered "treat as is." We have updated the outer rung of the doughnut chart in Figure 2 to reflect this evaluation. We have also modified the text within the Methods section to reflect this (page 7, paragraph 1): "The same physician panel re-evaluated *all 52 plans*." Within the Results section (page 11, paragraph 1): "During Round 2 evaluation *of all 52 plans,* there were no rejections nor major revisions (Figure 2)." Also, "The remaining *48* plans were 'treat as is.'"

3) **What is the difference between clinically treated plans and historically accepted plans in the 'Round 2' evaluation? What is the benefit of evaluating against both rather than one or the other?**
We thank the reviewer for this comment. To clarify this further, we have modified the Methods section (page 7, paragraph 2) as follows: "*Clinically treated plan denotes the patient-specific RT plan, which was delivered during the patient's treatment course. Within this context, evaluating Round 2 versus the clinically treated plan provides an individual, patient-level comparison of plan quality. Comparisons to historically accepted plans were based on summarized metrics captured from the entire 668 HN foundational library. Thus, Round 2 plan quality was assessed in the context of aggregate institutional experience with all 668 considered high-quality HN plans. Evaluating Round 2 plans in both situations more fully characterizes plan quality at both the patient-level and institutional experience- level. We* u*sed* Python 3.8 statistical software *for this analysis."*

4) **Having read the manuscript several times, I am unclear what makes the planning script 'virtual' and 'integrative' beyond a local development project name. I think it's defensible to refer to it as an automated planning script. Also, the authors should stick to either 'automated' or 'semi-automated' throughout.**
We appreciate this comment and agree with the reviewer regarding this inconsistency. We removed "semi" from the Introduction (page 4, paragraph 1) as follows: "Herein, we report the development of a*n* automated virtual integrative…". We removed "semi" from the Discussion (page 14, paragraph 1): "We developed and implemented a knowledge-based automated virtual integrative…"

5) **P5 line 43. What is the physics evaluation mentioned here? Is this purely the software input checking mentioned in the previous paragraph, or something else. Please clarify in the text.**
Thank you. We clarified that this refers to plan quality checks carried out as part of physics plan checks

6) **P5 line 48. Was this process mapping and risk scoring done as a multidisciplinary team?**
Yes it was a multi-disciplinary process. We clarified in the sentence.

7) **P5 line 53. Please give the citation for TG100 here, if referring to it.**

We apologize for this oversight. The citation for TG100 is now included in the Methods section (page 5, paragraph 2).

8) **P6, Patient selection. This section includes the general treatment planning details but is missing the dose grid size, dose calculation and optimization algorithm, and Eclipse version.**
Thank you these details were added.

9) **P7 line 13. I assume this means that some of the plans each week at this weekly panel, rather than the same plans were reviewed each week. Please clarify in the text, if it's the first option you can probably just remove 'weekly'.**
We agree with this clarification. We modified the Methods section (page 7, paragraph 1), which now reads: "Clinical plans underwent peer-review by a subspecialty panel of attending radiation oncologists."

10) **P10 lines 37-42. This seems either to belong in the methods section or is unnecessary repetition of methods.**
We appreciate this comment, and duplicative content within the first sentence has been removed from this Results section (page 10, paragraph 3). The Methods section (page 7, paragraph 1) now includes "*HN subsites were grouped by treatment paradigm and anatomic proximity."*

11) **P11 lines 15-22. This seems more like discussion than results. If it is results, I think the "frequently emphasized" needs to be quantitative.**
We thank the reviewer for this recommendation. We modified the Results section (page 10, paragraph 3) to include: *"Major revisions were limiting hot spots outside PTV, restricting hot spots within PTV to 105-110%, and improving target coverage."* The Discussion section (page 14, paragraph 3) now contains, "*Physicians frequently emphasized higher OAR prioritization. For instance, the clinical plan aggressively spared contralateral parotid further below the planning objective in a cT1N1 p16+ tonsil cancer, whereas AVI-planner less aggressively spared the contralateral parotid to meet constraints."*

12) **P11 lines 28-36. This seems either to belong in the methods section or is unnecessary repetition of methods.**
We appreciate this recommendation. In addition to the revisions listed above for Reviewer 2, Item 2, we also deleted duplicative content from this section within Results (page 11, paragraph 1).

13) **Figures 4 & 5. The symbols that presumably denote statistical significance need to be explained to the reader in the caption**.
We apologize for the confusion, and we would like to call the reviewer's attention to Figure Legends (page 21). We modified the legend for Figure 4 (page 21, paragraph 4) to include definition of statistical significance; however, no modifications were made for the symbol explanations. This legend now reads: "Red "x" denotes consensus thresholds [42]. *Statistical significance was achieved with $p<0.05$ on one-sided Kolmogrorov-Smirnov test.* Symbols along y-axis indicate statistically significant difference in OAR sparing between AVI-planner versus clinical plans: total cohort (panel A-circles), Oropharynx and p16+ SCC Unknown Primary (panel B-stars), Oral Cavity and Salivary (panel B- diamonds). Filled shapes indicate AVI-planner significantly improved sparing whereas unshaded symbols indicate clinical plan achieved *significantly* better sparing."

In responding to this comment, we also noted that the slide titles in Figures 1-5 failed to exactly match the titles listed within the figure legends. We apologize for this oversight. We have modified the figure legend titles and corresponding slides to match as follows:
"Figure 1: *Flow chart depicting write-enabled script d*evelopment and release process."
"Figure 2: Integrated pie and doughnut charts demonstrating clinical acceptability of AVI-planned cases *among* H&N subsites."
"Figure 3: Sample HN RT plans for *early-stage oropharynx and adjuvant Merkel cell carcinoma*. *Cases include* stage II…"
"Figure 4: Differential *normal tissue* sparing by clinical versus AVI-planner. *Comparisons* for all HN subsites (A)…"
"Figure 5: Box and whisker plot comparing dosimetrist interactive time between AVI-planner and manual optimization. *Significantly less time required for Round 2 AVI-planner* (n=51) *versus* manual optimization (n=10)."
"Table 1: Conformality and heterogeneity *(ICRU 83)* indexes *of* clinical and *Round 2* AVI-planner for high…"

We truly appreciate the opportunity to revise our manuscript and hope that you find it acceptable for publication.

Sincerely,
Co-authors

MICHIG AN MEDICINE
UNIVERSITY OF MICHIGAN

**Department of Radiation Oncology**

**University Hospital**
UH-B2C490
1500 E. Medical Center Dr. SPC 5010
Ann Arbor. MI 48109-5010
(734) 936-4300
(734) 763-7371 fax
www.med.umich.edu/ardonc

**Mid-Michigan Medical Center Alpena**
1501 West Chisholm
Alpena. MI 49707
(989) 356-7353
(989) 356-8118fax

**Metro Health University of Michigan Health**
5950 Metro Way
Wyoming. MI 49519
(616) 252-8180
(616) 252-8194 fax

**Mercy Health Lacks Cancer Center**
250 Cherry Street SE
Grand Rapids. MI 49503
(616) 685-6218
(616) 685-8918 fax

**Providence Park Hospital Southfield Campus**
22301 Foster Winter Drive Ste. 100
Southfield. MI 48075
(248) 849-3321
(248) 849-8448 fax

**Providence Park Hospital Novi Campus**
Assarian Cancer Center
47601 Grand River Avenue
Novi. MI 48374
(248) 465-4300
(248) 465-5471 fax

**Veterans Administration Medical Center**
2215 Fuller Road. 1148
Ann Arbor. MI 48105
(734) 845-3914
(734)845-3826 fax

3/21/2022

Dear Editor,

We are pleased to submit our second revision of our manuscript entitled

Development and Clinical Implementation of an Automated Virtual Integrative Planner for Radiation Therapy of Head and Neck Cancer

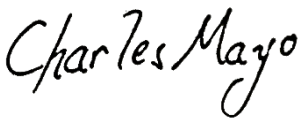
The original manuscript was submitted on 12/21/2021. It was assigned a number D-21-00367. The first revision was submitted on 1/14/2022 and assigned a different number D-21-00239R1. Evidently something went wrong in the submission process with the result of files being lost and perhaps a second group of reviewers involved in the evaluation.

The revision we are submitting already had responses and modifications in response to the reviewer comments from the original submissions D-21-00367. The revision we are submitting also includes responses and modifications in response to the reviewer comments generated from D-21-00239R1. Following instructions from the Managing Editor, Ms. Gayle (3/14/2022) we are attaching the revision files to D-21-00239R1.

We thank the editors and reviewers for their time. They have made the paper stronger and broader in scope. We thank the first set of reviewers for prompting us to make the source code available on public repository with the publication of the manuscript. This step will help the overall goal of promoting use and generalization of the approach to other researchers and institutions.

We have also uploaded as a supplemental document the file generated by ARO as part of the original submission. This provides the original reviewer comments and responses for reference.

Best Regards,



Charles Mayo PhD.
Professor of Radiation Oncology
University of Michigan

Advances in Radiation Oncology
March 18, 2022

Dear Editor,
Thank you and the reviewers for your careful consideration of our manuscript entitled, "Development and Clinical Implementation of an Automated Virtual Integrative Planner for Radiation Therapy of Head and Neck Cancer" and your offer to consider a revised version of the manuscript. We have the following responses to the reviewers' suggestions.

**Reviewer #1: The paper outlined the process of creating an automated virtual integrative knowledge-based planning application for head and neck RT. To do this the researchers utilized Eclipse's scripting software and referenced a catalog of previous HN cases to create an algorithm to complete the task automatically. Once they had their product they tested it on 52 cases, and then ran those results by a board of physicians to determine their quality. After their first round they made some changes to parameters in their program and tested it again against the same 52 cases and checked by the same board of physicians. Once they felt their product was satisfactory they implemented it in the clinic and claim to be using it ever since and have it working as expected.**

**Although, the process outlined in the paper is logical, however an oddity in their work was the relatively small testing pool, only 52 cases. It is understandable that not having that many cases to work with at a particular clinic, but I think it will be difficult for complex cases with multi-levels dosing prescription just to further test and refine their product.**

We would like to take this opportunity to build upon this reviewer's summary. In our discussion, we note that our clinic uses this software for warm start optimization for dosimetrists, who are then able to subsequently manually modify the HN RT plans. We do not use AVI-planner for retreatment or palliative cares. This algorithm plans up to 3  Similarly to any other data-driven planning algorithm, use of our script requires ongoing evaluation and adaptation within our clinic.

**Major concerns:**

1. **Perhaps overfitting the model with 668 cases (very high number of cases) and under validation with only 52 cases (small numbers)**

   We appreciate this feedback. For machine learning algorithms, we acknowledge that higher case numbers are necessary for training and validation. Overfitting is a characteristic of machine learning approaches. However, our AVI-planner model is not a machine learning algorithm. This has been clarified in the introduction (page 4, paragraph 1). This algorithm uses statistical analysis of DVH metrics, weighted experience scores (WES), and generalized evaluation metrics (GEM) from 668 historical clinically treated plans to assign optimization parameters. We then identified planning deficiencies by evaluating AVI-planner performance for 52 plans.

   Furthermore, to contextualize our parameter generation and testing set case numbers, we also reviewed relevant published autoplanning series. Compared to the literature, the number of cases we used to define the algorithm's parameters or to evaluate autoplans was comparable or larger as illustrated in the table below. Compared to other data driven

planning algorithms like Eclipse's RapidPlan and Pinnacle's Auto-Planning, the number of cases utilized in our work for model development and testing is higher.

| *Parameter Generation* | *Testing Set* | *Citation* |
|---|---|---|
| ***668*** | ***57*** | *This manuscript* |
| *57* | *23* | *Dumane VA, Tam J, Lo YC, Rosenzweig KE. RapidPlan for Knowledge-Based Planning of Malignant Pleural Mesothelioma. Pract Radiat Oncol. 2021 Mar-Apr;11(2):e219-e228. doi: 10.1016/j.prro.2020.06.003. Epub 2020 Jun 17. PMID: 32562788.* |
| *42* | *42 (not an independent set)* | *Moore KL, Brame RS, Low DA, Mutic S. Experience-based quality control of clinical intensity-modulated radiotherapy planning. Int J Radiat Oncol Biol Phys. 2011 Oct 1;81(2):545-51. doi: 10.1016/j.ijrobp.2010.11.030. Epub 2011 Jan 27. PMID: 21277097.* |
| *50* | *50* | *Giaddui T, Geng H, Chen Q, Linnemann N, Radden M, Lee NY, Xia P, Xiao Y. Offline Quality Assurance for Intensity Modulated Radiation Therapy Treatment Plans for NRG-HN001 Head and Neck Clinical Trial Using Knowledge-Based Planning. Adv Radiat Oncol. 2020 May 22;5(6):1342-1349. doi: 10.1016/j.adro.2020.05.005. PMID: 33305097; PMCID: PMC7718499.* |
| *10* | *10* | *Ouyang Z, Liu Shen Z, Murray E, Kolar M, LaHurd D, Yu N, Joshi N, Koyfman S, Bzdusek K, Xia P. Evaluation of auto-planning in IMRT and VMAT for head and neck cancer. J Appl Clin Med Phys. 2019 Jul;20(7):39-47. doi: 10.1002/acm2.12652. Epub 2019 Jul 4. PMID: 31270937; PMCID: PMC6612692.* |
| *83* | *20* | *Fogliata A, Reggiori G, Stravato A, Lobefalo F, Franzese C, Franceschini D, Tomatis S, Mancosu P, Scorsetti M, Cozzi L. RapidPlan head and neck model: the objectives and possible clinical benefit. Radiat Oncol. 2017 Apr 27;12(1):73. doi: 10.1186/s13014-017-0808-x. PMID: 28449704; PMCID: PMC5408433.* |
| *5* | *50 (not independent)* | *Krayenbuehl J, Norton I, Studer G, Guckenberger M. Evaluation of an automated knowledge based treatment planning system for head and neck. Radiat Oncol. 2015 Nov 10;10:226. doi: 10.1186/s13014-015-0533-2. PMID: 26555303; PMCID: PMC4641383.* |
| *30 and 60* | *15 and 15* | *Tol JP, Delaney AR, Dahele M, Slotman BJ, Verbakel WF. Evaluation of a knowledge-based planning solution for head and neck cancer. Int J Radiat Oncol Biol Phys. 2015 Mar 1;91(3):612-20. doi: 10.1016/j.ijrobp.2014.11.014. Epub 2015 Jan 30. PMID: 25680603.* |
| *0* | *30* | *Hansen CR, Bertelsen A, Hazell I, Zukauskaite R, Gyldenkerne N, Johansen J, Eriksen JG, Brink C. Automatic treatment planning improves the clinical quality of head and neck cancer treatment plans. Clin Transl Radiat Oncol. 2016 Sep 19;1:2-8. doi: 10.1016/j.ctro.2016.08.001. PMID: 29657987; PMCID: PMC5893480.* |

**2. No DVHs comparison shown wrt clinical plan**

As discussed in our Materials and Methods (pages 8, paragraph 3 and page 9) and results (pages 15-16), the weighted experience score (WES) quantitatively evaluates autoplan DVH

curves with respect to the DVHs of the 668 historical treated plans. We understand that a DVH may be more customary in this situation. Therefore, in Figure 5 we have added a typical representative DVH for a locally advanced oropharynx cancer for visual comparison. Within the results (page 15, paragraph 1), the text now reads, "*Given that oropharynx was the most prevalent HN subsite within both the foundational library and the validation cohort, we also selected a representative DVH from a locally advanced oropharynx cancer treated with definitive chemoradiation. This demonstrates typical DVH metrics from a case which was well represented in the model (Figure 5).*"

3. **Figure 3 - model plan's dose distribution looks much worse than clinical plan's even with round 2 optimization - also looking at 1 slide can can't be judged**

We wholeheartedly agree with the reviewer that the AVI-planner software does not generate HN RT plans that always exceed the caliber of HN RT plans generated by an experienced, subspecialized HN dosimetrist- this was not our intent. We aimed to rapidly create baseline plans which could be manually modified to improve consistency and reduce time, which we successfully accomplished. Clinically, the subtle conformality and heterogeneity differences would not increase risk of toxicity or detrimentally impact oncologic control. There was no sequential manual modification of the Round 2 plans in this evaluation; however, using this software in our clinic, dosimetrists can sequentially, manually modify the plans.

Similarly to the autosegmentation validation literature, it is critical to evaluate the best and worst case scenarios to reveal deficiencies in algorithm performance. We believe it is highly unlikely that any existing autoplanning algorithms are perfect or generate plans of higher quality than subspecialized dosimetrists for all cases. Furthermore, we would cautiously approach use of any algorithms where claims like these were made. Our goal for showing both cases was transparency with our audience about AVI-planner capabilities and limitations. We aimed to ensure that audiences understand for underrepresented cases, such as sinonasal or base of skull, the algorithm performance cannot be expected to be similar to a subspecialized dosimetrist. Within our Discussion section (page 19, paragraph 3), we note that underrepresented cases require more manual modification and oversight, whereas well-represented cases require less manual input.

We also agree that looking at 1 slide is not sufficient to compare plan quality between clinically treated versus automated plans. For this reason, we qualitatively (Figure 2) and quantitatively (Figure 4, Figure 5) assessed clinical acceptability and plan quality. We did modify Figure 3 to include a key for isodose levels per request of Reviewer #2 (see below). We updated the title and figure legend for Figure 3 to further clarify AVI-planner performance between typical, well-represented versus atypical, underrepresented subsites (page 25). We have also added a typical representative DVH for a locally advanced oropharynx cancer treated with definitive chemoradiation (Figure 5), in response to Reviewer #1 Major comment #2 above.

4. **What about the total number of MU comparison wrt Clinical plan, how about the modulation factor (increased by round 2), overloading too much MU with highly modulated plan could be detrimental in terms of patient safety due to MLC leakage and transmission**

We appreciate this comment, and we agree that avoiding anomalously high MUs is good practice. We have further investigated the number of monitor units among the plans generated by AVI-planner and clinically treated plans. The MU and plan complexity were significantly lower for AVI-planner than the clinically treated plans. We have updated our materials and methods (formula page 9, paragraph 1; statistical analysis page 11, paragraph 2) and results (page 13, paragraph 3) to reflect this.

5. **Is this model transformable to the other clinic(?)**
   We would like to call the reviewer's attention to the Discussion (page 19, paragraph 3) where we acknowledge the barriers to widespread adoption of this software. Given our standardized approach to treatment planning among our clinics, we are currently working to implement the model in our other affiliate sites. Also noted within our Discussion (page 18, paragraph 1), following a prior reviewer's suggestion, the source code for our software has been made publicly available.

6. **Round 1 vs Round 2: First, is this actually a machine learning algorithm or something else? Not clear. However, if it is truly a machine learning algorithm than a different set of patients should have been used during the Round 2 trials to prevent overtraining on the same dataset.**
   We apologize for this confusion. As noted in our response to Reviewer #1, Major #1 and #2 (above) as well as Reviewer #2, Major #1 (below), our introduction (page 4, paragraph 1) clarifies our algorithm is not a machine learning approach. The algorithm used a statistical approach to assign optimization parameters based upon the distribution of DVH metrics within the 668 plan library, which were treated between 2014-2019 (page 5, paragraph 1). None of the 52 validation cohort plans were incorporated into the algorithm. We have updated our materials and methods (page 7, paragraph 1) to reflect this.

7. **Usage of failure mode analysis is a good touch, although Not much mention is made as to how medium and high risk failure modes were mitigated, depending on what these failure modes were, this might impact the quality of research and how much this application can be trusted.**
   We would like to call the reviewer's attention to Supplementary table 2 with additional details describing mitigation strategiess for these high risk failure modes. We have further elaborated in the results section (page 13, paragraph 1). The higher and lower relative risk compared to manual process was customized to our clinic in each scenario, and in this case, we felt in made sense to mitigate failure modes relative to manual planning.

**Minor issues:**

1. **Beyond the syntax, the language did not reach the expectations I had for an article meant to be published as well.**
   We have grammatically refined the materials and methods section to clarify our approach and flow.

2. **It lacked a good flow. It is important to note, the paper could be restructure with better flow.**

We appreciate this comment and have modified both the content and flow of the materials and methods section, which now more closely align with the sequencing of our results section.

3. **Pg 1 Line 7: radiation treatment (RT):**
   We would like to draw the reviewer's attention to page 1, paragraph 1 where we previously assigned an abbreviation for "radiation (RT)." We therefore defer revising this sentence.

4. **Pg 7 Line 44: Unclear, confusing sentence, highlights overall low level of writing**
   This sentence has been removed, and the materials and methods section has been modified as above to address Reviewer 1 Minor Issues #1 and #2.

5. **Pg 8 Line 12: Syntax (.:)**
   This mistake has been corrected. The content now reads, "*Script modifications were as follows:"*
6. 
7. **Pg 8 Line 15: and prioritizations and prioritizations.**
   The duplicative content has been removed.

8. **Pg 8 Line 41: of those dose to those (?)**
   This section now reads, "*Subvolumes of PTV and OAR overlap were transformed into standardized segmented structures, and new constraints and priorities were added to enhance the algorithm's ability to more precisely control dose."*

9. **Figure 5: unnecessary.**
   Figure 5 and the corresponding figure legend have been removed. The results text (page 15, paragraph 3) now reads: "shorter for AVI-planner vs manually optimized plans, 2 vs 85 minutes respectively ($p<0.01$)." This Figure 5 was replaced with a DVH in response to Reviewer #1, Major #2 above.

**Reviewer #2: Major comments**

1. **Introduction: I would suggest to summaries the unique features and advantages of the proposed AVI-planner, especially compared to existing KBP algorithms.**
   We thank the reviewer for this comment. Though there are commercially available knowledge based planning products like Varian's RapidPlan and Pinnacle's Auto-Planner, the focus of our work is a uniquely designed algorithm and iterative evaluation methodology to compare automated plans in the context of historical norms, literature thresholds and among HN disease subsites. We have further characterized the unique features of our algorithm, though we would advocate that many more knowledge based planning algorithms exist in addition to RapidPlan and Auto-Planning. Furthermore, given the differences in how our algorithm was built, these commercially available algorithms are not an appropriate benchmark for our comparison.
   This AVI-planner algorithm is not a machine learning algorithm. AVI-planner uses statistical analyses from previously treated plan DVHs to inform the optimization of new

HN RT plans. We have updated the introduction (page 4, paragraph 1) now reads, "*The algorithm is not a machine learning approach. This algorithm was designed using the same treatment planning system tools applied by dosimetrists during the manual process and integrates historical optimization norms from prior plans. The AVI-planner algorithm uniquely generates optimization parameters based upon statistical analyses of DVH metrics from previously treated HN RT plans."*

2. **Page 3, L45, please describe more details about iterative learning?**
This sentence now reads: "Iterative learning, *a process incorporating manually driven feedback into model training,* improves automated HN plan quality [35]."

3. **What do you mean by "fill these gaps in quantitative data and detail"?**
Thank you for this comment. To clarify our message, we removed this content. The sentence now reads: "Herein, we report the development of an automated virtual integrative (AVI) planning algorithm." To characterize AVI-planner more succinctly, we also summarize the features of this software to address Reviewer # 2 Major Comment #1 listed above.

4. **Page 5, L57, "The priority score for each failure mode (a version of the relative risk priority number from TG-100) was assigned as high, medium or low" This is not consistent with Supplementary Table 2, which has only higher and lower for the relative risk.**
We apologize for the confusion, but would like to call the reviewer's attention to the legend of Supplementary Table 2. Here, we describe the failure modes were considered relative to the manual treatment planning process. As noted above for Reviewer #1, Major # 7, the higher and lower relative risk compared to manual process was customized to our clinic in each scenario, and in this case, we felt in made sense to mitigate failure modes relative to manual planning. We have updated the results section (page 13, paragraph ), which now reads: "*None of these failure modes were higher relative risk compared to the manual treatment planning process."*

5. **It is not clear to me how the AVI-planner was constructed and how the previous HN plans were used?**
We have provided further details in the materials and methods (page 5, paragraphs 1-3) regarding generation of the AVI-planner algorithm.

6. **Page 6, L30, how were the refinements done if not interactive?**
We would like to call the reviewer's attention to the materials and methods section (page 9) "*Write-enabled Script Refinement and Clinical Deployment*" where we describe the refinements to the software.

7. **Page 14, L19, what changes were made? Is it patient-specific changes or for all the patients? More details are needed.**
We appreciate this feedback. To avoid duplicative content between Materials & Methods and Results, we discuss the specific software refinements in the Materials and Methods section. This Results (page 14) sentence now reads, "*Software refinements were made to*

*the script in response to the Round 1 evaluation which included normal tissue constraints and priorities, dose-sculpting structures, segmented structures, and isocenter placement as discussed above in Methods section Write-enabled Script Refinement and Clinical Deployment.*"

8. **For those cases that were identified as 'treat as is' during Round 1, did the round 2 replanning further improve the plan quality?**
The purpose for Round 1 evaluation was to uncover model planning deficiencies and limitations which would consistently require additional manual dosimetrist input to be considered acceptable by a physician. This Round 1 evaluation step would be necessary to assess knowledge based planning models as well. We would like to call the reviewer's attention to figure 2 where all "treat as is" Round 1 remained "treat as is" following Round 2 evaluation. None of these plans were felt at risk for clinically relevant toxicity or detrimental oncologic outcome after Round 2 planning.

9. **Page 17, what interactions were done for the two methods (Eclipse optimizer and AVI planner)**
We have further elaborated upon the interactive time (page 16, paragraph 3). The sentence now reads, "*interactive time included all steps of the manual optimization such as segmentation structures, setting isocenter, ring and buffer dose sculpting structures, normal tissue optimization limits, target and OAR prioritization, setting the # arcs, optimization time"*

10. **The results section is too wordy and hard to follow, especially for the evaluation of OAR sparing. It would be better to put the comparison results in a table and rephrase this section to be more concise.**
Thank you for this comment. Given the number of OARs for consideration among all subsites, we agree this is a bulky section of our results. We feel this is one of the more important sections of the mnauscript. In order to consolidate, we have removed the text describing a per-plan difference. Please see our response to Reviewer #2, Major #12 below regarding Figure 4.

11. **Figures 1 and 5 provide limited information. Both can be easily explained in the context.**
We agree with the reviewer. Therefore, the original Figure 5 and the corresponding figure legend have been removed. The results text (page 17, paragraph 1) now reads: "shorter for AVI-planner vs manually optimized plans, 2 vs 85 minutes respectively ($p<0.01$)." Figure 5 is now a representative DVH in response to Reviewer #1, Major #3 as above.
We feel that Figure 1 serves as a helpful reference to orient our audience to the complex development process, and therefore we will keep Figure 1 as previously submitted.

12. **Figure 4 is a very busy figure. I think a tabulated table would be better.**
We thank the reviewer for this thoughtful comment. We have considered multiple iterations of tables and figures to concisely and clearly illustrate our complex analysis.

We refer the reader to Figure 4's legend (page 25) to clarify the comparisons, but we defer transforming this into a table.

**Minor comments**

1. **Figure 3, the dose level for each isodose line should be listed.**
   This figure now includes appropriately labeled isodose lines. We would also like to call the reviewer's attention to the Figure 3 legend (page 26) where the isodose lines are defined.

2. **Page 8, line 15, 'and prioritizations- and prioritizations' is duplicated.**
   The duplicative content has been removed.

We truly appreciate the opportunity to revise our manuscript and hope that you find it acceptable for publication.

Sincerely,
Co-authors

# Development and Clinical Implementation of an Automated Virtual Integrative Planner for Radiation Therapy of Head and Neck Cancer

Running title: "AVI-planner for head and neck cancer radiation"

Elizabeth M. Jaworski, MD, MS[1], Michelle L. Mierzwa, MD[1], Karen A. Vineberg, MS[1], John Yao, PhD[1], Jennifer L. Shah, MD[1], Caitlin A. Schonewolf, MD, MS[1], Dale Litzenberg, PhD, Laila A. Gharzai, MD, LLM[1], Martha M. Matuszak, PhD[1], Kelly C. Paradis, PhD[1], Ashley Dougherty, CMD[1], Pamela Burger, CMD[1], Daniel Tatro, CMD[1], George Spencer Arnould, CMD[1], Jean M. Moran, PhD[1], Choonik Lee, PhD[1], Avraham Eisbruch, MD[1], Charles S. Mayo, PhD[1]

[1]Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan

Corresponding Author:

Charles Mayo, PhD

University Hospital, B2C432

1500 East Medical Center Dr.

Ann Arbor, MI 48109-5010

cmayo@med.umich.edu

734-232-3837

**Statistician:** Charles S. Mayo, PhD cmayo@med.umich.edu

**Funding statement:** Funding provided by Varian Medical Systems.

**Disclosure statement:** Authors disclose the following: grant funding from Varian Medical Systems (JMM, CSM).

**Data sharing statement:** Research data are stored in an institutional repository and will be shared upon request to the corresponding author.

**Acknowledgements:** The authors would also like to acknowledge Steven Kronenberg for his assistance with the creation of the figures for this manuscript.

## Abstract

**Purpose**: Head and neck (HN) radiation (RT) treatment planning is complex and resource intensive. Deviations and inconsistent plan quality significantly impact clinical outcomes. We sought to develop a novel automated virtual integrative (AVI) knowledge-based planning application to reduce planning time, increase consistency, and improve baseline quality.

**Materials and Methods:** An in-house write-enabled script was developed from a library of 668 previously treated HN RT plans. Prospective hazard analysis was performed, and mitigation strategies were implemented before clinical release. The AVI-planner software was retrospectively validated in a cohort of 52 recent HN cases. A physician panel evaluated planning limitations during initial deployment, and feedback was enacted via software refinements. A final second set of plans was generated and evaluated. Kolmogorov-Smirnov (KS) test in addition to Generalized Evaluation Metric (GEM) and Weighted Experience Score (WES) were used to compare normal tissue sparing between final AVI-planner versus respective clinically treated and historically accepted plans. One-tailed T-test was used to compare the interactive time required for AVI-planner versus manual optimization.

**Results:** Initially, 86% of plans were acceptable to treat with 10% minor and 4% major revisions or rejection recommended. Variability was noted in plan quality among HN subsites, with high initial quality for oropharynx and oral cavity plans. Plans needing revisions were comprised of sinonasal, nasopharynx, p-16 negative SCC Unknown Primary or cutaneous primary sites. Normal tissue sparing varied within subsites, but AVI-planner significantly lowered mean larynx dose (median 18.5 Gy vs 19.7 Gy, $p<0.01$) compared to clinical plans. AVI-planner significantly reduced interactive optimization time (mean 2 vs 85 minutes, $p<0.01$).

**Conclusions:** AVI-planner reliably generated clinically acceptable RT plans for oral cavity, salivary, oropharynx, larynx and hypopharynx cancers. Physician driven iterative learning processes resulted in favorable evolution in HN RT plan quality with significant time savings, and improved consistency using AVI-planner.

## Introduction

Radiation therapy (RT) is a cornerstone of HN cancer treatment. Intensity-modulated radiation therapy (IMRT) has improved treatment accuracy and reduced RT-associated morbidity [1-10]. HN IMRT manual optimization is resource-intensive and variable, with heavy reliance upon physician and facility expertise [11-16]. HN IMRT implementation has been met with frequent treatment planning and quality assurance (QA) deviations, which are associated with worse outcomes [17-19]. Furthermore, the time required for HN IMRT planning must be considered in the context of survival advantages associated with minimizing total treatment time and time interval between consultation and starting treatment [20, 21]. HN RT delivered at high-accruing centers is associated with improved outcomes, though factors including travel burden and patients' resources influence access to these centers [22-24].

Automated planning has been developed to standardize treatment planning, maximize efficiency, improve plan quality, and mitigate geographic disparities by increasing access to high quality RT plans [13, 25]. Knowledge-based planning (KBP) models rely upon dosimetric and geometric experience from dose-volume histograms (DVH) of previously treated acceptable plans [25]. KBP benefits have been documented in various disease sites, including HN [26-34]. Iterative learning improves automated HN plan quality [35]. However, commercially available KBP algorithms are limited by smaller training datasets, lack of standardized inputs, and challenging user-interface for plan revision. Prior studies have characterized plan quality in cohorts of HN patients without regard for primary site, while others report achievements in only one subsite (e.g. oropharynx [36] or nasopharynx [31]). There is a paucity of data regarding automated planning algorithm performance among different HN sites.

Herein, we report an evaluation method used to fill these gaps in quantitative data and detail among different HN sites as part of developing an automated virtual integrative (AVI) planning algorithm. We sought to create preliminary automated HN RT plans for "warm start optimization" where dosimetrists continue optimization from the automated plan instead of starting each plan with a new manual process [37]. The algorithm was designed to act as a virtual dosimetrist, using the same treatment planning system tools applied by dosimetrists during the manual process and integrating the knowledge of the extensive history of prior plans and clinical processes. We describe the iterative learning process to address planning deficiencies noted for select primary sites. To our knowledge, this is the first investigation of a HN-specific automated planning algorithm whereby the identification of site-specific clinically-significant deficiencies drive autoplanner script refinements to improve overall RT plan quality.

**Methods:**

*Script Development, Hazard Analysis, and Clinical Release*

Our script release process is shown in Figure 1. The write-enabled script was developed to incorporate practice norms defined by a library of 668 previously treated HN RT plans collected at our institution between 2014-2019. This library was comprised of 31.3% oropharynx (n=209), 19.3% oral cavity (n=129), 14.7% larynx (n=98), 7.9% cutaneous (n=53), 6.6% salivary (n=44), 4.2% sinonasal (n=28), 3.7% nasopharynx (n=25), 3% Unknown Primary (n=20), 2.7% hypopharynx (n=18), 1.8% thyroid (n=12), 0.6% orbital or lacrimal (n=4), 4.2% "other" (n=28). Software inputs were standardized including nomenclature and complete sets of contoured organs at risk/planning target volumes (OAR/PTVs) with explicitly defined planning priorities and objectives. Within the foundational library, >90% of plans contained spinal cord, brainstem, bilateral cochlea, parotids, superior and inferior pharyngeal constrictors, oral cavity, esophagus, mandible, lips. When surgically present and clinically relevant, bilateral submandibular glands (SMG) were included in 75%, larynx in 81%, bilateral optic nerves, chiasm, eyes, and lenses were included in 18-25%, while only 11% included lacrimal glands (data not shown).

Physics evaluation of automated plans to pass quality plan check was the same as our routine clinical practice, which was then followed by a second phase of clinical evaluation. Before clinical use of the AVI-planner, a prospective hazard analysis was performed using a streamlined failure mode and effects analysis described by Paradis et al. [38]. A process map for clinical use of the script was generated with associated hazards (failure modes) from multidisciplinary feedback. The priority score for each failure mode (a version of the relative risk priority number from TG-100) was assigned as high, medium or low [39]. All failure modes with high or medium priority scores were mitigated before proceeding on to clinical deployment. The

AVI-planner software automatically placed an isocenter, segmented optimization structures, and generated beams and plan setup with full calculation. All plans were VMAT, calculated in Eclipse version 15.6, with the analytical anisotropic algorithm (AAA), using 0.25 cm grid size. Eclipse Scripting Application Programming Interface (ESAPI) enabled the integration of AVI-planner software with Eclipse (Varian Medical System, Palo Alto, CA).

The clinical version of ESAPI does not allow programmatic interaction with the optimizer to enable algorithmic replication of the dynamic manual processes for monitoring and adjusting of DVH metric values and priorities during optimization. This functionality is enabled only in the non-clinical, research version of ESAPI. Since our objective was an application that could be used clinically with the Food and Drug Administration (FDA) approved versions of ESAPI, we developed an application interface and algorithms so that the initial set of constrains provided high quality plans across for a wide range of HN subsites, without dynamic programmatic interaction with the optimizer.

*Write-enabled Script Refinement*

A 2-step refinement process, integrating physician evaluation, was used to refine the optimization constraint algorithm. The first iteration ("Round 1") reflected the first order "explicit" objectives for target and OAR coverage reflected in prescription document and the history of previously treated HN RT plans. The second refinement ("Round 2") used the plan evaluation context from Round 1 to reveal "implicit" second order objectives and priorities reflecting physician preferences for dose distribution details that may not be part of the prescription document.

In Round 1, a template of optimization constraints was defined from quantile analysis of the distribution specific structure-DVH metrics measured from the distribution values in our library of previously treated patients. Optimization constraints were set at the thresholds corresponding to the lower 30% of historic values. For structures evaluated based on Mean[Gy], optimization constraints corresponding to the lower 30% of historic values for D90%[Gy], D50%[Gy] and D10%[Gy] were used to enable differentially prioritizing portions receiving high dose due to proximity to PTV volumes (D10%[Gy]) vs portions which may receive a lower dose (D90%[Gy]).

The algorithm automates creation of structures used during optimization to control the shape of dose distributions. Clinical target structures (e.g. GTVs, CTVs, PTVs) and organs at risk (e.g. parotids, submandibular glands) are not modified by the algorithm. Optimization structures are created from unmodified clinical structures using the same operations used in manual creation of these structures (e.g. margins, Boolean operations) are used, but standardized by the algorithm so that all plans are consistent. Optimization structures created include sub-volumes of target volumes based on overlap with margins to enable specificity of compromise in clinician directed trade-offs between target coverage and OAR sparing. Dose sculpting structures included rings were created around the structures and used to use the normal tissue objective to conform prescription isodose lines to the corresponding PTV volumes. Optimization structures segmented high dose PTV sub volumes out of larger, lower dose PTV volumes.

In development cycles preceding the physician team evaluation in Round 1, the algorithm was iteratively optimized on the subsets of HN patients by the development team of physics, dosimetry and software developer. Prior to physician inspection, all parameters in the application were fixed and plans were generated for all patients in the cohort for physicians to inspect

"Round 1" planning used the validation cohort described below. "Round 1" indicates optimization via the initially released AVI-planner script with minor manual edits. Following Round 1 evaluation, physician feedback was incorporated via software modifications in three categories.:

1) Normal tissue optimization constraints and prioritizations- and prioritizations – Mapping between priorities listed in prescription objectives identified in the user interface and optimization constraint priorities used during the optimization were refined to improve alignment with physician's intent.Dose values for optimization constraints may be adjusted.

2) Dose-sculpting structures- Shaping of the dose distribution in areas not delineated by organs at risk are controlled by first creating dose sculpting structures (e.g. rings, buffers between PTVs, etc) with defined spatial relationships between targets and organs at risk, and second controlling the dose distributions in those structures with optimization constraints.

3) Segmentation of areas with PTV and OAR overlap- adding standardized structures with known spatial relationships between specific PTV and OAR volumes provides finer control of those dose to those areas in the optimizer.

*Patient Selection*

This study was IRB exempt (HUM 00126332) for quality improvement. AVI-planner in Round 1 optimization was retrospectively validated within a cohort of 52 HN cancer patients treated between 2019-2020. We included oral cavity, oropharynx, larynx, hypopharynx, cutaneous, sinonasal, and salivary primaries to account for anatomy and OARs, adjuvant vs definitive RT, target dose, and fractionation. Institutional dose-escalation or de-escalation

protocol patients were included. We excluded hypofractionated and palliative patients. Simulation CT scans were performed on a Philips Brilliance big-bore 16 slice scanner (Koninklijke Philips N.V., Amsterdam, Netherlands) using 3 mm slices. Patients were scanned head-first, supine with IV contrast and immobilized in 5-point thermoplastic masks. Intact and postoperative boost and elective CTV contours were delineated referencing published guidelines [40, 41] with a 3 mm PTV margin. Dosimetrists manually optimized clinical plans using Eclipse (Varian Medical System, Palo Alto, CA), which were delivered on Varian TrueBeam or Clinac linear accelerators with 120 leaf MLC using 6-MV photons with 2-4 VMAT arcs.

*Plan Evaluation and Statistical Analysis*

Clinical plans underwent peer-review by a subspecialty panel of attending radiation oncologists. Institutional protocols specified prioritization of target coverage and objectives for OAR sparing (Supplementary Table 1). To identify AVI-planner limitations consistently requiring additional manual input for "warm start optimization," the physician panel evaluated clinical acceptability of "Round 1" AVI-planner cases. HN subsites were grouped by treatment paradigm and anatomic proximity. These plans were "rejected" if the plan was unsafe and unsalvageable despite reoptimization. "Major revisions" indicated a high perceived risk of either 1) a clinically relevant toxicity due to exceeded OAR constraints or 2) risk of recurrence from target under-coverage. Plans with "minor revisions" were safe with room for improvement in conformality, heterogeneity, or target coverage. The highest quality plans were deemed "treat as is." Physician feedback from Round 1 was addressed per "Write-enabled Script Model Refinement." All 52 cases were then replanned with the AVI-planner script without manual modifications and labeled "Round 2." The same physician panel re-evaluated all 52 plans.

Beyond stand-alone clinical acceptability, Round 2 AVI-planner quality was compared to 1) clinically treated plans 2) historically accepted plans and 3) literature-based thresholds [42]. Clinically treated plan denotes the patient-specific RT plan, which was delivered during the patient's treatment course. Within this context, evaluating Round 2 versus the clinically treated plan provides an individual, patient-level comparison of plan quality. Comparisons to historically accepted plans were based on summarized metrics captured from the entire 668 HN foundational library. Thus, Round 2 plan quality was assessed in the context of aggregate institutional experience with all 668 considered high-quality HN plans. Evaluating Round 2 plans in both situations more fully characterizes plan quality at both the patient-level and institutional experience- level. We used Python 3.8 statistical software for this analysis. A one-sided, Kolmogorov-Smirnov (KS) test was used to determine if the distribution of AVI-planner OAR mean or D0.1cc values was higher or lower than clinically treated plans of the validation cohort. The distribution of per plan differences was analyzed. AVI-planner values were compared to literature based thresholds [42] using a normal distribution, with matched cardinality, centered on each threshold with a 0.5 Gy standard deviation as the reference distribution using a t-test for mean value difference.

To compare AVI-planner to historic plans, constraint metrics within the algorithm were derived from 668 previously treated plans using the previously described Generalized Evaluation Metric (GEM) and Weighted Experience Score (WES) described by XXX et al. [43]. GEM compares DVH metrics to constraints and historical values, which are cast onto a sigmoidal curve with scale of 0 to 1, where GEM = 0.5 if the constraint was met and 0.95 when 95% of historical values were lower than the current plan's value. WES ranks the DVH curves with respect to historical values, on a 0 to 1 scale, with values weighted according to historic

variability. WES correlates with NTCP but rises sooner with respect to dose, correlating with physician preferences to drive doses below NTCP thresholds.

VRxGy[%] was used to assess coverage at the prescribed dose for each dose level. The ICRU Conformality index (CI) [26, 44] was calculated for PTV_High, PTV_Low and PTV_Mid00 volume

$$CI_{ICRU} = \frac{Body{:}\,VRx[cc]}{PTV{:}\,Volume[cc]}$$

Dose heterogeneity within PTV volumes was assessed using ICRU 83 $HI_1$ [45].

$$HI_1 = \frac{(D2\%[Gy] - D98\%[Gy])}{D50\%[Gy]}$$

These were calculated for the PTV subvolumes, not overlapping with volumes at prescribed doses as PTV_High, PTV!_Low and PTV!_Mid00 in TG-263 nomenclature.

Descriptive statistics were utilized to evaluate dosimetric endpoints and time required for treatment planning. One-tailed T-test was used to compare the interactive time required for AVI-planning versus manual planning.

After the evaluation process described, AVI-planner was deployed to the clinic. In our clinic this deployment is done in a staged process, with testing carried out first by the chief dosimetrist and another dosimetrist. The initial use of the application is carefully monitored by the physicians and physicists stakeholders in the initial limited clinical release for proper functioning and introduction of hazards. After validated in this monitored process, the script was

then deployed without change and is in routine use in the clinic. Requests for additional improvements and features are monitored and incorporated into future development cycles.

## Results

*Failure Mode and Effects Analysis (FMEA)*

Before clinical deployment, 12 failure modes were identified relating to contour generation (7), plan creation (1), treatment field generation (2), plan optimization (1) and plan approval (1) (Supplementary Table 2). Five were higher relative risk. Failure modes and associated mitigations are shown in Supplementary Table 2. Several code modifications were prompted by FMEA. These included detailed analysis of structure volumes at the beginning and end of the algorithm to identify changes made, checks at entry of the script algorithm that PTV and organ at risk volumes are approved and cannon be edited, enforcement of naming conventions for structure sets, course and plans to minimize risk of unintentional use of an automated plan.

*Validation and Clinical Implementation: Round 1*

We retrospectively validated AVI-planner in 52 patients, which consisted of mostly men (69%) with locally advanced oropharynx (40%) or oral cavity and salivary (31%) cancers (Figure 2, Supplementary Table 3). Definitive intent organ-preservation RT comprised the majority of plans (58%) with a median of 70 Gy (range 54-80 Gy) in 35 fractions (range 27-35 fx). 62% received concurrent chemotherapy.

Overall, 86% of Round 1 plans were safe to treat; however, we identified variability in plan quality among different HN subsites (Figure 2). All oropharynx and p16+ SCC Unknown Primary (21/21 plans), larynx and hypopharynx (7/7 plans) were "treat as is." Similarly, most oral cavity and salivary cases (14/16 plans; 87.5%) required no revisions. This contrasts with the frequency of major revisions or rejections recommended for sinonasal, nasopharynx and p16-

negative SCC Unknown Primary (1/6 plans; 17%) and cutaneous (1/2 plans; 50%- Figure 2). Minor revisions were increasing conformality and reducing heterogeneity. Major revisions were limiting hot spots outside PTV, restricting hot spots within PTV to 105-110%, and improving target coverage. Sample Round 1 isodose distributions are shown for both a definitive early stage p16+ base of tongue cancer considered "treat as is" (Figure 3B), compared to "major revisions" for an adjuvantly treated malar cheek Merkel cell carcinoma (Figure 3E).Changes were made to the script in response to the Round 1 evaluation. These included addition to the start panel of the user interface of structure volumes and volume fraction of overlap with PTV structures. This supplemental information aids in customizing relative prioritization DVH metrics for structures. Buffer structures to control low and intermediate dose levels between PTVs and in the posterior neck region were modified. Detection of atypical target locations and shifting from our standard isocenter location algorithm isocenter to optimize use of the central 0.5 cm MLC leaves was added in response to the one plan rejected the second plan requiring major revision in Round 1. For those 2 cases the PTV volumes were ipsilateral, involving the skull and upper neck. To shorten development time in refining algorithm performance, parameters were placed in an external configuration file, to significantly limit the scope of changes that require recompiling the code. The change to use of planning parameters configuration file also supports eventual extension of the script to other clinics enabling facilitating local customizations.

*Clinical Reassessment: Round 2*

During Round 2 evaluation of all 52 plans, there were no rejections nor major revisions (Figure 2). Minor revisions were recommended for 1 oral cavity (6.3%) and 3 sinonasal or nasopharynx or p16-negative SCC Unknown Primary plans (50%). The remaining 48 plans were

"treat as is." Minor revisions in Round 2 focused on improving conformality, or more aggressive sparing of spinal cord, optics and contralateral orbit or salivary structures. This evolution in quality is evident for the adjuvantly treated Merkel cell carcinoma (Figure 3D-F).

Conformality and dose heterogeneity for all PTV levels were similar between Round 2 AVI-planner and clinical cases (Table 1). To evaluate patterns of OAR sparing in Round 2 AVI-planner, we compared the entire distribution of mean or D0.1cc dose values between AVI-planner cases versus clinically treated or historically accepted plan values (Figure 4A). Considering all 52 plans, the contralateral parotid dose distribution was higher with AVI-planner compared to clinically treated plans (median 25 vs 23 Gy, $p<0.01$) with a difference of 1.9 Gy per plan, with higher doses compared to historic plans (WES 0.50 vs 0.42, $p<0.01$). Similarly, inferior pharyngeal constrictor muscles had higher distribution of mean dose in AVI-planner versus clinical plans (median 21 Gy vs 19 Gy, $p<0.01$) with a 1 Gy per plan difference. Compared to historically accepted values, the dose to inferior constrictors was higher (WES 0.53 vs 0.35, $p<0.01$), and narrowly exceeded our constraint (GEM 0.52). Conversely, AVI-planner lowered the dose to ipsilateral SMG (62 Gy vs 65 Gy, $p=0.04$) though this was not clinically relevant (GEM >0.90) [43]. AVI-planner lowered the distribution of mean dose to the larynx compared to clinical (median 19 Gy vs 20 Gy, $p<0.01$) and historical plans (WES 0.29 vs 0.44, $p<0.01$) with a per plan difference of 1.3 Gy. Brainstem D0.1cc from AVI-planner was lower than clinical plans (median 28 Gy vs 32 Gy, $p<0.01$) with a per-plan difference of 5 Gy. Spinal cord distribution of D0.1cc was lower in AVI-planner as compared to historic plans (WES 0.44 vs 0.63, $p<0.01$), but similar to clinical plans (median 36.4 Gy vs 36.7 Gy, $p=0.9$). Distribution of dose to optic nerves, chiasm, eyes, contralateral SMG, superior pharyngeal constrictors, oral

cavity, mandible and esophagus were similar among AVI-planner cases, clinical and historic plans (Figure 4A).

For oropharynx or p16+ SCC Unknown Primary (n=21; Figure 4B), the distribution of mean larynx dose was significantly lower with AVI-planner versus clinical plans (median 18 vs 20 Gy, $p<0.01$) or historical plans (WES 0.28 vs 0.44, $p<0.01$), which was clinically relevant (GEM 0.46). Esophagus and ipsilateral parotid were spared equally among AVI-planner, clinical and historic plans. Distribution of mean dose to contralateral SMG was higher for AVI-planner compared to clinical plans (median 36 Gy vs 33 Gy, $p=0.02$) and historical plans (WES 0.44 vs 0.41, $p=0.02$), corresponding to 0.13 Gy per plan difference. Neither clinical nor AVI-planner met constraints for relevant sparing (GEM 0.62 and 0.57). Contralateral parotid distribution of mean dose was higher for AVI-planner compared to clinical plans (median 25 vs 23 Gy, $p=0.047$), but similar to historically accepted plans (WES 0.50 vs 0.43, $p=0.1$). Superior and inferior pharyngeal constrictors received higher dose with AVI-planner and exceeded constraints compared to historical controls ($p<0.01$). Oral cavity distribution of mean dose was higher with AVI-planner versus clinical (37 vs 33 Gy, $p=0.04$) and historic plans (WES 0.64 vs 0.53, $p=0.04$; Figure 4B).

OAR doses were similar between clinical and Round 2 AVI-planner for the remaining HN subsites (Figure 4B and 4C). Of note, the oral cavity/salivary contralateral parotid distribution of mean dose was significantly higher for AVI-planner compared to clinical plans (median 26 Gy vs 23 Gy, $p<0.01$), and historic plans (WES 0.53 vs 0.42, $p<0.01$) and did not meet constraints (GEM 0.56) (Figure 4B). Two cutaneous plans did not reach the 3 plan threshold required for formal comparison.

*Time Study*

We compared Eclipse optimizer interactive time for 10 recent manual plans versus interactive time with AVI-planner software for 51 of the validation cohort patients. Of the 10 manually optimized plans, 70% were oropharynx (n=7), while 30% were comprised of oral cavity (n=1), thyroid (n=1), and Unknown Primary (n=1). Mean time for manual interaction time was shorter for AVI-planner vs manually optimized plans, 2 vs 85 minutes respectively ($p<0.01$; Figure 5).

## Discussion

We developed and implemented a knowledge-based automated virtual integrative software to facilitate HN treatment planning. Initially, we identified inconsistent plan quality among different HN subsites. Following iterative software adaptations, we noted favorable evolution in target coverage, heterogeneity and OAR sparing. This software exceeded our "warm start optimization" goal and rapidly created clinically acceptable plans without manual adjustments for many HN subsites. We have published the source code for AVI-Planner at a GitHub repository (http://xxxx.xxxx.xxxx ) to promote use and development of automated planning.

Regarding clinical acceptability of automated HN plans, we found 86% of Round 1 plans were treat as is, which is comparable to 88% by Radiation Planning Assistant [46]. Our script was developed from a diverse training dataset, capturing unique nuances and planning considerations. The inconsistent site-specific plan quality likely resulted from limited experience within the foundational library. Our heterogeneous library accrued over 5 years, but there were fewer cutaneous (7.9%), sinonasal (4.2%) and nasopharynx (3.7%) plans. Improvements in both conformality and heterogeneity were shown for prostate and cervix cancer after refining Varian's RapidPlan default settings [47], but studies detailing specific software refinements and evolution in plan quality among multiple subsites are limited for HN [35, 46].

Physicians frequently emphasized higher OAR prioritization. For instance, the clinical plan aggressively spared contralateral parotid further below the planning objective in a cT1N1 p16+ tonsil cancer, whereas AVI-planner less aggressively spared the contralateral parotid to meet constraints. Given OAR constraint heterogeneity, we compared the Round 2 AVI-planner results to consensus thresholds [42]. Structure laterality is reported, however the distinction of

ipsilateral or contralateral relative to the target is less readily available. AVI-planner achieved lower contralateral parotid doses (median 25 Gy) than the 26 Gy threshold ($p < 0.01$). AVI-planner lowered dose to optic structures, eyes, brainstem, spinal cord, esophagus, inferior constrictors and mandible compared to accepted thresholds. Larynx doses achieved in AVI-planner cases were lower than thresholds (35 Gy, $p<0.01$), and values with automation approaches by Fogliata et al. (24.8 ± 5.2 Gy) or Ouyang et al. (28.4 Gy) [26, 27]. This highlights the importance of benchmarking automation achievements against literature values, historic norms, and the validation subset.

Secondarily, the failure modes addressed during development can be found in manual planning, suggesting these hazards already exist and may be more likely to happen without the software. Thus, automated planning does not obviate standard clinical and physics QA. Similar to Wang et al. [48], inconsistencies in standardized OAR prioritization affected the performance of our model. In line with time-savings noted by other groups for autoplanned nasopharynx [31, 49, 50] and oropharynx cancers [33, 36], we confirmed time savings with AVI-planner.

Limitations of this work include this software is integrated only with Eclipse for 30-35 fraction plans. Given the revisions required for sinonasal, nasopharynx, and cutaneous sites, this software should be used cautiously near the skull base. Three target dose levels are currently supported, but additional dose levels require manual editing. Dosimetrists must also ensure the relevance of automated decisions. For instance, planners must remain vigilant about modifying the isocenter location or number of arcs for a unilateral target. Physicians must explicitly address planning preferences in the planning directive. For example, in a locally advanced maxillary sinus cancer requiring adjuvant RT following an orbital exenteration, aggressively sparing the remaining contralateral orbit and lacrimal gland may take precedence over PTV coverage. The

user-friendly interface contains the same tools used in manual optimization, allowing real-time modification by dosimetrists, compared to fully automated optimization which must run to completion before permitting revision. However, these standardized optimization parameters likely differ from personalized approaches of experienced dosimetrists. Therefore, additional time may be required for revisions. Inter-institutional heterogeneity in delineated OARs, inconsistent OAR contouring, and variability in constraints and prioritization are barriers to widespread adoption of automated planning.

To our knowledge, this is the first report identifying HN primary site-specific variability in automated plan quality, which favorably evolved with physician input. Our work cautions against interpreting that automated planning achievements are universal among HN subsites. This is relevant for clinics that would ideally employ one planning algorithm for all HN cases, instead of separate optimization algorithms for each HN subsite. We are not advocating this software in lieu of skilled dosimetrists or treatment at high-accruing centers. However, in settings of limited resources, increased demand, urgent starts or reduced subspecialized dosimetrists, AVI-planner software can be easily integrated into workflows to increase availability of high quality HN RT plans.

Furthermore, our institutional adoption of AVI-planner into routine practice has expanded the number of dosimetrists able to rapidly generate high quality HN plans. We plan to release AVI-planner to our affiliate sites to improve plan quality and uniformity, we are also extending this standardized approach to other disease sites including lung and prostate. In the future, automated planning will facilitate adaptive RT planning. Future software upgrades may incorporate gEUD, update the foundational library, and focus on well-lateralized cases near the skin surface. Application programming interfaces that enable clinics to programmatically

automate all parts of the treatment planning process, give clinics the tools they need to increase efficiency and consistency in plan quality in their process workflows. To promote these clinical improvements it is highly desirable for manufactures to provide application programming interfaces that give users at minimum the same capabilities they have in manual operations to algorithmically interact with optimizers, dose calculation engines, reference points, and scheduling capabilities.

## References


1. Hawkins, P.G., et al., *Organ-Sparing in Radiotherapy for Head-and-Neck Cancer: Improving Quality of Life.* Semin Radiat Oncol, 2018. **28**(1): p. 46-52.
2. Nutting, C.M., et al., *Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial.* Lancet Oncol, 2011. **12**(2): p. 127-36.
3. Pow, E.H., et al., *Xerostomia and quality of life after intensity-modulated radiotherapy vs. conventional radiotherapy for early-stage nasopharyngeal carcinoma: initial report on a randomized controlled clinical trial.* Int J Radiat Oncol Biol Phys, 2006. **66**(4): p. 981-91.
4. Kam, M.K., et al., *Prospective randomized study of intensity-modulated radiotherapy on salivary gland function in early-stage nasopharyngeal carcinoma patients.* J Clin Oncol, 2007. **25**(31): p. 4873-9.
5. Dirix, P. and S. Nuyts, *Evidence-based organ-sparing radiotherapy in head and neck cancer.* Lancet Oncol, 2010. **11**(1): p. 85-91.
6. Lee, N., et al., *Intensity-modulated radiation therapy in head and neck cancers: an update.* Head Neck, 2007. **29**(4): p. 387-400.
7. Eisbruch, A., et al., *Parotid gland sparing in patients undergoing bilateral head and neck irradiation: techniques and early results.* Int J Radiat Oncol Biol Phys, 1996. **36**(2): p. 469-80.
8. Murdoch-Kinch, C.A., et al., *Dose-effect relationships for the submandibular salivary glands and implications for their sparing by intensity modulated radiotherapy.* Int J Radiat Oncol Biol Phys, 2008. **72**(2): p. 373-82.
9. Feng, F.Y., et al., *Intensity-modulated chemoradiotherapy aiming to reduce dysphagia in patients with oropharyngeal cancer: clinical and functional results.* J Clin Oncol, 2010. **28**(16): p. 2732-8.
10. Beadle, B.M., et al., *Improved survival using intensity-modulated radiation therapy in head and neck cancers: a SEER-Medicare analysis.* Cancer, 2014. **120**(5): p. 702-10.
11. Boero, I.J., et al., *Importance of Radiation Oncologist Experience Among Patients With Head-and-Neck Cancer Treated With Intensity-Modulated Radiation Therapy.* J Clin Oncol, 2016. **34**(7): p. 684-90.
12. Lee, C.C., et al., *Survival rate in nasopharyngeal carcinoma improved by high caseload volume: a nationwide population-based study in Taiwan.* Radiat Oncol, 2011. **6**: p. 92.
13. Cilla, S., et al., *Template-based automation of treatment planning in advanced radiotherapy: a comprehensive dosimetric and clinical evaluation.* Sci Rep, 2020. **10**(1): p. 423.
14. Batumalai, V., et al., *How important is dosimetrist experience for intensity modulated radiation therapy? A comparative analysis of a head and neck case.* Pract Radiat Oncol, 2013. **3**(3): p. e99-e106.
15. Moore, K.L., et al., *Experience-based quality control of clinical intensity-modulated radiotherapy planning.* Int J Radiat Oncol Biol Phys, 2011. **81**(2): p. 545-51.
16. Nelms, B.E., et al., *Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems.* Pract Radiat Oncol, 2012. **2**(4): p. 296-305.

17. Eisbruch, A., et al., *Multi-institutional trial of accelerated hypofractionated intensity-modulated radiation therapy for early-stage oropharyngeal cancer (RTOG 00-22).* Int J Radiat Oncol Biol Phys, 2010. **76**(5): p. 1333-8.
18. Zhong, H., et al., *The Impact of Clinical Trial Quality Assurance on Outcome in Head and Neck Radiotherapy Treatment.* Front Oncol, 2019. **9**: p. 792.
19. Peters, L.J., et al., *Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02.* J Clin Oncol, 2010. **28**(18): p. 2996-3001.
20. Graboyes, E.M., et al., *Association of Treatment Delays With Survival for Patients With Head and Neck Cancer: A Systematic Review.* JAMA Otolaryngol Head Neck Surg, 2019. **145**(2): p. 166-177.
21. Rosenthal, D.I., et al., *Importance of the treatment package time in surgery and postoperative radiation therapy for squamous carcinoma of the head and neck.* Head Neck, 2002. **24**(2): p. 115-26.
22. Wuthrick, E.J., et al., *Institutional clinical trial accrual volume and survival of patients with head and neck cancer.* J Clin Oncol, 2015. **33**(2): p. 156-64.
23. Naghavi, A.O., et al., *Patient choice for high-volume center radiation impacts head and neck cancer outcome.* Cancer Med, 2018. **7**(10): p. 4964-4979.
24. George, J.R., S.S. Yom, and S.J. Wang, *Combined modality treatment outcomes for head and neck cancer: comparison of postoperative radiation therapy at academic vs nonacademic medical centers.* JAMA Otolaryngol Head Neck Surg, 2013. **139**(11): p. 1118-26.
25. Hussein, M., et al., *Automation in intensity modulated radiotherapy treatment planning-a review of recent innovations.* Br J Radiol, 2018. **91**(1092): p. 20180270.
26. Ouyang, Z., et al., *Evaluation of auto-planning in IMRT and VMAT for head and neck cancer.* J Appl Clin Med Phys, 2019. **20**(7): p. 39-47.
27. Fogliata, A., et al., *RapidPlan head and neck model: the objectives and possible clinical benefit.* Radiat Oncol, 2017. **12**(1): p. 73.
28. Krayenbuehl, J., et al., *Evaluation of an automated knowledge based treatment planning system for head and neck.* Radiat Oncol, 2015. **10**: p. 226.
29. Tol, J.P., et al., *Evaluation of a knowledge-based planning solution for head and neck cancer.* Int J Radiat Oncol Biol Phys, 2015. **91**(3): p. 612-20.
30. Gintz, D., et al., *Initial evaluation of automated treatment planning software.* J Appl Clin Med Phys, 2016. **17**(3): p. 331-346.
31. Giaddui, T., et al., *Offline Quality Assurance for Intensity Modulated Radiation Therapy Treatment Plans for NRG-HN001 Head and Neck Clinical Trial Using Knowledge-Based Planning.* Adv Radiat Oncol, 2020. **5**(6): p. 1342-1349.
32. Hansen, C.R., et al., *Automatic treatment planning improves the clinical quality of head and neck cancer treatment plans.* Clin Transl Radiat Oncol, 2016. **1**: p. 2-8.
33. Kusters, J., et al., *Automated IMRT planning in Pinnacle : A study in head-and-neck cancer.* Strahlenther Onkol, 2017. **193**(12): p. 1031-1038.
34. Krayenbuehl, J., et al., *Planning comparison of five automated treatment planning solutions for locally advanced head and neck cancer.* Radiat Oncol, 2018. **13**(1): p. 170.
35. Fogliata, A., et al., *RapidPlan knowledge based planning: iterative learning process and model ability to steer planning strategies.* Radiat Oncol, 2019. **14**(1): p. 187.

36. Kamima, T., et al., *Multi-institutional evaluation of knowledge-based planning performance of volumetric modulated arc therapy (VMAT) for head and neck cancer.* Phys Med, 2019. **64**: p. 174-181.
37. Ahunbay, E.E., O. Ates, and X.A. Li, *An online replanning method using warm start optimization and aperture morphing for flattening-filter-free beams.* Med Phys, 2016. **43**(8): p. 4575.
38. Paradis, K.C., et al., *The Fusion of Incident Learning and Failure Mode and Effects Analysis for Data-Driven Patient Safety Improvements.* Pract Radiat Oncol, 2020.
39. Huq, M.S., et al., *The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation therapy quality management.* Med Phys, 2016. **43**(7): p. 4209.
40. Biau, J., et al., *Selection of lymph node target volumes for definitive head and neck radiation therapy: a 2019 Update.* Radiother Oncol, 2019. **134**: p. 1-9.
41. Gregoire, V., et al., *Delineation of the primary tumour Clinical Target Volumes (CTV-P) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO, TROG consensus guidelines.* Radiother Oncol, 2018. **126**(1): p. 3-24.
42. Lee, A.W., et al., *International Guideline on Dose Prioritization and Acceptance Criteria in Radiation Therapy Planning for Nasopharyngeal Carcinoma.* Int J Radiat Oncol Biol Phys, 2019. **105**(3): p. 567-580.
43. XXX.
44. Baltas, D., et al., *A conformal index (COIN) to evaluate implant quality and dose specification in brachytherapy.* Int J Radiat Oncol Biol Phys, 1998. **40**(2): p. 515-24.
45. *ICRU report 83 Prescribing, recording, and reporting photon-beam intensity-modulated radiation therapy (IMRT).* J ICRU 10:35–36, 2010.
46. Olanrewaju, A., et al., *Clinical Acceptability of Automated Radiation Treatment Planning for Head and Neck Cancer Using the Radiation Planning Assistant.* Pract Radiat Oncol, 2021. **11**(3): p. 177-184.
47. Hussein, M., et al., *Clinical validation and benchmarking of knowledge-based IMRT and VMAT treatment planning in pelvic anatomy.* Radiother Oncol, 2016. **120**(3): p. 473-479.
48. Wang, Y., B.J.M. Heijmen, and S.F. Petit, *Knowledge-based dose prediction models for head and neck cancer are strongly affected by interorgan dependency and dataset inconsistency.* Med Phys, 2019. **46**(2): p. 934-943.
49. Chang, A.T.Y., et al., *Comparison of Planning Quality and Efficiency Between Conventional and Knowledge-based Algorithms in Nasopharyngeal Cancer Patients Using Intensity Modulated Radiation Therapy.* Int J Radiat Oncol Biol Phys, 2016. **95**(3): p. 981-990.
50. Hu, J., et al., *Quantitative Comparison of Knowledge-Based and Manual Intensity Modulated Radiation Therapy Planning for Nasopharyngeal Carcinoma.* Front Oncol, 2020. **10**: p. 551763.

**Figure Legends**

Figure 1: Flow chart depicting write-enabled script development and release process.

Figure 2. Integrated pie and doughnut charts demonstrating clinical acceptability of AVI-planned cases among H&N subsites. Central chart (blue) shows plan frequency by subsite (n=52). Innermost doughnut chart shows "Round 1" clinical acceptability. Outermost doughnut chart "Round 2" shows evolution in acceptability for 7 plans initially requiring revisions.

Figure 3. Sample HN RT plans for early-stage oropharynx and adjuvant Merkel cell carcinoma. Cases include stage II cT2N2M0 p16+ squamous cell carcinoma of the right base of tongue treated with definitive chemoradiation to 70 Gy in 35 fx (A-C); and stage III pT1 pN1a(sn) Merkel Cell carcinoma treated adjuvantly to 54 Gy in 30 fx (D-F). Left panels (A,D) Clinically Treated Plan. Middle panels (B) Round 1 AVI-planner "treat as is;" (E) Round 1 AVI-planner "major revision" due to conformality and 119% hotspot outside PTV. Right panels (C, F) Round 2 following AVI-planner upgrades. Isodose lines (absolute dose, Gy) show 75 (light green), 70 (white), 65 (pink), 60 (red), 54 (yellow), 51 (green), 45 (orange), 40 (purple), 30 (cyan), 25 (green), 20 (dark blue).

Figure 4: Differential normal tissue sparing by clinical versus AVI-planner. Comparisons for all HN subsites (A); oropharynx, p16+ SCC Unknown Primary, Oral Cavity, Salivary, Larynx, Hypopharynx (B); Sinonasal, Nasopharynx, and p16-negative SCC Unknown Primary (C). OARs are listed on y-axis with corresponding metric. Dose (Gy) on x-axis. Box plots of clinical (blue) or Round 2 AVI-planner (yellow) provide median, IQR, minimum, and maximum doses. Red "x" denotes consensus thresholds [42]. Statistical significance was achieved with $p<0.05$ on one-sided Kolmogrorov-Smirnov test. Symbols along y-axis indicate statistically significant

difference in OAR sparing between AVI-planner versus clinical plans: total cohort (panel A-circles), Oropharynx and p16+ SCC Unknown Primary (panel B-stars), Oral Cavity and Salivary (panel B- diamonds). Filled shapes indicate AVI-planner significantly improved sparing whereas unshaded symbols indicate clinical plan achieved significantly better sparing.

Figure 5: Box and whisker plot comparing dosimetrist interactive time between AVI-planner and manual optimization. Significantly less time required for Round 2 AVI-planner (n=51) versus manual optimization (n=10). * indicates $p<0.01$

Table 1. Conformality and heterogeneity (ICRU 83) indexes of clinical and Round 2 AVI-planner for high, intermediate and low PTV.

## Abstract

**Purpose**: Head and neck (HN) radiation (RT) treatment planning is complex and resource intensive. Deviations and inconsistent plan quality significantly impact clinical outcomes. We sought to develop a novel automated virtual integrative (AVI) knowledge-based planning application to reduce planning time, increase consistency, and improve baseline quality.

**Materials and Methods:** An in-house write-enabled script was developed from a library of 668 previously treated HN RT plans. Prospective hazard analysis was performed, and mitigation strategies were implemented before clinical release. The AVI-planner software was retrospectively validated in a cohort of 52 recent HN cases. A physician panel evaluated planning limitations during initial deployment, and feedback was enacted via software refinements. A final second set of plans was generated and evaluated. Kolmogorov-Smirnov (KS) test in addition to Generalized Evaluation Metric (GEM) and Weighted Experience Score (WES) were used to compare normal tissue sparing between final AVI-planner versus respective clinically treated and historically accepted plans. ~~One-tailed~~ T-test was used to compare the interactive time, complexity, and monitor units ~~required~~ for AVI-planner versus manual optimization.

**Results:** Initially, 86% of plans were acceptable to treat with 10% minor and 4% major revisions or rejection recommended. Variability was noted in plan quality among HN subsites, with high initial quality for oropharynx and oral cavity plans. Plans needing revisions were comprised of sinonasal, nasopharynx, p-16 negative SCC Unknown Primary or cutaneous primary sites. Normal tissue sparing varied within subsites, but AVI-planner significantly lowered mean larynx dose (median 18.5 Gy vs 19.7 Gy, $p<0.01$) compared to clinical plans. AVI-planner significantly reduced interactive optimization time (mean 2 vs 85 minutes, $p<0.01$).

**Conclusions:** AVI-planner reliably generated clinically acceptable RT plans for oral cavity, salivary, oropharynx, larynx and hypopharynx cancers. Physician driven iterative learning processes resulted in favorable evolution in HN RT plan quality with significant time savings, and improved consistency using AVI-planner.

## Introduction

Radiation therapy (RT) is a cornerstone of HN cancer treatment. Intensity-modulated radiation therapy (IMRT) has improved treatment accuracy and reduced RT-associated morbidity [1-10]. HN IMRT manual optimization is resource-intensive and variable, with heavy reliance upon physician and facility expertise [11-16]. HN IMRT implementation has been met with frequent treatment planning and quality assurance (QA) deviations, which are associated with worse outcomes [17-19]. Furthermore, the time required for HN IMRT planning must be considered in the context of survival advantages associated with minimizing total treatment time and time interval between consultation and starting treatment [20, 21]. HN RT delivered at high-accruing centers is associated with improved outcomes, though factors including travel burden and patients' resources influence access to these centers [22-24].

Automated planning has been developed to standardize treatment planning, maximize efficiency, improve plan quality, and mitigate geographic disparities by increasing access to high quality RT plans [13, 25]. Knowledge-based planning (KBP) models rely upon dosimetric and geometric experience from dose-volume histograms (DVH) of previously treated acceptable plans [25]. KBP benefits have been documented in various disease sites, including HN [26-34]. Iterative learning, a process incorporating manually driven feedback into model training, improves automated HN plan quality [35]. However, commercially available KBP algorithms are limited by smaller training datasets, lack of standardized inputs, ~~and~~ challenging user-interface for plan revision, and limited ability to customize commercial algorithms to fit specific clinical needs. Script-based approaches like ours enable clinic-specific customization. Prior studies have characterized plan quality in cohorts of HN patients without regard for primary site, while others

report achievements in only one subsite (e.g. oropharynx [36] or nasopharynx [31]). There is a paucity of data regarding automated planning algorithm performance among different HN sites.

Herein, we report ~~an evaluation method used to fill these gaps in quantitative data and detail among different HN sites as part of developing~~the development of an automated virtual integrative (AVI) planning algorithm. The algorithm is not a machine learning approach. This algorithm was designed using the same treatment planning system tools applied by dosimetrists during the manual process and integrates historical optimization norms from prior plans. The AVI-planner algorithm uniquely generates optimization parameters based upon statistical analyses of DVH metrics from previously treated HN RT plans. We sought to create preliminary automated HN RT plans for "warm start optimization" where dosimetrists continue optimization from the automated plan instead of starting each plan with a new manual process [37]. ~~The algorithm was designed to act as a virtual dosimetrist, using the same treatment planning system tools applied by dosimetrists during the manual process and integrating the knowledge of the extensive history of prior plans and clinical processes.~~ We describe the iterative learning process to address planning deficiencies noted for select primary sites. To our knowledge, this is the first investigation of a HN-specific automated planning algorithm whereby the identification of site-specific clinically-significant deficiencies drive autoplanner script refinements to improve overall RT plan quality.

Formatted: Highlight

**Methods:**

*Script Development and ~~,~~ Hazard Analysis~~, and Clinical Release~~*

Our script release process is shown in Figure 1. The write-enabled script was developed to incorporate practice norms defined by a library of 668 previously treated HN RT plans collected at our institution between 2014-2019. This library was comprised of 31.3% oropharynx (n=209), 19.3% oral cavity (n=129), 14.7% larynx (n=98), 7.9% cutaneous (n=53), 6.6% salivary (n=44), 4.2% sinonasal (n=28), 3.7% nasopharynx (n=25), 3% Unknown Primary (n=20), 2.7% hypopharynx (n=18), 1.8% thyroid (n=12), 0.6% orbital or lacrimal (n=4), 4.2% "other" (n=28). Software inputs were standardized including nomenclature and complete sets of contoured organs at risk/planning target volumes (OAR/PTVs) with explicitly defined planning priorities and objectives. Within the foundational library, >90% of plans contained spinal cord, brainstem, bilateral cochlea, parotids, superior and inferior pharyngeal constrictors, oral cavity, esophagus, mandible, lips. When surgically present and clinically relevant, bilateral submandibular glands (SMG) were included in 75%, larynx in 81%, bilateral optic nerves, chiasm, eyes, and lenses were included in 18-25%, while only 11% included lacrimal glands (data not shown).

~~In development cycles~~During development, the AVI-planner algorithm statistically evaluated DVH parameters from the 668 plan library, which then informed optimization parameters. Optimization constraints were defined as less than 30% of historic values. ~~preceding the physician team evaluation in Round 1, t~~A team of physicists, dosimetrists and software developers then used the~~he a~~ algorithm ~~was~~ to iteratively optimize~~d on the~~ a subset of 20~~subsets of~~ HN patients ~~by the development team of physics, dosimetry and software developer~~. ~~,~~None of

the 20 HN plans were included in the physician Round 1 evaluation. Prior to Round 1 evaluation (see below)~~physician inspection~~, all planning parameters in the algorithm~~application~~ were finalized for physician evaluation~~.~~. Based upon standardized input t~~fixed and plans were generated for all patients in the cohort for physicians to inspect~~

~~The algorithm automates creation of structures used during optimization to control the shape of dose distributions. Clinical t~~argets ~~structures~~ and OARs, ~~(e.g. GTVs, CTVs, PTVs) and organs at risk (e.g. parotids, submandibular glands) are not modified~~ the algorithm created a full set of ~~by the algorithm. O~~optimization structures ~~are created from unmodified clinical structures using the same operations used in manual creation of these structures (e.g. margins,~~using typical margin and ~~B~~ boolean operations~~.) are used, but standardized by the algorithm so that all plans are consistent.~~ Optimization structures ~~created include~~included sub-volumes of overlapping OAR and target structures, as well as high dose PTV subvolumes segmented from lower PTV volumes. Dose s~~volumes based on overlap with margins to enable specificity of compromise in clinician directed trade-offs between target coverage and OAR sparing. Dose s~~culpting ~~structures included~~ rings were ~~created around the structures and used to use~~used by the normal tissue objective to conform prescription isodose lines~~. to the corresponding PTV volumes. Optimization structures segmented high dose PTV sub volumes out of larger, lower dose PTV volumes.~~

~~Physics evaluation of automated plans to pass quality plan check was the same as our routine clinical practice, which was then followed by a second phase of clinical evaluation. Before clinical use of the AVI-planner, a prospective hazard analysis was performed using a streamlined failure mode and effects analysis described by Paradis et al. [38]. A process map for clinical use of the script was generated with associated hazards (failure modes) from~~

~~multidisciplinary feedback. The priority score for each failure mode (a version of the relative risk priority number from TG-100) was assigned as high, medium or low [39]. All failure modes with high or medium priority scores were mitigated before proceeding on to clinical deployment.~~ The AVI-planner software automatically placed an isocenter, segmented optimization structures, and generated beams and plan setup with full calculation. All plans were VMAT, calculated in Eclipse version 15.6, with the analytical anisotropic algorithm (AAA), using 0.25 cm grid size. Eclipse Scripting Application Programming Interface (ESAPI) enabled the integration of AVI-planner software with Eclipse (Varian Medical System, Palo Alto, CA). The non-clinical, research version of ESAPI mimicked manual optimization and allowed interaction with the optimizer during optimization. However,

~~T~~the clinical ESAPI version ~~of ESAPI~~ di~~does~~ not allow this ~~programmatic~~ interaction. ~~with the optimizer to enable algorithmic replication of the dynamic manual processes for monitoring and adjusting of DVH metric values and priorities during optimization. This functionality is enabled only in the non-clinical, research version of ESAPI.~~ Since our objective was designing software ~~an application that~~compatible ~~could be used clinically~~ with the Food and Drug Administration (FDA) approved ESAPI versions ~~of ESAPI~~, our ~~we developed an application~~ interface and algorithm~~s so that the initial set of constrains provided~~generated ~~high quality~~ HN plans ~~across for a wide range of HN subsites~~which could be sequentially, manually modified after optimization. Optimization with our AVI-planner algorithm did not allow for dynamic real-time~~, without dynamic programmatic~~ interaction with the optimizer.

Routine physics quality plan check was employed for the automated plans, which then proceeded onto a second phase of clinical evaluation. Before clinical use, a prospective hazard analysis was performed using a streamlined failure mode and effects analysis described by

Paradis et al. [38]. A process map for clinical use of the script was generated with associated hazards (failure modes) from multidisciplinary feedback. The priority score for each failure mode (a version of the relative risk priority number from TG-100) was assigned as high, medium, or low [39]. All failure modes with high or medium priority scores were mitigated before proceeding to plan evaluation and clinical deployment.

~~*Write-enabled Script Refinement*~~

~~A 2-step refinement process, integrating physician evaluation, was used to refine the optimization constraint algorithm. The first iteration ("Round 1") reflected the first order "explicit" objectives for target and OAR coverage reflected in prescription document and the history of previously treated HN RT plans. The second refinement ("Round 2") used the plan evaluation context from Round 1 to reveal "implicit" second order objectives and priorities reflecting physician preferences for dose distribution details that may not be part of the prescription document.~~

~~In Round 1, a template of optimization constraints was defined from quantile analysis of the distribution specific structure-DVH metrics measured from the distribution values in our library of previously treated patients. Optimization constraints were set at the thresholds corresponding to the lower 30% of historic values. For structures evaluated based on Mean[Gy], optimization constraints corresponding to the lower 30% of historic values for D90%[Gy], D50%[Gy] and D10%[Gy] were used to enable differentially prioritizing portions receiving high dose due to proximity to PTV volumes (D10%[Gy]) vs portions which may receive a lower dose (D90%[Gy]).~~

~~The algorithm automates creation of structures used during optimization to control the shape of dose distributions. Clinical target structures (e.g. GTVs, CTVs, PTVs) and organs at risk (e.g. parotids, submandibular glands) are not modified by the algorithm. Optimization structures are created from unmodified clinical structures using the same operations used in manual creation of these structures (e.g. margins, Boolean operations) are used, but standardized by the algorithm so that all plans are consistent. Optimization structures created include sub-volumes of target volumes based on overlap with margins to enable specificity of compromise in clinician directed trade-offs between target coverage and OAR sparing. Dose sculpting structures included rings were created around the structures and used to use the normal tissue objective to conform prescription isodose lines to the corresponding PTV volumes. Optimization structures segmented high dose PTV sub-volumes out of larger, lower dose PTV volumes.~~

~~In development cycles preceding the physician team evaluation in Round 1, the algorithm was iteratively optimized on the subsets of HN patients by the development team of physics, dosimetry and software developer. Prior to physician inspection, all parameters in the application were fixed and plans were generated for all patients in the cohort for physicians to inspect~~

~~"Round 1" planning used the validation cohort described below. "Round 1" indicates optimization via the initially released AVI planner script with minor manual edits. Following Round 1 evaluation, physician feedback was incorporated via software modifications in three categories.:~~

~~1) Normal tissue optimization constraints and prioritizations and prioritizations – Mapping between priorities listed in prescription objectives identified in the user interface and optimization constraint priorities used during the optimization were refined to improve alignment with physician's intent.Dose values for optimization constraints may be adjusted.~~

~~2) Dose-sculpting structures- Shaping of the dose distribution in areas not delineated by organs at risk are controlled by first creating dose sculpting structures (e.g. rings, buffers between PTVs, etc) with defined spatial relationships between targets and organs at risk, and second controlling the dose distributions in those structures with optimization constraints.~~

~~3) Segmentation of areas with PTV and OAR overlap- adding standardized structures with known spatial relationships between specific PTV and OAR volumes provides finer control of those dose to those areas in the optimizer.~~

Formatted: Highlight

*Patient Selection*

This study was IRB exempt (HUM ~~00126332~~xxxxxxx) for quality improvement. AVI-planner in Round 1 optimization was retrospectively validated within a cohort of 52 HN cancer patients treated between 2019-2020. None of these 52 plans were included within the foundational 668 plan library. We included oral cavity, oropharynx, larynx, hypopharynx, cutaneous, sinonasal, and salivary primaries to account for anatomy and OARs, adjuvant vs definitive RT, target dose, and fractionation. Institutional dose-escalation or de-escalation protocol patients were included. We excluded hypofractionated and palliative patients. Simulation CT scans were performed on a Philips Brilliance big-bore 16 slice scanner (Koninklijke Philips N.V., Amsterdam, Netherlands) using 3 mm slices. Patients were scanned head-first, supine with IV contrast and immobilized in 5-point thermoplastic masks. Intact and postoperative boost and elective CTV contours were delineated referencing published guidelines [40, 41] with a 3 mm PTV margin. Dosimetrists manually optimized clinical plans using Eclipse

(Varian Medical System, Palo Alto, CA), which were delivered on Varian TrueBeam or Clinac linear accelerators with 120 leaf MLC using 6-MV photons with 2-4 VMAT arcs.

*Plan Evaluation ~~and Statistical Analysis~~*

Clinical plans underwent peer-review by a subspecialty panel of attending radiation oncologists. Institutional protocols specified prioritization of target coverage and objectives for OAR sparing (Supplementary Table 1). To identify AVI-planner limitations consistently requiring additional manual input for "warm start optimization," the physician panel evaluated clinical acceptability of "Round 1" AVI-planner cases. HN subsites were grouped by treatment paradigm and anatomic proximity. These plans were "rejected" if the plan was unsafe and unsalvageable despite reoptimization. "Major revisions" indicated a high perceived risk of either 1) a clinically relevant toxicity due to exceeded OAR constraints or 2) risk of recurrence from target under-coverage. Plans with "minor revisions" were safe with room for improvement in conformality, heterogeneity, or target coverage. The highest quality plans were deemed "treat as is." Physician feedback from Round 1 was addressed per "Write-enabled Script ~~Model~~ Refinement." All 52 cases were then replanned with the AVI-planner script without manual modifications and labeled "Round 2." The same physician panel re-evaluated all 52 plans.

Beyond stand-alone clinical acceptability, Round 2 AVI-planner quality was compared to 1) clinically treated plans 2) historically accepted plans and 3) literature-based thresholds [42]. Clinically treated plan denotes the patient-specific RT plan, which was delivered during the patient's treatment course. Within this context, evaluating Round 2 versus the clinically treated plan provides an individual, patient-level comparison of plan quality. Comparisons to historically accepted plans were based on summarized metrics captured from the entire 668 HN foundational library. Thus, Round 2 plan quality was assessed in the context of aggregate institutional

experience with all 668 considered high-quality HN plans. Evaluating Round 2 plans in both situations more fully characterizes plan quality at both the patient-level and institutional experience- level. ~~We used Python 3.8 statistical software for this analysis. A one-sided, Kolmogorov-Smirnov (KS) test was used to determine if the distribution of AVI-planner OAR mean or D0.1cc values was higher or lower than clinically treated plans of the validation cohort. The distribution of per-plan differences was analyzed. AVI-planner values were compared to literature based thresholds [42] using a normal distribution, with matched cardinality, centered on each threshold with a 0.5 Gy standard deviation as the reference distribution using a t-test for mean value difference.~~

To compare AVI-planner to historic plans, constraint metrics within the algorithm were derived from 668 previously treated plans using the previously described Generalized Evaluation Metric (GEM) and Weighted Experience Score (WES) described by Mayo et al. [43]. GEM compares DVH metrics to constraints and historical values, which are cast onto a sigmoidal curve with scale of 0 to 1, where GEM = 0.5 if the constraint was met and 0.95 when 95% of historical values were lower than the current plan's value. WES ranks the DVH curves with respect to historical values, on a 0 to 1 scale, with values weighted according to historic variability. WES correlates with NTCP but rises sooner with respect to dose, correlating with physician preferences to drive doses below NTCP thresholds.

VRxGy[%] was used to assess coverage at the prescribed dose for each dose level. The ICRU Conformality index (CI) [26, 44] was calculated for PTV_High, PTV_Low and PTV_Mid00 volume

$$CI_{ICRU} = \frac{Body{:}\,VRx[cc]}{PTV{:}\,Volume[cc]}$$

Dose heterogeneity within PTV volumes was assessed using ICRU 83 $HI_1$ [45].

$$HI_1 = \frac{(D2\%[Gy] - D98\%[Gy])}{D50\%[Gy]}$$

These were calculated for the PTV subvolumes, not overlapping with volumes at prescribed doses as PTV_High, PTV!_Low and PTV!_Mid00 in TG-263 nomenclature.

We collected total monitor units (MU) per plan for the 52 patient cohort as well as calculated complexity described by Younge et al [46] as below.

$$M = \frac{1}{MU} \sum_{i=1}^{N} MU_i \, x \, \frac{y_i}{A_i}$$

~~Descriptive statistics were utilized to evaluate dosimetric endpoints and time required for treatment planning. One-tailed T-test was used to compare the interactive time required for AVI-planning versus manual planning.~~

*Write-enabled Script Refinement and Clinical Deployment*

Iterative learning occurred by a two-step process, which used physician feedback to refine the optimization algorithm. The first iteration of plans reflected the explicitly stated prescription planning objectives using statistical data gained from the 668 HN plan library ("Round 1"). In Round 1, a template of optimization constraints was defined by quantile analysis of DVH metrics within our plan library of 668 previously treated patients. For structures evaluated by Mean[Gy], the constraint corresponded with the lower 30% of historic values for

D90%[Gy], D50%[Gy] and D10%[Gy]. This enabled prioritization of portions receiving lower dose further away from PTV (D90%[Gy]) as compared to portions receiving high dose in close proximity to PTV volumes (D10%[Gy]). “Round 1” planning used the validation cohort described below, and “Round 1” indicates optimization with the initially released AVI-planner script and minor manual edits. After Round 1 evaluation, we discovered additional implicit physician preferences and expectations which were not stated in the prescription planning documentation. The algorithm was modified in several ways to incorporate physician feedback. This modified, refined algorithm was subsequently used to generate refined plans (“Round 2”). To shorten development time required to refine algorithm performance, these modified parameters were placed in an external configuration file. This limited the scope of changes that required recompiling the code, and also facilitates more rapid customization in the future when releasing this script to other clinics. Script modifications were as follows:

1) Normal tissue constraints and priorities – Instead of limiting the level of priority to 1, 2, or 3, we included additional more granular priority levels (i.e priority 1, 1.5, 2, 2.5, 3, etc) to better align with physicians’ intent. Instead of a fixed constraint value, the algorithm was modified to allow increasing or decreasing a given constraint.

2) Dose-sculpting structures- Automatically generated rings and buffers were added to increase conformality and minimize dose in non-target and non-OAR normal tissues (i.e minimize low and intermediate dose within the base of the neck).

3) Segmented structures- Subvolumes of PTV and OAR overlap were transformed into standardized segmented structures, and new constraints and priorities were added to enhance the algorithm’s ability to more precisely control dose. For example, areas of parotid and PTV

overlap were segmented out, and a higher priority for sparing was placed upon the non-overlapping ipsilateral parotid.

4) Isocenter placement- In response to major revision or rejected plans, we modified ~~These included addition to the start panel of the user interface of structure volumes and volume fraction of overlap with PTV structures. This supplemental information aids in customizing relative prioritization DVH metrics for structures. Buffer structures to control low and intermediate dose levels between PTVs and in the posterior neck region were modified.~~ the algorithm's ability to detect unilateral or atypical~~Detection of atypical~~ target location~~s~~ and subsequently adapt isocenter placement ~~shifting from our standard isocenter location algorithm isocenter~~was added. This~~ to optimize~~maximized use of the central 0.5 cm MLC leaves. ~~was added in response to the one plan rejected the second plan requiring major revision in Round 1. For those 2 cases the PTV volumes were ipsilateral, involving the skull and upper neck.~~

~~To shorten development time in refining algorithm performance, parameters were placed in an external configuration file, to significantly limit the scope of changes that require recompiling the code. The change to use of planning parameters configuration file also supports eventual extension of the script to other clinics enabling facilitating local customizations.~~After the evaluation process described, AVI-planner was deployed to the clinic as a staged process. ~~. In our clinic this deployment is done in a staged process, with testing carried out first by the chief dosimetrist and another dosimetrist.~~ The initial use of the application was~~is~~ carefully monitored by ~~the~~ physician~~s~~ and physicist~~s~~ stakeholders in the initial limited clinical release for proper functioning and introduction of hazards. ~~After validated in this monitored process,~~Following validation, the script was then deployed without change~~s~~. The script is routinely used in clinic, though dosimetrists regularly manually modify these automated plans with physician input ~~and is~~

~~in routine use in the clinic~~. Requests for additional improvements and features are monitored and incorporated into future development cycles.

*Statistical Analysis*

We used Python 3.8 statistical software for this analysis. A one-sided, Kolmogorov-Smirnov (KS) test was used to determine if the distribution of AVI-planner OAR mean or D0.1cc values was higher or lower than clinically treated plans of the validation cohort. The distribution of per plan differences was analyzed. AVI-planner values were compared to literature based thresholds [42] using a normal distribution, with matched cardinality, centered on each threshold with a 0.5 Gy standard deviation as the reference distribution using a t-test for mean value difference. Descriptive statistics were utilized to evaluate dosimetric endpoints and time required for treatment planning. One-tailed T-test was used to compare the interactive time required for AVI-planning versus manual planning. Two-tailed T-test was used for comparing total monitor units (MUs) and complexity between AVI-planner and clinical plans.

Field Code Changed

## Results

*Failure Mode and Effects Analysis (FMEA)*

Before clinical deployment, 12 failure modes were identified relating to contour generation (7), plan creation (1), treatment field generation (2), plan optimization (1) and plan approval (1) (Supplementary Table 2). None of these failure modes were higher relative risk compared to the manual treatment planning process.~~Five were higher relative risk.~~ A detailed summary of the ~~F~~failure modes and associated mitigations is~~are~~ shown in Supplementary Table 2. Several code modifications were prompted by FMEA. These included detailed analysis of structure volumes at the beginning and end of the algorithm to identify changes made, checks at entry of the script algorithm that PTV and organ at risk volumes are approved and canno~~t~~ be edited, enforcement of naming conventions for structure sets, course and plans to minimize risk of unintentional use of an automated plan.

*Validation and Clinical Implementation: Round 1*

We retrospectively validated AVI-planner in 52 patients, which consisted of mostly men (69%) with locally advanced oropharynx (40%) or oral cavity and salivary (31%) cancers (Figure 2, Supplementary Table 3). Definitive intent organ-preservation RT comprised the majority of plans (58%) with a median of 70 Gy (range 54-80 Gy) in 35 fractions (range 27-35 fx). 62% received concurrent chemotherapy.

Overall, 86% of Round 1 plans were safe to treat; however, we identified variability in plan quality among different HN subsites (Figure 2). All oropharynx and p16+ SCC Unknown Primary~~–~~ (21/21 plans), larynx and hypopharynx (7/7 plans) were "treat as is." Similarly, most oral cavity and salivary cases (14/16 plans; 87.5%) required no revisions. This contrasts with the

frequency of major revisions or rejections recommended for sinonasal, nasopharynx and p16-negative SCC Unknown Primary (1/6 plans; 17%) and cutaneous (1/2 plans; 50%- Figure 2). Minor revisions were increasing conformality and reducing heterogeneity. Major revisions were limiting hot spots outside PTV, restricting hot spots within PTV to 105-110%, and improving target coverage. Sample Round 1 isodose distributions are shown for both a definitive early stage p16+ base of tongue cancer considered "treat as is" (Figure 3B), compared to "major revisions" for an adjuvantly treated malar cheek Merkel cell carcinoma (Figure 3E). Software refinements~~Changes~~ were made to the script in response to the Round 1 evaluation which included normal tissue constraints and priorities, dose-sculpting structures, segmented structures, and isocenter placement as discussed above in Methods section *Write-enabled Script Refinement and Clinical Deployment*. ~~These included addition to the start panel of the user interface of structure volumes and volume fraction of overlap with PTV structures. This supplemental information aids in customizing relative prioritization DVH metrics for structures. Buffer structures to control low and intermediate dose levels between PTVs and in the posterior neck region were modified. Detection of atypical target locations and shifting from our standard isocenter location algorithm isocenter to optimize use of the central 0.5 cm MLC leaves was added in response to the one plan rejected the second plan requiring major revision in Round 1. For these 2 cases the PTV volumes were ipsilateral, involving the skull and upper neck. To shorten development time in refining algorithm performance, parameters were placed in an external configuration file, to significantly limit the scope of changes that require recompiling the code. The change to use of planning parameters configuration file also supports eventual extension of the script to other clinics enabling facilitating local customizations.~~

*Clinical Reassessment: Round 2*

During Round 2 evaluation of all 52 plans, there were no rejections nor major revisions (Figure 2). Minor revisions were recommended for 1 oral cavity (6.3%) and 3 sinonasal or nasopharynx or p16-negative SCC Unknown Primary plans (50%). The remaining 48 plans were "treat as is." Minor revisions in Round 2 focused on improving conformality, or more aggressive sparing of spinal cord, optics and contralateral orbit or salivary structures. This evolution in quality is evident for the adjuvantly treated Merkel cell carcinoma (Figure 3D-F).

Conformality and dose heterogeneity for all PTV levels were similar between Round 2 AVI-planner and clinical cases (Table 1). The number of MU per plan was significantly lower for AVI-planner Round 2 (mean 619.7 ± 69.7 MU) compared to the clinically treated plan (693.5 ± 219.4 MU; p=0.03). Similarly, AVI-planner generated less complex plans as compared to clinically treated plans (mean complexity score 0.13 ± 0.02 vs 0.14 ± 0.03; p<0.01). To evaluate patterns of OAR sparing in Round 2 AVI-planner, we compared the entire distribution of mean or D0.1cc dose values between AVI-planner cases versus clinically treated or historically accepted plan values (Figure 4A). Given that oropharynx was the most prevalent HN subsite within both the foundational library and the validation cohort, we also selected a representative DVH from a locally advanced oropharynx cancer treated with definitive chemoradiation. This demonstrates typical DVH metrics from a case which was well represented in the model (Figure 5).

Considering all 52 plans, the contralateral parotid dose distribution was higher with AVI-planner compared to clinically treated plans (median 25 vs 23 Gy, p<0.01) ~~with a difference of 1.9 Gy per plan~~, with higher doses compared to historic plans (WES 0.50 vs 0.42, p<0.01). I~~Similarly, i~~nferior pharyngeal constrictor muscles had higher distribution of mean dose in AVI-

planner versus clinical plans (median 21 Gy vs 19 Gy, p<0.01) ~~with a 1 Gy per plan difference~~, with higher doses compared to historic plans~~. Compared to historically accepted values, the dose to inferior constrictors was higher~~ (WES 0.53 vs 0.35, p<0.01), and narrowly exceeded our constraint (GEM 0.52). Conversely, AVI-planner lowered the dose to ipsilateral SMG (62 Gy vs 65 Gy, p=0.04) though this was not clinically relevant (GEM >0.90) [43]. AVI-planner lowered the distribution of mean dose to the larynx compared to clinical (median 19 Gy vs 20 Gy, p<0.01) and historical plans (WES 0.29 vs 0.44, p<0.01) ~~with a per plan difference of 1.3 Gy~~. Brainstem D0.1cc from AVI-planner was lower than clinical plans (median 28 Gy vs 32 Gy, p<0.01) ~~with a per plan difference of 5 Gy~~. Spinal cord distribution of D0.1cc was lower in AVI-planner as compared to historic plans (WES 0.44 vs 0.63, p<0.01), but similar to clinical plans (median 36.4 Gy vs 36.7 Gy, p=0.9). Distribution of dose to optic nerves, chiasm, eyes, contralateral SMG, superior pharyngeal constrictors, oral cavity, mandible, and esophagus were similar among AVI-planner cases, clinical and historic plans (Figure 4A).

For oropharynx or p16+ SCC Unknown Primary (n=21; Figure 4B), the distribution of mean larynx dose was significantly lower with AVI-planner versus clinical plans (median 18 vs 20 Gy, p<0.01) or historical plans (WES 0.28 vs 0.44, p<0.01), which was clinically relevant (GEM 0.46). Esophagus and ipsilateral parotid were spared equally among AVI-planner, clinical and historic plans. Distribution of mean dose to contralateral SMG was higher for AVI-planner compared to clinical plans (median 36 Gy vs 33 Gy, p=0.02) and historical plans (WES 0.44 vs 0.41, p=0.02)~~, corresponding to 0.13 Gy per plan difference~~. Neither clinical nor AVI-planner met constraints for relevant sparing (GEM 0.62 and 0.57). Contralateral parotid distribution of mean dose was higher for AVI-planner compared to clinical plans (median 25 vs 23 Gy, p=0.047), but similar to historically accepted plans (WES 0.50 vs 0.43, p=0.1). Superior and

inferior pharyngeal constrictors received higher dose with AVI-planner and exceeded constraints compared to historical controls ($p<0.01$). Oral cavity distribution of mean dose was higher with AVI-planner versus clinical (37 vs 33 Gy, $p=0.04$) and historic plans (WES 0.64 vs 0.53, $p=0.04$; Figure 4B).

OAR doses were similar between clinical and Round 2 AVI-planner for the remaining HN subsites (Figure 4B and 4C). Of note, the oral cavity/salivary contralateral parotid distribution of mean dose was significantly higher for AVI-planner compared to clinical plans (median 26 Gy vs 23 Gy, $p<0.01$), and historic plans (WES 0.53 vs 0.42, $p<0.01$) and did not meet constraints (GEM 0.56) (Figure 4B). Two cutaneous plans did not reach the 3 plan threshold required for formal comparison.

*Time Study*

We compared Eclipse optimizer interactive time for 10 recent manual plans versus interactive time with AVI-planner software for 51 of the validation cohort patients. This interactive time included all steps of the manual optimization such as segmentation structures, setting isocenter, ring and buffer dose sculpting structures, normal tissue optimization limits, target and OAR prioritization, setting the # arcs, optimization time. Of the 10 manually optimized plans, 70% were oropharynx (n=7), while 30% were comprised of oral cavity (n=1), thyroid (n=1), and Unknown Primary (n=1). Mean time for manual interaction time was shorter for AVI-planner vs manually optimized plans, 2 vs 85 minutes respectively ($p<0.01$~~); Figure 5)~~.

## Discussion

We developed and implemented a knowledge-based automated virtual integrative software to facilitate HN treatment planning. Initially, we identified inconsistent plan quality among different HN subsites. Following iterative software adaptations, we noted favorable evolution in target coverage, heterogeneity and OAR sparing. This software exceeded our "warm start optimization" goal and rapidly created clinically acceptable plans without manual adjustments for many HN subsites. We have published the source code for AVI-Planner at a GitHub repository (http://xxxx.xxxx.xxxx ) to promote use and development of automated planning.

Regarding clinical acceptability of automated HN plans, we found 86% of Round 1 plans were treat as is, which is comparable to 88% by Radiation Planning Assistant [47]. Our script was developed from a diverse training dataset, capturing unique nuances and planning considerations. The inconsistent site-specific plan quality likely resulted from limited experience within the foundational library. Our heterogeneous library accrued over 5 years, but there were fewer cutaneous (7.9%), sinonasal (4.2%) and nasopharynx (3.7%) plans. Improvements in both conformality and heterogeneity were shown for prostate and cervix cancer after refining Varian's RapidPlan default settings [48], but studies detailing specific software refinements and evolution in plan quality among multiple subsites are limited for HN [35, 47].

Physicians frequently emphasized higher OAR prioritization. For instance, the clinical plan aggressively spared contralateral parotid further below the planning objective in a cT1N1 p16+ tonsil cancer, whereas AVI-planner less aggressively spared the contralateral parotid to meet constraints. Given OAR constraint heterogeneity, we compared the Round 2 AVI-planner results to consensus thresholds [42]. Structure laterality is reported, however the distinction of

ipsilateral or contralateral relative to the target is less readily available. AVI-planner achieved lower contralateral parotid doses (median 25 Gy) than the 26 Gy threshold ($p < 0.01$). AVI-planner lowered dose to optic structures, eyes, brainstem, spinal cord, esophagus, inferior constrictors and mandible compared to accepted thresholds. Larynx doses achieved in AVI-planner cases were lower than thresholds (35 Gy, $p<0.01$), and values with automation approaches by Fogliata et al. (24.8 ± 5.2 Gy) or Ouyang et al. (28.4 Gy) [26, 27]. This highlights the importance of benchmarking automation achievements against literature values, historic norms, and the validation subset.

Secondarily, the failure modes addressed during development can be found in manual planning, suggesting these hazards already exist and may be more likely to happen without the software. Thus, automated planning does not obviate standard clinical and physics QA. Similar to Wang et al. [49], inconsistencies in standardized OAR prioritization affected the performance of our model. In line with time-savings noted by other groups for autoplanned nasopharynx [31, 50, 51] and oropharynx cancers [33, 36], we confirmed time savings with AVI-planner. AVI-planner generated less complex plans ($p<0.01$) with fewer MU ($p=0.03$) compared to the clinically treated plans.

Limitations of this work include this software is integrated only with Eclipse for 30-35 fraction plans. Given the revisions required for sinonasal, nasopharynx, and cutaneous sites, this software should be used cautiously near the skull base. Three target dose levels are currently supported, but additional dose levels require manual editing. Dosimetrists must also ensure the relevance of automated decisions. For instance, planners must remain vigilant about modifying the isocenter location or number of arcs for a unilateral target. Physicians must explicitly address planning preferences in the planning directive. For example, in a locally advanced maxillary

sinus cancer requiring adjuvant RT following an orbital exenteration, aggressively sparing the remaining contralateral orbit and lacrimal gland may take precedence over PTV coverage. The user-friendly interface contains the same tools used in manual optimization, allowing real-time modification by dosimetrists, compared to fully automated optimization which must run to completion before permitting revision. However, these standardized optimization parameters likely differ from personalized approaches of experienced dosimetrists. Therefore, additional time may be required for revisions. Inter-institutional heterogeneity in delineated OARs, inconsistent OAR contouring, and variability in constraints and prioritization are barriers to widespread adoption of automated planning.

To our knowledge, this is the first report identifying HN primary site-specific variability in automated plan quality, which favorably evolved with physician input. Our work cautions against interpreting that automated planning achievements are universal among HN subsites. This is relevant for clinics that would ideally employ one planning algorithm for all HN cases, instead of separate optimization algorithms for each HN subsite. We are not advocating this software in lieu of skilled dosimetrists or treatment at high-accruing centers. However, in settings of limited resources, increased demand, urgent starts or reduced subspecialized dosimetrists, AVI-planner software can be easily integrated into workflows to increase availability of high quality HN RT plans.

Furthermore, our institutional adoption of AVI-planner into routine practice has expanded the number of dosimetrists able to rapidly generate high quality HN plans. We plan to release AVI-planner to our affiliate sites to improve plan quality and uniformity, we are also extending this standardized approach to other disease sites including lung and prostate. In the

future, automated planning will facilitate adaptive RT planning. Future software upgrades may incorporate gEUD, update the foundational library, and focus on well-lateralized cases near the skin surface. Application programming interfaces that enable clinics to programmatically automate all parts of the treatment planning process, give clinics the tools they need to increase efficiency and consistency in plan quality in their process workflows. To promote these clinical improvements it is highly desirable for manufactures to provide application programming interfaces that give users at minimum the same capabilities they have in manual operations to algorithmically interact with optimizers, dose calculation engines, reference points, and scheduling capabilities.

**Formatted:** Indent: First line: 0.5"

## References


1. Hawkins, P.G., et al., *Organ-Sparing in Radiotherapy for Head-and-Neck Cancer: Improving Quality of Life.* Semin Radiat Oncol, 2018. **28**(1): p. 46-52.
2. Nutting, C.M., et al., *Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial.* Lancet Oncol, 2011. **12**(2): p. 127-36.
3. Pow, E.H., et al., *Xerostomia and quality of life after intensity-modulated radiotherapy vs. conventional radiotherapy for early-stage nasopharyngeal carcinoma: initial report on a randomized controlled clinical trial.* Int J Radiat Oncol Biol Phys, 2006. **66**(4): p. 981-91.
4. Kam, M.K., et al., *Prospective randomized study of intensity-modulated radiotherapy on salivary gland function in early-stage nasopharyngeal carcinoma patients.* J Clin Oncol, 2007. **25**(31): p. 4873-9.
5. Dirix, P. and S. Nuyts, *Evidence-based organ-sparing radiotherapy in head and neck cancer.* Lancet Oncol, 2010. **11**(1): p. 85-91.
6. Lee, N., et al., *Intensity-modulated radiation therapy in head and neck cancers: an update.* Head Neck, 2007. **29**(4): p. 387-400.
7. Eisbruch, A., et al., *Parotid gland sparing in patients undergoing bilateral head and neck irradiation: techniques and early results.* Int J Radiat Oncol Biol Phys, 1996. **36**(2): p. 469-80.
8. Murdoch-Kinch, C.A., et al., *Dose-effect relationships for the submandibular salivary glands and implications for their sparing by intensity modulated radiotherapy.* Int J Radiat Oncol Biol Phys, 2008. **72**(2): p. 373-82.
9. Feng, F.Y., et al., *Intensity-modulated chemoradiotherapy aiming to reduce dysphagia in patients with oropharyngeal cancer: clinical and functional results.* J Clin Oncol, 2010. **28**(16): p. 2732-8.
10. Beadle, B.M., et al., *Improved survival using intensity-modulated radiation therapy in head and neck cancers: a SEER-Medicare analysis.* Cancer, 2014. **120**(5): p. 702-10.
11. Boero, I.J., et al., *Importance of Radiation Oncologist Experience Among Patients With Head-and-Neck Cancer Treated With Intensity-Modulated Radiation Therapy.* J Clin Oncol, 2016. **34**(7): p. 684-90.
12. Lee, C.C., et al., *Survival rate in nasopharyngeal carcinoma improved by high caseload volume: a nationwide population-based study in Taiwan.* Radiat Oncol, 2011. **6**: p. 92.
13. Cilla, S., et al., *Template-based automation of treatment planning in advanced radiotherapy: a comprehensive dosimetric and clinical evaluation.* Sci Rep, 2020. **10**(1): p. 423.
14. Batumalai, V., et al., *How important is dosimetrist experience for intensity modulated radiation therapy? A comparative analysis of a head and neck case.* Pract Radiat Oncol, 2013. **3**(3): p. e99-e106.
15. Moore, K.L., et al., *Experience-based quality control of clinical intensity-modulated radiotherapy planning.* Int J Radiat Oncol Biol Phys, 2011. **81**(2): p. 545-51.
16. Nelms, B.E., et al., *Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems.* Pract Radiat Oncol, 2012. **2**(4): p. 296-305.

17. Eisbruch, A., et al., *Multi-institutional trial of accelerated hypofractionated intensity-modulated radiation therapy for early-stage oropharyngeal cancer (RTOG 00-22).* Int J Radiat Oncol Biol Phys, 2010. **76**(5): p. 1333-8.
18. Zhong, H., et al., *The Impact of Clinical Trial Quality Assurance on Outcome in Head and Neck Radiotherapy Treatment.* Front Oncol, 2019. **9**: p. 792.
19. Peters, L.J., et al., *Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02.* J Clin Oncol, 2010. **28**(18): p. 2996-3001.
20. Graboyes, E.M., et al., *Association of Treatment Delays With Survival for Patients With Head and Neck Cancer: A Systematic Review.* JAMA Otolaryngol Head Neck Surg, 2019. **145**(2): p. 166-177.
21. Rosenthal, D.I., et al., *Importance of the treatment package time in surgery and postoperative radiation therapy for squamous carcinoma of the head and neck.* Head Neck, 2002. **24**(2): p. 115-26.
22. Wuthrick, E.J., et al., *Institutional clinical trial accrual volume and survival of patients with head and neck cancer.* J Clin Oncol, 2015. **33**(2): p. 156-64.
23. Naghavi, A.O., et al., *Patient choice for high-volume center radiation impacts head and neck cancer outcome.* Cancer Med, 2018. **7**(10): p. 4964-4979.
24. George, J.R., S.S. Yom, and S.J. Wang, *Combined modality treatment outcomes for head and neck cancer: comparison of postoperative radiation therapy at academic vs nonacademic medical centers.* JAMA Otolaryngol Head Neck Surg, 2013. **139**(11): p. 1118-26.
25. Hussein, M., et al., *Automation in intensity modulated radiotherapy treatment planning-a review of recent innovations.* Br J Radiol, 2018. **91**(1092): p. 20180270.
26. Ouyang, Z., et al., *Evaluation of auto-planning in IMRT and VMAT for head and neck cancer.* J Appl Clin Med Phys, 2019. **20**(7): p. 39-47.
27. Fogliata, A., et al., *RapidPlan head and neck model: the objectives and possible clinical benefit.* Radiat Oncol, 2017. **12**(1): p. 73.
28. Krayenbuehl, J., et al., *Evaluation of an automated knowledge based treatment planning system for head and neck.* Radiat Oncol, 2015. **10**: p. 226.
29. Tol, J.P., et al., *Evaluation of a knowledge-based planning solution for head and neck cancer.* Int J Radiat Oncol Biol Phys, 2015. **91**(3): p. 612-20.
30. Gintz, D., et al., *Initial evaluation of automated treatment planning software.* J Appl Clin Med Phys, 2016. **17**(3): p. 331-346.
31. Giaddui, T., et al., *Offline Quality Assurance for Intensity Modulated Radiation Therapy Treatment Plans for NRG-HN001 Head and Neck Clinical Trial Using Knowledge-Based Planning.* Adv Radiat Oncol, 2020. **5**(6): p. 1342-1349.
32. Hansen, C.R., et al., *Automatic treatment planning improves the clinical quality of head and neck cancer treatment plans.* Clin Transl Radiat Oncol, 2016. **1**: p. 2-8.
33. Kusters, J., et al., *Automated IMRT planning in Pinnacle : A study in head-and-neck cancer.* Strahlenther Onkol, 2017. **193**(12): p. 1031-1038.
34. Krayenbuehl, J., et al., *Planning comparison of five automated treatment planning solutions for locally advanced head and neck cancer.* Radiat Oncol, 2018. **13**(1): p. 170.
35. Fogliata, A., et al., *RapidPlan knowledge based planning: iterative learning process and model ability to steer planning strategies.* Radiat Oncol, 2019. **14**(1): p. 187.

36. Kamima, T., et al., *Multi-institutional evaluation of knowledge-based planning performance of volumetric modulated arc therapy (VMAT) for head and neck cancer*. Phys Med, 2019. **64**: p. 174-181.
37. Ahunbay, E.E., O. Ates, and X.A. Li, *An online replanning method using warm start optimization and aperture morphing for flattening-filter-free beams.* Med Phys, 2016. **43**(8): p. 4575.
38. Paradis, K.C., et al., *The Fusion of Incident Learning and Failure Mode and Effects Analysis for Data-Driven Patient Safety Improvements.* Pract Radiat Oncol, 2020.
39. Huq, M.S., et al., *The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation therapy quality management.* Med Phys, 2016. **43**(7): p. 4209.
40. Biau, J., et al., *Selection of lymph node target volumes for definitive head and neck radiation therapy: a 2019 Update.* Radiother Oncol, 2019. **134**: p. 1-9.
41. Gregoire, V., et al., *Delineation of the primary tumour Clinical Target Volumes (CTV-P) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO, TROG consensus guidelines.* Radiother Oncol, 2018. **126**(1): p. 3-24.
42. Lee, A.W., et al., *International Guideline on Dose Prioritization and Acceptance Criteria in Radiation Therapy Planning for Nasopharyngeal Carcinoma.* Int J Radiat Oncol Biol Phys, 2019. **105**(3): p. 567-580.
43. Mayo, C.S., et al., *Incorporating big data into treatment plan evaluation: Development of statistical DVH metrics and visualization dashboards.* Adv Radiat Oncol, 2017. **2**(3): p. 503-514.
44. Baltas, D., et al., *A conformal index (COIN) to evaluate implant quality and dose specification in brachytherapy.* Int J Radiat Oncol Biol Phys, 1998. **40**(2): p. 515-24.
45. *ICRU report 83 Prescribing, recording, and reporting photon-beam intensity-modulated radiation therapy (IMRT).* J ICRU 10:35–36, 2010.
46. Younge, K.C., et al., *Predicting deliverability of volumetric-modulated arc therapy (VMAT) plans using aperture complexity analysis.* J Appl Clin Med Phys, 2016. **17**(4): p. 124-131.
47. Olanrewaju, A., et al., *Clinical Acceptability of Automated Radiation Treatment Planning for Head and Neck Cancer Using the Radiation Planning Assistant.* Pract Radiat Oncol, 2021. **11**(3): p. 177-184.
48. Hussein, M., et al., *Clinical validation and benchmarking of knowledge-based IMRT and VMAT treatment planning in pelvic anatomy.* Radiother Oncol, 2016. **120**(3): p. 473-479.
49. Wang, Y., B.J.M. Heijmen, and S.F. Petit, *Knowledge-based dose prediction models for head and neck cancer are strongly affected by interorgan dependency and dataset inconsistency.* Med Phys, 2019. **46**(2): p. 934-943.
50. Chang, A.T.Y., et al., *Comparison of Planning Quality and Efficiency Between Conventional and Knowledge-based Algorithms in Nasopharyngeal Cancer Patients Using Intensity Modulated Radiation Therapy.* Int J Radiat Oncol Biol Phys, 2016. **95**(3): p. 981-990.
51. Hu, J., et al., *Quantitative Comparison of Knowledge-Based and Manual Intensity Modulated Radiation Therapy Planning for Nasopharyngeal Carcinoma.* Front Oncol, 2020. **10**: p. 551763.

1

2

1

2

Formatted: Line spacing: Double

30

**Figure Legends**

Figure 1~~:~~: Flow chart depicting write-enabled script development and release process.

Figure 2~~.~~: Highlighting the importance of disease subsite-specific AVI-planner algorithm performance. We show an ~~I~~integrated pie and doughnut chart~~s~~ demonstrating clinical acceptability of AVI-planned cases among H&N subsites. Central chart (blue) shows plan frequency by subsite (n=52). Innermost doughnut chart shows "Round 1" clinical acceptability. Outermost doughnut chart "Round 2" shows evolution in acceptability for 7 plans initially requiring revisions.

Figure 3~~.~~: Identifying the strengths and limitations of automated planning among typical versus atypical HN cases. We show s~~S~~ample ~~HN RT~~ plans for early-stage oropharynx and adjuvant Merkel cell carcinoma.~~ – Cases~~ Top panel shows ~~include~~ a stage II cT2N2M0 p16+ squamous cell carcinoma of the right base of tongue treated with definitive chemoradiation to 70 Gy in 35 fx (A-C), which is considered a typical case and well-represented within the model. ~~; and~~The bottom panel shows a stage III pT1 pN1a(sn) Merkel Cell carcinoma treated adjuvantly to 54 Gy in 30 fx (D-F), which is considered atypical and underrepresented within the model. Left panels (A,D) Clinically Treated Plan. Middle panels (B) Round 1 AVI-planner "treat as is;" (E) Round 1 AVI-planner "major revision" due to conformality and 119% hotspot outside PTV.~~–~~ Right panels (C, F) Round 2 following AVI-planner upgrades. Isodose lines (absolute dose, Gy) show 75 (light green), 70 (white), 65 (pink), 60 (red), 54 (yellow), 51 (green), 45 (orange), 40 (purple), 30 (cyan), 25 (green), 20 (dark blue).

Figure 4~~:~~: Differential normal tissue sparing by clinical versus AVI-planner. Comparisons for all HN subsites (A); oropharynx, p16+ SCC Unknown Primary, Oral Cavity, Salivary, Larynx,

Hypopharynx (B); Sinonasal, Nasopharynx, and p16-negative SCC Unknown Primary (C). OARs are listed on y-axis with corresponding constraint~~metric~~. Dose (Gy) on x-axis. Box plots of clinical (blue) or Round 2 AVI-planner (yellow) provide median, IQR, minimum, and maximum doses. Red "x" denotes consensus thresholds [42]. The difference in normal tissue sparing between AVI-planner and clinical plan is typically small compared to the difference in relation to established thresholds. Statistical significance was achieved with $p<0.05$ on one-sided Kolmogrorov-Smirnov test. Symbols along y-axis indicate statistically significant difference in OAR sparing between AVI-planner versus clinical plans: total cohort (panel A-circles), Oropharynx and p16+ SCC Unknown Primary (panel B-stars), Oral Cavity and Salivary (panel B- diamonds). Filled shapes indicate AVI-planner significantly improved sparing whereas unshaded symbols indicate clinical plan achieved significantly better sparing.

Figure 5: Representative typical dose-volume histogram for a cT4N1M0 p16+ squamous cell carcinoma of the left tonsil treated definitively to 70 Gy, comparing manual plan (squares) versus AVI-planner (triangles). X-axis is Dose (Gy), Y-axis volume (%).~~Figure 5: Box and whisker plot comparing dosimetrist interactive time between AVI-planner and manual optimization. Significantly less time required for Round 2 AVI-planner (n=51) versus manual optimization (n=10). * indicates p<0.01~~

Table 1. Conformality and heterogeneity (ICRU 83) indexes of clinical and Round 2 AVI-planner for high, intermediate, and low PTV.

## Abstract

**Purpose**: Head and neck (HN) radiation (RT) treatment planning is complex and resource intensive. Deviations and inconsistent plan quality significantly impact clinical outcomes. We sought to develop a novel automated virtual integrative (AVI) knowledge-based planning application to reduce planning time, increase consistency, and improve baseline quality.

**Materials and Methods:** An in-house write-enabled script was developed from a library of 668 previously treated HN RT plans. Prospective hazard analysis was performed, and mitigation strategies were implemented before clinical release. The AVI-planner software was retrospectively validated in a cohort of 52 recent HN cases. A physician panel evaluated planning limitations during initial deployment, and feedback was enacted via software refinements. A final second set of plans was generated and evaluated. Kolmogorov-Smirnov (KS) test in addition to Generalized Evaluation Metric (GEM) and Weighted Experience Score (WES) were used to compare normal tissue sparing between final AVI-planner versus respective clinically treated and historically accepted plans. T-test was used to compare the interactive time, complexity, and monitor units for AVI-planner versus manual optimization.

**Results:** Initially, 86% of plans were acceptable to treat with 10% minor and 4% major revisions or rejection recommended. Variability was noted in plan quality among HN subsites, with high initial quality for oropharynx and oral cavity plans. Plans needing revisions were comprised of sinonasal, nasopharynx, p-16 negative SCC Unknown Primary or cutaneous primary sites. Normal tissue sparing varied within subsites, but AVI-planner significantly lowered mean larynx dose (median 18.5 Gy vs 19.7 Gy, $p<0.01$) compared to clinical plans. AVI-planner significantly reduced interactive optimization time (mean 2 vs 85 minutes, $p<0.01$).

**Conclusions:** AVI-planner reliably generated clinically acceptable RT plans for oral cavity, salivary, oropharynx, larynx and hypopharynx cancers. Physician driven iterative learning processes resulted in favorable evolution in HN RT plan quality with significant time savings, and improved consistency using AVI-planner.

## Introduction

Radiation therapy (RT) is a cornerstone of HN cancer treatment. Intensity-modulated radiation therapy (IMRT) has improved treatment accuracy and reduced RT-associated morbidity [1-10]. HN IMRT manual optimization is resource-intensive and variable, with heavy reliance upon physician and facility expertise [11-16]. HN IMRT implementation has been met with frequent treatment planning and quality assurance (QA) deviations, which are associated with worse outcomes [17-19]. Furthermore, the time required for HN IMRT planning must be considered in the context of survival advantages associated with minimizing total treatment time and time interval between consultation and starting treatment [20, 21]. HN RT delivered at high-accruing centers is associated with improved outcomes, though factors including travel burden and patients' resources influence access to these centers [22-24].

Automated planning has been developed to standardize treatment planning, maximize efficiency, improve plan quality, and mitigate geographic disparities by increasing access to high quality RT plans [13, 25]. Knowledge-based planning (KBP) models rely upon dosimetric and geometric experience from dose-volume histograms (DVH) of previously treated acceptable plans [25]. KBP benefits have been documented in various disease sites, including HN [26-34]. Iterative learning, a process incorporating manually driven feedback into model training, improves automated HN plan quality [35]. However, commercially available KBP algorithms are limited by smaller training datasets, lack of standardized inputs, challenging user-interface for plan revision, and limited ability to customize commercial algorithms to fit specific clinical needs. Script-based approaches like ours enable clinic-specific customization. Prior studies have characterized plan quality in cohorts of HN patients without regard for primary site, while others

report achievements in only one subsite (e.g. oropharynx [36] or nasopharynx [31]). There is a paucity of data regarding automated planning algorithm performance among different HN sites.

Herein, we report the development of an automated virtual integrative (AVI) planning algorithm. The algorithm is not a machine learning approach. This algorithm was designed using the same treatment planning system tools applied by dosimetrists during the manual process and integrates historical optimization norms from prior plans. The AVI-planner algorithm uniquely generates optimization parameters based upon statistical analyses of DVH metrics from previously treated HN RT plans. We sought to create preliminary automated HN RT plans for "warm start optimization" where dosimetrists continue optimization from the automated plan instead of starting each plan with a new manual process [37]. We describe the iterative learning process to address planning deficiencies noted for select primary sites. To our knowledge, this is the first investigation of a HN-specific automated planning algorithm whereby the identification of site-specific clinically-significant deficiencies drive autoplanner script refinements to improve overall RT plan quality.

**Methods:**

*Script Development and Hazard Analysis*

Our script release process is shown in Figure 1. The write-enabled script was developed to incorporate practice norms defined by a library of 668 previously treated HN RT plans collected at our institution between 2014-2019. This library was comprised of 31.3% oropharynx (n=209), 19.3% oral cavity (n=129), 14.7% larynx (n=98), 7.9% cutaneous (n=53), 6.6% salivary (n=44), 4.2% sinonasal (n=28), 3.7% nasopharynx (n=25), 3% Unknown Primary (n=20), 2.7% hypopharynx (n=18), 1.8% thyroid (n=12), 0.6% orbital or lacrimal (n=4), 4.2% "other" (n=28). Software inputs were standardized including nomenclature and complete sets of contoured organs at risk/planning target volumes (OAR/PTVs) with explicitly defined planning priorities and objectives. Within the foundational library, >90% of plans contained spinal cord, brainstem, bilateral cochlea, parotids, superior and inferior pharyngeal constrictors, oral cavity, esophagus, mandible, lips. When surgically present and clinically relevant, bilateral submandibular glands (SMG) were included in 75%, larynx in 81%, bilateral optic nerves, chiasm, eyes, and lenses were included in 18-25%, while only 11% included lacrimal glands (data not shown).

During development, the AVI-planner algorithm statistically evaluated DVH parameters from the 668 plan library, which then informed optimization parameters. Optimization constraints were defined as less than 30% of historic values. A team of physicists, dosimetrists and software developers then used the algorithm to iteratively optimize a subset of 20 HN patients. None of the 20 HN plans were included in the physician Round 1 evaluation. Prior to Round 1 evaluation (see below), all planning parameters in the algorithm were finalized for physician evaluation. Based upon standardized input targets and OARs, the algorithm created a full set of optimization structures using typical margin and boolean operations. Optimization

structures included sub-volumes of overlapping OAR and target structures, as well as high dose PTV subvolumes segmented from lower PTV volumes. Dose sculpting rings were used by the normal tissue objective to conform prescription isodose lines. The AVI-planner software automatically placed an isocenter, segmented optimization structures, and generated beams and plan setup with full calculation. All plans were VMAT, calculated in Eclipse version 15.6, with the analytical anisotropic algorithm (AAA), using 0.25 cm grid size. Eclipse Scripting Application Programming Interface (ESAPI) enabled the integration of AVI-planner software with Eclipse (Varian Medical System, Palo Alto, CA). The non-clinical, research version of ESAPI mimicked manual optimization and allowed interaction with the optimizer during optimization. However, the clinical ESAPI version did not allow this interaction. Since our objective was designing software compatible with the Food and Drug Administration (FDA) approved ESAPI versions, our interface and algorithm generated HN plans which could be sequentially, manually modified after optimization. Optimization with our AVI-planner algorithm did not allow for dynamic real-time interaction with the optimizer.

Routine physics quality plan check was employed for the automated plans, which then proceeded onto a second phase of clinical evaluation. Before clinical use, a prospective hazard analysis was performed using a streamlined failure mode and effects analysis described by Paradis et al. [38]. A process map for clinical use of the script was generated with associated hazards (failure modes) from multidisciplinary feedback. The priority score for each failure mode (a version of the relative risk priority number from TG-100) was assigned as high, medium, or low [39]. All failure modes with high or medium priority scores were mitigated before proceeding to plan evaluation and clinical deployment.

*Patient Selection*

This study was IRB exempt (HUM 00126332) for quality improvement. AVI-planner in Round 1 optimization was retrospectively validated within a cohort of 52 HN cancer patients treated between 2019-2020. None of these 52 plans were included within the foundational 668 plan library. We included oral cavity, oropharynx, larynx, hypopharynx, cutaneous, sinonasal, and salivary primaries to account for anatomy and OARs, adjuvant vs definitive RT, target dose, and fractionation. Institutional dose-escalation or de-escalation protocol patients were included. We excluded hypofractionated and palliative patients. Simulation CT scans were performed on a Philips Brilliance big-bore 16 slice scanner (Koninklijke Philips N.V., Amsterdam, Netherlands) using 3 mm slices. Patients were scanned head-first, supine with IV contrast and immobilized in 5-point thermoplastic masks. Intact and postoperative boost and elective CTV contours were delineated referencing published guidelines [40, 41] with a 3 mm PTV margin. Dosimetrists manually optimized clinical plans using Eclipse (Varian Medical System, Palo Alto, CA), which were delivered on Varian TrueBeam or Clinac linear accelerators with 120 leaf MLC using 6-MV photons with 2-4 VMAT arcs.

*Plan Evaluation*

Clinical plans underwent peer-review by a subspecialty panel of attending radiation oncologists. Institutional protocols specified prioritization of target coverage and objectives for OAR sparing (Supplementary Table 1). To identify AVI-planner limitations consistently requiring additional manual input for "warm start optimization," the physician panel evaluated clinical acceptability of "Round 1" AVI-planner cases. HN subsites were grouped by treatment paradigm and anatomic proximity. These plans were "rejected" if the plan was unsafe and unsalvageable despite reoptimization. "Major revisions" indicated a high perceived risk of either 1) a clinically relevant toxicity due to exceeded OAR constraints or 2) risk of recurrence from

target under-coverage. Plans with “minor revisions” were safe with room for improvement in conformality, heterogeneity, or target coverage. The highest quality plans were deemed “treat as is.” Physician feedback from Round 1 was addressed per “Write-enabled Script Refinement.” All 52 cases were then replanned with the AVI-planner script without manual modifications and labeled “Round 2.” The same physician panel re-evaluated all 52 plans.

Beyond stand-alone clinical acceptability, Round 2 AVI-planner quality was compared to 1) clinically treated plans 2) historically accepted plans and 3) literature-based thresholds [42]. Clinically treated plan denotes the patient-specific RT plan, which was delivered during the patient’s treatment course. Within this context, evaluating Round 2 versus the clinically treated plan provides an individual, patient-level comparison of plan quality. Comparisons to historically accepted plans were based on summarized metrics captured from the entire 668 HN foundational library. Thus, Round 2 plan quality was assessed in the context of aggregate institutional experience with all 668 considered high-quality HN plans. Evaluating Round 2 plans in both situations more fully characterizes plan quality at both the patient-level and institutional experience- level.

To compare AVI-planner to historic plans, constraint metrics within the algorithm were derived from 668 previously treated plans using the previously described Generalized Evaluation Metric (GEM) and Weighted Experience Score (WES) described by Mayo et al. [43]. GEM compares DVH metrics to constraints and historical values, which are cast onto a sigmoidal curve with scale of 0 to 1, where GEM = 0.5 if the constraint was met and 0.95 when 95% of historical values were lower than the current plan’s value. WES ranks the DVH curves with respect to historical values, on a 0 to 1 scale, with values weighted according to historic

variability. WES correlates with NTCP but rises sooner with respect to dose, correlating with physician preferences to drive doses below NTCP thresholds.

VRxGy[%] was used to assess coverage at the prescribed dose for each dose level. The ICRU Conformality index (CI) [26, 44] was calculated for PTV_High, PTV_Low and PTV_Mid00 volume

$$CI_{ICRU} = \frac{Body{:}\,VRx[cc]}{PTV{:}\,Volume[cc]}$$

Dose heterogeneity within PTV volumes was assessed using ICRU 83 $HI_1$ [45].

$$HI_1 = \frac{(D2\%[Gy] - D98\%[Gy])}{D50\%[Gy]}$$

These were calculated for the PTV subvolumes, not overlapping with volumes at prescribed doses as PTV_High, PTV!_Low and PTV!_Mid00 in TG-263 nomenclature.

We collected total monitor units (MU) per plan for the 52 patient cohort as well as calculated complexity described by Younge et al [46] as below.

$$M = \frac{1}{MU} \sum_{i=1}^{N} MU_i \, x \, \frac{y_i}{A_i}$$

*Write-enabled Script Refinement and Clinical Deployment*

Iterative learning occurred by a two-step process, which used physician feedback to refine the optimization algorithm. The first iteration of plans reflected the explicitly stated

prescription planning objectives using statistical data gained from the 668 HN plan library ("Round 1"). In Round 1, a template of optimization constraints was defined by quantile analysis of DVH metrics within our plan library of 668 previously treated patients. For structures evaluated by Mean[Gy], the constraint corresponded with the lower 30% of historic values for D90%[Gy], D50%[Gy] and D10%[Gy]. This enabled prioritization of portions receiving lower dose further away from PTV (D90%[Gy]) as compared to portions receiving high dose in close proximity to PTV volumes (D10%[Gy]). "Round 1" planning used the validation cohort described below, and "Round 1" indicates optimization with the initially released AVI-planner script and minor manual edits. After Round 1 evaluation, we discovered additional implicit physician preferences and expectations which were not stated in the prescription planning documentation. The algorithm was modified in several ways to incorporate physician feedback. This modified, refined algorithm was subsequently used to generate refined plans ("Round 2"). To shorten development time required to refine algorithm performance, these modified parameters were placed in an external configuration file. This limited the scope of changes that required recompiling the code, and also facilitates more rapid customization in the future when releasing this script to other clinics. Script modifications were as follows:

1) Normal tissue constraints and priorities – Instead of limiting the level of priority to 1, 2, or 3, we included additional more granular priority levels (i.e priority 1, 1.5, 2, 2.5, 3, etc) to better align with physicians' intent. Instead of a fixed constraint value, the algorithm was modified to allow increasing or decreasing a given constraint.

2) Dose-sculpting structures- Automatically generated rings and buffers were added to increase conformality and minimize dose in non-target and non-OAR normal tissues (i.e minimize low and intermediate dose within the base of the neck).

3) Segmented structures- Subvolumes of PTV and OAR overlap were transformed into standardized segmented structures, and new constraints and priorities were added to enhance the algorithm's ability to more precisely control dose. For example, areas of parotid and PTV overlap were segmented out, and a higher priority for sparing was placed upon the non-overlapping ipsilateral parotid.

4) Isocenter placement- In response to major revision or rejected plans, we modified the algorithm's ability to detect unilateral or atypical target location and subsequently adapt isocenter placement was added. This maximized use of the central 0.5 cm MLC leaves.

After the evaluation process described, AVI-planner was deployed to the clinic as a staged process. The initial use of the application was carefully monitored by physician and physicist stakeholders in the initial limited clinical release for proper functioning and introduction of hazards. Following validation, the script was then deployed without changes. The script is routinely used in clinic, though dosimetrists regularly manually modify these automated plans with physician input. Requests for additional improvements and features are monitored and incorporated into future development cycles.

*Statistical Analysis*

We used Python 3.8 statistical software for this analysis. A one-sided, Kolmogorov-Smirnov (KS) test was used to determine if the distribution of AVI-planner OAR mean or D0.1cc values was higher or lower than clinically treated plans of the validation cohort. The distribution of per plan differences was analyzed. AVI-planner values were compared to literature based thresholds [42] using a normal distribution, with matched cardinality, centered on each threshold with a 0.5 Gy standard deviation as the reference distribution using a t-test for

mean value difference. Descriptive statistics were utilized to evaluate dosimetric endpoints and time required for treatment planning. One-tailed T-test was used to compare the interactive time required for AVI-planning versus manual planning. Two-tailed T-test was used for comparing total monitor units (MUs) and complexity between AVI-planner and clinical plans.

## Results

*Failure Mode and Effects Analysis (FMEA)*

Before clinical deployment, 12 failure modes were identified relating to contour generation (7), plan creation (1), treatment field generation (2), plan optimization (1) and plan approval (1) (Supplementary Table 2). None of these failure modes were higher relative risk compared to the manual treatment planning process. A detailed summary of the failure modes and associated mitigations is shown in Supplementary Table 2. Several code modifications were prompted by FMEA. These included detailed analysis of structure volumes at the beginning and end of the algorithm to identify changes made, checks at entry of the script algorithm that PTV and organ at risk volumes are approved and cannot be edited, enforcement of naming conventions for structure sets, course and plans to minimize risk of unintentional use of an automated plan.

*Validation and Clinical Implementation: Round 1*

We retrospectively validated AVI-planner in 52 patients, which consisted of mostly men (69%) with locally advanced oropharynx (40%) or oral cavity and salivary (31%) cancers (Figure 2, Supplementary Table 3). Definitive intent organ-preservation RT comprised the majority of plans (58%) with a median of 70 Gy (range 54-80 Gy) in 35 fractions (range 27-35 fx). 62% received concurrent chemotherapy.

Overall, 86% of Round 1 plans were safe to treat; however, we identified variability in plan quality among different HN subsites (Figure 2). All oropharynx and p16+ SCC Unknown Primary (21/21 plans), larynx and hypopharynx (7/7 plans) were “treat as is.” Similarly, most oral cavity and salivary cases (14/16 plans; 87.5%) required no revisions. This contrasts with the

frequency of major revisions or rejections recommended for sinonasal, nasopharynx and p16-negative SCC Unknown Primary (1/6 plans; 17%) and cutaneous (1/2 plans; 50%- Figure 2). Minor revisions were increasing conformality and reducing heterogeneity. Major revisions were limiting hot spots outside PTV, restricting hot spots within PTV to 105-110%, and improving target coverage. Sample Round 1 isodose distributions are shown for both a definitive early stage p16+ base of tongue cancer considered "treat as is" (Figure 3B), compared to "major revisions" for an adjuvantly treated malar cheek Merkel cell carcinoma (Figure 3E). Software refinements were made to the script in response to the Round 1 evaluation which included normal tissue constraints and priorities, dose-sculpting structures, segmented structures, and isocenter placement as discussed above in Methods section *Write-enabled Script Refinement and Clinical Deployment*.

*Clinical Reassessment: Round 2*

During Round 2 evaluation of all 52 plans, there were no rejections nor major revisions (Figure 2). Minor revisions were recommended for 1 oral cavity (6.3%) and 3 sinonasal or nasopharynx or p16-negative SCC Unknown Primary plans (50%). The remaining 48 plans were "treat as is." Minor revisions in Round 2 focused on improving conformality, or more aggressive sparing of spinal cord, optics and contralateral orbit or salivary structures. This evolution in quality is evident for the adjuvantly treated Merkel cell carcinoma (Figure 3D-F).

Conformality and dose heterogeneity for all PTV levels were similar between Round 2 AVI-planner and clinical cases (Table 1). The number of MU per plan was significantly lower for AVI-planner Round 2 (mean 619.7 ± 69.7 MU) compared to the clinically treated plan (693.5 ± 219.4 MU; p=0.03). Similarly, AVI-planner generated less complex plans as compared to clinically treated plans (mean complexity score 0.13 ± 0.02 vs 0.14 ± 0.03; $p<0.01$). To evaluate

patterns of OAR sparing in Round 2 AVI-planner, we compared the entire distribution of mean or D0.1cc dose values between AVI-planner cases versus clinically treated or historically accepted plan values (Figure 4A). Given that oropharynx was the most prevalent HN subsite within both the foundational library and the validation cohort, we also selected a representative DVH from a locally advanced oropharynx cancer treated with definitive chemoradiation. This demonstrates typical DVH metrics from a case which was well represented in the model (Figure 5).

Considering all 52 plans, the contralateral parotid dose distribution was higher with AVI-planner compared to clinically treated plans (median 25 vs 23 Gy, $p<0.01$), with higher doses compared to historic plans (WES 0.50 vs 0.42, $p<0.01$). Inferior pharyngeal constrictor muscles had higher distribution of mean dose in AVI-planner versus clinical plans (median 21 Gy vs 19 Gy, $p<0.01$), with higher doses compared to historic plans (WES 0.53 vs 0.35, $p<0.01$), and narrowly exceeded our constraint (GEM 0.52). Conversely, AVI-planner lowered the dose to ipsilateral SMG (62 Gy vs 65 Gy, $p=0.04$) though this was not clinically relevant (GEM >0.90) [43]. AVI-planner lowered the distribution of mean dose to the larynx compared to clinical (median 19 Gy vs 20 Gy, $p<0.01$) and historical plans (WES 0.29 vs 0.44, $p<0.01$). Brainstem D0.1cc from AVI-planner was lower than clinical plans (median 28 Gy vs 32 Gy, $p<0.01$). Spinal cord distribution of D0.1cc was lower in AVI-planner as compared to historic plans (WES 0.44 vs 0.63, $p<0.01$), but similar to clinical plans (median 36.4 Gy vs 36.7 Gy, $p=0.9$). Distribution of dose to optic nerves, chiasm, eyes, contralateral SMG, superior pharyngeal constrictors, oral cavity, mandible, and esophagus were similar among AVI-planner cases, clinical and historic plans (Figure 4A).

For oropharynx or p16+ SCC Unknown Primary (n=21; Figure 4B), the distribution of mean larynx dose was significantly lower with AVI-planner versus clinical plans (median 18 vs 20 Gy, p<0.01) or historical plans (WES 0.28 vs 0.44, p<0.01), which was clinically relevant (GEM 0.46). Esophagus and ipsilateral parotid were spared equally among AVI-planner, clinical and historic plans. Distribution of mean dose to contralateral SMG was higher for AVI-planner compared to clinical plans (median 36 Gy vs 33 Gy, p=0.02) and historical plans (WES 0.44 vs 0.41, p=0.02). Neither clinical nor AVI-planner met constraints for relevant sparing (GEM 0.62 and 0.57). Contralateral parotid distribution of mean dose was higher for AVI-planner compared to clinical plans (median 25 vs 23 Gy, p=0.047), but similar to historically accepted plans (WES 0.50 vs 0.43, p=0.1). Superior and inferior pharyngeal constrictors received higher dose with AVI-planner and exceeded constraints compared to historical controls (p<0.01). Oral cavity distribution of mean dose was higher with AVI-planner versus clinical (37 vs 33 Gy, p=0.04) and historic plans (WES 0.64 vs 0.53, p=0.04; Figure 4B).

OAR doses were similar between clinical and Round 2 AVI-planner for the remaining HN subsites (Figure 4B and 4C). Of note, the oral cavity/salivary contralateral parotid distribution of mean dose was significantly higher for AVI-planner compared to clinical plans (median 26 Gy vs 23 Gy, p<0.01), and historic plans (WES 0.53 vs 0.42, p<0.01) and did not meet constraints (GEM 0.56) (Figure 4B). Two cutaneous plans did not reach the 3 plan threshold required for formal comparison.

*Time Study*

We compared Eclipse optimizer interactive time for 10 recent manual plans versus interactive time with AVI-planner software for 51 of the validation cohort patients. This interactive time included all steps of the manual optimization such as segmentation structures,

setting isocenter, ring and buffer dose sculpting structures, normal tissue optimization limits, target and OAR prioritization, setting the # arcs, optimization time. Of the 10 manually optimized plans, 70% were oropharynx (n=7), while 30% were comprised of oral cavity (n=1), thyroid (n=1), and Unknown Primary (n=1). Mean time for manual interaction time was shorter for AVI-planner vs manually optimized plans, 2 vs 85 minutes respectively ($p<0.01$).

## Discussion

We developed and implemented a knowledge-based automated virtual integrative software to facilitate HN treatment planning. Initially, we identified inconsistent plan quality among different HN subsites. Following iterative software adaptations, we noted favorable evolution in target coverage, heterogeneity and OAR sparing. This software exceeded our "warm start optimization" goal and rapidly created clinically acceptable plans without manual adjustments for many HN subsites. We have published the source code for AVI-Planner at a GitHub repository (http://xxxx.xxxx.xxxx ) to promote use and development of automated planning.

Regarding clinical acceptability of automated HN plans, we found 86% of Round 1 plans were treat as is, which is comparable to 88% by Radiation Planning Assistant [47]. Our script was developed from a diverse training dataset, capturing unique nuances and planning considerations. The inconsistent site-specific plan quality likely resulted from limited experience within the foundational library. Our heterogeneous library accrued over 5 years, but there were fewer cutaneous (7.9%), sinonasal (4.2%) and nasopharynx (3.7%) plans. Improvements in both conformality and heterogeneity were shown for prostate and cervix cancer after refining Varian's RapidPlan default settings [48], but studies detailing specific software refinements and evolution in plan quality among multiple subsites are limited for HN [35, 47].

Physicians frequently emphasized higher OAR prioritization. For instance, the clinical plan aggressively spared contralateral parotid further below the planning objective in a cT1N1 p16+ tonsil cancer, whereas AVI-planner less aggressively spared the contralateral parotid to meet constraints. Given OAR constraint heterogeneity, we compared the Round 2 AVI-planner results to consensus thresholds [42]. Structure laterality is reported, however the distinction of

ipsilateral or contralateral relative to the target is less readily available. AVI-planner achieved lower contralateral parotid doses (median 25 Gy) than the 26 Gy threshold ($p < 0.01$). AVI-planner lowered dose to optic structures, eyes, brainstem, spinal cord, esophagus, inferior constrictors and mandible compared to accepted thresholds. Larynx doses achieved in AVI-planner cases were lower than thresholds (35 Gy, $p<0.01$), and values with automation approaches by Fogliata et al. (24.8 ± 5.2 Gy) or Ouyang et al. (28.4 Gy) [26, 27]. This highlights the importance of benchmarking automation achievements against literature values, historic norms, and the validation subset.

Secondarily, the failure modes addressed during development can be found in manual planning, suggesting these hazards already exist and may be more likely to happen without the software. Thus, automated planning does not obviate standard clinical and physics QA. Similar to Wang et al. [49], inconsistencies in standardized OAR prioritization affected the performance of our model. In line with time-savings noted by other groups for autoplanned nasopharynx [31, 50, 51] and oropharynx cancers [33, 36], we confirmed time savings with AVI-planner. AVI-planner generated less complex plans ($p<0.01$) with fewer MU ($p=0.03$) compared to the clinically treated plans.

Limitations of this work include this software is integrated only with Eclipse for 30-35 fraction plans. Given the revisions required for sinonasal, nasopharynx, and cutaneous sites, this software should be used cautiously near the skull base. Three target dose levels are currently supported, but additional dose levels require manual editing. Dosimetrists must also ensure the relevance of automated decisions. For instance, planners must remain vigilant about modifying the isocenter location or number of arcs for a unilateral target. Physicians must explicitly address planning preferences in the planning directive. For example, in a locally advanced maxillary

sinus cancer requiring adjuvant RT following an orbital exenteration, aggressively sparing the remaining contralateral orbit and lacrimal gland may take precedence over PTV coverage. The user-friendly interface contains the same tools used in manual optimization, allowing real-time modification by dosimetrists, compared to fully automated optimization which must run to completion before permitting revision. However, these standardized optimization parameters likely differ from personalized approaches of experienced dosimetrists. Therefore, additional time may be required for revisions. Inter-institutional heterogeneity in delineated OARs, inconsistent OAR contouring, and variability in constraints and prioritization are barriers to widespread adoption of automated planning.

To our knowledge, this is the first report identifying HN primary site-specific variability in automated plan quality, which favorably evolved with physician input. Our work cautions against interpreting that automated planning achievements are universal among HN subsites. This is relevant for clinics that would ideally employ one planning algorithm for all HN cases, instead of separate optimization algorithms for each HN subsite. We are not advocating this software in lieu of skilled dosimetrists or treatment at high-accruing centers. However, in settings of limited resources, increased demand, urgent starts or reduced subspecialized dosimetrists, AVI-planner software can be easily integrated into workflows to increase availability of high quality HN RT plans.

Furthermore, our institutional adoption of AVI-planner into routine practice has expanded the number of dosimetrists able to rapidly generate high quality HN plans. We plan to release AVI-planner to our affiliate sites to improve plan quality and uniformity, we are also extending this standardized approach to other disease sites including lung and prostate. In the future, automated planning will facilitate adaptive RT planning. Future software upgrades may

incorporate gEUD, update the foundational library, and focus on well-lateralized cases near the skin surface. Application programming interfaces that enable clinics to programmatically automate all parts of the treatment planning process, give clinics the tools they need to increase efficiency and consistency in plan quality in their process workflows. To promote these clinical improvements it is highly desirable for manufactures to provide application programming interfaces that give users at minimum the same capabilities they have in manual operations to algorithmically interact with optimizers, dose calculation engines, reference points, and scheduling capabilities.

## References


1. Hawkins, P.G., et al., *Organ-Sparing in Radiotherapy for Head-and-Neck Cancer: Improving Quality of Life.* Semin Radiat Oncol, 2018. **28**(1): p. 46-52.
2. Nutting, C.M., et al., *Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial.* Lancet Oncol, 2011. **12**(2): p. 127-36.
3. Pow, E.H., et al., *Xerostomia and quality of life after intensity-modulated radiotherapy vs. conventional radiotherapy for early-stage nasopharyngeal carcinoma: initial report on a randomized controlled clinical trial.* Int J Radiat Oncol Biol Phys, 2006. **66**(4): p. 981-91.
4. Kam, M.K., et al., *Prospective randomized study of intensity-modulated radiotherapy on salivary gland function in early-stage nasopharyngeal carcinoma patients.* J Clin Oncol, 2007. **25**(31): p. 4873-9.
5. Dirix, P. and S. Nuyts, *Evidence-based organ-sparing radiotherapy in head and neck cancer.* Lancet Oncol, 2010. **11**(1): p. 85-91.
6. Lee, N., et al., *Intensity-modulated radiation therapy in head and neck cancers: an update.* Head Neck, 2007. **29**(4): p. 387-400.
7. Eisbruch, A., et al., *Parotid gland sparing in patients undergoing bilateral head and neck irradiation: techniques and early results.* Int J Radiat Oncol Biol Phys, 1996. **36**(2): p. 469-80.
8. Murdoch-Kinch, C.A., et al., *Dose-effect relationships for the submandibular salivary glands and implications for their sparing by intensity modulated radiotherapy.* Int J Radiat Oncol Biol Phys, 2008. **72**(2): p. 373-82.
9. Feng, F.Y., et al., *Intensity-modulated chemoradiotherapy aiming to reduce dysphagia in patients with oropharyngeal cancer: clinical and functional results.* J Clin Oncol, 2010. **28**(16): p. 2732-8.
10. Beadle, B.M., et al., *Improved survival using intensity-modulated radiation therapy in head and neck cancers: a SEER-Medicare analysis.* Cancer, 2014. **120**(5): p. 702-10.
11. Boero, I.J., et al., *Importance of Radiation Oncologist Experience Among Patients With Head-and-Neck Cancer Treated With Intensity-Modulated Radiation Therapy.* J Clin Oncol, 2016. **34**(7): p. 684-90.
12. Lee, C.C., et al., *Survival rate in nasopharyngeal carcinoma improved by high caseload volume: a nationwide population-based study in Taiwan.* Radiat Oncol, 2011. **6**: p. 92.
13. Cilla, S., et al., *Template-based automation of treatment planning in advanced radiotherapy: a comprehensive dosimetric and clinical evaluation.* Sci Rep, 2020. **10**(1): p. 423.
14. Batumalai, V., et al., *How important is dosimetrist experience for intensity modulated radiation therapy? A comparative analysis of a head and neck case.* Pract Radiat Oncol, 2013. **3**(3): p. e99-e106.
15. Moore, K.L., et al., *Experience-based quality control of clinical intensity-modulated radiotherapy planning.* Int J Radiat Oncol Biol Phys, 2011. **81**(2): p. 545-51.
16. Nelms, B.E., et al., *Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems.* Pract Radiat Oncol, 2012. **2**(4): p. 296-305.

17. Eisbruch, A., et al., *Multi-institutional trial of accelerated hypofractionated intensity-modulated radiation therapy for early-stage oropharyngeal cancer (RTOG 00-22).* Int J Radiat Oncol Biol Phys, 2010. **76**(5): p. 1333-8.
18. Zhong, H., et al., *The Impact of Clinical Trial Quality Assurance on Outcome in Head and Neck Radiotherapy Treatment.* Front Oncol, 2019. **9**: p. 792.
19. Peters, L.J., et al., *Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02.* J Clin Oncol, 2010. **28**(18): p. 2996-3001.
20. Graboyes, E.M., et al., *Association of Treatment Delays With Survival for Patients With Head and Neck Cancer: A Systematic Review.* JAMA Otolaryngol Head Neck Surg, 2019. **145**(2): p. 166-177.
21. Rosenthal, D.I., et al., *Importance of the treatment package time in surgery and postoperative radiation therapy for squamous carcinoma of the head and neck.* Head Neck, 2002. **24**(2): p. 115-26.
22. Wuthrick, E.J., et al., *Institutional clinical trial accrual volume and survival of patients with head and neck cancer.* J Clin Oncol, 2015. **33**(2): p. 156-64.
23. Naghavi, A.O., et al., *Patient choice for high-volume center radiation impacts head and neck cancer outcome.* Cancer Med, 2018. **7**(10): p. 4964-4979.
24. George, J.R., S.S. Yom, and S.J. Wang, *Combined modality treatment outcomes for head and neck cancer: comparison of postoperative radiation therapy at academic vs nonacademic medical centers.* JAMA Otolaryngol Head Neck Surg, 2013. **139**(11): p. 1118-26.
25. Hussein, M., et al., *Automation in intensity modulated radiotherapy treatment planning-a review of recent innovations.* Br J Radiol, 2018. **91**(1092): p. 20180270.
26. Ouyang, Z., et al., *Evaluation of auto-planning in IMRT and VMAT for head and neck cancer.* J Appl Clin Med Phys, 2019. **20**(7): p. 39-47.
27. Fogliata, A., et al., *RapidPlan head and neck model: the objectives and possible clinical benefit.* Radiat Oncol, 2017. **12**(1): p. 73.
28. Krayenbuehl, J., et al., *Evaluation of an automated knowledge based treatment planning system for head and neck.* Radiat Oncol, 2015. **10**: p. 226.
29. Tol, J.P., et al., *Evaluation of a knowledge-based planning solution for head and neck cancer.* Int J Radiat Oncol Biol Phys, 2015. **91**(3): p. 612-20.
30. Gintz, D., et al., *Initial evaluation of automated treatment planning software.* J Appl Clin Med Phys, 2016. **17**(3): p. 331-346.
31. Giaddui, T., et al., *Offline Quality Assurance for Intensity Modulated Radiation Therapy Treatment Plans for NRG-HN001 Head and Neck Clinical Trial Using Knowledge-Based Planning.* Adv Radiat Oncol, 2020. **5**(6): p. 1342-1349.
32. Hansen, C.R., et al., *Automatic treatment planning improves the clinical quality of head and neck cancer treatment plans.* Clin Transl Radiat Oncol, 2016. **1**: p. 2-8.
33. Kusters, J., et al., *Automated IMRT planning in Pinnacle : A study in head-and-neck cancer.* Strahlenther Onkol, 2017. **193**(12): p. 1031-1038.
34. Krayenbuehl, J., et al., *Planning comparison of five automated treatment planning solutions for locally advanced head and neck cancer.* Radiat Oncol, 2018. **13**(1): p. 170.
35. Fogliata, A., et al., *RapidPlan knowledge based planning: iterative learning process and model ability to steer planning strategies.* Radiat Oncol, 2019. **14**(1): p. 187.

36. Kamima, T., et al., *Multi-institutional evaluation of knowledge-based planning performance of volumetric modulated arc therapy (VMAT) for head and neck cancer*. Phys Med, 2019. **64**: p. 174-181.
37. Ahunbay, E.E., O. Ates, and X.A. Li, *An online replanning method using warm start optimization and aperture morphing for flattening-filter-free beams.* Med Phys, 2016. **43**(8): p. 4575.
38. Paradis, K.C., et al., *The Fusion of Incident Learning and Failure Mode and Effects Analysis for Data-Driven Patient Safety Improvements.* Pract Radiat Oncol, 2020.
39. Huq, M.S., et al., *The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation therapy quality management.* Med Phys, 2016. **43**(7): p. 4209.
40. Biau, J., et al., *Selection of lymph node target volumes for definitive head and neck radiation therapy: a 2019 Update.* Radiother Oncol, 2019. **134**: p. 1-9.
41. Gregoire, V., et al., *Delineation of the primary tumour Clinical Target Volumes (CTV-P) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO, TROG consensus guidelines.* Radiother Oncol, 2018. **126**(1): p. 3-24.
42. Lee, A.W., et al., *International Guideline on Dose Prioritization and Acceptance Criteria in Radiation Therapy Planning for Nasopharyngeal Carcinoma.* Int J Radiat Oncol Biol Phys, 2019. **105**(3): p. 567-580.
43. Mayo, C.S., et al., *Incorporating big data into treatment plan evaluation: Development of statistical DVH metrics and visualization dashboards.* Adv Radiat Oncol, 2017. **2**(3): p. 503-514.
44. Baltas, D., et al., *A conformal index (COIN) to evaluate implant quality and dose specification in brachytherapy.* Int J Radiat Oncol Biol Phys, 1998. **40**(2): p. 515-24.
45. *ICRU report 83 Prescribing, recording, and reporting photon-beam intensity-modulated radiation therapy (IMRT).* J ICRU 10:35–36, 2010.
46. Younge, K.C., et al., *Predicting deliverability of volumetric-modulated arc therapy (VMAT) plans using aperture complexity analysis.* J Appl Clin Med Phys, 2016. **17**(4): p. 124-131.
47. Olanrewaju, A., et al., *Clinical Acceptability of Automated Radiation Treatment Planning for Head and Neck Cancer Using the Radiation Planning Assistant.* Pract Radiat Oncol, 2021. **11**(3): p. 177-184.
48. Hussein, M., et al., *Clinical validation and benchmarking of knowledge-based IMRT and VMAT treatment planning in pelvic anatomy.* Radiother Oncol, 2016. **120**(3): p. 473-479.
49. Wang, Y., B.J.M. Heijmen, and S.F. Petit, *Knowledge-based dose prediction models for head and neck cancer are strongly affected by interorgan dependency and dataset inconsistency.* Med Phys, 2019. **46**(2): p. 934-943.
50. Chang, A.T.Y., et al., *Comparison of Planning Quality and Efficiency Between Conventional and Knowledge-based Algorithms in Nasopharyngeal Cancer Patients Using Intensity Modulated Radiation Therapy.* Int J Radiat Oncol Biol Phys, 2016. **95**(3): p. 981-990.
51. Hu, J., et al., *Quantitative Comparison of Knowledge-Based and Manual Intensity Modulated Radiation Therapy Planning for Nasopharyngeal Carcinoma.* Front Oncol, 2020. **10**: p. 551763.

**Figure Legends**

Figure 1: Flow chart depicting write-enabled script development and release process.

Figure 2: Highlighting the importance of disease subsite-specific AVI-planner algorithm performance. We show an integrated pie and doughnut chart demonstrating clinical acceptability of AVI-planned cases among H&N subsites. Central chart (blue) shows plan frequency by subsite (n=52). Innermost doughnut chart shows "Round 1" clinical acceptability. Outermost doughnut chart "Round 2" shows evolution in acceptability for 7 plans initially requiring revisions.

Figure 3: Identifying the strengths and limitations of automated planning among typical versus atypical HN cases. We show sample plans for early-stage oropharynx and adjuvant Merkel cell carcinoma. Top panel shows a stage II cT2N2M0 p16+ squamous cell carcinoma of the right base of tongue treated with definitive chemoradiation to 70 Gy in 35 fx (A-C), which is considered a typical case and well-represented within the model. The bottom panel shows a stage III pT1 pN1a(sn) Merkel Cell carcinoma treated adjuvantly to 54 Gy in 30 fx (D-F), which is considered atypical and underrepresented within the model. Left panels (A,D) Clinically Treated Plan. Middle panels (B) Round 1 AVI-planner "treat as is;" (E) Round 1 AVI-planner "major revision" due to conformality and 119% hotspot outside PTV. Right panels (C, F) Round 2 following AVI-planner upgrades. Isodose lines (absolute dose, Gy) show 75 (light green), 70 (white), 65 (pink), 60 (red), 54 (yellow), 51 (green), 45 (orange), 40 (purple), 30 (cyan), 25 (green), 20 (dark blue).

Commented [JE(1)]: Rework why we are showing this

Figure 4: Differential normal tissue sparing by clinical versus AVI-planner. Comparisons for all HN subsites (A); oropharynx, p16+ SCC Unknown Primary, Oral Cavity, Salivary, Larynx,

Hypopharynx (B); Sinonasal, Nasopharynx, and p16-negative SCC Unknown Primary (C). OARs are listed on y-axis with corresponding constraint. Dose (Gy) on x-axis. Box plots of clinical (blue) or Round 2 AVI-planner (yellow) provide median, IQR, minimum, and maximum doses. Red "x" denotes consensus thresholds [42]. The difference in normal tissue sparing between AVI-planner and clinical plan is typically small compared to the difference in relation to established thresholds. Statistical significance was achieved with $p<0.05$ on one-sided Kolmogrorov-Smirnov test. Symbols along y-axis indicate statistically significant difference in OAR sparing between AVI-planner versus clinical plans: total cohort (panel A-circles), Oropharynx and p16+ SCC Unknown Primary (panel B-stars), Oral Cavity and Salivary (panel B- diamonds). Filled shapes indicate AVI-planner significantly improved sparing whereas unshaded symbols indicate clinical plan achieved significantly better sparing.

Figure 5: Representative typical dose-volume histogram for a cT4N1M0 p16+ squamous cell carcinoma of the left tonsil treated definitively to 70 Gy, comparing manual plan (squares) versus AVI-planner (triangles). X-axis is Dose (Gy), Y-axis volume (%).

Table 1. Conformality and heterogeneity (ICRU 83) indexes of clinical and Round 2 AVI-planner for high, intermediate, and low PTV.

Table

Table 1. Conformality and heterogeneity (ICRU 83) indexes of clinical and Round 2 AVI-planner for high, intermediate and low PTV.

| | PTV_High | | | PTV_Mid | | | PTV_Low | | |
|---|---|---|---|---|---|---|---|---|---|
| | Clinical Plan | AVI-planner | p-value | Clinical Plan | AVI-planner | p-value | Clinical Plan | AVI-planner | p-value |
| Conformality Index | 1.1 ± 0.7 | 1.2 ± 0.7 | 0.7 | 1.3 ± 0.4 | 1.4 ± 0.5 | 0.7 | 1.3 ± 0.2 | 1.3 ± 0.2 | 0.8 |
| Heterogeneity Index | 1.1 ± 0.04 | 1.1 ± 0.03 | 0.5 | 1.1 ± 0.04 | 1.1 ± 0.04 | 0.4 | 1.2 ± 0.1 | 1.2 ± 0.1 | 0.4 |

Table 1. Conformality and heterogeneity (ICRU 83) indexes of clinical and Round 2 AVI-planner for high, intermediate and low PTV.

| | PTV_High | | | PTV_Mid | | | PTV_Low | | |
|---|---|---|---|---|---|---|---|---|---|
| | Clinical Plan | AVI Planner | p-value | Clinical Plan | AVI Planner | p-value | Clinical Plan | AVI Planner | p-value |
| Conformality Index | 1.1 ± 0.7 | 1.2 ± 0.7 | 0.7 | 1.3 ± 0.4 | 1.4 ± 0.5 | 0.7 | 1.3 ± 0.2 | 1.3 ± 0.2 | 0.8 |
| Heterogeneity Index | 1.1 ± 0.04 | 1.1 ± 0.03 | 0.5 | 1.1 ± 0.04 | 1.1 ± 0.04 | 0.4 | 1.2 ± 0.1 | 1.2 ± 0.1 | 0.4 |

Figure 1: Flow chart depicting write-enabled script development and release process.

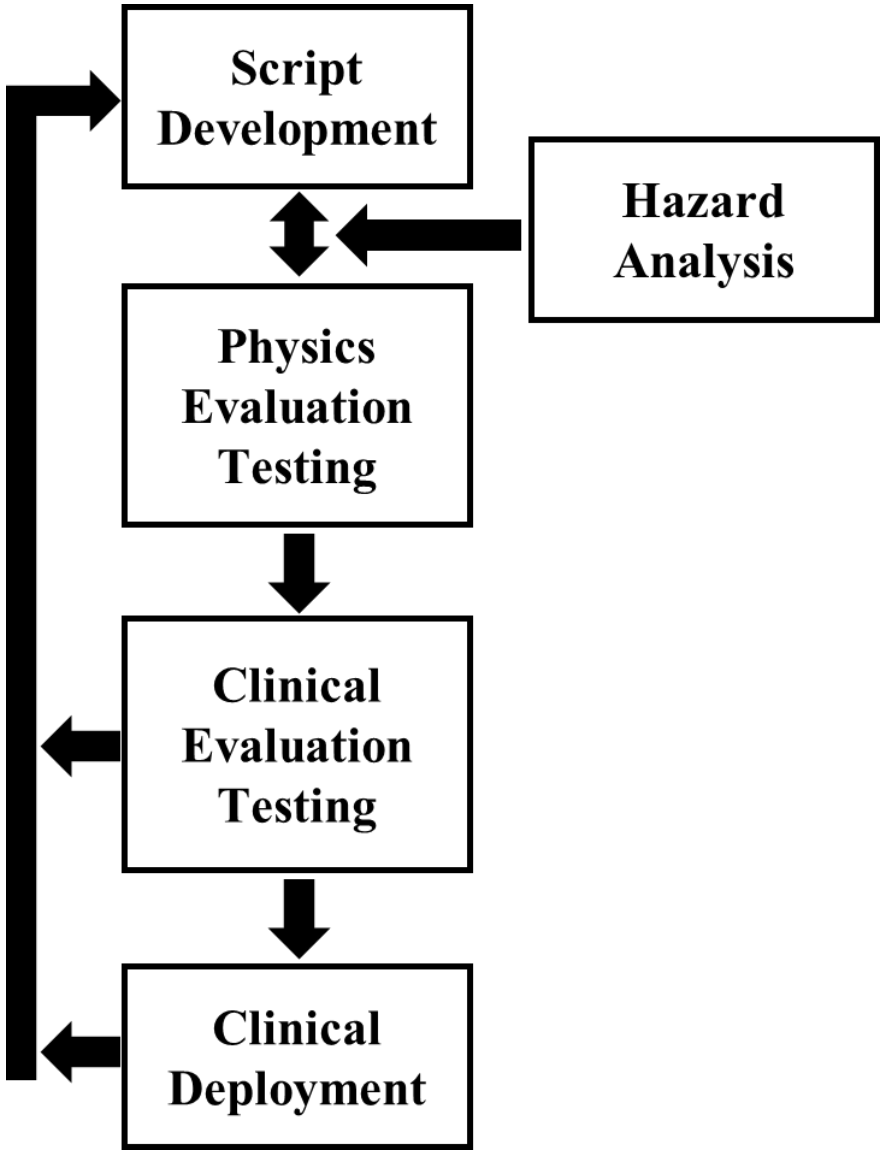
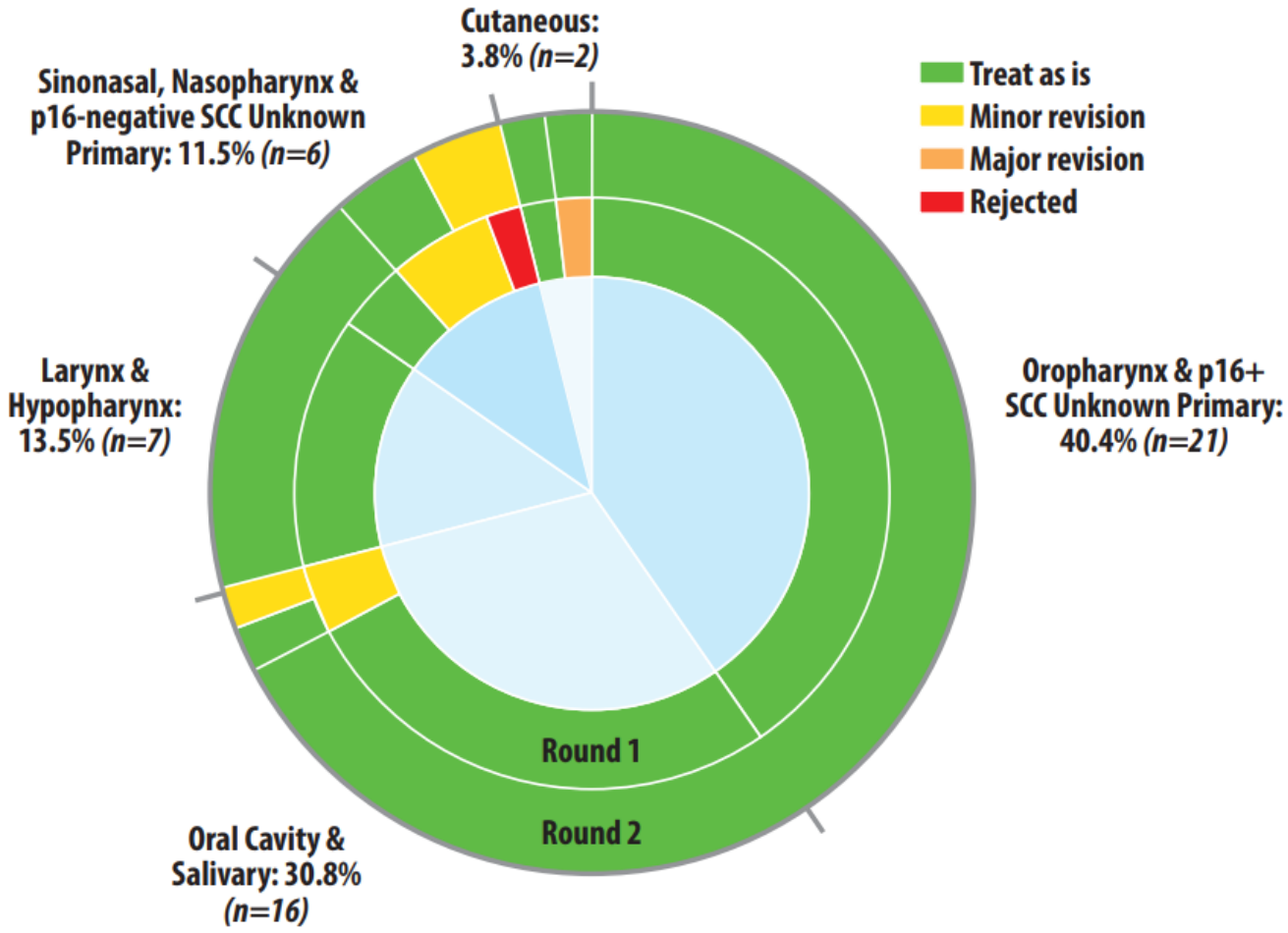
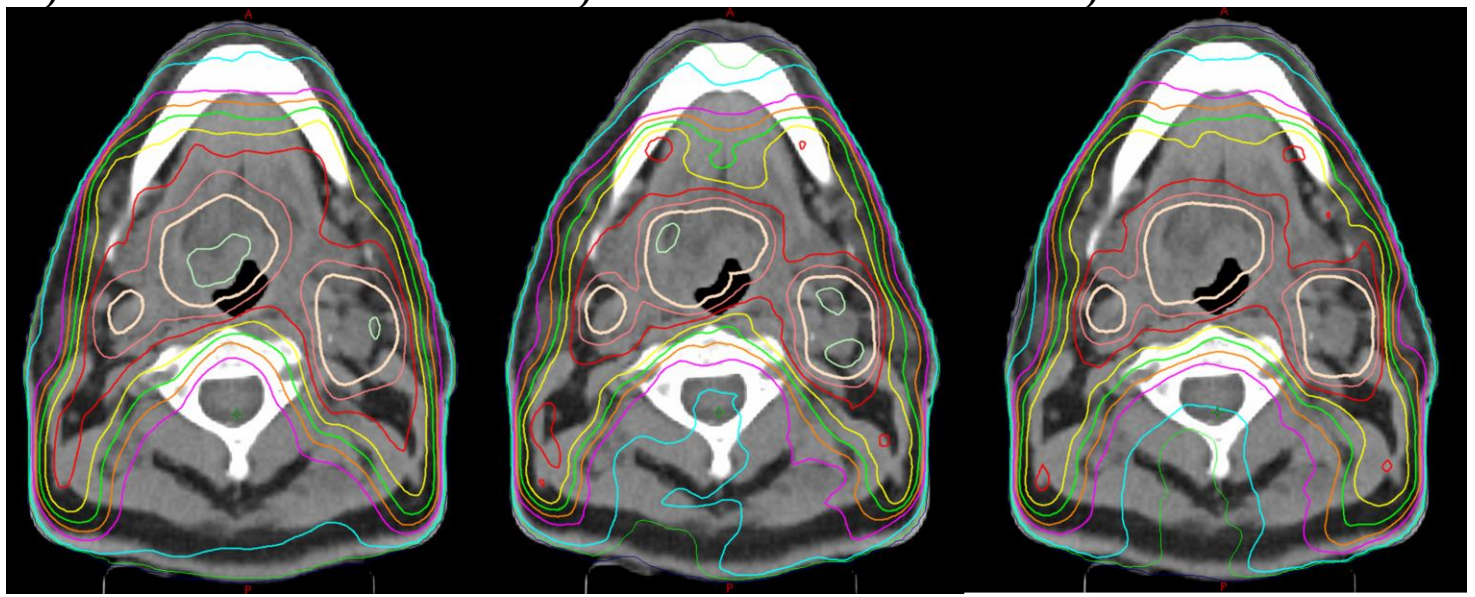


Figure 2: Integrated pie and doughnut charts demonstrating clinical acceptability of AVI-planned cases among H&N subsites.

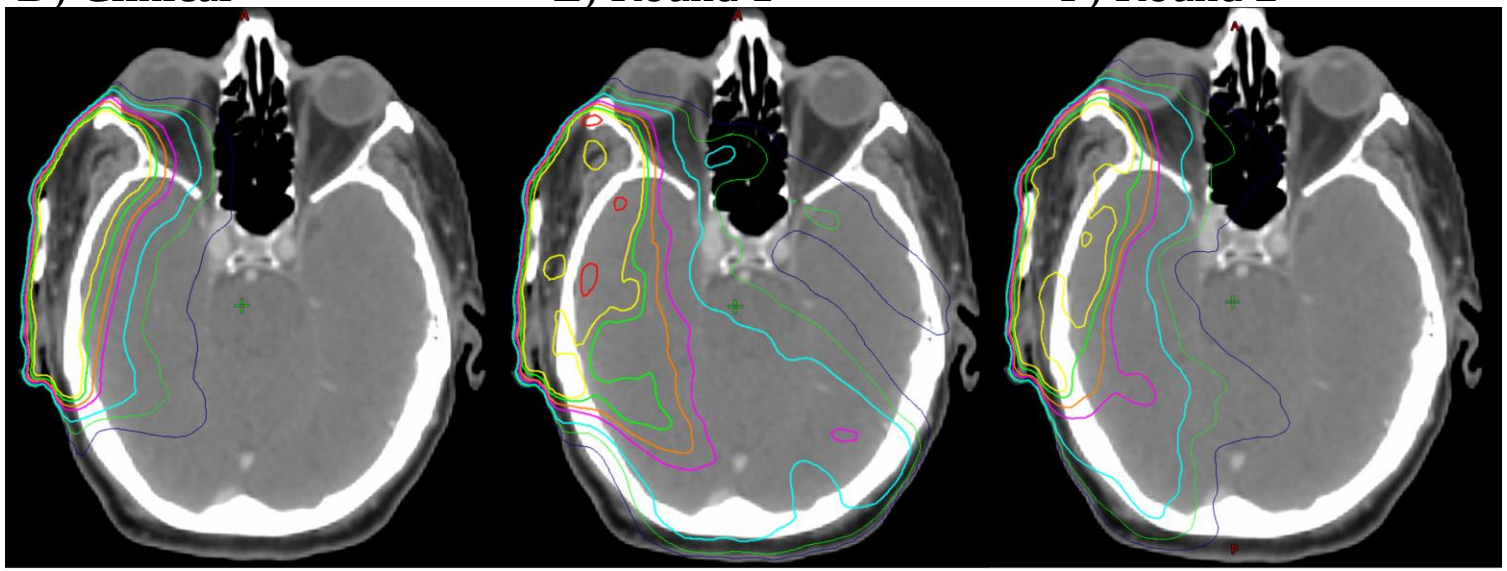


# Figure 3: Sample H&N RT plans for early-stage oropharynx and adjuvant Merkel cell carcinoma.

A) Clinical B) Round 1 C) Round 2



D) Clinical E) Round 1 F) Round 2



# Figure 4: Differential normal tissue sparing by clinical versus AVI-planner.

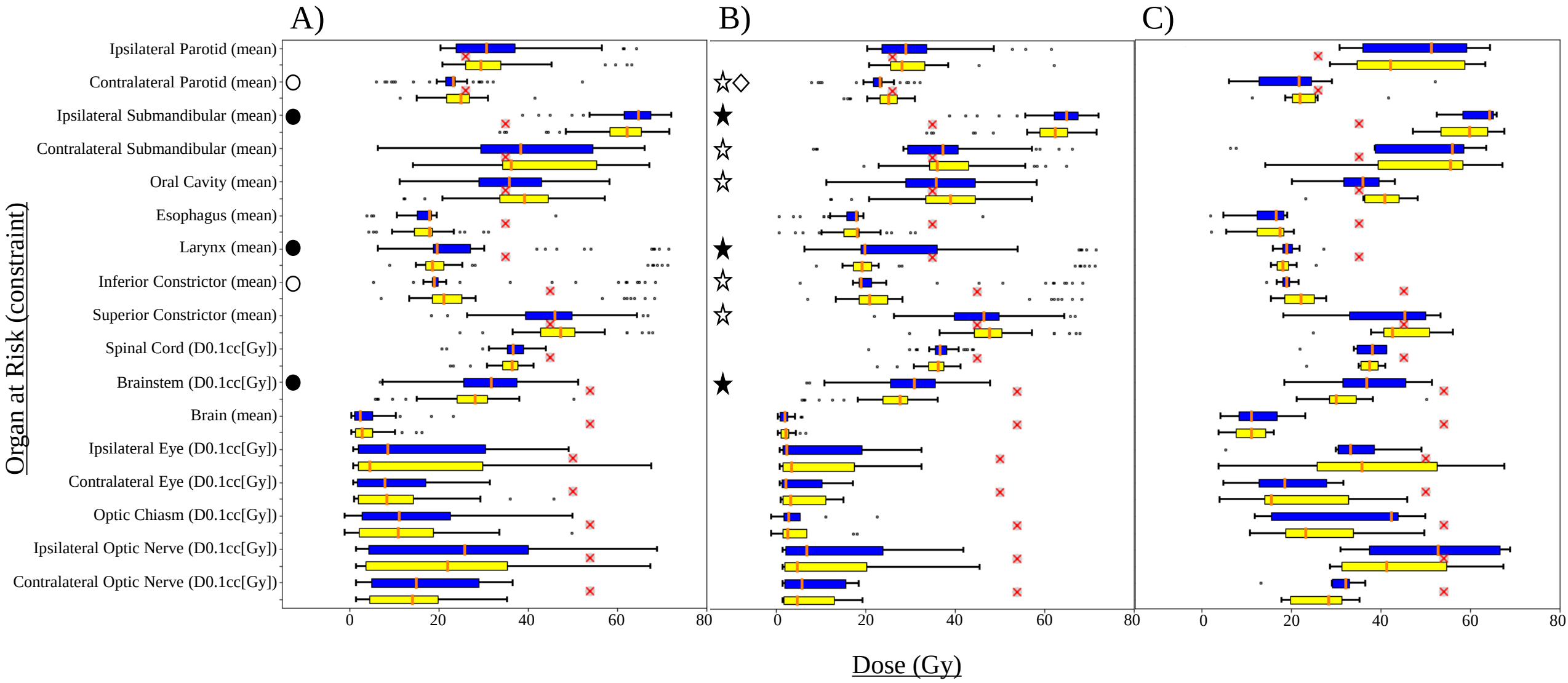


Figure 5: Box and whisker plot comparing dosimetrist interactive time between AVI-planner and manual optimization.

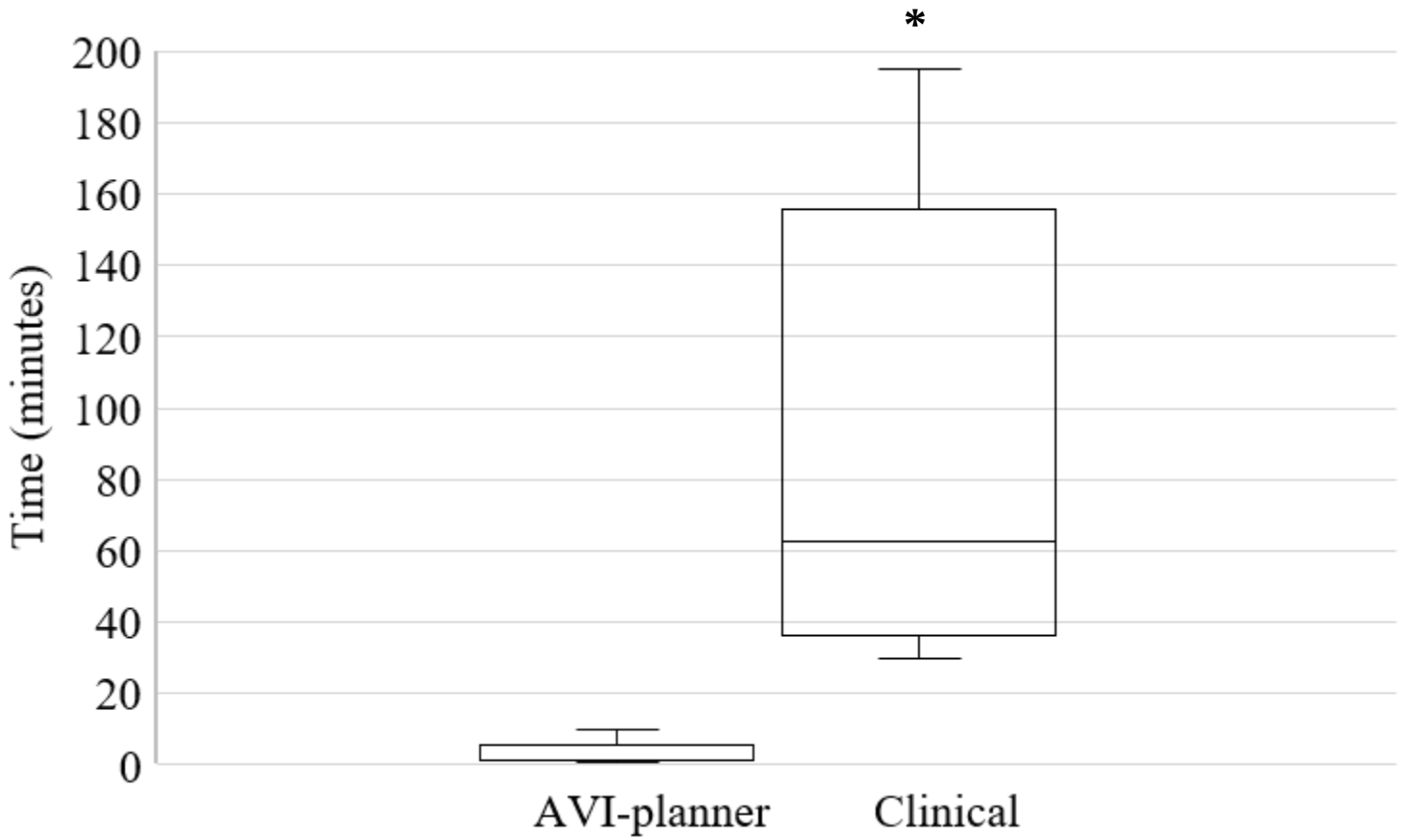


Figure 1

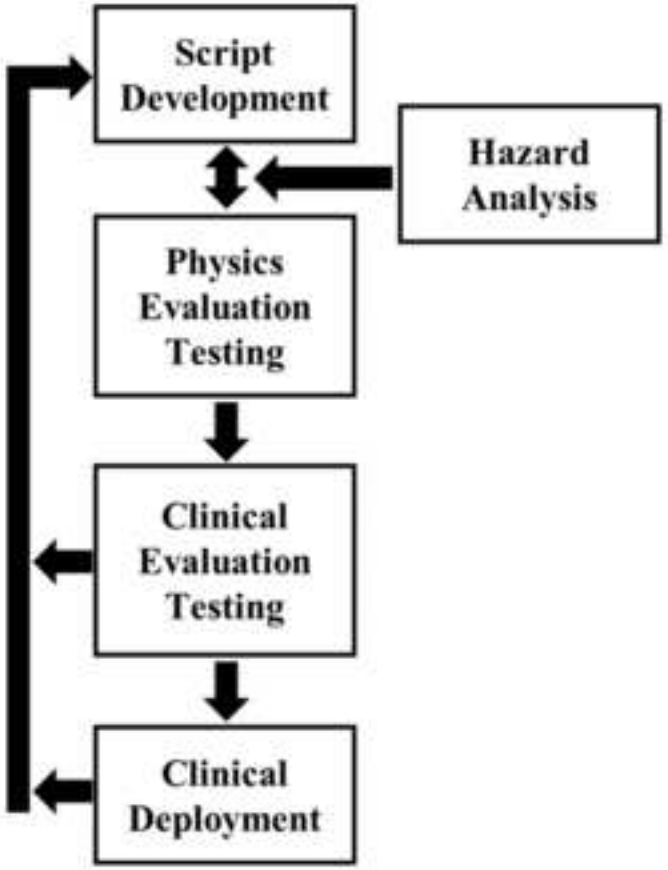


Figure 2

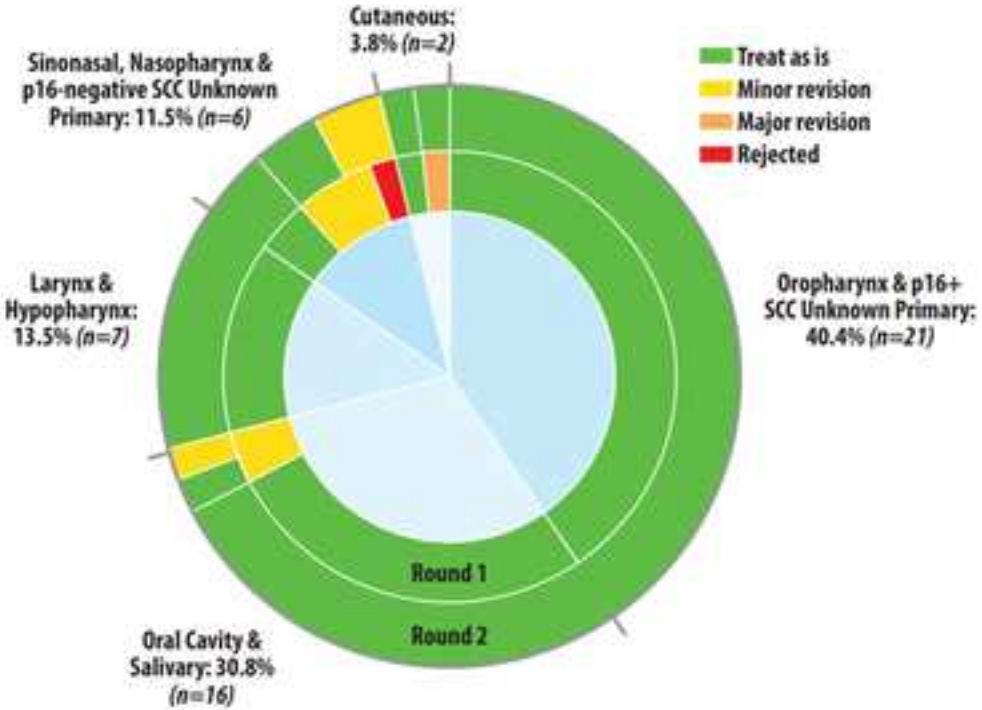


Figure 3

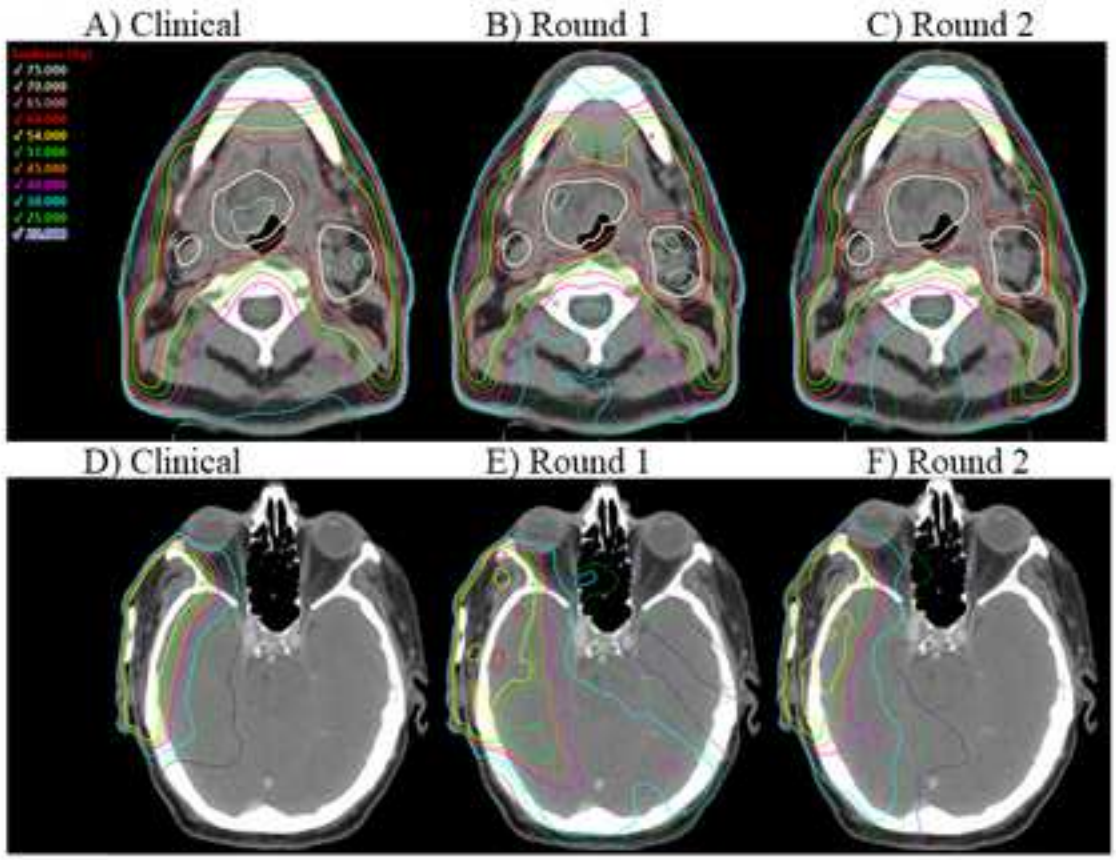


Figure 4

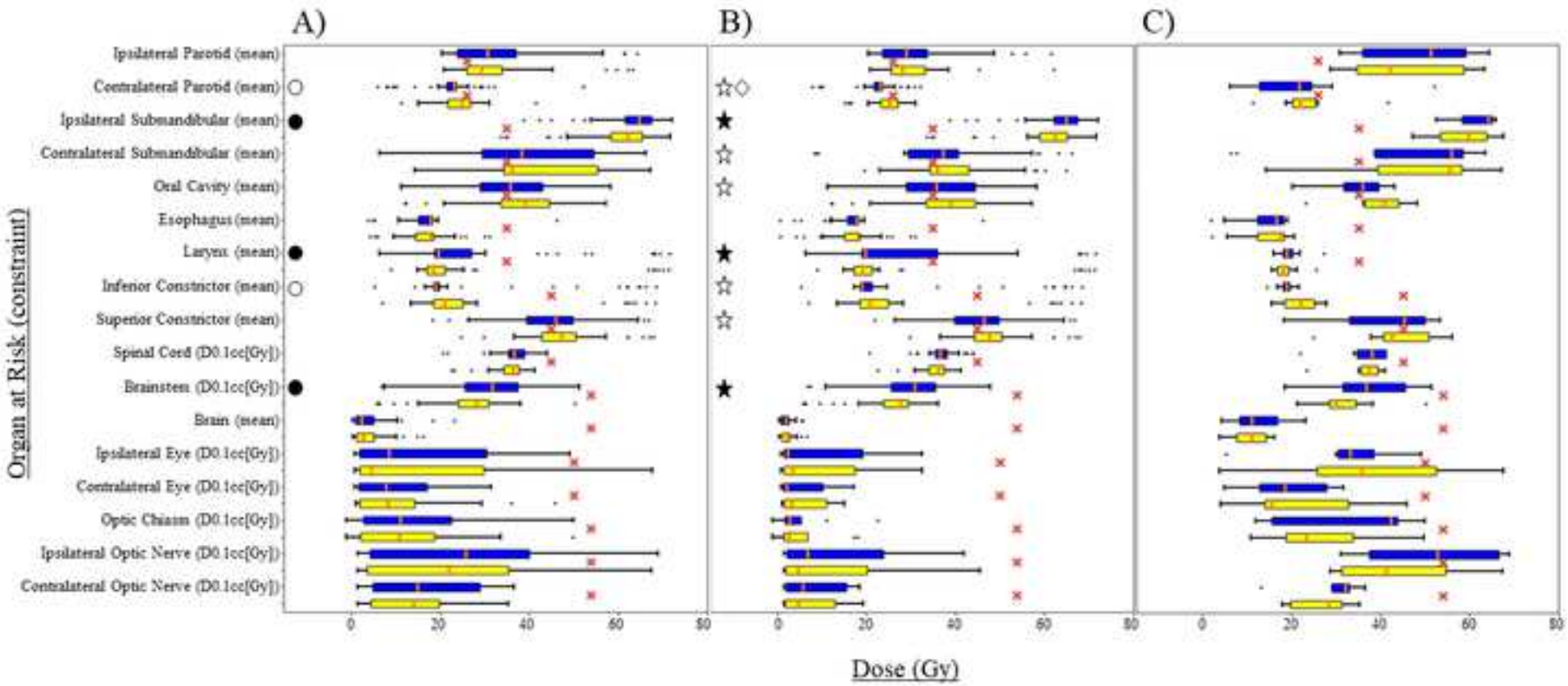
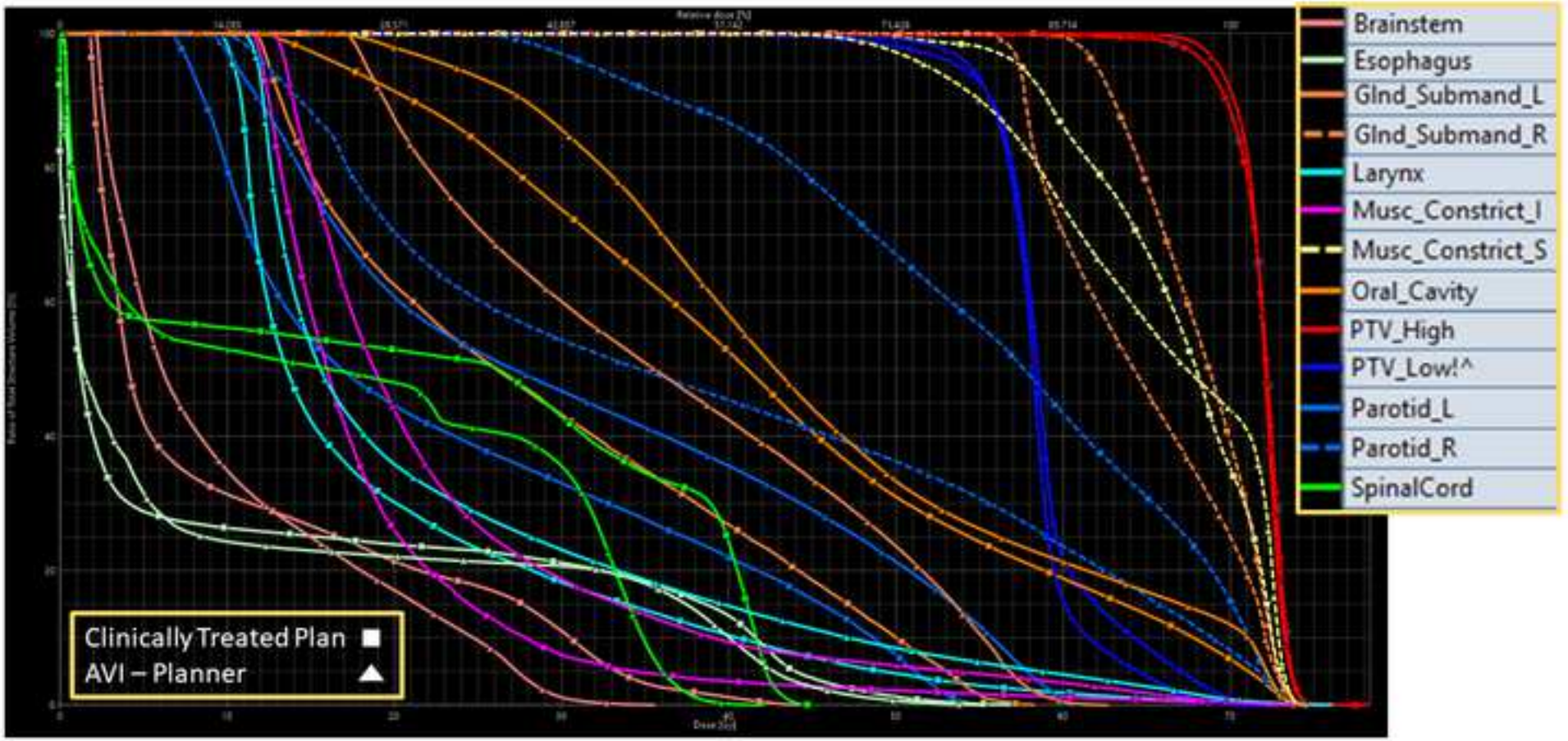


Figure 5



**Declaration of interests**

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: