


Can the Student Outperform the Master? A Plan Comparison Between Pinnacle Auto-Planning and Eclipse knowledge-Based RapidPlan Following a Prostate-Bed Plan Competition

April Smith, MS¹, Andrew Granatowicz, MS¹, Cole Stoltenberg, MS¹, Shuo Wang, PhD¹, Xiaoying Liang, PhD², Charles A. Enke, MD¹, Andrew O. Wahl, MD¹, Sumin Zhou, PhD¹, and Dandan Zheng, PhD¹ 

Abstract

Purpose: Pinnacle Auto-Planning and Eclipse RapidPlan are 2 major commercial automated planning engines that are fundamentally different: Auto-Planning mimics real planners in the iterative optimization, while RapidPlan generates static dose objectives from estimations predicted based on a prior knowledge base. This study objectively compared their performances on intensity-modulated radiotherapy planning for prostate fossa and lymphatics adopting the plan quality metric used in the 2011 American Association of Medical Dosimetrists Plan Challenge. **Methods:** All plans used an identical intensity-modulated radiotherapy beam setup and a simultaneous integrated boost prescription (68 Gy/56 Gy to prostate fossa/lymphatics). Auto-Planning was used to retrospectively plan on 20 patients, which were subsequently employed as the library to build an RapidPlan model. To compare the 2 engines' performances, a test set including 10 patients and the Plan Challenge patient was planned by both Auto-Planning (master) and RapidPlan (student) without manual intervention except for a common dose normalization and evaluated using the plan quality metric that included 14 quantitative submetrics ranging over target coverage, spillage, and organ at risk doses. Plan quality metric scores were compared between the Auto-Planning and RapidPlan plans using the Mann-Whitney *U* test. **Results:** There was no significant difference between the overall performance of the 2 engines on the 11 test cases ($P = .509$). Among the 14 submetrics, Auto-Planning and RapidPlan showed no significant difference on most submetrics except for 2. On the Plan Challenge case, Auto-Planning scored 129.9 and RapidPlan scored 130.3 out of 150, as compared with the average score of 116.9 ± 16.4 (range: 58.2-142.5) among the 125 Plan Challenge participants. **Conclusion:** Using an innovative study design, an objective comparison has been conducted between 2 major commercial automated inverse planning engines. The 2 engines performed comparably with each other and both yielded plans at par with average human planners. Using a constant-performing planner (Auto-Planning) to train and to compare, RapidPlan was found to yield plans no better than but as good as its library plans.

Keywords

automation, treatment planning, Auto-Plan, RapidPlan, KBP

¹ Radiation Oncology, University of Nebraska Medical Center, Omaha, NE, USA

² University of Florida Proton Therapy Institute, Jacksonville, FL, USA

Corresponding Author:

Dandan Zheng, PhD, Department of Radiation Oncology, University of Nebraska Medical Center, Fred & Pamela Buffett Cancer Center, 986861 Nebraska Medical Center, Omaha, NE 68198-6861, USA.

Email: sabrinadan@gmail.com



Abbreviations

AAMD, American Association of Medical Dosimetrists; AP, Auto-Planning; DVH, dose–volume histogram; IMRT, intensity-modulated radiotherapy; KBP, knowledge-based planning; OARs, organs at risk; PQM, plan quality metric; PTV, planning target volume; RP, RapidPlan;

Received: January 07, 2019; Revised: March 15, 2019; Accepted: April 15, 2019.

Introduction

Treatment planning is an essential step of radiotherapy and is conventionally performed manually by professionally trained medical dosimetrists and physicists through time- and effort-consuming iterative steps. In the past decade, there have been intensive research and development of methods to automate treatment planning, especially inverse treatment planning. Several of these state-of-the-art algorithms are now available in commercial treatment planning systems, of which Pinnacle Auto-Planning (AP) and Eclipse RapidPlan (RP) are the 2 frontrunners. Although these 2 automation algorithms were both designed to improve the treatment planning efficiency and reduce the planner dependence of plan quality,^{1–9} they operate through 2 fundamentally different mechanisms. RapidPlan is one commercial implementation of knowledge-based planning (KBP) algorithms. This type of algorithm utilizes statistical or machine learning methods to mine the historical planning data and build predictive models to estimate the expected dose–volume histograms (DVHs) for organs at risk (OARs) on new patients. So for this type of algorithm, a library of successful plans is required as the initial input to configure the DVH estimation models, and the configured models then automatically add planning objectives to conduct plan optimization. In contrast, AP does not require a prior library of successful plans, but uses instead the iterative approach of progressive optimization that mimics the steps a skilled human planner would take, such as creating rings, hotspot or coldspot regions of interest, and planning structures from the overlaps between the targets and OARs to iteratively fine-tune the target coverage and OAR sparing results. Despite these differences, both engines have been embraced by their users since their clinical introductions and numerous studies have been reported on implementing as well as utilizing these new tools.^{1–21}

Although the rapid growing body of literature has proven the clinical utilities of both AP and RP in improving planning efficiency and reducing plan quality variability among planners with varying levels of experience, a natural question to ask is, Which one works better? To the best of our knowledge, there has not been such a comparison between these 2 distinct algorithms. One may get a possible glimpse in the recent report of a comparison conducted by Wu *et al* on oropharyngeal cancer inverse planning using the separately developed AP technique and an in-house KBP model.²² In that study, they concluded that the 2 algorithms performed comparably by reviewing planning target volume (PTV) coverage and major OAR sparing and reporting plan reviewer (radiation oncologist) preference.

However, the authors acknowledged that because their comparison results are specific to the in-house KBP, they may not be applicable to the other KBP approaches (eg, RP). Also, because the KBP models and the AP technique were independently developed by 2 different institutions in their study, the human inputs could have heavily biased the comparison. Therefore, a performance comparison between AP and RP, especially one that is objective, is still lacking. In addition, for KBP algorithms like RP, although it is generally believed that the performance is determined and limited by the qualities of the library plans used as the knowledge base, this theory has never been verified due to the performance variability among human planners and even within the same planner on different cases. Therefore, in our study, we applied an innovative study design and an established evaluation metric that is objective and quantitative to address 2 critical questions: (1) Which one performs better, AP or RP? (2) Can RP outperform the planner whose plans are used as the knowledge base to configure the model?

To do this, our study adopted the planning objectives and evaluation metrics used in the 2011 American Association of Medical Dosimetrists (AAMD) Plan Challenge.²³ In this planning competition, a common data set together with very specific planning objectives were given to a population of treatment planners to design radiotherapy plans for prostate fossa and lymphatics with a simultaneous integrated boost. A “plan quality metric” (PQM) with 14 submetrics that address coverage, conformity, spillage, and OAR dose to bladder and rectum was then used to score the plans and rank the planners.²³ In our study, both AP and RP were objectively compared using this PQM on a cohort of postprostatectomy patients including the patient used for the Plan Challenge. To ensure a rigorous comparison, the same AP technique was used to generate the library plans for training the RP model, and no manual adjustment was performed on the AP or RP plans except for a common normalization. Using this innovative study design, we sought to exploit the inherent differences between these 2 engines, asking the question, can the student (RP) outperform the master (AP)?

Materials and Methods

Overall Study Design

As described in Figure 1, our study used an innovative workflow to objectively design the comparison and adopted the quantitative PQM used in the 2011 AAMD Plan Challenge to definitively perform the comparison. Under the approval of our

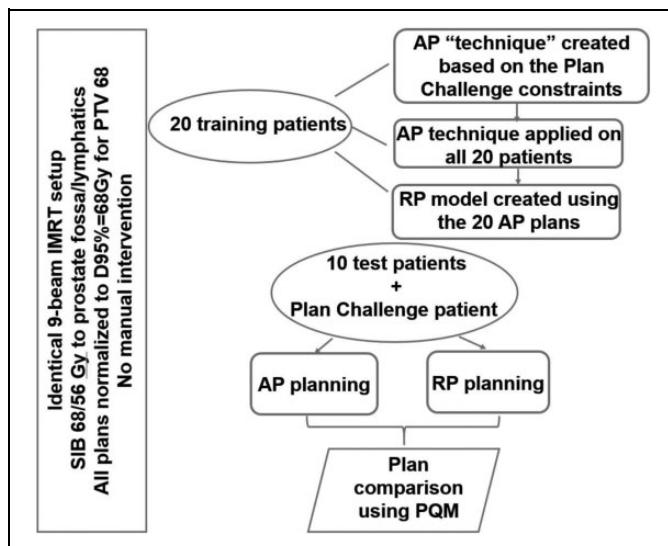


Figure 1. A schematic of our innovative study design.

institutional review board, this retrospective study involved the clinical images and structure sets of 30 postprostatectomy patients and the Plan Challenge patient. Auto-Planning was employed to generate intensity-modulated radiotherapy (IMRT) plans on 20 patients based on a “technique” developed and tested to produce reasonable DVH’s based on the constraints outlined in the AAMD Plan Challenge study. These 20 plans were subsequently used to configure the RP model. Finally, 10 other patients as well as the Plan Challenge patient were planned by both the AP technique and the RP model in parallel. The plan quality of the 2 parallel plans was quantitatively compared on each patient using the PQM. More details will be described in the following sections.

Patient Selection and Treatment Planning

Under the approval of our institutional review board, 30 patients were selected from consecutively treated patients who previously received radiotherapy to the prostate fossa and lymphatics from 2016 to 2018 in our institution. Because the treatment plans of this study were planned to deliver on a TrueBeamSTx (Varian Medical Systems, Palo Alto) that was commonly commissioned for the 2 treatment planning systems, patients requiring treatment fields larger than the 22 cm maximum Y-jaw field allowed by the TrueBeamSTx’s HDMLC were excluded. The 30 patients were then randomly divided into a modeling cohort of 20 patients and a testing cohort of 10 patients. All patients were replanned for a total dose of 56 Gy to the lymphatics with an integrated boost to the prostate fossa to 68 Gy, following the fractionation scheme of 34 fractions used in the 2011 AAMD Plan Challenge. An identical beam arrangement with 9 equally spaced coplanar 6 MV IMRT beams was used for all plans. In our study, AutoPlanning (AP) planning used Pinnacle v9.10 (Philips Medical Systems, Fitchburg) on a Smart Enterprise server and RapidPlan (RP) planning used Eclipse v13.6 (Varian Medical Systems) on a 7-FAS server.

Using the first 5 patients of the 20 modeling-cohort patients, a “technique” (a prerequisite of AP; a list of preset optimization dose goals for AP defined by the user)⁵ was developed for Pinnacle AP automated treatment planning engine (Philips Medical Systems) to yield plans trying to meet the dose objectives specified by the Plan Challenge.²³ The technique was developed based on the PQM dose objectives²³ for the varying anatomy but without driving to achieve an optimal technique for every patient. The AP technique was then used to generate plans for all 20 patients in the modeling or training cohort. In clinical practice, a user would most likely impose manual interventions in the optimization process or run the AP process more than once to improve the plan quality; however in our study, in order to ensure a rigorous comparison, all plans were from one round of AP planning and had no manual intervention.

These 20 AP plans were then exported to Eclipse (Varian Medical Systems), normalized to 100% 68-Gy prescription dose coverage to 95% of the PTV 68 volume and used to train an RP dose–volume histogram (DVH) estimation model. The same normalization was also performed on all other plans for the testing cohort that will be described below, and the normalization was always performed in Eclipse (Pinnacle AP plans were exported to Eclipse before normalization) to eliminate volume calculation and dose rounding differences between the 2 treatment planning systems.

Finally, the Pinnacle AP technique and the Eclipse RP model were used in parallel to generate plans on the 10 testing-cohort patients and the Plan Challenge patient. Again, no manual intervention in the plan generation was imposed on these plans, except for the final dose normalization performed in Eclipse.

Scoring

Both the AP and RP plans for the 10 testing-cohort patients and the Plan Challenge patient were scored using PlanIQ (Sun Nuclear Corporation, Melbourne) and applying the PQM scoring algorithm specified in the Plan Challenge.²³ The PQM has an ideal score of 150 points summed from 14 submetrics each with a unique score function. Many of the submetrics have a score function that is linear in nature, but some are quadratic or a combination of gradients. For further information on the scoring algorithm as well as other details of the Plan Challenge, please refer to the paper of Nelms *et al.*²³

Evaluation

On the 11 test cases (10 in the testing cohort and the Plan Challenge patient), the total PQM score as well as the score for each of the 14 submetrics were compared between the corresponding AP and RP plans using a 2-tailed Mann-Whitney *U* test. If a significant difference (defined as $P < .05$) was detected, a 1-tailed Mann-Whitney *U* test was then conducted. The total PQM scores of the AP and RP plans on the Plan Challenge patient were also compared with the

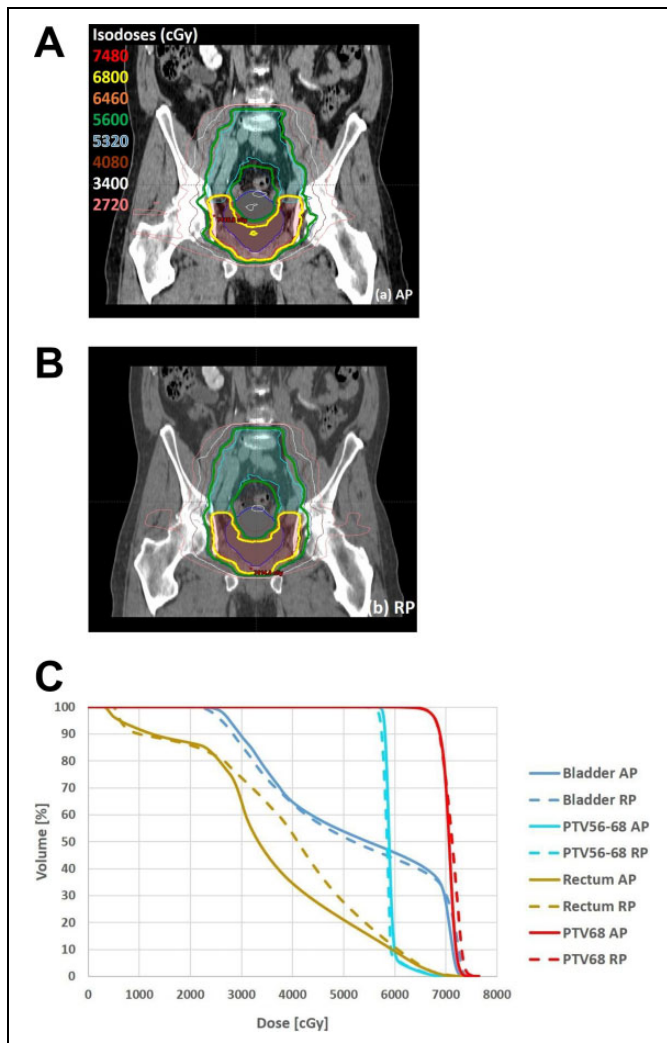


Figure 2. Isodose distributions (A, B) and the DVH (C) plots comparing the AP and RP plans on one example patient (patient 3). On the isodose distributions, the targets are shown in color wash (red is PTV 68 and cyan is PTV 56-68). AP indicates Auto-Planning; DVH, dose-volume histogram; PTV, planning target volume; RP, RapidPlan.

reported scores achieved by the 2011 AAMD Plan Challenge participants.²³

Results

Total PQM Comparison Between AP and RP

The isodose distribution and DVHs are shown for one example patient in Figure 2 to compare the AP and RP plans. Both plans appeared to have a dose distribution conformal to the targets and sparing the OARs. The maximum dose (hotspot) in the RP plan was slightly higher than the AP plan (112.5% vs 110.2%). On all 11 test cases, AP outsourced RP on 7 cases and RP outsourced AP on the remaining 4 (Figure 3). On the total PQM score, there was no significant difference between the AP and RP plans ($P = .509$).

Submetric Comparison Between AP and RP

Table 1 lists the average scores as well as the comparison P values for the 14 submetrics from the AP and RP plans. The scoring differences for most submetrics were insignificant, except for the following 2: (1) dose (Gy) covering highest 0.03 cc of PTV 68 and (2) percent of the (PTV 56-PTV 68) volume ≥ 58.8 Gy (ie, percent above 105% of 56 Gy). The first submetric describes the in-target hotspot, and for this AP outsourced RP ($P = .01$ in the 1-tailed Mann-Whitney U test) with an average score of 6.2 *versus* 3.7 (the full score for this submetric is 10), indicating a better target dose uniformity of the AP plans. The second submetric describes the percent of the (PTV 56-PTV 68) volume that receives above 105% of 56 Gy, and RP outsourced AP with an average score of 2.5 *versus* 0.4 ($P = .01$ and the full score is 10). It is worth noting that all plans were normalized in Eclipse to have 95% of PTV 68 receive the prescription dose (68 Gy), which would by definition yield a full score of 30 on the first submetric shown in the table. However, the scores for this submetric were not always perfect and that was due to the small calculation (interpolation) differences between Eclipse and PlanIQ on doses and volumes. On these plans, the PTV 56 coverage requirement, 95% of the volume receiving 56 Gy, was always met with the normalization we used. Therefore, the scores for these 2 submetrics were not compared.

Score Comparison Among AP, RP, and Human Planners on the Plan Challenge Case

From the report by Nelms *et al*, a total of 140 plans were received by the Plan Challenge from which 125 plans were accepted after rejecting extra plans from the same planner and plans with impractical treatment times.²³ Among the 125 planners, the majority were certified medical dosimetrists (81%) and with a high self-reported planner confidence level of 5 (34%) or 4 (47%) on a scale of 1-beginner to 5-master. Pinnacle (56%) and Eclipse (32%) also consisted the vast majority of the treatment planning systems used by the 125 Plan Challenge participants. As can be seen from Figure 4, both AP and RP performed at par with average human planners that participated in the Plan Challenge, with the AP and RP plan PQM scores comparing favorably with all human participant scores as well as the scores of the participants using either treatment planning systems. Specifically, out of the perfect score of 150, AP scored 129.9, RP scored 130.3, and the average score of all participants was 116.9.

Treatment Planning Efficiency

Both AP and RP are highly automated and efficient. In our experiment, only one round of AP or RP planning was performed involving no human intervention on the optimization. With our hardware/software combination for each treatment planning system, on average for each case, the total treatment planning time was about 30 minutes for AP and 10 minutes for RP. The AP treatment planning time included about 2 minutes

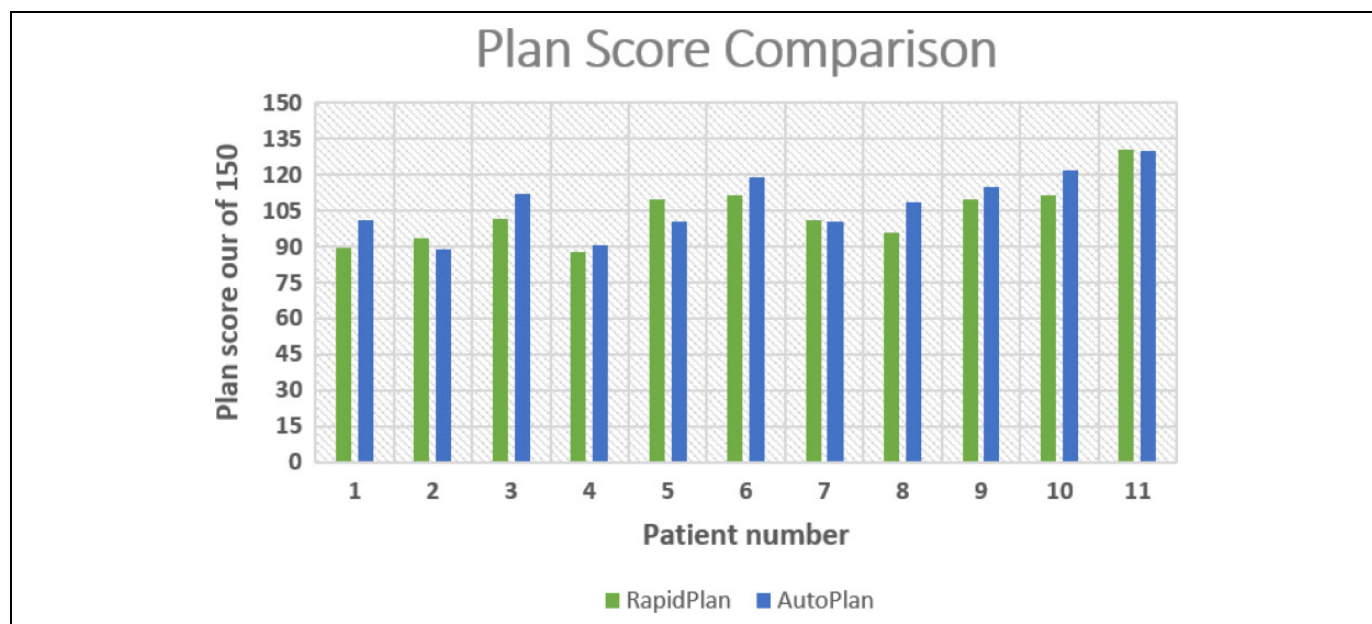


Figure 3. A comparison of the total PQM scores of the corresponding RP versus AP plans on the 11 test cases. The perfect score is 150 and patient #11 is the Plan Challenge case. AP indicates Auto-Planning; PQM, plan quality metric; RP, RapidPlan.

Table 1. Comparison of the Individual Scoring Submetrics Between AP and RP Plans on the 11 Test Cases.^a

Metric Description	<i>P</i> Value	Average AP Score	Average RP Score
[PTV 68] V68 >95%	NA	29.5	28.6
[PTV 56] V56 >95%	NA	30	30
[Prostate bed] V68 >99%	.49	9.9	10
[PTV 68] D0.03 cc <78.2 Gy	.01	6.2	3.7
Rectum V68 <10 cc	.94	4.4	4.5
[PTV 56-PTV 68] V58.8 <45%	.01	0.4	2.5
68 Gy spillage <50 cc	.06	4.7	6.5
Rectum V65 <35%	.76	9.2	9.3
Rectum V40 <45%	.94	3.8	2.9
Bladder V65 <40%	.76	2.4	2.8
Global max location	.49	4.5	4.1
Rectum serial slice evaluation	.16	-2.7	6.4
PTV 68 conformation number >0.5	.2	4.1	4
Bladder V40 <70%	.7	1.4	1.2

Abbreviations: AP, Auto-Planning; NA, not applicable; PQM, plan quality metric; RP, RapidPlan.

^aReported are the average scores as well as the *P* values (2 tailed for insignificant differences and 1 tailed for significant differences).

of human inputs, spent mostly in assigning unmatched structures. The RP treatment planning time included about 5 minutes of human inputs, spent mostly in adding beams, setting prescriptions, assigning unmatched structures, and waiting for the DVH estimation to generate.

Discussion

Using an innovative workflow to objectively design the comparison and adopting the quantitative PQM scoring system

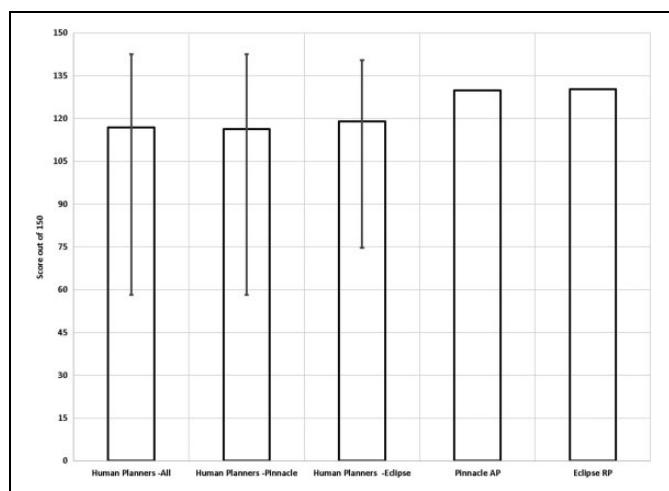


Figure 4. Comparison of the total PQM scores of the AP and RP plans on the Plan Challenge case with the human planner scores from the Plan Challenge. The average scores and score ranges were plotted for all human participants as well as all Pinnacle planners and all Eclipse planners. AP indicates Auto-Planning; RP, RapidPlan.

used in the Plan Challenge to definitively perform the comparison, our study showed that (1) as previously believed, the performance of RP, a KBP algorithm, is determined by the quality of the plans used as the knowledge base. Although the student (RP) did not outperform its master (AP), it performed as well as its master. (2) Although there were some plan quality characteristics differences, there was no significant difference between the overall performances of the 2 automated planning engines in our experiment. Both automated engines performed at par with the average human planners on the Plan Challenge case. (3) Both engines provided consistent

plan quality and an efficient treatment planning workflow that reduces the manual planning time (which was estimated to be at least an hour per plan).

On the third item described above, it is worth noting that when using the 20 training plans generated by AP to configure the RP model, no dosimetric outlier was identified, whereas in our clinical practice configuring RP models with clinically accepted historical plans, outliers have always been identified, which would then require re-planning before they could be used in the plan library for model configuration. This indicates that AP is able to produce plans with more consistent plan quality on different patients than the real planners could. Comparing the planning time between AP and RP, it was easy to notice that the total treatment planning time used by AP was much longer than RP. This could be attributed to many factors. Intrinsically, AP consists of many iterative optimization cycles and hence could take longer while the RP optimization is not iterative in nature. But at the same time, the reported planning time here was also directly related to our specific software and hardware combination for each treatment planning system as well as the maximum number of iterations we set in AP. Although the total treatment planning time for AP was longer than RP in our experiment, the human interaction time or manual planning time was shorter. This was mostly due to the fact the prescription and beam setup was defined in the AP technique and hence automatic, whereas for RP these were manual steps. Utilizing presets such as a plan template could easily further reduce the RP manual planning time. Also for both treatment planning systems, a substantial part of the manual planning time was spent in assigning unmatched structures because some of the clinical structure labels did not match the ones specified by the automated planning engines. Standardizing the clinical naming conventions could help reduce this part of manual interaction time and further automate the process for both engines. It is worth noting that although many recommended to iteratively replan the dosimetric outlier library plans for model refinement when configuring an RP model,^{24,25} a study by Delaney *et al* found that dosimetric outliers may only start to degrade the model performance when they reach a high percentage of total training plans and conversely, a small number of outliers may even lead to an improved model performance.²⁶ This seemingly counterintuitive finding was due to the fact that RP always places the OAR line objective at the inferior boundary so that a few “bad” plans will not compromise the line objective for the corresponding OAR but the line objectives for other OARs may in effect get set stricter by these “bad” plans. So in our case, the dosimetric consistency of the AP training plans, while helpful in performing an objective comparison of 2 systems, may not lead to the optimal clinical performance of RP.

In our experiment, both AP and RP plans competed favorably on the Plan Challenge case when compared with the plans by the real participants, indicating these tools, when properly used, can generate treatment plans comparable to an average planner. On the other hand, there are also some caveats about these results. First, although fully automated treatment

planning was used in our experiment, the performances of these automated planning engines were also the direct results of the human experience, such as setting up the AP technique. Second, the improvements of the treatment planning systems not related to the automated engines such as the optimization algorithms could have also favored the automated plans since these plans used more recent versions of the software (ie, Pinnacle v9.10 released in 2014 and Eclipse v13.6 released in 2015) than what were available to and used by the Plan Challenge participants in 2011. Third, our experiment was limited to a specific planning task and may not be generalized to the planning of all disease sites. Finally, while providing an objective way of quantitatively evaluating the plan quality for the comparison, using a single quantitative PQM to represent a complex multifaceted clinical evaluation of many competing factors may have its limitations, although they were partially offset by the detailed inspection of all submetrics.

Our study conducted a well-controlled experiment by involving only the automated planning engines and applying no human intervention in the treatment planning process. In clinical practice, such human interactions are routinely used to further improve the plan quality. For example, the performance of RP depends not only on the quality of DVH predictions but also on how the model is configured to turn these predictions into objectives for the plan optimization. In our work, we did not attempt to optimize the second part but instead used the default line objectives generated by the model for the OARs, as in this way less subjective bias would be introduced. In addition, because the target objectives were user set, the poorer target homogeneity in the RP plans could also reflect the suboptimal setting of these objectives in the model or suboptimal performance of the optimizer in attempting to meet these objectives together with the OAR line objectives. If such target hotspot doses are encountered in the clinical practice and deemed unacceptable, then the planner would either reoptimize the particular plan using manual structures or refine the target objectives in RP model configuration for a general trend change. Also, as shown by one of the submetrics where a significant difference was found between AP and RP, in our clinical experience, we have also observed that target conformity sometimes needs further improvement with the AP setting described in this article, so the planner often works one extra cycle of optimization or add a manual structure in the optimization to achieve a better plan. Such skilled human inputs have also been reported in the numerous studies applying these engines on diverse clinical inverse optimization problems.^{1-20,22,27-29} Interestingly, Janssen *et al* applied the DVH estimation for a plan quality audit on clinical AP plans that had been manually refined after AP if deemed necessary and identified some suboptimal plans that could be further improved.¹² This study indicated that they may not always produce Pareto-optimal plans, despite the success of these automated planning engines in producing plans at or near clinical acceptance and driving the DVHs for OARs beyond the specified dose goals. Obviously realizing this, the vendors have also been working to further improve these automated engines. For example, multicriteria optimization

has just been introduced into the newest version of Eclipse, and PlanIQ has been utilized by the newest version of Pinnacle to give anatomy-specific updates on the dose goals for AP.

Although our controlled experiment has provided useful insight on objectively comparing the 2 automated planning engines, the readers should also be aware of some limitations of our study. First of all, our study was conducted on IMRT plans for prostate fossa and lymphatics with simultaneous integrated boost, whether the conclusion could be generalized to other disease sites or planning techniques was not probed. Also, ours was a relatively small cohort study, using a total of 31 cases. The effect of including more plans for RP modeling or increasing the number of test plans for comparison was not studied. As the RP model was trained on a minimum required 20 cases, the RP model performance could have been limited by the model quality instead of the RP itself. However, the RP model seemed to be of reasonable quality judged by the vendor-supplied model training results as well as the good RP performances on the 11 test cases. Possibly benefitted from the dosimetric consistency of the AP plans and the well-dispersed anatomical distributions, the 20 training cases were able to train a reasonable model using the relatively simple training methods in RP.

Conclusion

Using an innovative study design, an objective comparison has been conducted between 2 major commercial automated inverse planning engines. The 2 engines performed comparably with each other and both yielded plans at par with average human planners. Using a constant-performing planner (AP) to train and to compare, RP was found to yield plans no better than, but as good as, its library plans.

Authors' Note

This study has been approved by the University of Nebraska Medical Center institutional review board (IRB # 128-18-EP). Informed consent was not required for the proposed retrospective medical record analysis.

Acknowledgments

The authors would like to thank Dr Benjamin E. Nelms of ProKnow for generously providing consultation and software support for this study. We would also like to thank Philips Medical Systems for the Auto-Planning research license and Sun Nuclear Corporation for the PlanIQ research license.


Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article.

ORCID iD

Dandan Zheng, PhD  <https://orcid.org/0000-0003-2259-1633>

References

- Berry SL, Ma R, Boczkowski A, Jackson A, Zhang P, Hunt M. Evaluating inter-campus plan consistency using a knowledge based planning model. *Radiother Oncol.* 2016;120(2):349-355.
- Chang ATY, Hung AWM, Cheung FWK, 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):981-990.
- Fogliata A, Belosi F, Clivio A, et al. On the pre-clinical validation of a commercial model-based optimisation engine: application to volumetric modulated arc therapy for patients with lung or prostate cancer. *Radiother Oncol.* 2014;113(3):385-391.
- Fogliata A, Nicolini G, Bourcier C, et al. Performance of a knowledge-based model for optimization of volumetric modulated arc therapy plans for single and bilateral breast irradiation. *PLoS One.* 2015;10(12): e0145137.
- Gintz D, Latifi K, Caudell J, et al. Initial evaluation of automated treatment planning software. *J Appl Clin Med Phys.* 2016;17(3): 331-346.
- Hansen CR, Bertelsen A, Hazell I, et al. Automatic treatment planning improves the clinical quality of head and neck cancer treatment plans. *Clin Transl Radiat Oncol.* 2016;1:2-8.
- Hazell I, Bzdusek K, Kumar P, et al. Automatic planning of head and neck treatment plans. *J Appl Clin Med Phys.* 2016;17(1): 272-282.
- Kusters JMAM, Bzdusek K, Kumar P, et al. Automated IMRT planning in pinnacle: a study in head-and-neck cancer. *Strahlenther Onkol.* 2017;193(12):1031-1038.
- Nawa K, Haga A, Nomoto A, et al. Evaluation of a commercial automatic treatment planning system for prostate cancers. *Med Dosim.* 2017;42(3):203-209.
- Chen H, Wang H, Gu H, et al. Study for reducing lung dose of upper thoracic esophageal cancer radiotherapy by auto-planning: volumetric-modulated arc therapy vs intensity-modulated radiation therapy. *Med Dosim.* 2018;43(3):243-250.
- Hussein M, South CP, Barry MA, et al. Clinical validation and benchmarking of knowledge-based IMRT and VMAT treatment planning in pelvic anatomy. *Radiother Oncol.* 2016;120(3): 473-479.
- Janssen TM, Kusters M, Wang Y, et al. Independent knowledge-based treatment planning QA to audit Pinnacle AutoPlanning. *Radiother Oncol.* 2018;133:198-204.
- Krayenbuehl J, Zamburlini M, Ghandour S, et al. Planning comparison of five automated treatment planning solutions for locally advanced head and neck cancer. *Radiat Oncol.* 2018;13(1):225. doi: 10.1186/s13014-018-1176-x.
- Kubo K, Monzen H, Ishii K, et al. Dosimetric comparison of RapidPlan and manually optimized plans in volumetric modulated arc therapy for prostate cancer. *Phys Med.* 2017;44:199-204.
- Pogson EM, Aruguman S, Hansen CR, et al. Multi-institutional comparison of simulated treatment delivery errors in ssIMRT,

- manually planned VMAT and AutoPlan-VMAT plans for nasopharyngeal radiotherapy. *Phys Med*. 2017;42:55-66.
16. Rice A, Zoller I, Kocos K, et al. The implementation of RapidPlan in predicting deep inspiration breath-hold candidates with left-sided breast cancer. *Med Dosim*. In press.
 17. Scaggion A, Fusella M, Roggio A, et al. Reducing inter- and intra-planner variability in radiotherapy plan output with a commercial knowledge-based planning solution. *Phys Med*. 2018;53:86-93.
 18. Schubert C, Waletzko O, Weiss C, et al. Intercenter validation of a knowledge based model for automated planning of volumetric modulated arc therapy for prostate cancer. The experience of the German RapidPlan consortium. *PLoS One*. 2017;12(5): e0178034.
 19. Speer S, Klein A, Kober L, Weiss A, Yohannes I, Bert C. Automation of radiation treatment planning: evaluation of head and neck cancer patient plans created by the pinnacle(3) scripting and auto-planning functions. *Strahlenther Onkol*. 2017;193(8): 656-665.
 20. Wang S, Zheng D, Zhang C, et al. Automatic planning on hippocampal avoidance whole-brain radiotherapy. *Med Dosim*. 2017; 42(1):63-68.
 21. Wu H, Jiang F, Yue H, Zhang H, Wang K, Zhang Y. Applying a RapidPlan model trained on a technique and orientation to another: a feasibility and dosimetric evaluation. *Radiat Oncol*. 2016;11(1):108. doi:10.1186/s13014-016-0684-9.
 22. Wu B, Kusters M, Kunze-Busch M, et al. Cross-institutional knowledge-based planning (KBP) implementation and its performance comparison to Auto-Planning Engine (APE). *Radiother Oncol*. 2017;123(1):57-62.
 23. Nelms BE, Robinson G, Markham J, et al. Variation in external beam treatment plan quality: an inter-institutional study of planners and planning systems. *Pract Radiat Oncol*. 2012;2(4): 296-305.
 24. Fogliata A, Wang PM, Belosi F, et al. Assessment of a model based optimization engine for volumetric modulated arc therapy for patients with advanced hepatocellular cancer. *Radiat Oncol*. 2014;9:236. doi:10.1186/s13014-014-0236-0.
 25. Fogliata A, Nicolini G, Clivio A, et al. A broad scope knowledge based model for optimization of VMAT in esophageal cancer: validation and assessment of plan quality among different treatment centers. *Radiat Oncol*. 2015;10:220. doi:10.1186/s13014-015-0530-5.
 26. Delaney AR, Tol JP, Dahele M, Cuijpers J, Slotman BJ, Verbakel WF. Effect of dosimetric outliers on the performance of a commercial knowledge-based planning solution. *Int J Radiat Oncol Biol Phys*. 2016;94(3):469-477.
 27. Alpuche Aviles JE, Cordero Marcos MI, Sasaki D, Sutherland K, Kane B, Kuusela E. Creation of knowledge-based planning models intended for large scale distribution: minimizing the effect of outlier plans. *J Appl Clin Med Phys*. 2018;19(3): 215-226.
 28. Cilla S, Ianiro A, Macchia G, et al. EP-1885: evaluation of pinnacle automated VMAT planning for complex pelvic treatments. *Radiother Oncol*. 2018;127:S1020-S1021.
 29. Fusella M, Scaggion A, Pivato N, Rossato MA, Zorz A, Paiusco M. Efficiently train and validate a RapidPlan model through APQM scoring. *Med Phys*. 2018;45(6):2611-2619.