Title:
Use of planning metrics software for automated feedback to radiotherapy students

Short Title:
Planning Metrics RT Education

Authors:
Pete Bridge, Mark Warren, Marie Pagett

Institution:
Directorate of Medical Imaging and Radiotherapy
School of Health Sciences
University of Liverpool
Liverpool L69 3BX

Corresponding Author:
Pete Bridge
Directorate of Medical Imaging and Radiotherapy
School of Health Sciences
University of Liverpool
Liverpool L69 3BX
Tel: 01517958366
Email pete.bridge@liverpool.ac.uk

Conflicts of Interest Notification:
There are no conflicts of interest

Abstract:

Background and purpose:

Pre-registration teaching of radiotherapy planning in a non-clinical setting should allow students the opportunity to develop clinical decision making skills. Students frequently struggle with their ability to prioritise and optimise multiple objectives when producing a clinically acceptable plan. Emerging software applications providing quantitative assessment of plan quality are designed for clinical use but may have value for teaching these skills. This project aimed to evaluate the potential value of automated feedback to second year BSc (Hons) Radiotherapy students.

Materials and Methods:

All 26 students studying a pre-registration radiotherapy planning module were provided with automated prediction of relative feasibility for left lung tumour planning targets by planning metrics software. Students were also provided with interim quantitative reports during the development of their plan. Student perceptions of the software were gathered using an anonymous questionnaire. Independent blinded marking of plans was performed after module completion and analysed for correlation with software-assigned marks.

Results:

25 plans were utilised for marking comparison and 16 students submitted feedback relating to the software. Overall student feedback was positive regarding the software. A “strong” Spearman Rank Order Correlation (r_s = 0.7165) was evident between human and computer marks (p= 0.000055).

Conclusions:

Automated software is capable of providing useful feedback to students as a teaching aid, in particular with regard to relative feasibility of goals. The strong correlation between human and computer marks suggests a role in benchmarking or moderation; however the narrow scope of assessment parameters suggests value as an adjunct and not a replacement to human marking.

Keywords:

Automated feedback, clinical education, radiotherapy planning, undergraduate education

Use of planning metrics software for automated feedback to radiotherapy students

Introduction

Practical experience of radiotherapy planning and incorporation of these skills into module assessments is a common adjunct to formal examination of radiotherapy students' planning knowledge and understanding. Students frequently struggle with the high-level decision making which underpins their development of a clinically acceptable plan; particularly the extent to which they can prioritise and optimise multiple objectives. For example the overriding need to cover a target volume and surrounding margin of tissue with a high dose can lead to high dose in adjacent structures which can be challenging to avoid. These situations are commonly faced by radiotherapy clinicians; the recent development[1] of a decision support tool for plan comparison illustrated the highly complex nature of this. Providing objective feedback regarding each of the frequently contradictory objectives found in treatment planning is challenging yet vital to ensure this does not overshadow student learning of dosimetric principles and process.

There has long been keen interest in developing valid metrics for assessment of radiotherapy plan quality[2]. There are several emerging planning metrics software applications [3,4] that offer three main tools that could help to provide useful feedback. At the pre-planning stage, these programs can interrogate CT and structure datasets to provide a prediction of the extent to which plan objectives can be achieved[5] as seen in Figure 1. During plan evaluation and optimisation quantitative measures can be assigned to a variety of objectives in order to provide a rapid overview of plan quality across a range of metrics. Finally, completed individual plans can be quantitatively assessed with a score against a range of individually weighted planning objectives.

Although these applications are designed for clinical use as plan evaluation tools there is potential academic value in providing automated feedback to students regarding plan quality. Automated feedback use has been reported in medical education studies ranging from simple online multiple choice tests[6] to clinical competency essay marking[7]. It has also been consistently used to good effect in the field of computer coding education[8] where users submit their code and receive feedback designed to identify aspects that need improving. Planning metrics software works in a similar manner by providing a rapid overview of student performance across a range of parameters to highlight the most challenging aspects and focus efforts accordingly. Since the software is also capable of pre-assessing a dataset in order to predict the extent to which plan objectives can be achieved, this offers additional value as a formative teaching tool by providing students with a measure of expectation in relation to planning goals. The additional capacity of the software to automatically assign a "mark" for a student's plan suggests the potential for use as summative assessment. Although replacing a human examiner's qualitative assessment of a plan is controversial, planning metrics software could provide additional summative feedback on assessments to complement human marking. From a formative perspective the software could provide students with useful additional "on demand" feedback on their planning skills and optimise tutor support time during scheduled teaching.

This project aimed to evaluate the feasibility and value of software-assisted feedback to pre-registration radiotherapy students as they gain planning understanding and skills.

Methods and Materials

An evaluation license for PlanIQ v2.1 (Sun Nuclear Corporation, Florida) was utilised for provision of feedback. Reports from this planning metrics software were made available to all students in Year 2 of the BSc Radiotherapy Course at the University of Liverpool. Students were invited to participate in the evaluation project and were advised that provision of feedback and data was voluntary and that all data was anonymous in nature. The University Ethics Committee provided approval for the project.

All students planned the same lower left lobe non-small cell lung tumour to a target dose of 66Gy and were provided with target outlines. Students were guided to outline the Organs at Risk (OAR). A range of planning goals was provided to the students and also input into the software as a plan evaluation algorithm. Goals comprised a mixture of parameters relating to both target coverage and OAR doses. These were drawn from reported studies[9], trial protocols and local clinical practice and included a mixture of easy, challenging and impossible targets. Table 1 summarises these goals.

Students were provided with a preliminary assessment of the relative difficulty of achieving the range of goals that had been generated by the software during one of the teaching sessions. They were able to request as many interim reports of their plan performance as they wished to inform their plan development prior to submission. These requests were verbal during scheduled practical sessions with an immediate report generated. Outside timetabled teaching sessions, students were able to email a request for a report with a maximum turnaround time of 24 hours. Students were also provided with a report based on a complex intensity-modulated radiotherapy (IMRT) plan for the same patient for comparison.

Data collection was conducted in two phases. In Phase One, at the end of the module and after submission of the formal plan evaluation assessment, students were invited to provide their feedback on the value of the software. Data from consenting students were collected using a paper-based survey tool comprising a mixture of Likert-style question stems and open questions.

Phase two entailed independent marking of student plans for comparison with software generated marks. An experienced marker assessed the clinical acceptability of each completed student plan using the criteria outlined in Table 2. The primary criteria assessed the dose distribution only, and were used for direct comparison of human and automated marking. The secondary criteria assessed the student's understanding of clinical plan production by considering their use of beam modifiers, shielding and angle selection. Two scores were produced for assessment against the PlanIQ software: a Dose Distribution Score (a mean percentage score of all the primary criteria), and an Overall Score (a mean percentage of both primary and secondary criteria). Both scores were analysed for statistical significance. These marks were not used as module summative grades; for this module student marks were assigned for plan evaluation only and not plan generation.

Analysis of the student feedback data was descriptive in nature with Likert responses being collated. Student responses to the open question answers were grouped by themes for triangulation and interpretation purposes; findings arising from this qualitative data are the subject of a separate

paper. The human and automated plan marks were subjected to correlation analysis; anomalies in mark assignation were investigated in order to determine explanations for divergence.

After submission and plan marking had been completed and ratified, individual feedback was generated using the software to provide students with an indication of how they performed against the class mean, minimum and maximum across the range of objectives.

Results

Student usage

All students made use of the software at least once and the total number of reports generated across the BSc cohort was 33. Consent was provided for a total of 25 plans to be utilised for the summative marking comparison. Out of these a total of 16 students (61.5%) submitted feedback relating to the tool.

Cohort metric results

Table 1 summarises class performance against the full range of planning goals within the software. It can be seen that in general students struggled with target coverage; a common issue with lung plans. There was little variance on most of the target metrics with the "Planning Target Volume" (PTV) maximum dose of 107.5% having the greatest and also being the most challenging. As expected across a diverse cohort, there was a large difference in student performance across the various OAR metrics. The spinal cord "Planning Organ at Risk Volume" (PRV) and Oesophagus in particular saw a large variance with a wide range of doses within these. Lung and heart doses were relatively easily achieved with only the challenging Heart "V25Gy" goal being impossible due to tumour and heart proximity. Table 3 compares the software predictions for each goal with cohort achievement. It is interesting to note the failure of the software to recognise the challenge associated with target coverage in the thorax. It also predicted difficulties in achieving target maximum and lung dose limits which were not a problem for the cohort.

Student feedback

Overall student feedback was positive regarding the software as seen in Table 4; 75% of responses indicated that the software should be used in the future. Students felt that the software particularly helped them to understand their goals for the plan with only 6% of responses disagreeing. Students were less enthusiastic about the role of the feedback provided by the software with 50% of them agreeing that the feedback helped them to plan better and understand planning principles.

Automated marking results

A "strong" Spearman Rank Order Correlation (r_s = 0.7165) was calculated between the human "Dose Distribution" score and computer marked score (p= 0.000055). Figure 2 illustrates this data as a

scatter plot; it is evident that there are some outliers with the 2 lowest scores attributed by the human marker and further investigation into reasons for these points is ongoing. The human marker had also provided assessment on additional and less quantifiable parameters relating to ease and reproducibility of setup in an "Overall" score. Figure 3 illustrates the effect of these additional objectives on the correlation of marks; it can be seen that the outliers have been eliminated but the overall correlation is weaker (r_s = 0.5601; p=00362).

Discussion

Resource Implications

Generation of planning metric reports was time consuming within the study but this was due to the licensing agreement for the evaluation which restricted usage to a single laptop. If the software were to be deployed across the University network for student access then this would drastically reduce instructor input. The software does offer the potential to reduce instructor time demands by providing students with individualised feedback. This can in turn provide a structure for instructor intervention and make practical sessions more efficient.

Pre-planning feedback

Some of the goals were clearly easily achieved while others were impossible; especially those arising from reduced scatter contribution and lack of charged particle equilibrium in lung tissue. The feedback from students indicated the value of the pre-planning "feasibility prediction" in identifying which parameters they would be expected to achieve and which would be insurmountable obstacles. This in turn prompted useful discussion in classes about the relative importance of different parameters and underpinning physical principles explaining any challenges arising.

Planning performance

The difference in variance in relation to the Cord PRV and Oesophagus doses was interesting with some students clearly making an effort to not only meet the maximum dose limit constraints but also further reduce dose where possible. It is important to consider that the decision-making process of expert clinical practitioners is not fully understood, and their variance in surpassing objectives is not known. Recent studies[10] comparing expert planners against automated solutions suggest that clinical decision-making may not adhere directly to predefined quantitative parameters. Attempts to assess student performance must therefore reflect this variation in practice and it may therefore be advantageous for student assessments to challenge assumptions in practice and apply radiobiological principles to their decision making. Future study gathering student feedback on their planning decisions would provide valuable insight with regards to this.

Summative marking

The strong correlation between the marks assigned independently by the human marker and the software was encouraging and at least indicates a good level of internal reliability for the human marker. Indeed the software could have potential roles in an assessment benchmarking exercise or moderation activities. In terms of summative assessment, however, it was clear that the human marker had also based their full assessment on less quantifiable parameters relating to ease and reproducibility of setup. The effect of these can be seen in Figure 3 where the outliers have been eliminated. This may indicate that these students had compensated for poor attainment of some quantitative objectives by exhibiting good planning practices. Use of automated software to assign a summative assessment mark is clearly an oversimplification. It may, however, have a role in providing additional marker support by providing a summary of achievement in relation to key parameters.

Pedagogical Implications

Although the software does provide a good overview of student performance which can aid their formative development there are some pedagogical issues. In particular it is important that students learn essential plan evaluation skills including slice-by-slice visual checks, accurate interpretation of dose-volume histograms and the more subtle "holistic" evaluation including clinical decision making. There is a danger that over-reliance on numeric output will reduce student engagement with these core skills and future use of planning metrics software will need to ensure that students understand the complementary nature of this tool rather than depending on it entirely.

Conclusions

This study has demonstrated that automated software is capable of providing students with useful guidance in relation to a range of radiotherapy planning parameters. As a formative tool, the software can help students to focus on achievable and challenging objectives and provide a rapid summary of their performance. The software has potential value as a teaching aid to provide additional student support and thus optimise tutor time. Care must be taken to ensure use of the tool does not inhibit development of core plan evaluation skills and it is recommended that it only be adopted in later stages of the Course with more complex planning to aid students who have already demonstrated these skills. Summative assessment can be provided by the software and this correlates well with human marking; this should be used as an adjunct and not a replacement to ensure a more holistic planning approach is adopted by students and tutors alike.

Acknowledgements

The authors would like to acknowledge the kind support of Sun Nuclear Corporation (Florida) for the provision of a temporary free "PlanIQ" planning metrics software license for evaluation purposes.

Conflict of Interest Statement

A temporary free “PlanIQ” planning metrics software license was provided for evaluation purposes by Sun Nuclear Corporation (Florida). The company had no direct input into study design, data collection, analysis or writing up.

References

1. Brodin NP, Maraldo MV, Aznar MC et al. Interactive decision-support tool for risk-based radiation therapy plan comparison for Hodgkin Lymphoma. Int J Radiat Oncol Biol Phys 2014; 88(2): 433-445.

2. Moore KL, Brame RS, Low DA, Mutic S. Quantitative metrics for assessing plan quality. Semin Radiat Oncol 2012; 22(1): 62-69.

3. Holloway LC, Miller J, Kumar S, Whelan BM, Vinod SK. Comp Plan: A computer program to generate dose and radiobiological metrics from dose-volume histogram files. Med Dosim 2012; 37(3): 305–309.

4. Zhao B, Joiner MC, Orton CG, Burmeister J. SABER: A new software tool for radiotherapy treatment plan evaluation. Med Phys 2010; 37(11): 5586-5592.

5. Crowe SB, Kairn T, Kenny J, Knight RT, Hill B, Langton CM, Trapp JV. Treatment plan complexity metrics for predicting IMRT pre-treatment quality assurance results. Australas Phys Eng S 2014; 37 (3): 475-482.

6. Mitra NK, Barua A. Effect of online formative assessment on summative performance in integrated musculoskeletal system module. BMC Med Educ 2015: 15(29): 1-7.

7. Latifi S, Gierl MJ, Boulais AP, De Champlain AF. Using automated scoring to evaluate written responses in English and French on a high-stakes clinical competency examination. Eval Health Prof 2016; 39(1): 100-113.

8. Alemán JLF. Automated assessment in a programming tools course. IEEE T Educ 2011; 54(4): 576-581.

9. Marks LB, Bentzen SM, Deasy JO et al. QUANTEC: Organ-specific paper: Radiation dose–volume effects in the lung. Int J Radiat Oncol Biol Phys 2010; 76(3): S70-S76.

10. Voet PJW, Dirkx MLP, Breedveld S, Fransen D, Levendag PC, Heijmen BJM. Towards fully automated multicriteria plan generation: a prospective clinical study. Int J Radiat Oncol Biol Phys 2013; 8(3): 866-872.

Table 1: Cohort performance against planning objectives

Volume	Goal	Cohort Mean	Cohort Worst	Cohort Best	Variance
PTV	Max < 70.95Gy	68.5	70.8	67.1	1.2
PTV	D2% < 70.95Gy	67.2	68.8	66.3	0.6
PTV	D2% < 69.3Gy	67.2	68.8	66.3	0.6
PTV	D98% > 62.69Gy	61.5	60.9	62.9	0.2
CTV	D99% > 65.34Gy	63.1	62.3	64.1	0.2
CTV	D98% > 62.69Gy	63.4	62.6	64.3	0.2
Heart	V40Gy < 30%	14.2	21.5	9.2	7.5
Heart	V30Gy < 40%	18.5	24.1	13.3	4.9
Heart	V25Gy < 10%	20.7	26.8	17	5.2
Rt Lung	V30Gy < 15%	0	0	0	0.0
Lungs	V20Gy < 30%	20.1	21.9	18	0.9
Lungs	V20Gy < 35%	20.1	21.9	18	0.9
Cord PRV	Max dose < 45Gy	18.2	40.6	3.9	67.6
Oesophagus	Max dose <50Gy	25.4	47.8	14.9	57.9

Table 2: Human marking objectives

Marking Criteria	Assessment Parameters
Primary: Conformity to PTV	Visual inspection of 95% and 90% isodose line.
Primary: PTV Heterogeneity	Visual inspection of 105% and 100% isodoes within the PTV.
Primary: OAR doses	DVH reading of canal PRV, lung V20Gy and heart V30Gy and V40Gy. Visual inspection of 85% and 65% isodose lines in relation to contoured OAR.
Primary: Dose to other tissue	Visual inspection of isodose lines in other healthy tissue not contoured as an OAR.
Secondary: Collimation	Assess clinical acceptability of MLC and jaw positions
Secondary: Wedges	Check whether wedges contributed to or hindered the planning goals
Secondary: Gantry	Visual inspection that beam angles were optimised to avoid unnecessary healthy tissue exposure
Secondary: Weighting	Check that beam weighting was optimal

Table 3: Feasibility prediction accuracy

Volume	Goal	Prediction	Cohort Achievement
PTV	Max < 70.95Gy	Challenging	25
PTV	D2% < 70.95Gy	Challenging	25
PTV	D2% < 69.3Gy	Challenging	25
PTV	D98% > 62.69Gy	Probable	1
CTV	D99% > 65.34Gy	Probable	0
CTV	D98% > 62.69Gy	Probable	24
Heart	V40Gy < 30%	Probable	25
Heart	V30Gy < 40%	Probable	25
Heart	V25Gy < 10%	Challenging	0
Rt Lung	V30Gy < 15%	Probable	25
Lungs	V20Gy < 30%	Challenging	25
Lungs	V20Gy < 35%	Challenging	25
Cord PRV	Max dose < 45Gy	Probable	25
Oesophagus	Max dose <50Gy	Probable	25

Table 4: Student feedback summary

Likert Stem	SD	D	N	A	SA
The software helped me to understand my goals for the plan	0	1	2	10	3
The formative feedback about my plan helped me to develop my plan better	0	3	5	4	4
Feedback from the software helped me with my understanding of planning principles	0	2	6	7	1
I would recommend this software be used to support future planning tasks	0	2	2	7	5

KEY: SD = Strongly Disagree, D = Disagree, N = Neutral, A = Agree, SA = Strongly Agree

Figure 1: Screenshot from PlanIQ showing feasibility prediction of a range of parameters for the plan

Metric ID	PlanIQ Result	Goal [1]	Goal [1] Feasibility
[PTV] Max dose (Gy)	66.0010	< 70.95	Challenging
[PTV] D[2.0%] (Gy)	66.0010	< 70.95	Challenging
[PTV] D[2.0%] (Gy)	66.0010	< 69.3	Challenging
[PTV] D[98.0%] (Gy)	66.0010	> 59.4	Probable
[CTV] D[99.0%] (Gy)	66.0010	> 62.7	Probable
[CTV] D[98.0%] (Gy)	66.0010	< 89.1	Probable
[HEART] V[40.0Gy] (%)	2.8310	< 30	Probable
[HEART] V[30.0Gy] (%)	3.0534	< 40	Probable
[HEART] V[25.0Gy] (%)	3.7571	< 30	Challenging
[RT LUNG] V[30.0Gy] (%)	0.0000	< 15	Probable
[LUNGS] V[20.0Gy] (%)	9.6612	< 30	Challenging
[LUNGS] V[20.0Gy] (%)	9.6612	< 35	Challenging
[CORD PRV] Max dose (Gy)	0.4615	< 45	Probable
[OESOPHAGUS] Max dose (Gy)	8.7159	< 50	Probable

Figure 2: Scatter plot of Human "Dose Distribution" score and "Computer" marks (primary objectives only in human marking)

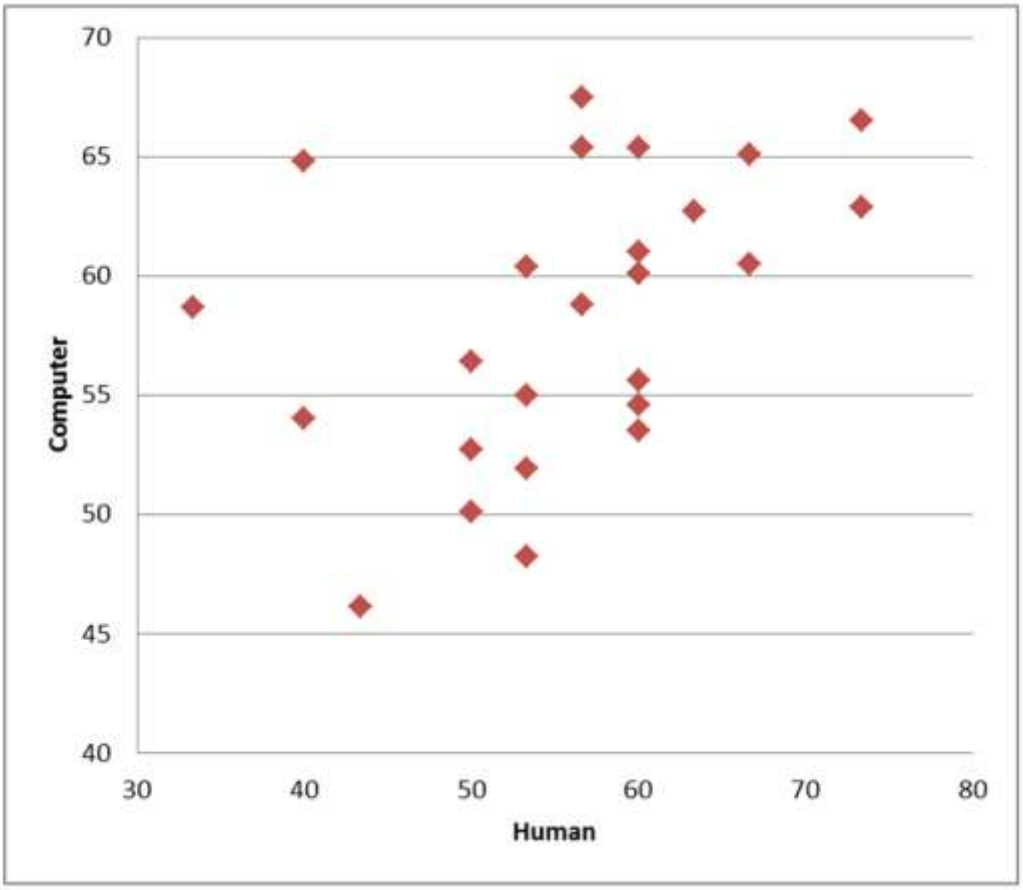

Figure 3: Scatter plot of Human "Dose Distribution" score and "Computer" marks (includes both primary and secondary criteria in human marking)