Head and neck plan quality: where we are and does it matter?

B Nelms¹, J Caudell², D Gintz² and V Feygelman²
¹Canis Lupus LLC, Merrimac, Wisconsin 53561, USA
²Moffitt Cancer Center, Tampa, Florida 33612, USA

E-mail: vladimir.feygelman@moffitt.org

Abstract. Radiation treatment planning for the Head and Neck (HN) cancers is challenging and complex, and here we investigate various aspects of HN plan quality. We present the results of an international plan quality study, in which a web-based platform (ProKnow) was used to analyse a total of 238 plans submitted from 34 countries, all for the same dataset, anatomy, and quantified, objective plan scoring algorithm. All treatment planning systems (TPS) studied were able to produce high quality plans: 6 of 6 TPS models had scores in the top 25%, and 4 of those in the top 10%. However, all TPS and modalities also showed substantial variability in plan quality across all planners, suggesting needs for training and/or automation. Concerning current automation solutions, both commercially available approaches (AutoPlanning and Knowledge-Based Planning) are so far unable to outperform experienced human planners when tight target dose coverage and homogeneity are required. Target dose homogeneity is not to be discounted, as it may be linked with the rate of complications such as reactive gastrostomy tube placement after oropharynx chemoradiation.

1. Introduction
In this work, we review the state of affairs on a variety of topics pertinent to the quality of Head and Neck (HN) radiation treatment planning. We cover (i) the results of a recent, international study that collected and measured metric-based plan quality as well as dosimetric accuracy (measured vs. calculated) for a challenging case; (ii) early experiences with planning automation, including published results from various groups including ours; and (iii) potential link between target dose homogeneity and the rate of clinical complications.

2. Treatment planning study

2.1. Methods
A web-based platform (ProKnow, http://www.proknowsystems.com) was used to run the study. A published HN patient CT from MPPG 5a [1] was provided with pre-contoured targets and organs-at-risk (OAR), and a set of dosimetric metrics and corresponding objectives. The plan included a primary target and bilateral elective neck node CTV treated using simultaneous integrated boost to three dose levels. Uploaded plans were scored in real-time and with full transparency using an objective, composite scoring formalism [2]. A total of 21 metrics were extracted, scored function-wise, and summed for each plan, resulting in the composite Plan Quality Metric score (PQM). Target dose coverage and homogeneity made up 63% of the possible total, with the remainder accounting for OAR sparing. The relative weight of the different dose objectives, while clinically realistic, is by necessity somewhat subjective, as it reflects preferences of the study team. However, once the weights
are fixed, the rest of the scoring is applied objectively to all participants, who knew the details up front and were left with the task of maximizing their score as they deemed fit after the minimum constraints and objectives were met. The participants were free to choose the treatment planning system (TPS), treatment modality, beam energy, and planning techniques, provided they were clinically viable and practical. Interactive histograms and scatterplots of all metrics are available to the public via the ProKnow website.

In addition to generating histograms of PQM distributions, statistical comparisons between different sub-groups were performed. The median PQMs were compared with non-parametric Kruskal-Wallis ANOVA followed by Dunn’s multiple comparisons test (GraphPad Software, La Jolla California USA) between the following treatment modalities: C-arm linac VMAT and IMRT, helical tomotherapy (HTT), and protons. All treatment planning systems were included. For IMRT, segmented and dynamic plans were scored together, as were scattered and scanned proton beams. P-value below 0.05 was considered an indication of statistical significance. In addition, other descriptive statistics per TPS were recorded, such as the minimum and maximum PQM and the standard deviations.

For standard linacs, the median VMAT vs. IMRT PQM were additionally pair-wise compared per TPS by a Mann-Whitney non-parametric test when a reasonable number of IMRT plans were available (i.e. Eclipse and Monaco). Next, the median photon PQM scores were compared with the Kruskal-Wallis ANOVA test between five different TPS.

The median number of monitor units and/or estimated beam-on time were compared between the photon modalities: VMAT, IMRT, and HTT. Spearman correlation coefficient $r$ was computed between VMAT and IMRT PQM and corresponding total MU. Similarly, a possible correlation was explored between the VMAT, IMRT, and HTT PQM and estimated beam-on time.

Finally, the “pre-treatment” dosimetric QA results were requested and analysed in terms of the method employed and $\gamma$-analysis passing rates, with the 3% (global)/3 mm/10% low dose threshold, 3% (global)/2mm/10%, and 2% (local)/2 mm/20% criteria combinations.

### 2.2. Results
A total of 238 plans were submitted from 34 countries. 39% were from the USA. In terms of treatment modalities, VMAT made up 73% of the submissions, followed by fixed gantry IMRT (18%), tomotherapy (6%), and protons (3%).

The participation by TPS and overall distribution of the PQM are shown in figures 1 and 2, respectively.

The overall median VMAT (N=173) score was 135.6 vs. 130.1 for IMRT (N=43). The overall maximum PQM of 146.9 was achieved with a 3-arc VMAT plan generated by the Eclipse TPS, requiring only 642 total MU. The ANOVA comparison between different modalities showed no statistically significant differences. No significant differences were also found between the median VMAT vs. IMRT PQM for the individual TPS. The results of non-parametric ANOVA comparisons of median PQM for five different TPS with the sufficiently high number of entries are presented in Table 1 along with some descriptive statistics. All TPS were able to produce high quality plans: 6 of 6 models had scores in the top 25%, and 4 of those in the top 10%. Statistical significance was only demonstrated for Monaco’s PQMs, which were lower pair-wise comparisons than all other systems except HTT, reaching high significance (p<0.005) against Pinnacle and RayStation. The results of comparisons between the efficiency (total MU or time) of different photon modalities are presented in table 2. The median number of MU for VMAT was statistically significantly lower than for IMRT (p<0.0001), and the same holds for beam-on time. VMAT beam-on time was significantly less than HTT (p<0.0001), and so was IMRT time (p<0.0001).
Figure 1. Participation by TPS.

Figure 2. Overall PQM score distribution.

Table 1. Descriptive statistics of photon PQM for different TPS and p-values for the median score pair-wise tests between TPS.

|         | N  | Min | Median | Max  | SD  | p-value    |
|---------|----|-----|--------|------|-----|------------|
| Eclipse | 120| 86.0| 133.6  | 146.9| 12.5| >0.9       |
| HTT     | 15 | 100.8| 131.1  | 145.8| 12.7| 0.03       |
| Monaco  | 36 | 99.1 | 125.9  | 141.5| 11.8| >0.9       |
| Pinnacle| 34 | 114.5| 138.8  | 144.6| 7.8 | >0.9       |
| RayStation | 26 | 88.9 | 138.5  | 145.8| 13.5| >0.9       |

Table 2. Monitor units and estimated beam-on times for different photon modalities

| Plan efficiency metric | MU | Time (min) |
|------------------------|----|------------|
| Statistical            | VMAT | IMRT | VMAT | IMRT | HTP |
| N                      | 172 | 42 | 153 | 19 | 15 |
| Min                    | 308 | 433 | 1.7 | 2.6 | 5.1 |
| Median                 | 596 | 977 | 3.0 | 5.0 | 9.18 |
| Max                    | 1616 | 3159 | 4.8 | 8.1 | 27.1 |
| SD                     | 198 | 696 | 0.6 | 1.3 | 6.1 |

No correlation was found between the PQM and MU for either VMAT or IMRT (max $r < 0.46$).

Over 95% of the reported (77) QA tests used either a true (65%) or collapsed single angle (31%) composite measurements. The results support setting aggressive accuracy goals in terms of TPS algorithm commissioning. The median 2% (local normalization)/2mm gamma-analysis passing rate was 90%, with the top quartile of performers showing passing rates over 95%.

The most important finding in this section is that all TPS and modalities showed substantial variability in plan quality distribution, suggesting needs for planner training and/or automation. The latter is the subject of the following discussion.

3. Attempts at automation

3.1. AutoPlanning

AutoPlanning (AP) is offered in Pinnacle TPS (Philips). It attempts to mimic the steps of a human treatment planner while performing iterative optimization based on dynamically created regions of interest. Technically, it does not require prior knowledge of similar plans. When quantitatively compared to manually designed plans with the emphasis on target dose coverage and homogeneity, similar to Section 2, AP on average scored lower than the human planner [3]. While excelling in
sparing OARs, the software lacks reliable tools to improve target coverage parameters to the levels required by our institution, such as <105% single voxel hot spot.

3.2. Knowledge-Based Planning
Knowledge-Based Planning (KBP) is a feature in Eclipse TPS (Varian). It tries to predict achievable dose-volume histograms based on the library of the previously accepted plans that have similar relative geometry of the targets and OARs. We are not aware of any published study demonstrating that a library of previous high-quality HN plans would ensure equally high-quality output. Studies such as [4] used too loose target coverage/homogeneity criteria to draw conclusions in that regard.

3.3. Possible clinical implications
PQM scoring in the HN plan challenge and our AP evaluation [3] emphasized target dose homogeneity. To our knowledge, no study directly connecting IMRT/VMAT target dose homogeneity with complication rates exists. However, there is some published evidence that the rates of reactive gastrostomy tube placement after oropharynx chemoradiation are substantially lower when a strict 105% hot spot limit is enforced [5], compared to a study [6] presumably based on less strict RTOG-type [7] objectives that allow, as a minor deviation, up to 117% target dose to 1 cc of tissue. This observation is particularly important given that the results similar to [6] are used to justify more expensive radiotherapy modalities in HN cancer [8], without fully exhausting the capabilities of photon treatments with conventional linacs [9].

4. Conclusions
Dosimetric quality of HN plans remains highly variable and operator-dependent. The long-term solution for this problem may be automation. However, in the realm of commercially available TPS, automated solutions are so far not competitive with humans when tight target dose coverage and homogeneity are required. There is some published evidence, admittedly far short of unequivocal, that such requirements may be contributing to reduction in complication rates.

5. References
[1] Smilowitz J B et al 2015 J. Appl. Clin. Med. Phys. 16 14-34
[2] Nelms B E et al 2012 Pract. Radiat. Oncol. 2 296-305
[3] Gintz D et al 2016 J. Appl. Clin. Med. Phys. 17 331-46
[4] Tol J P 2015 Int. J. Radiat. Oncol. Biol. Phys. 91 612-20
[5] Strom T et al 2013 JAMA Otolaryngology - Head & Neck Surgery 139 1242-6
[6] Setton J et al 2015 Cancer 121 294-301
[7] RTOG 1016. https://www.rtog.org/ClinicalTrials/ProtocolTable/StudyDetails.aspx?study=1016
[8] Blanchard P et al 2015 Radiother. Oncol. 120 48-55
[9] Baldock C 2009 J. Phys.: Conf. Ser. 164 012002