From Preference Learning to Clinical Insight: Interpreting Expert Treatment Plan Evaluation

Jiahan Zhang[1], Yang Lei[1], James Tam[1], Alan Yu[1], Sining Chen[1], Robert Samstein[1], Kenneth Rosenzweig[1], Ming Chao[1], Tian Liu[1], Junyi Xia[1]
[1]Icahn School of Medicine at Mount Sinai, New York, NY

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

ABSTRACT

Plan evaluation is a critical component in the treatment planning workflow. Yet, objective assessment remains challenging because expert assessment relies on complex, multidimensional tradeoffs that are not fully captured by predefined dose-volume constraints. This study aims to quantitatively interpret expert treatment plan evaluation to improve transparency and reproducibility.

INTRODUCTION

Objective treatment plan evaluation requires balancing competing dosimetric objectives through complex and implicit clinical tradeoffs. A data-driven pairwise preference-based scoring model has shown the ability to approximate expert plan evaluation. Yet, the internal decision logic of such models has remained largely unexplored. We aim to interpret a treatment plan scoring model by quantifying learned dosimetric tradeoffs, identifying sources of inter-expert disagreement, and assessing the stability of feature prioritization. This study provides a foundation for planner education and automated plan evaluation.

METHODS AND MATERIALS

We studied a planner preference-aligned plan evaluation framework, which learns from expert-labeled and LLM-generated pairwise plan preferences based on clinically relevant dose-volume metrics. The overall workflow is shown in Fig. 1. The model was trained using 1,240 locally advanced non-small cell lung cancer plan pairs.

To interpret the learned plan score function, three analyses were performed. First, we perturbed individual dosimetric features while keeping the plan score constant to quantify tradeoffs between competing objectives. Second, SHAP-based feature importance was compared for 35 expert-reviewed plan pairs. Plans with unanimous expert preference and plans with inter-expert disagreement were compared to determine the sources of ambiguity. Third, feature ranks were assessed across all plans to evaluate the consistency of objective priority across treatment plans.

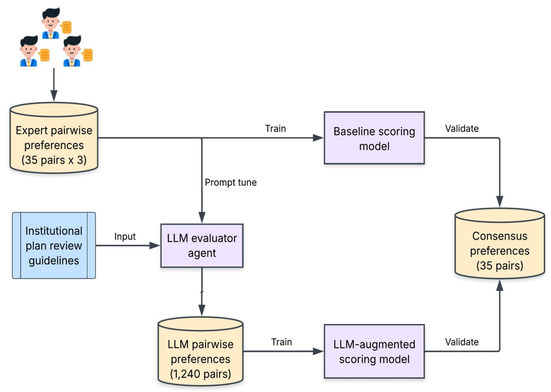

Figure 1. Overall workflow

RESULTS

SHAP values represent the relative contribution of all features to the model prediction results for individual cases. Figure 2 shows the SHAP values of the 10 features used in this study. Note that PTV coverage is not evaluated because all plans were normalized to the same PTV coverage. From the figure, it is clear that lung V20Gy and mean heart dose were the most influential predictors, and esophagus and cord D0.03cc were also important considerations.

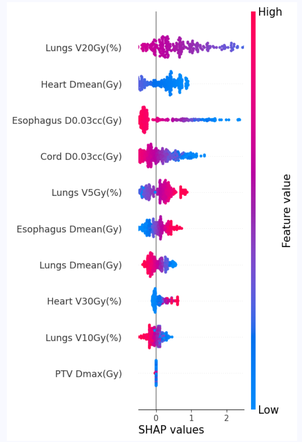

Figure 2. SHAP values of ten dosimetric features

Tradeoff sensitivity analysis was performed by perturbing individual dosimetric features while maintaining the same plan score by adjusting a secondary feature. Fifty plans were randomly selected for the analysis. Figure 3 shows the tradeoffs between four pairs of dosimetric objectives. The tradeoff relation is approximately linear.

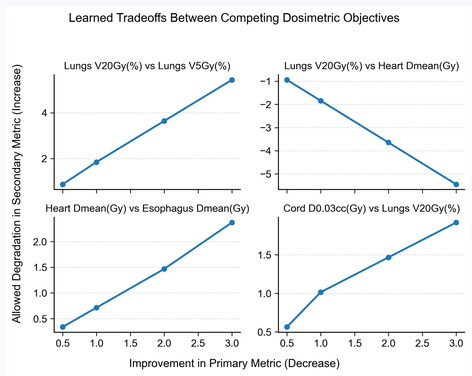

Figure 3. Tradeoff sensitivity curves between four pairs of dosimetric objectives. Each curve illustrates the amount of degradation in a secondary metric required to offset improvements in a primary metric to keep plan quality score constant.

The SHAP value differences between preferred and non-preferred plans were used to define a feature dominance ratio:$D = max_j|\Delta\varphi_j|/\sum_{j=1}^{M}|\Delta\varphi_j|$, where M is the total number of features and $|\Delta\varphi_j|$ is the difference between the SHAP values of the two plans. As shown in Fig. 4, the difference between consensus plans and disagreement plans is not statistically significant (p=0.38, Wilcoxon rank-sum test).

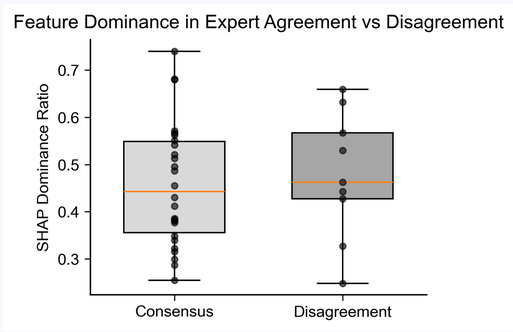

Figure 4. Learned tradeoffs between competing dosimetric objectives

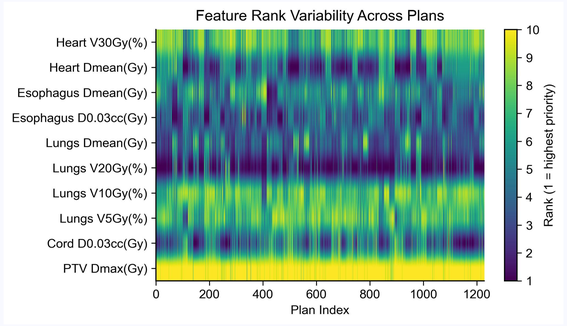

Figure 5. Learned tradeoffs between competing dosimetric objectives

Feature rank stability was assessed by ranking absolute SHAP values for each plan. The feature importance ranks across plans were used to distinguish universally prioritized objectives from context-dependent features. Figure 5 shows that lung V20Gy is consistently ranked high. Heart and lung mean doses are also ranked high for most cases. This indicates the decision-making process is consistent across plans.

DISCUSSION

By leveraging a small set of expert judgments to anchor the model, and using the LLM to generate a much larger preference dataset, we achieve a scoring function that closely approximates human evaluators while requiring minimal expert time. The results confirmed that higher-dose metrics such as lung V20Gy and heart mean dose dominate expert reasoning, consistent with findings from RTOG 0617 and clinical toxicity literature.

The proposed framework is highly flexible, allowing adaptations to institutional preferences. The scoring model can be build using institution-specific requirements in the form of a prompt. The majority of the workflow, including plan generation, LLM plan evaluation, and plan scoring agent training, are automated. The only requirement needed from the human experts is a small set of pairwise comparison results for validating the LLM plan evaluation agent. The offers a scalable solution that can be applied to most clinics, even ones with limited resources. Overall, the proposed hybrid framework offers a practical and powerful path toward consistent, expert-aligned plan quality assessment.

CONCLUSIONS

This study provides insight into clinically meaningful tradeoffs and plan evaluation priorities. The expert preference-based plan evaluation method supports transparent plan evaluation and establishes a foundation for an interpretable and automated treatment planning system.

CONTACT

Jiahan Zhang
Icahn School of Medicine at Mount Sinai
New York, NY
Jiahan.zhang@mountsinai.org
212-241-9071