Quantifying Differences between Clinical Practices using Natural Language Processing for AI Model Generalization in Radiation Oncology

Samuel Luk, PhD[1,2]; Shaotai Hu[1]; Kaivalya Bhatt[1]; Petros Kalendralis, PhD[3]; Alan Kalet, PhD[4]
[1]Boston University, [2]Boston Medical Center, [3]Maastricht University Medical Centre, [4]University of Washington

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

ABSTRACT

Purpose: Many Artificial intelligence (AI) applications have been developed recently to improve the quality and efficiency of radiotherapy processes. Despite the success of image-based applications, AI models trained with alphanumeric clinical data suffer from the lack of generalizability due to differences in clinical practices. In this multi-institutional study, we aim to quantify differences between clinical data from different radiation oncology (RO) practices using natural language processing and create a quantity that can inform the potential usability of an AI model in a clinic without going through the complete validation process.

Methods: Anonymized RO datasets are extracted from three US institutions and one European institution. Each clinical dataset was converted into multidimensional vectors using Word2Vec model. The generated word vectors are then shared, and weighted cosine similarity (Csim) between the vectors were calculated to quantify differences in prescription, plan complexity, and treatment approaches of various anatomic tumor locations across different institutions.

Results: For similar practices, Csims would range between 0.8-0.9, and for more diverging practices, Csims would be lower. For example, for Lung prescription patterns, the Csims between US institutions would range above 0.9, while the Csims between the European institution and US institutions are ranging from 0.67 to 0.78, indicating an observable difference between the datasets.

Conclusion: This study shows that Csim can capture differences between clinical datasets caused by differences in clinical practices. With the outlined vector space modeling framework, AI researchers can create suitable corpora for comparison and correlate Csim with model performance for validation and QA purposes.

CONTACT

Samuel Luk
Boston University/Boston Medical Center
Boston, MA
sammhluk@bu.edu

INTRODUCTION

To improve the quality and efficiency of radiotherapy processes, many artificial intelligence (AI) applications have been developed recently. These AI applications have shown promising ability to handle tasks involving medical images such as auto-segmentation. However, AI models that are developed utilizing alphanumerical clinical data usually do not have the same generalizability, potentially due to disparity in clinical practices. To help researchers assess the usability of a newly developed AI model using alphanumerical data in a clinic without going through the complete validation process, we propose a comprehensive method to quantify variations between clinical practices that can correlate to potential model performance using natural language processing (NLP). We will showcase the methodology by answering two questions:

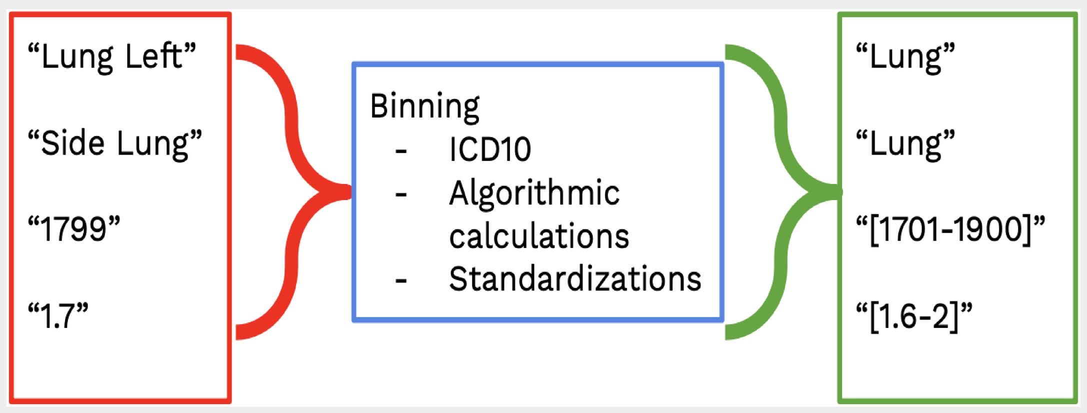
1) **how similar are the prescription patterns between different institutions on different tumor locations?**
2) **what is the difference in plan complexity between different treatment planning techniques in different institutions?**

METHODS AND MATERIALS

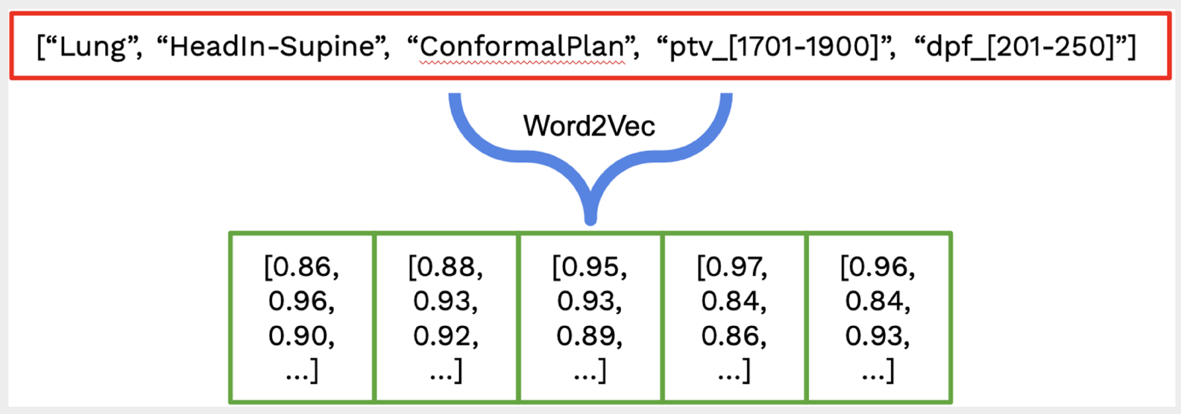
The participating institutions include Boston Medical Center (MA, USA), University of Vermont Medical Center (VT, USA), University of Washington (WA, USA) and Maastro Clinic (Netherlands). Each institution are located at different geographic locations, using different technologies and having different patient sizes. A summary of the differences at the time of study is listed in Table 1.

The NLP workflow to convert clinical data to multi-dimensional vectors using Word2Vec and quantify the differences between different clinical practices using cosine similarity (Csim) is illustrated below.

1. Wrangling and Binning Raw



2. Corpus Generation –> VSM Embeddings



RESULTS

Differences in clinical practice between participating institutions are observed. Figure 1 shows an example with the prescription patterns of various tumor locations are plotted for all clinics. For example, we observed that breast patients in participating US clinics receive an additional 10 Gy boost treatment, while it is less common in Maastro.

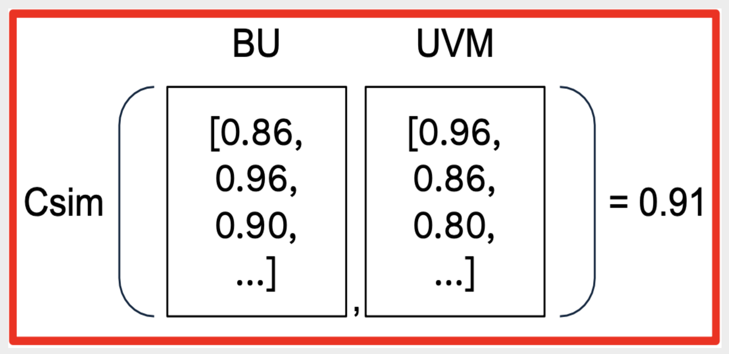
In Figure 2, we illustrate the ability of the Word2Vec model to quantify similarity/difference between clinical practices by pairing term-frequency plots with the computed Csim. In this example, lung prescription parameters (Q1) are compared and when Csim is high (0.94), the frequency plots are very similar and when it is low (0.68), there are more meaningful difference between the two clinics, say the more frequent use of IMRT in Maastro. In general, we observed that for similar practices, Csims would range between 0.8-0.9, and for more diverging practices, Csims would be lower. However, this is tumor site and parameters dependent. Also note that Csim captures also the correlation between the selected parameters instead of just term frequency.

Figure 3 shows a combined term-frequency plot and the corresponding Csims of VMAT treatments in all participating clinics with plan complexity parameters (MU/cGy, MU/deg and dose per fraction). It is clear that all clinics planned VMAT treatments similarly despite having different TPS. All Csims are above 0.8.

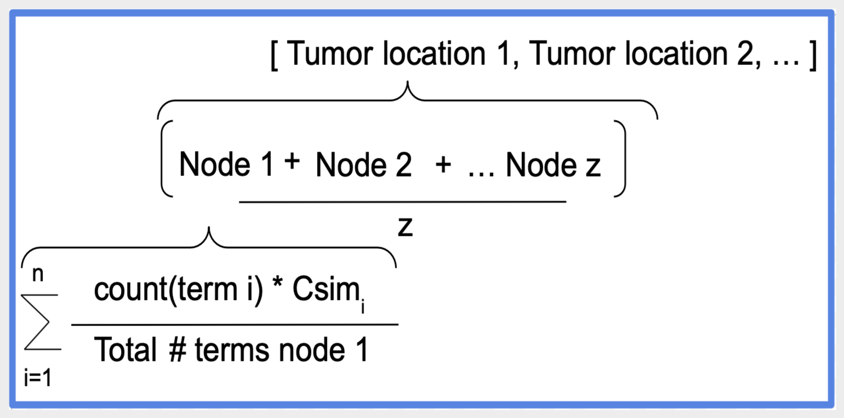
Clinic	Where	Linacs	TPS	OIS	Size
BMC	US	Varian	Eclipse	Aria	S/M
UVMMC	US	Elekta	Pinnacle	Mosaiq	S/M
UWMC	US	Elekta	RS/MC	Mosaiq	L
MC	EU	Varian	Eclipse	Aria	L

Table 1. Geographic locations, technologies and sizes of the participating institutions during time of study.

3. Cosine Similarity (Csims) Calculation



4. Apply Term Frequency-Based Weights



5. Final Csims Per Tumor Location

Clinic 1 vs Clinic 2 Weighted Csims

Lung: 0.85	Metastasis: 0.76
GU: 0.94	Breast: 0.87
GI: 0.76	GYN: 0.89

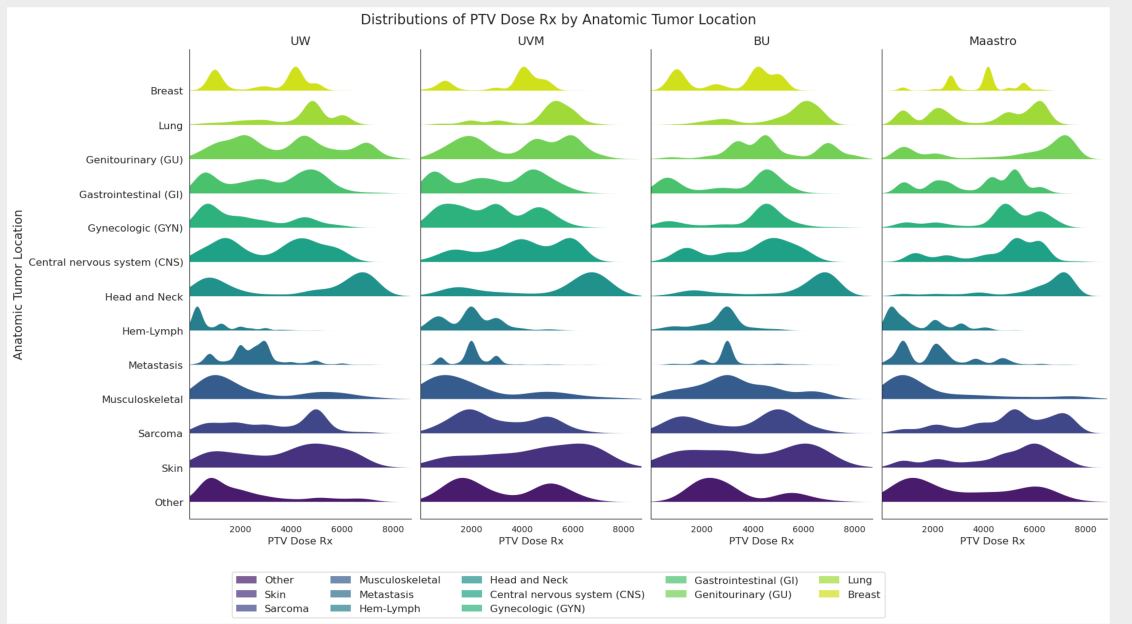


Figure 1. Prescription dose patterns of various tumor locations in different clinics

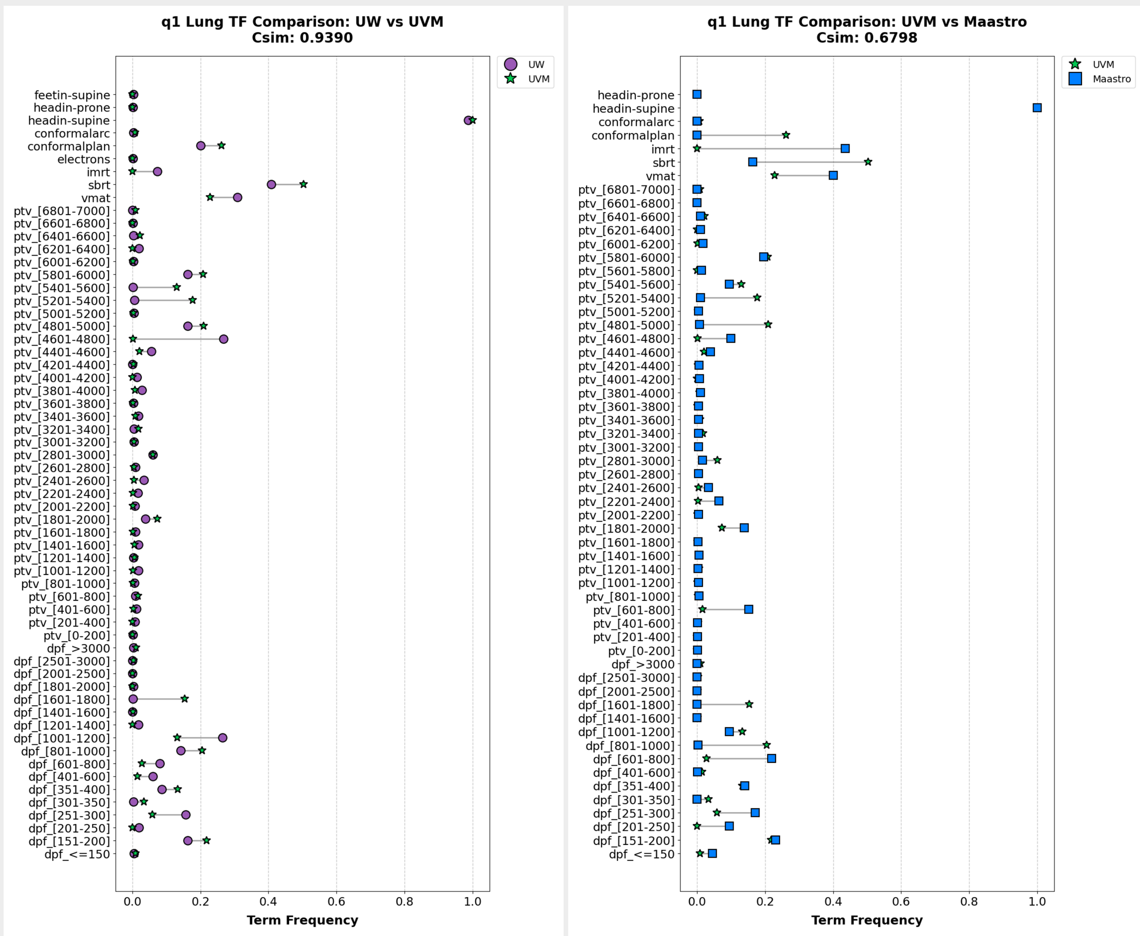


Figure 2. Term-frequency plots of prescription parameters of Lung treatments between UW/UVM (right) and UVM/Maastro (left) to illustrate the difference in practices when Csim is high/low.

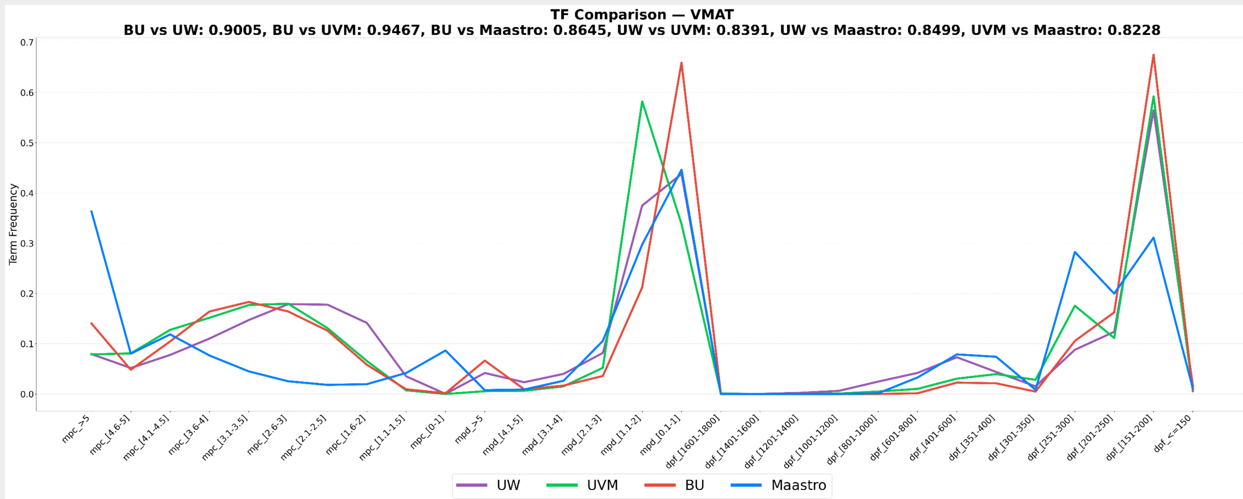


Figure 3. Combined term-frequency plot and corresponding Csim of VMAT treatments in all participating clinics with parameters relating to treatment plan complexity.

DISCUSSION & CONCLUSIONS

We have demonstrated that the NLP workflow can effectively captures differences and reveals insights on clinical practice disparities between institutions of varying demographics, technologies and size, and represents it quantitatively with Csims. A single number, Csims, can potentially be correlated with AI model performance and used as a simple and quick method to validate AI models and a continuous quality assurance of model performance utilizing alphanumerical parameters in different clinics. Next step of the project is to correlate the Csim with AI model performance in a multi-institutional study.

REFERENCES

1. S. M. H. Luk, E. C. Ford, M. H. Phillips, and A. M. Kalet, "Improving the Quality of Care in Radiation Oncology using Artificial Intelligence," *Clin. Oncol* 2022
2. P. Kalendralis *et al.*, "Automatic quality assurance of radiotherapy treatment plans using Bayesian networks: A multi-institutional study," *Front. Oncol.*, 2023
3. Kapoor et.al. 'Infrastructure tools to support an effective Radiation Oncology Learning Health System', JACMP 2023

BOSTON MEDICAL CENTER
BOSTON UNIVERSITY