An NLP-Based Collaborative Intelligence Tool to Support Precision Medicine

S. DIETERICH, Ph.D.[1], P. ARORA[2] and S. AZGHADI, MD[1]
1 UC Davis Medical Center, Sacramento, CA, USA
2 UC Davis, Davis, CA, USA

UC DAVIS COMPREHENSIVE CANCER CENTER Radiation Oncology

JULY 12–16 2020 VIRTUAL JOINT AAPM | COMP MEETING EASTERN TIME [GMT-4] Improving Health QUALITY. Increasing Global IMPACT

INTRODUCTION

To practice evidence based medicine, Radiation Oncologists have to memorize hundreds of treatment recommendations based on a decision tree including patients' pathology, cancer staging, age, co-morbidities and other factors[1].

As new clinical trial results are published, studies have shown multi-year delays to move new best evidence into routine clinical practice[2-4]. For example, more than 30% of women with N2/N3 breast cancer in the US still do not receive the recommended post-mastectomy radiation, increasing the local recurrence rate by 20%-27% and reducing overall survival by 8-9%[5].

Clinicians in small rural clinics in the US and in developing countries face additional barriers to implement new evidence, because they often do not have resources to access peer-reviewed literature or attend conferences to update their clinical expertise.

AIM

Our aim to identify and engineer an application feature space for an NLP topic model to classify and visualize new peer-reviewed clinical trials publications according to the classification system developed by Hansen et.al.[1]. In consultation with our physicians, the focus will be on early stage breast cancer trials, because there are both a significant number of trial publications and standard treatment approaches accepted throughout the Radiation Oncology community to use as gold standards to compare our results with. Based on initial findings, we can expand our study focus to other disease sites.

METHOD

The Literature Review System is divided into four containers:

1) **Front End:** This is written in Typescript and uses react framework, and serves views to the user. It allows user to interact with the system by adding new papers, rating the saved papers and viewing treatment recommendations.
2) **Database:** It stores all information about the user, papers saved by the user and the paper recommendations made by the model.
3) **Backend:** The server acts as a gateway between the frontend and the database, mostly making database queries and forwarding the response to the front end.
4) **Data Pipeline:** This makes request to PubMed database looking for the papers relevant to the treatment as requested by the user.

In order to display the treatment recommendations in order of their relevance, Principal Component analysis was used. This helped in making clusters of papers which had high frequency of related terms. The core papers from the Handbook of Evidence Based Radiation Oncolgy were marked and their centre of mass was found. This point was assumed to be the point of most relevance. The papers from that cluster were ranked in order of the shortest distance from the centre of mass, which gave the order of relevance.

RESULTS

Early stage breast cancer was chosen as test scenario because of prolific clinical trial data and mostly unambiguous treatment recommendations,. A user interface, predictive literature search model, and automated data gathering subsystem were built. User-rating of suggested papers was implemented to improve modeling. Papers were evaluated based on similarity through term-frequency and inverse-document frequency analytics. Heuristic similarities were used to augment TD-IDF scores. We used Natural Language Processing techniques on abstracts from a subset of papers for primary component analysis with the goal to build similarity clusters to improve the model.

The initial deployment of the model-based literature search component showed that the ML algorithm needs to be augmented by NLP-based clustering analysis to improve filtering of relevant information. The accuracy of the NLP analysis is limited by paywall restrictions of peer-reviewed publications, which currently limits utility of the decision support tool for physicians with limited subscription resources. E.g. at UC Davis, all full text non-open-access publications from Elsevier have to be requested through the library, limiting our ability to expand the NLP analysis beyond the abstracts. **Preprint publication of clinical trial data could overcome this barrier.**

Principal component analysis suggests that papers cluster around treatment techniques (IORT vs. IMRT/VMAT), which will allow model finetuning based on the technology available to the physician user of the decision model.

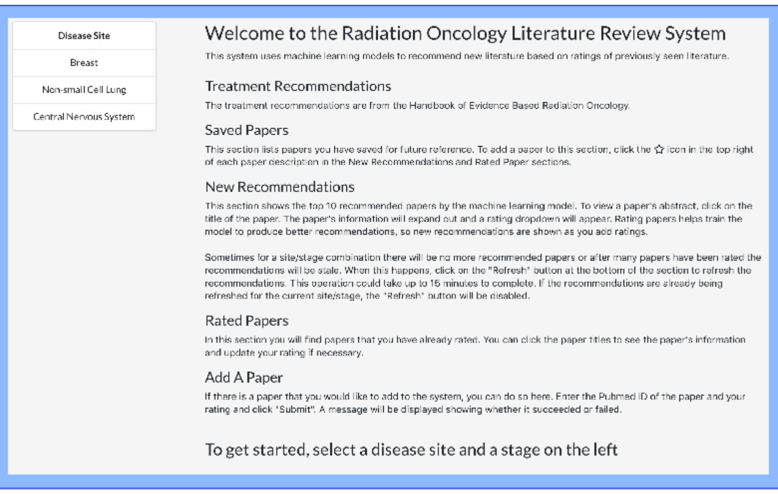

Figure 1: Decision Support Landing Page with current features. In the next software version, new recommendations will combine user ratings with the results from the primary component analysis.

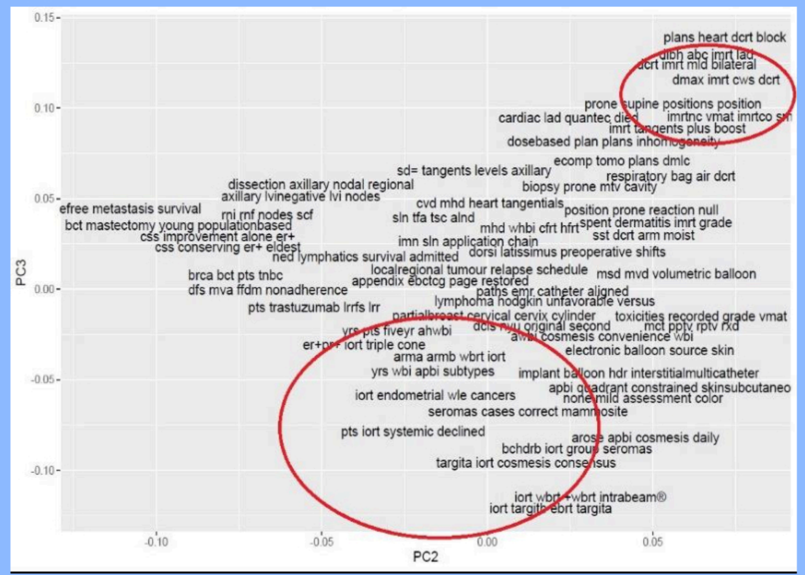

Figure 2: Principal component analysis showing the clustering of IORT vs. VMAT/IMRT technology use. This allows finetuning of the user-facing software interface to reduce visibility of studies using technology not available to the radiation oncologist

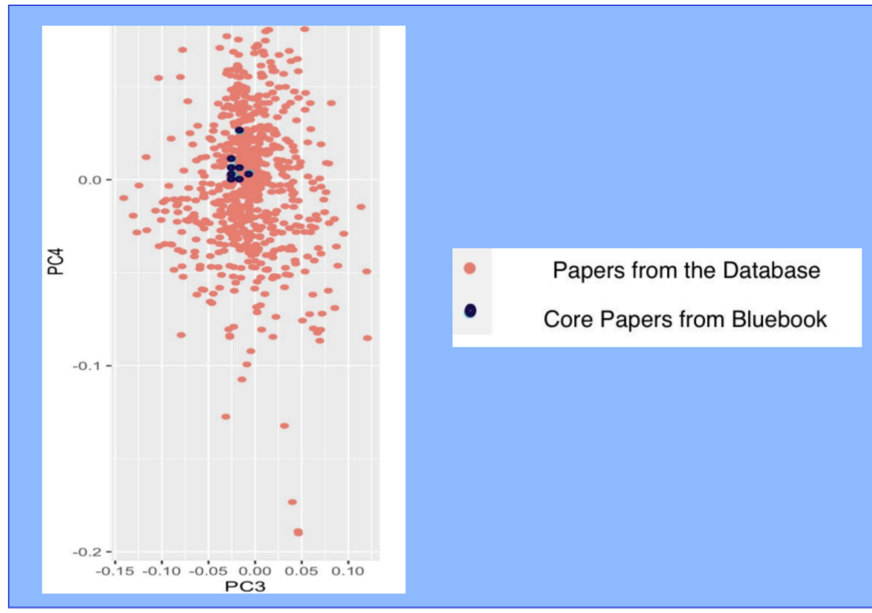

Figure 3: Principal component analysis showing the location of the "gold standard" reference papers for early stage breast cancer from Hansen et.al.[1] in blue.

CONCLUSIONS

This feasibility study of a decision support system has demonstrated that machine-learning algorithms combined with natural language processing can provide clinical radiation oncologists with a relevant selection of peer-reviewed clinical trial publications tailored to an individual patient's pathology and staging. This reduces mental load and could shorten current multi-year delays of implementing evidence-based recommendations clinical practice.

We are further investigating the potential for using text mining and natural language processing techniques for automated "reading" and classification of new research and mapping relevant studies to appropriate diagnostic factors. We expect that this could be accomplished by developing a language model capable of classifying the referenced literature according to factors such as study design, cohort factors, type of cancer, etc. There are several possible approaches to this problem, including techniques such as entity extraction and matching/classifying, PCA classification based on TF-IDF weightings, and using LDA topic model probabilities to develop topic covariant models that can be mapped to the structured data represented in the current diagnostic criteria.

Our next step will be the development and testing of these and/or other methods in order to establish an effective, proven method of programmatically mapping newer research to the existing diagnostic framework.

ACKNOWLEDGEMENTS

We thank Amy Studer from the UC Davis library for her assistance with the literature search. Pamela Reynolds from the UC Davis Data Science Initiative provided support for the R-based NLP analysis. The first version of software was produced by UCD ECS193 students (Andy Wu, Scott Reichelt, Robert Tucker).

REFERENCES

[1] E.K. Hansen, M. Roach, *Handbook of evidence-based radiation oncology.* (Springer, 2010).
[2] R. Grol, J.J.T.I. Grimshaw, "From best evidence to best practice: effective implementation of change in patients' care," **362**, 1225-1230 (2003).
[3] J. Ploeg, B. Davies, N. Edwards, W. Gifford, P.E.J.W.o.E.B.N. Miller, "Factors influencing best-practice guideline implementation: lessons learned from administrators, nursing staff, and project leaders," **4**, 210-219 (2007).
[4] G.J.S.m.j. Szulanski, "Exploring internal stickiness: Impediments to the transfer of best practice within the firm," **17**, 27-43 (1996).
[5] Q.D. Chu, G. Caldito, J.K. Miller, B. Townsend, "Postmastectomy Radiation for N2/N3 Breast Cancer: Factors Associated with Low Compliance Rate," Journal of the American College of Surgeons **220**, 659-669

CONTACT INFORMATION

AUTHOR CONTACT: sdieterich@ucdavis.edu