Presented at the
Practical Big Data Workshop 2023

AI-powered curation tools for radiotherapy (RT) planning data: structure name standardization and intended target dose inference

Julie Shade Ph.D., Pranav Lakshminarayanan, Michael Bowers, David Murray Ph.D., Peter Hoban Ph.D., Todd McNutt Ph.D. FAAPM, MAY 2023

Introduction: For RT treatment planning data to be used to develop machine learning models that predict organ at risk (OAR) doses using plan geometry and target prescribed doses, two curation steps are essential: (i) standardizing OAR nomenclature across training and validation datasets, and (ii) ensuring that target volume(s) and associated intended doses are accurately identified (intended target/dose; ITD). The same requirement exists for datasets to be subjected to analytics.

Objectives: The goal of this study was to build a cloud-based (Microsoft Azure Synapse) RT data curation platform to perform matching of incoming OAR names to TG-263 compliant nomenclature and to perform ITD identification. As plans are curated, the algorithms are intended to be re-trained with curated data.

OAR Naming Standardization: A single-institution dataset containing 164614 structures with 8855 unique structure names was labeled with 465 TG-263 term/qualifier (PRV, resident, etc.) labels, across Head and Neck (H&N), Prostate, Pancreas, and Thoracic disease sites. The data were augmented by randomly introducing typos, transposing words, and removing spaces. Fasttext natural language processing text classification models were developed and achieved testing macro-averaged F1-scores of 0.981 [95% CI: 0.978, 0.984], 0.969 [0.965, 0.973], 0.990 [0.988, 0.993], and 0.983 [0.980, 0.986], respectively.

ITD Algorithm: The ITD algorithm first calculated dose volume histograms (DVHs) for all unions of structures identified as possible targets. Unsupervised agglomerative clustering was used to identify distinct targets. A regression model was then used to infer the prescribed dose(s) using DVH features.

Case Study: An external dataset of 334 DICOM RT plans was de-identified and automatically curated using Synapse pipelines. The dataset contained 26833 structures (4767 unique names). 4586/4767 unique names were not present in the unaugmented development set; 3913 (85.32%) were matched correctly. Curation of the dataset, including manual review of naming and ITD, took under 4 hours.

Conclusions: We have developed comprehensive tools for automatic data curation that accurately infer TG-263 structure names, treatment targets, and prescriptions, enabling advanced analytics and ML model development.

Oncospace, Inc.
1812 Ashland Avenue
Baltimore, MD, 21205
+1-443-773-2271

LinkedIn | Twitter
www.Oncospace.com
Information@Oncospace.com