considered ($P < 0.01$). When evaluating OARs individually, MC showed significantly higher ratings for brainstem, esophagus, larynx, eyes, optic nerves, while lips, whereas parotids, acoustics, and lenses were indistinguishable.

Conclusion: Simple 3D architectures consistently outcompete more complex networks by quantitative measures. Qualitative assessment for clinical acceptability may not agree with quantitative performance, especially when the entire range of OARs is evaluated.

Abstract 2152 – Table 1

	DLC		MC		T-test
OARs	mean	SD	mean	SD	*P*-val
All	3.2	0.9	3.8	0.8	< 0.001
Mandible	3.8	1.3	4.3	0.8	0.11
Parotid	4.0	1.1	4.0	1.0	1
Acoustics	3.0	0.8	3.0	0.7	0.64
Brachial plexus	3.8	0.7	3.7	0.8	0.34
Eye	3.0	0.8	3.6	0.8	< 0.001
Optic nerve	3.0	0.7	4.0	0.7	< 0.001
Chiasm	2.5	1.1	2.8	1.3	0.29

Author Disclosure: J. Marsilla: None. J. Kim: None. S. Kim: None. D. Tkachuk: None. A.J. Hope: Honoraria; AstraZeneca. Travel Expenses; Elekta, Inc. B. Haibe-Kains: Canada Research Chair in Computational Pharmacogenomics; Canada Research Chair in Computational Pharmacogenomics.

2153

Initial Plan Quality Evaluation Using a Novel AI-Driven Planning System and Paradigm for Adaptive Head and Neck Patients

N. Nasser,[1,2] J.J. Caudell,[2] E.G. Moros,[2] V. Feygelman,[2] and G. Redler[2]; [1]*University of South Florida, Tampa, FL,* [2]*H. Lee Moffitt Cancer Center and Research Institute, Department of Radiation Oncology, Tampa, FL*

Purpose/Objective(s): An innovative ring gantry CBCT-guided adaptive radiotherapy system enables efficient planning and online adaptation to account for interfraction changes. Initial planning strategies and plan quality for head and neck (H&N) patients using the artificial intelligence (AI) plan optimizer for this system are evaluated.

Materials/Methods: Nine previously treated H&N patients are used in this study. Planning target volumes (PTVs) have prescribed dose levels: $PTV_{high} = 70.0/69.3/66.0$, $PTV_{med} = 63.0/60$, and $PTV_{low} = 56.0/54.1/54.0/52.8$ Gy. Clinical planning CT and structures are used for AI plan generation with the novel paradigm based on clinical goal prioritization. The system automatically optimizes 5 candidates plans: 7/9/12 equidistant field IMRT and 2/3 full arc VMAT. Three planning approaches are tested: 1) input physician goals; 2) prioritize PTV coverage by excluding organ at risk (OAR) goals when OAR intersects PTVs; 3) generate and assign goals to OAR subvolumes cropped from PTVs. Utility of generalized helper structures for controlling hot spot location and low dose spillage is investigated. The best of each patient's 5 candidate AI plans (AcurosXB) is compared to clinical (collapsed cone) plan using NRG guidelines.

Results: Strategies 1 and 2 consistently result in PTV under-coverage, high PTV hot spots, and failed OAR goals when intersecting PTVs. Strategy 3 provides acceptable PTV coverage and OAR dosimetry comparable to clinical plans. Helper structures (posterior block, distal OAR subvolumes, and $PTV_{med/low}$ cropped from PTV_{high}) were found to be necessary for shaping dose distributions similarly to clinical plans. Optimal plan type varied: 12 (n = 5), 9 (n = 2), 7 (n = 1) field IMRT and 3 arc VMAT (n = 1). Comparison of AI and clinical plans is based on using strategy (3) with helper structures. Compared to clinical plans, for all PTVs, average $D_{99\%}$, $D_{95\%}$, and D_{max} are higher by 3.1% (103.6% vs. 100.5%), 1.3% (102.2 vs. 100.9%), and 3.3% (108.8% vs. 105.5%), respectively, but still within NRG guidelines. However, global hot spots may fall outside of PTVs and are higher by 5%. Hot spot differences may be from differences in dose calculation algorithms. The spinal cord, brainstem, parotid, oral cavity, mandible and thyroid doses are below NRG guidelines for both plans, with AI plan doses slightly lower. Larynx, Pharynx and submandibular doses on average are higher than the NRG guidelines due to PTV proximity for both plans, with AI plan doses slightly higher.

Conclusion: The AI plan optimizer for this adaptive platform utilizes a novel planning paradigm based on clinical goals rather than direct optimization parameters and can efficiently generate H&N treatment plans. AI plans are comparable to clinical plans with slightly better PTV coverage and lower OAR doses. However, issues with higher and spatially undesirable calculated hot spots remain. All plans meet NRG guidelines.

Author Disclosure: N. Nasser: Research Grant; Varian Medical Systems, Inc. J.J. Caudell: Research Grant; Varian Medical Systems. Honoraria; Varian Medical Systems. Consultant; Varian Medical Systems. E.G. Moros: Research Grant; Varian Medical Systems. Promote research in the AAPM; American Association of Physicists in Medicine. Promoting research in the AAPM; American Association of Physicists in Medicine. V. Feygelman: Research Grant; Varian Medical Systems, Inc. G. Redler: Research Grant; Varian Medical Systems, Inc.

2154

Deep Learning and Harmonization of Multi-Institutional Data for Automated Gross Tumor and Nodal Segmentation for Oropharyngeal Cancer

D. Plana,[1] P.K. Guntaka,[2] J.M. Qian,[3,4] P. Zhou,[1] A. Hosny,[3,5] D.N. Margalit,[5,6] J.D. Schoenfeld,[5,6] R.B. Tishler,[5,6] R.I. Haddad,[5] R. Uppaluri,[5] B. Haibe-Kains,[7] H. Aerts,[3] and B.H. Kann[3,5]; [1]*Harvard Medical School, Boston, MA,* [2]*Harvard School of Dental Medicine, Boston, MA,* [3]*Artificial Intelligence in Medicine (AIM) Program, Brigham and Women's Hospital, Harvard Medical School, Boston, MA,* [4]*BWH/DFCI/MGH - Harvard Radiation Oncology Program, Boston, MA,* [5]*Dana-Farber Cancer Institute, Boston, MA,* [6]*Department of Radiation Oncology, Brigham and Women's Hospital, Boston, MA,* [7]*Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada*

Purpose/Objective(s): Automated tumor segmentation for oropharyngeal cancer (OPC) has the potential to improve treatment planning, response assessment, and clinical translation of imaging-based biomarkers. Deep learning has shown promise for cancer imaging segmentation, but performance for OPC tumors has been suboptimal with studies generally limited to small single-institution settings. In this study, we curated and harmonized multiple heterogeneous, multi-institutional datasets to develop and validate computed tomography (CT)-based, deep learning models for total gross tumor volume (GTV), primary (GTVp), and nodal (GTVn) segmentations.

Materials/Methods: Data was obtained from The Cancer Imaging Archive (TCIA) and included 1228 CT simulation scans from OPC patients treated with definitive radiotherapy collected from 2003-2014 from four institutions that included original gross tumor volumes (GTV) as delineated by the treating radiation oncologist. GTVs were manually reviewed and ground-truth labels were harmonized to include distinct GTVp and GTVn labels. Cases were split randomly, such that 70% of cases were used for model training, 15% for tuning, and 15% for independent performance testing. Utilizing a modified 3D UNET-based architecture, models were trained and tuned to predict total GTV, GTVp and GTVn. Model performance was assessed by measuring precision, recall, and dice score coefficient (DSC) when comparing deep learning-generated to ground truth volumes in the independent test set. Patients without contrast-enhanced CT were excluded from the GTVp model a priori, given the importance of contrast in delineating primary tumor.

Results: Algorithms' median performance on test sets with 95% confidence intervals. GTVp test set limited to contrast-enhanced scans. Within the total dataset (n = 1228), median age was 59, cases were 82% male, and HPV status was 49% positive, 16% negative, and 35% unknown. Tumor staging was 45% T3-4, 54% T1-T2, and 1% T0/X. Nodal staging was 75% N2-N3 and 25% N0-1. Overall, the GTVn model had the highest performance (median DSC: 0.76; 95% CI: 0.73, 0.77), followed by total GTV (median DSC: 0.71; 95% CI: 0.68, 0.73), and GTVp (median DSC 0.68; 95% CI: 0.63, 0.71).

Conclusion: Deep learning with multiple harmonized data sources can yield effective models for OPC primary and nodal segmentation using CT alone. The utility of these models will depend on the clinical use case and will be explored on further investigation, though current model performance metrics, particularly for nodal segmentation, are likely adequate for prospective testing in clinical and research applications.

Abstract 2154 – Table 1

	Dice Score	Precision	Recall
GTV (n = 142)	0.71 (0.68, 0.73)	0.76 (0.72, 0.79)	0.66 (0.63, 0.72)
GTVn (n = 142)	0.76 (0.73, 0.77)	0.82 (0.76, 0.84)	0.73 (0.69, 0.77)
GTVp (n = 112)	0.68 (0.63, 0.71)	0.71 (0.68, 0.75)	0.71 (0.64, 0.76)

Author Disclosure: D. Plana: None. P.K. Guntaka: None. J.M. Qian: None. P. Zhou: None. A. Hosny: None. D.N. Margalit: None. J.D. Schoenfeld: Research Grant; BMS, Merck, Regeneron. Consultant; Debiopharm, Tilos, Catenion, LEK, Immunitas, STIMIT, Astellas. Advisory Board; BMS, Debiopharm, AstraZeneca, Nanobiotix, Immunitas. Travel Expenses; BMS, Debiopharm. Stock Options; Immunitas; NCI Match Subprotocol Z1D. R.B. Tishler: None. R.I. Haddad: None. R. Uppaluri: Merck. B. Haibe-Kains: None. H. Aerts: None. B.H. Kann: None.

2155

Automatic Extracapsular Extension Identification in Head and Neck Cancer Using Deep Neural Network with Local-Global Information

Y. Wang,[1] T.V. Thomas,[2] W.N. Duggar,[2] P.R. Roberts,[2] R.T. Gatewood,[3] L. Bian,[1] and H. Wang[1]; [1]*Mississippi State University, Mississippi State, MS,* [2]*Department of Radiation Oncology, University of Mississippi Medical Center, Jackson, MS,* [3]*University of Mississippi Medical Center, JACKSON, MS*

Purpose/Objective(s): The extracapsular extension (ECE) is a strong predictor of patients' survival outcomes with head and neck squamous cell carcinoma (HNSCC). ECE occurs when metastatic tumor cells within the lymph node break through the nodal capsule into surrounding tissues. It is crucial to identify the occurrence of ECE as it changes staging and management for the patients. Current clinical ECE detection relying on radiologists' visual identification is extremely labor-intensive, time-consuming, and error-prone, and consequently, pathologic confirmation is required. Therefore, we aim to perform ECE identification automatically using a deep learning-based technique to analyze the presence or absence of ECE and correlate that with gold standard histopathological findings.

Materials/Methods: This research proposes a novel deep learning method to detect ECE from 3D computed tomography (CT) scans. The proposed network has multi-scale input that captures both local and global information. The two types of inputs are fitted into two pathways with deep convolutional neural networks (DCNNs), 3D CNN baseline and 3D DenseNet. A sliding-cube approach is applied to extract small paired 3D patches from patient data with different scales at data preparation. The network will classify the patient based on patch-level classification results. Different training scenarios are designed for the experimental test.

Results: Based on five-fold cross-validation, the experimental results have demonstrated that our model can identify ECE and non-ECE patients, specifically in the patch-level. We have achieved the ECE detection with 96.92% accuracy 98.84% AUC. Detailed results are shown in the Table 1.

Conclusion: The study demonstrates the ability to use a deep learning-based method for ECE direct detection. The proposed local-global network can help capture sufficient features and achieve high classification accuracy. The outcome of this research is expected to promote the implementation of artificial intelligence for ECE identification for head and neck cancer diagnosis in the radiology computer vision field.

Abstract 2155 – Table 1: ECE Classification Results With Different Deep Learning Models

	3D DenseNet		3D Baseline CNN		
DL Model	Single-input local	Multi-input	Single-input local	Single-input global	Multi-input
Training accuracy	0.9044	1.0000	0.9020	0.9947	0.9925
Validation accuracy	0.6918	0.9532	0.8540	0.9160	0.9692
Training AUC	0.9367	1.0000	0.9458	0.9990	0.9961
Validation AUC	0.7152	0.9824	0.9055	0.9562	0.9884

Author Disclosure: Y. Wang: None. T.V. Thomas: Employee; University of Mississippi Medical Center; AMA, ASTRO. W.N. Duggar: None. P.R. Roberts: None. R.T. Gatewood: None. L. Bian: None. H. Wang: None.

2156

Patient-Reported Outcome+ Platform for Remote Symptom Management, Featuring Automated Triage System

X. Tang, N. Hacker, A. Danish, and S. Chen; *Liyfe, Inc., New York, NY*

Purpose/Objective(s): Despite the promising results, symptom monitoring is not widely implemented in the clinic. This is largely due to the additional staffing that is needed to triage the increased patients' symptom reporting. We herein develop a PRO+ platform for remote symptom management, where the plus sign stands for automated triage pathways. This includes AI and evidence-based triage chatbot and severity classification algorithm, to reduce manual triage needs.

Materials/Methods: 22 most common symptoms were included in the triage system. For each symptom, triage pathways were built in chatbot format. Typically, a triage starts with onset questions, like "when did it start?", "how many episodes?", followed by PRO CTCAE evaluation, how daily life was affected, home medication and remedies patients used, associated symptoms, etc. For most questions, multiple choices were identified as potential answers for patients to choose from. I.e., 10 scales were provided for pain evaluation. Some questions offer sharing pictures or free text. A decision tree-based classification algorithm was developed to classify the reported symptoms to non-urgent, urgent, and emergent situations. Three metrics were used for ongoing evaluation during development. 1) Patient engagement rate. Volunteer patients used the system to report symptoms. If a chatbot conversation was initiated and finished, it was counted as a successful engagement. 2) Chatbot triage efficiency. Nurses reviewed the reported symptom details and marked the situations that enough information was collected to provide patients clinical advice. Corresponding percentage was calculated. Although we expect that there will always be cases that additional communication is needed to triage patients, we want to collect enough information to triage most straightforward cases. 3) Classification accuracy. The classification algorithm results were compared with manual results by nurses, and discrepancy was reported.

Results: The PRO+ platform was built. Patients can select any symptom to report. Once a symptom is selected, the chatbot will carry a conversation with patient to collect necessary information for triage. The chat bot conversation is adapted real-time to the reported symptoms and conversation. 150 volunteer patients reported symptoms, and the engagement rate was 91%. 80% of the time the chatbot had collected enough information for nurses to provide clinical advice. The severity classification discrepancy was 95%.

Conclusion: A PRO+ platform was built for symptom management. It has automated triage pathways that can potentially allow clinics to provide remote symptom monitoring with reduced manual triage needs. More