reference contour. Calculations were performed in R using the RadOnc package and manually segmented contours were approved by an expert radiation oncologist with 19 years of experience.

Results: A total of 112 patients were identified for inclusion after manual review of treatment planning CTs. Contours were generated using two atlas/model-based segmentation products and a deep learning segmentation method. The mean DSC for the prostate, bladder, rectum was calculated for each segmentation method (Table 1). Deep learning segmentation outperformed model-based methods for all structures with the highest mean DSC but still had significant disagreement with manually segmented structures. Hip arthroplasty in particular reduced overall performance more than other anatomical edge cases, followed by prostatic hypertrophy.

Conclusion: Anatomic edge cases present a challenging and relevant consideration in clinical implementation of autosegmentation software.

Author Disclosure: A. Kanwar: None. C. Claunch: None. S. Rana: None. B. Merz: None. A.Y. Hung: Employee; Providence Sacred Heart Hospital of Spokane. Member; American College of Radiology. Advisor; American College of Radiology. R.F. Thompson: Data Science Institute; American College of Radiology.

83

Artificial Intelligence Guided Physician Directive Improves Head and Neck Planning Quality and Practice Uniformity: A Prospective Study

M. Mashayekhi, R. McBeth, D. Nguyen, S.B. Jiang, and M.H. Lin; *Department of Radiation Oncology University of Texas Southwestern Medical Center, Dallas, TX*

Purpose/Objective(s): Heterogeneity of clinical practice is a common challenge, especially for head and neck cancer. Our in-house built artificial intelligence (AI) dose prediction model holds promise to predict the best possible dose distribution without going through planning process. Our prior single physician study demonstrated that hybrid of physician intelligence and artificial intelligence achieved better plan than solely using AI estimated directive or physician directive. In this work, we implemented hybrid directive approach in four head and neck physicians with different practice styles and prospectively studied the improvement in head and neck planning quality and practice uniformity.

Materials/Methods: Definitive and post-operative AI models were fully integrated with treatment planning system via application programming interface (API). The physicians run the prediction upon the completion of contour to preview the predicted 3D dose and the uncertainty estimation. As a hybrid approach, physicians compare the AI estimation with their own directive to make final decision on the directives for treatment planning. A group of 60 retrospective cases and 60 prospective cases were included in this study. The physician directive before AI guidance (PD) and the AI guided physician directive (AG-PD) were compared as well as the achieved plans. A dose difference of 3 Gray (Gy) was considered clinically significant.

Results: Among the four H&N physicians, there are three different practice styles and led to a large variation in PD before AI employed. There is only 51% frequency that resulted plan achieved within 3 Gy from the PD. With the AI-guidance, the variations among the physicians significantly improved. Despite the fact that the AG-PD are more mostly strict than the PD, the frequency that the resulted plan achieved within 3 Gy from the AG-PD significantly improved to 86%. The improvements were mostly statistically significant ($P < 0.05$) at the organs at risk that physicians frequently make planning tradeoff, such as oral cavity, larynx, and salivary glands. (Table 1)

Conclusion: AI guided physician decision support improved the uniformity of practice, directive achievability, and the final plan metrics in a significant percentage of patients.

Abstract 83 — Table 1

OAR	MD Directive		Achieved Plan	
	Before AI (Gy)	After AI (Gy)	Before AI (Gy)	After AI (Gy)
Spinal Cord Max	39 ± 6 *	33 ± 2	36 ± 5 *	30 ± 3
Constrictor Mean	34 ± 22	21 ± 1	34 ± 23	19 ± 1
Larynx Mean	31 ± 9 *	20 ± 4	33 ± 14 *	20 ± 3
Oral Cavity Mean	21 ± 5	19 ± 3	25 ± 14	21 ± 3
SMG_CL Mean	34 ± 4	25 ± 13	33 ± 10	25 ± 13

* Statistical significance of the dose comparison, $P < .05$

Author Disclosure: M. Mashayekhi: None. R. McBeth: None. D. Nguyen: None. S.B. Jiang: Partnership; TRIO Inc, TMIT. Patent/License Fees/Copyright; Manteia Technologies. M. Lin: None.

84

Human-Level Precision Upper Abdominal OAR Contouring With Anatomically Preserving Deep Learning During Magnetic Resonance Imaging Guided Adaptive Radiotherapy (MRgRT)

G. Gungor,[1] M. Michalet,[2] A. Lombard,[3] T. Roque,[3] B. Atalar,[4] B. Temur,[4] I. Serbez,[4] D. Azria,[2] L. de Vitry,[3] O. Riou,[2] N. Paragios,[3] E. Ozyar,[1] and P. Fenoglietto[2]; *[1]Department of Radiation Oncology, Acibadem MAA University, Maslak Hospital, Istanbul, Turkey, [2]Montpellier Cancer Institute, Montpellier, France, [3]Therapanacea, Paris, France, [4]Acibadem MAA University, Maslak Hospital, Istanbul, Turkey*

Purpose/Objective(s): Magnetic resonance imaging guided radiotherapy (MRgRT) offers the ability of daily treatment adaptation: a game changer for various cancers. Contouring of organs at risk (OAR) during adaptation is time-consuming and lacks reproducibility across physicians, hampering the accuracy of high precision MRgRT and diminishing its adoption potential. Artificial intelligence (AI) can accelerate and homogenize OAR delineation. This study aims at (i) assessing the reproducibility of clinicians OAR delineation, (ii) comparing the precision between clinical experts (CEs) and AI based contours (AC) and (iii) evaluating the clinical benefit of AI tools for treatment standardization.

Materials/Methods: For the case of low field abdominal MR-based daily treatment adaptation, transfer learning was applied on a CE/FDA-cleared deep learning solution. Models were re-trained using 270 retrospectively selected annotated fractions samples treated with a MR-LINAC at two European cancer care excellence centers. Validation was performed using 2 cohorts of (i) 15 double-blindly annotated patients and (ii) a random 50/50 mix of 30 CEs and AI based annotations. Contours of 8 OARs (right/left kidneys, stomach, liver, duodenum, inferior vena cava, bowel and, abdominal aorta) were scored by 3 CEs as A/ acceptable, B/ acceptable after minor corrections, and C/not acceptable.

Results: The average interobserver variability among the 8 OARs in terms of DICE score coefficient (DSC) was 84.38% with the highest and lowest scores being observed for stomach (95%) and bowel (68%), respectively. The average DSC between CEs and AI annotations was 85.88% with the left/right kidneys (94%) and the duodenum/vena cava (76%) depicting the highest and lowest values, respectively. CE and AI produced annotations scored as A for 89.36% and 71.89% and were considered acceptable (A+B) for 100% and 92.49% of the cases, respectively. AI solutions seem to suffer in organs whit significant discrepancies across CEs for top and the bottom slices.

Conclusion: The results show that AI-driven contours are clinically useable in most cases. Disagreement between experts reflect the subjectivity of scoring. Objective metrics should be used in complement.

Abstract 84 – Table 1: DSC and clinical acceptability comparison between CE and AI based annotations

	Dice (%)		Accept (A+B) (%)	
Organ	CE	AI	CE	AI
Aorta	85	85	100	100
Duodenum	75	76	100	92.6
Bowel	68	78	100	53.7
Left Kidney	92	94	100	100
Liver	91	92	100	100
Right Kidney	92	94	100	100
Stomach	95	92	100	95.5
Vena Cava	77	76	100	98.1

Author Disclosure: G. Gungor: None. M. Michalet: None. A. Lombard: Stock Options; Therapanacea. T. Roque: None. B. Atalar: None. B. Temur: None. I. Serbez: None. D. Azria: None. L. de Vitry: None. O. Riou: None. N. Paragios: Stock; Therapanacea. E. Ozyar: Research Grant; ViewRay. P. Fenoglietto: None.

85

Accurate Prostate Cancer Detection and Segmentation Using Non-Local Mask R-CNN With Histopathological Ground Truth

Z. Dai,[1] I. Jambor,[2] P. Taimen,[3] M. Pantelic,[4] M.A. Elshaikh,[5] A. Dabaja,[6] C. Rogers,[6] O. Ettala,[7] P. Boström,[8] H. Aronen,[9] H. Merisaari,[3] and N. Wen[1]; *[1]Department of Radiation Oncology, Henry Ford Health System, Detroit, MI, [2]Department of Radiology, University of Turku, Turku, Finland, [3]Institute of Biomedicine, University of Turku, Turku, Finland, [4]Henry Ford Health System, Detroit, MI, [5]Department of Radiation Oncology, Henry Ford Cancer Institute, Detroit, MI, [6]Vattikuti Urology Institute, Henry Ford Health System, Detroit, MI, [7]Department of Urology, Turku University Hospital, Turku, Finland, [8]Department of Clinical Medicine, University of Turku, Turku, Finland, [9]Department of Diagnostic Radiology, University of Turku, Turku, Finland*

Purpose/Objective(s): We aim to develop deep learning (DL) models to accurately detect and segment intraprostatic lesions (IL) on biparametric MRI (bp-MRI).

Materials/Methods: Three patient cohorts with ground truth IL delineated on different modalities were collected. 158 patients from two datasets had suspicious ILs delineated based on bp-MRI: 97 patients were from PROSTATEx-2 Challenge with biopsy result independent from bp-MRI based delineation, 61 patients were from IMPROD clinical Trial with biopsy done for each delineation; 64 patients from IMPROD clinical Trial had ILs identified and delineated by using whole mount prostatectomy specimen sections as reference standard; 40 private patients were unlabeled. We proposed a non-local Mask R-CNN to improve segmentation accuracy by addressing the imperfect registration issue between MRI modalities. We also proposed to post aggregate 2D predictions to estimate IL volumes within the whole prostatic gland and keep top-5 3D predictions for each patient. In order to explore the small dataset problem, we employed different learning techniques including transfer learning and semi-supervised learning with pseudo labelling. We experimented with two label selection strategies to see how they affected model performance. The first strategy kept only one prediction by referring to biopsy result, in order to minimize false positives; while the second strategy kept all top-5 predictions. 3D top-5 detection rate, dice similarity coefficient (DSC), 95 percentile Hausdorff Distance (95 HD, mm) and true positive ratio (TPR) were our evaluation metrics. We compared DL model prediction with prostatectomy-based ground truth delineation to accurately evaluate the boundary and malignancy level. We separately evaluated ILs of all Gleason Grade Group (GGG) and clinically significant ILs (GGG > 2).

Results: Main results are summarized in Table 1.

Conclusion: Our proposed method demonstrates state-of-art performance in the detection and segmentation of ILs and shows great effectiveness for clinically significant ILs.

Abstract 85 – Table 1: Results of models

Model	BMP	NMP	fNMBP	S1-1	S2-3
All GGG					
Detection Rate	68.3%	78.0%	75.6%	80.5%	80.5%
DSC	0.429 ± 0.165	0.504 ± 0.165	0.543 ± 0.159	0.548 ± 0.165	0.513 ± 0.191
95 HD	7.65 ± 4.50	6.66 ± 3.64	6.28 ± 3.47	5.72 ± 3.17	5.81 ± 2.89
TPR	0.558 ± 0.268	0.589 ± 0.222	0.625 ± 0.190	0.613 ± 0.193	0.749 ± 0.190
GGG ≥ 2					
Detection Rate	89.5%	89.5%	89.5%	94.7%	84.2%
DSC	0.469 ± 0.171	0.516 ± 0.176	0.579 ± 0.155	0.604 ± 0.135	0.631 ± 0.122
95 HD	8.10 ± 5.26	7.51 ± 4.09	6.72 ± 3.73	6.36 ± 3.44	6.16 ± 3.08
TPR	0.539 ± 0.270	0.544 ± 0.217	0.593 ± 0.196	0.580 ± 0.190	0.746 ± 0.165

The number of ILs in GGG 1, 1+, 2, 2+, 3, 3+, and 4 were 4, 2, 16, 1, 7, 3, and 8 respectively. BMP, baseline Mask R-CNN; NMP, non-local Mask R-CNN; fNMBP, non-local Mask R-CNN trained with bp-MRI based delineation then fine-tuned with prostatectomy-based delineation; S1-1, strategy 1 at 1st iteration; S2-3, strategy 2 at 3rd iteration. Statistically significant ($\alpha = 0.05$), H_0 = model performs worse than BMP.

Author Disclosure: Z. Dai: None. I. Jambor: None. P. Taimen: None. M. Pantelic: None. M.A. Elshaikh: None. A. Dabaja: None. C. Rogers: None. O. Ettala: None. P. Boström: None. H. Aronen: None. H. Merisaari: None. N. Wen: Employee; William Beaumont Hospital. Research Grant; American Cancer Society. Honoraria; Varian Medical System.

86

Improving GI Toxicity Models Through Deep Learning-Based Segmentation and Biomechanical Model-Based Dose Accumulation

M.M. McCulloch,[1] G. Cazoulat,[1] B.M. Anderson,[1] E. Kirimli,[2] B. Rigaud,[1] S. Gryshkevych,[3] S. Svensson,[4] A.N. Ohrt,[1] J. Ohrt,[5] N. Chopra,[1] R. Mathew,[2] M. Zaid,[2] D. Elganainy,[6] P. Balter,[1] E.J. Koay,[2] and K.K. Brock[1]; *[1]The University of Texas MD Anderson Cancer Center, Houston, TX, [2]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, [3]RaySearch Laboratories, AB, Stockholm, Sweden, [4]RaySearch Laboratories AB, Stockholm, Sweden, [5]MD Anderson Cancer Center, Houston, TX, [6]Dana-Farber Cancer Institute, Boston, MA*

Purpose/Objective(s): Current GI toxicity models are limited by the uncertainty in the estimated delivered dose used to develop the models. The purpose of this study is to assess the benefit of harnessing deep learning and biomechanical models to improve the understanding of the dose-toxicity relationship.

Materials/Methods: Retrospective dose accumulation was conducted on 75 patients with primary and metastatic liver cancer treated with external beam radiotherapy guided by daily CT-on-rails (CTOR), based on a novel deformable image registration (DIR) approach that applies biomechanical model-based and intensity-based DIR inside and outside the liver, respectively. The liver was auto-segmented on all images using a clinically validated 2D deep learning model. On a sub-cohort, the combination of stomach and duodenum region of interest (GI ROI) was auto-contoured using a 3D UNet model trained independently on 102 patients. These contours were compared to physician-drawn contours on CTOR, using Dice similarity coefficient (DSC), mean distance to agreement (DTA), Hausdorff distance (HD), and planned dose metrics. Intra-observer variability was evaluated. Doses were converted to equivalent dose in 2Gy fractions (EQD2) for analysis ($\alpha/\beta = 3.5$). Differences between planned and delivered dose were calculated, as well as GI ROI daily dose variation. Toxicity probability was calculated based on previously developed accumulated dose normal tissue complication probability (NTCP) models for duodenum and stomach.

Results: Average (SD) DSC, DTA, and HD between the deep learning model and physician-drawn GI ROI were 0.9 (0.1), 0.2 (0.1) cm, and 2.4