Title

A knowledge-based organ dose prediction tool for brachytherapy treatment planning of patients with cervical cancer

Journal

Brachytherapy, 19(5)

ISSN

1538-4721

Authors

Yusufaly, Tahir I
Kallis, Karoline
Simon, Aaron
et al.

Publication Date

2020-09-01

DOI

10.1016/j.brachy.2020.04.008

A knowledge-based organ dose prediction tool for brachytherapy treatment planning of cervical cancer patients

Abbreviations: BT = brachytherapy, DVH = dose-volume histogram, EBRT = external beam radiotherapy, GYN = gynecologic, HRCTV = high-risk clinical target volume, IMRT = intensity-modulated radiation therapy, OAR = organ-at-risk, T&O = tandem-and-ovoid

Purpose: To explore knowledge-based organ-at-risk (OAR) dose estimation for intracavitary brachytherapy planning for cervical cancer. Using established external-beam knowledge-based dose-volume histogram (DVH) estimation methods, we sought to predict bladder, rectum, and sigmoid D_{2cc} for tandem-and-ovoid treatments.

Methods: 136 loco-regionally advanced cervical cancer patients treated with 456 (356:100 training:validation ratio) CT-based tandem-and-ovoid brachytherapy fractions were analyzed. Single fraction prescription doses were 5.5-8 Gy with dose criteria for the high-risk clinical target volume (HRCTV), bladder, rectum and sigmoid. DVH estimations were obtained by subdividing training set OARs into HRCTV boundary distance sub-volumes and computing cohort-averaged differential DVHs. Full DVH estimation was then performed on the training and validation sets. Model performance was quantified by $\Delta D_{2cc}=D_{2cc}$(actual)-D_{2cc}(predicted) (mean and standard deviation). ΔD_{2cc} between training and validation sets were compared with a Student's t-test ($p<0.01$ significant). Categorical variables (physician, fraction-number, total fractions, case complexity) that might explain model variance were examined using an ANOVA test (Bonferroni-corrected $p<0.01$ threshold).

Results: Training set deviations were bladder ΔD_{2cc} = -0.04$\pm$0.61 Gy, rectum ΔD_{2cc} = 0.02$\pm$0.57 Gy and sigmoid ΔD_{2cc} = -0.05$\pm$0.52 Gy. Model predictions on validation set did not statistically differ: bladder ΔD_{2cc} = -0.02$\pm$0.46 Gy ($p=0.80$), rectum ΔD_{2cc} = -0.007$\pm$0.47 Gy ($p=0.53$), and sigmoid ΔD_{2cc} = -0.07$\pm$0.47 Gy ($p=0.70$). The only significant categorical variable was attending physician for bladder and rectum $\Delta D_{2cc.}$

Conclusion: A simple boundary distance-driven knowledge-based DVH estimation exhibited promising results in predicting critical brachytherapy dose metrics. Future work will

examine the utility of these predictions for quality control and automated brachytherapy planning.

Key Words: knowledge-based planning, cervical cancer, dose predictions, machine learning, quality control, treatment planning

Introduction

Brachytherapy (BT) is an essential component of cervical cancer treatment, which has been linked to improved pelvic control and disease-free survival (1,2). Image guidance allows for tumor dose escalation and normal tissue sparing, by enabling applicator and subsequent source positioning to be tailored to individual tumor features and anatomy. High-quality gynecologic (GYN) BT requires a skilled, coordinated multi-disciplinary team to carry out labor-intensive workflows (3). Currently, clinicians rely on their BT experience, serial on-treatment pelvic exams, and images from earlier time points to make decisions about applicator type prior to the procedure. Quality assessment for BT treatment plans is challenging, as patient anatomy, applicator choice, implant quality, and source loading pattern can all affect tumor coverage and organ-at-risk (OAR) sparing. Currently there are no standardized tools to assist practitioners in troubleshooting cases that do not achieve dosimetric goals. Furthermore, utilization of BT for cervical cancer is declining, some of which could be due to the requirements of the sophisticated workflow (4). This comes at a cost to patients, as a lack of BT is associated with reduced cause-specific and overall survival (5). Additionally, BT remains the standard of care, as demonstrated by a recent study that evaluated the use of stereotactic body radiation therapy (SBRT) in place of BT for cervical cancer treatment and closed early due to concerns for toxicity (6).

Computational prediction of achievable dosimetric parameters could increase clinical efficiency, improve treatment quality, and expand the accessibility and utilization of BT. Population-based guidelines, and current protocols such as the ongoing clinical trial EMBRACE II recommended dose constraints (7), provide clinicians with static plan quality metrics against which to assess individual BT treatment plan quality. While useful for ensuring patients do not exceed critical normal tissue limits, such guidelines are not patient-specific and only provide

limits. In analogous situations of external beam radiotherapy (EBRT) where only static population-based limits are used as plan quality guidance, high degrees of plan quality variability and excess dose to normal tissues have been observed (8,9).

Machine learning methods have been applied to EBRT to automate human-driven processes through a technique known as knowledge-based planning (KBP) (10–14). Knowledge-based dose estimation models are trained on large datasets of prior treatments and provide patient-specific dosimetric predictions for new patients. Automated planning with KBP is accomplished using patient-specific dose predictions to guide plan optimization. These approaches have not been systematically translated to the unique challenges of intracavitary GYN BT, where dosimetry is highly constrained by the implanted applicator and the degrees of freedom for dose modulation are reduced. The purpose of this work was to explore the accuracy of knowledge-based OAR dose estimation for high-dose rate (HDR) BT treatment of cervical cancer. Using established external-beam knowledge-based DVH estimation methods (11), we sought to accurately predict bladder, rectum, and sigmoid D_{2cc} for standard tandem-and-ovoid (T&O) treatments. These dose estimations are an important precursor and step towards knowledge-based planning in BT.

To our knowledge, this work is the first application of knowledge-based dose estimation to GYN intracavitary BT where models are based only on contours of organs and target. This approach could facilitate multi-institutional data-driven quality control, and increase BT utilization by giving clinicians objective assurances that their dose distributions are of high quality.

Materials and Methods

Model specification

The mathematical framework employed in this work for DVH estimation of OARs is closely related to an approach developed for intensity-modulated radiation therapy (IMRT) (11). The planning datasets consist of structure sets for the OARs, OAR_{ij}, for i=1...N cases (where "case" refers to a single fraction in a patient's course of treatment) and j=1…M OARs, and corresponding structure-specific dose matrices $D[x, y, z; i, j]$, where (x, y, z) is the 3D-position of a point that lies within the j^{th} structure. To facilitate equivalence we normalize dose in the i^{th} case to its prescription dose, $\widetilde{D}[x, y, z; i, j] = D[x, y, z; i, j] / D_{Rx_i}$.

Our model is built on the boundary distance feature that quantifies the minimum distance r between any OAR volume element and the high-risk clinical target volume (HRCTV) target (11). The primary assumption of this model is that the probability that a voxel in the j^{th} organ will take a dose value between $\widetilde{D}$ and $\widetilde{D}+\Delta\widetilde{D}$ is given by $p_j(\widetilde{D}; r)\Delta\widetilde{D}$, with $\int_0^\infty p_j(\widetilde{D}; r)\, d\widetilde{D} = 1$. That this probability distribution is a function of r implies that any two points equidistant from their respective HRCTV boundaries within the same organ will have the same normalized dose distribution with respect to the prescription dose.

The practical consequences of these assumptions are that the dose-distance data of each of the N cases can be pooled to generate an enlarged dataset used to estimate the 'ensemble-averaged' dose-distance kernels $p_j(\widetilde{D}; r)$ for each of the *j=1-3* OARs (bladder, rectum, and sigmoid). Then, for a new case i_{N+1}, once we extract the differential volume of each OAR as a function of r, $dV_{i_{N+1}j}/dr$, these kernels can be used to predict the differential DVH

$$V'^{pred}_{i_{N+1}}\left(\widetilde{D}_j\right)=\int dr\left(\frac{dV_{i_{N+1}j}}{dr}\right)p_j(\widetilde{D};r)\ .$$

In turn, this is transformed into the cumulative DVH,

$$DVH^{pred}_{i_{N+1}}\left(\widetilde{D}_j\right)=V^{total}_{i_{N+1}j}-\int_0^{\widetilde{D}_j} dD\,V'^{pred}_{i_{N+1}}\left(\widetilde{D}\right)\ ,$$

where $V^{total}_{i_{N+1}j}$ is the total volume of the j^{th} OAR for the case. From this predicted cumulative DVH, D_{2cc} metrics of the bladder, rectum, and sigmoid are extracted, since D_{2cc} is currently the only standardized OAR DVH metric that is used to evaluate clinical plans.

Model training and validation

136 loco-regionally advanced cervical cancer patients, over a six-year period (2012-2018, UCSD IRB Project #181609), treated with *N*=456 (356:100 training:validation ratio) T&O CT-guided BT fractions were analyzed retrospectively in an integrated training-validation workflow as illustrated in Figure 1. The 100 case validation set was composed of all 5 fractions of 10 patients completely independent of the training set, and 50 cases randomly sampled from the remaining 126 patients (consisted of 2 treatment fractions from 9 patients, and single fractions from 33 patients, for a total of 42 patients). The purpose of this was to evaluate model performance on both totally independent patients, as well as independent treatment fractions from patients used to train the model.

Single-fraction prescription doses were 4-8 Gy. Plans were created using institutional dose criteria for external beam + BT equivalent dose in 2 Gy fractions (EQD2s), which were originally based on the 2011 update to the ABS HDR BT guidelines for locally advanced cervical cancer (15), and later updated to incorporate soft constraints from the ongoing EMBRACE-II trial (7). Hard planning constraints include high-risk clinical target volume

(HRCTV) D_{90}>85 Gy, bladder D_{2cc}<90 Gy, rectum D_{2cc}<75 Gy, and sigmoid D_{2cc}<75 Gy. Soft planning aims (recommended but not required) include bladder D_{2cc}<80 Gy, rectum D_{2cc}<65 Gy, and sigmoid D_{2cc}<70 Gy. The HRCTV contour included residual disease at the time of BT and the whole cervix. Our planning process consisted of the following steps. First, the T&O dwell positions were set to a standard loading pattern, and then normalized to deliver prescription to point A. Then radiation oncologists manually adjusted dwell positions or dragged isodose lines to achieve target coverage while minimizing dose to OARs. During this tuning process, EQD2 values were evaluated on a spreadsheet to ensure planning objectives were met.

DVH estimation models were obtained by subdividing OARs into HRCTV boundary distance sub-volumes (extending from overlapping with the HRCTV to 10.6 cm radially from the HRCTV) and computing an ensemble-averaged differential DVH estimate from the training set sub-volumes. Full DVH estimation was performed on all cases in the training and validation sets by applying OAR sub-volume DVH models to each fraction's OARs. The proposed framework was implemented in the form of in-house extensions to MIM (version 9.6.3, MIM Software Inc., Cleveland, Ohio, USA). The DVH predictions for any new case take less than 10 seconds.

Model performance was quantified by analyzing the residual ΔD_{2cc}=¿ Actual D_{2cc} – Predicted D_{2cc}, where D2cc is the absolute OAR dose for a single BT fraction. Standard deviation over these residuals was taken as a measure of model error, as has been done for prior EBRT KBP studies (10–13,16). Goodness-of-fit was measured by the Pearson correlation coefficient R and the variance of the ΔD_{2cc} distribution. We chose to report most analysis and figures in absolute dose, as absolute dose is more commonly used to evaluate OARs during treatment planning, and thus is more clinically meaningful. To ensure this assumption was valid, metrics were also computed for relative dose (i.e. dose normalized to prescription for that BT fraction).

Consistency between the distributions of ΔD_{2cc} for training and validation sets was checked with an unpaired Student's t-test ($p<0.01$ significance threshold).

Variance reduction via incorporation of continuous geometric features

In an attempt to uncover possible anatomic variability not captured by the boundary distance approach, we identified a preliminary list of eleven geometric features suspected to have additional predictive power, including:

1) HRCTV volume
2) An anterior-posterior asymmetry metric $AP_{asymmetry}$, defined as the furthest posterior distance of the HRCTV boundary from the tandem minus the furthest anterior distance of the HRCTV boundary from the tandem
3) Nine additional geometric features measuring OAR orientation relative to the base of the HRCTV. The centroid of the inferior-most HRCTV slice was defined as the origin, and the closest 2cc to the HRCTV was identified for each of the three OARs. The rationale for this orientation feature is that the inferior-most slice of the HRCTV serves as a surrogate for the top of the ovoids (the ovoids can contribute to dose deviations that might not be captured by the HRCTV-driven model) and the closest 2cc of the OARs likely correspond with the structure's highest dose values. The vector connecting HRCTV base to the OAR's closest 2cc is then decomposed into the radial distance ρ, azimuth φ, and height (superior-inferior distance) z in the redefined coordinate system. A negative z indicates the OAR's closest approach is inferior to the HRCTV and therefore near the ovoids.

Some of these features are depicted in Figure 2. From this candidate feature set $F_{candidate}$, we sought to identify the subset of predictor features $F_{predictors} \subseteq F_{candidate}$ that could explain error in

our initial model. We performed stepwise regression, iteratively adding and removing candidate features to $F_{predictors}$, performing least squares multiple linear regression of ΔD_{2cc} on $F_{predictors}$, and continuing until only the candidate features with statistically significant predictive power remained in the predictor set ($p<0.01$ threshold and Bonferroni corrections for multiple hypothesis testing). This stepwise regression analysis was performed on all training cases, and the end result was a linear model of ΔD_{2cc} as a function of a few significant variables. In order to determine whether these variables could improve predictive accuracy, the linear model was applied as a correction to predict D_{2cc} in the validation dataset.

Additional attempts at variance reduction via discrete categorical stratifications

In addition to the aforementioned set of candidate continuous features, we also considered discrete categories $C_{candidate}$ that could potentially explain and reduce variance via stratification. Our list of candidate categories fell into five classes, with the exact breakdown of our datasets by categories listed in Table 1:

1) The chronological fraction number of a case within a patient's treatment, which defines five distinct groups
2) The total number of prescribed fractions of the BT treatment
3) The tumor stage
4) The 'brachytype', a variable that attempts to capture case complexity. Although only T&O fractions were included in the training and validation datasets, some patients were treated with other applicators and/or needles for at least one other fraction. We suspect that the use of needles or other applicators for some fractions might indicate more challenging anatomy, and wanted to determine whether this affected model predictions. We defined three distinct groups: group 1 corresponding to cases from patients who underwent

entirely T&O treatments, group 2 to patients who had either a tandem-and-ring (T&R) or tandem-and-cylinder (T&C) implant at some point during treatment, and group 3 to patients who received supplemental needles for at least one fraction.

5) The attending radiation oncologist, which resulted in five different groupings for this dataset.

For each categorical variable considered, we tested for group-dependent differences in the combined training and validation cohort in the distribution of ΔD_{2cc} via an ANOVA test with a post-hoc Tukey's B analysis ($p<0.01$ significance threshold for group-specific variation with Bonferroni corrections for multiple hypothesis testing). All statistical data analysis was performed using MATLAB (R2018a, MathWorks, Inc., Natick, Massachusetts, USA).

Results

Actual bladder D_{2cc}values for the combined training and validation cohort displayed a [minimum:maximum] range of [1.65Gy:7.30Gy], with a mean ± standard deviation of 4.64±1.03Gy. The corresponding statistics for rectum D_{2cc} were [1.12 Gy:6.50 Gy] and 3.58±0.93 Gy, and for sigmoid D_{2cc}, they were [1.63Gy:6.58Gy] and 3.88±0.86Gy. Average DVHs for each OAR, for both actual clinical plans and predictions, are shown in Figure 3.

The model predicted D_{2cc} to bladder, rectum and sigmoid to within 0.46-0.61 Gy, as quantified by standard deviation (see Figure 4). Model accuracy did not statistically differ between the validation and training datasets for bladder ($p=0.80$), rectum ($p=0.53$) or sigmoid ($p=0.70$). When the validation dataset was separated into totally independent patients (group 1) and independent treatment fractions from patients used to train the model (group 2), performance

metrics were similar (mean ± standard deviation of ΔD_{2cc} = 0.06 ± 0.45 Gy, -0.02 ± 0.42 Gy and -0.07 ± 0.47 Gy for bladder, rectum and sigmoid for group 1; and -0.08 ± 0.75 Gy, -0.05 ± 0.43 Gy and -0.10 ± 0.50 Gy for group 2). Model accuracy did not significantly differ between validation groups ($p>0.25$). Mean and standard deviation over ΔD_{2cc} for dose normalized to prescription for training (validation) were bladder = -0.51 ± 9.43% (0.13 ± 9.38%), rectum = 0.36 ± 8.84% (-0.46 ± 7.05%), and sigmoid = -0.75 ± 8.05% (-1.14 ± 7.77%). Multiplying these numbers by a 6-7 Gy prescription (the most common prescriptions of our dataset), these numbers are similar to those obtained for absolute dose.

The results of the categorical stratification analyses on the combined training and validation cohorts for various group variables are reported in Table 2. The ANOVA and post-hoc analyses revealed that ΔD_{2cc} values significantly varied between radiation oncologists for bladder ($p<0.001$) and rectum ($p<0.001$) (see Figure 5). No significant differences were found between any other stratifications. To test whether inter-practitioner differences in OAR dose might be related to differences in target coverage, we ran an ANOVA and post-hoc analysis for D90, dose to 90% of the HRCTV, normalized to prescription. As indicated in Figure 4, while there are physician-dependent differences in coverage, these differences are not clearly related to the corresponding differences seen for ΔD_{2cc} metrics.

The stepwise linear regression highlighted up to one significant variable correcting each organ model. Bladder ΔD_{2cc} was correlated with $AP_{asymmetry}$, rectum ΔD_{2cc} was correlated with z_{rectum}, and sigmoid ΔD_{2cc} was correlated with none of the analyzed features (regression equations and adjusted R-squared shown in Supplementary Material, Table S1). The results of the stepwise multiple linear regression suggest that certain geometric features have nonzero predictive

correlation with ΔD_{2cc}, but ultimately these corrections made on average very modest ~0.02 Gy improvements in ΔD_{2cc} prediction accuracy as quantified by the standard deviation (σ ¿ and correlation coefficient R (Table 3).

Discussion

Machine learning is thus far relatively unexplored in the BT realm. One study applied machine learning to automate planning for prostate low dose rate BT (17). For HDR BT, several recent papers have investigated the use of advanced computational methods in multi-objective optimization criteria (18–20), and multiple pilot studies (21–23) have successfully automated various aspects of treatment planning. Damato *et al* (24) developed simple mathematical models to predict bladder and rectum D_{2cc} for interstitial GYN BT, using a dataset of 20 patients. However, to our knowledge this is the first study that applies machine learning to patient-specific dosimetric prediction in intracavitary GYN BT. In contrast to the interstitial GYN work (24), we had a much larger patient group and thus were able to validate our model on an independent dataset, and our models include only geometric inputs.

As shown in Figure 4, our knowledge-based DVH estimation system predicts $D_{2cc}$to OARs to within 0.46-0.61 Gy standard deviation. This amounts to a ±0.9-1.2 Gy 95% confidence interval for each BT fraction. Model performance did not significantly differ between the training and two validation sets, indicating that the model is not skewed towards patients included in the training set, and does not suffer from overfitting. Although the entire OAR DVH is not considered in current cervical brachytherapy practice and thus was not a focus of this work, the model did predict reasonable DVHs on average, as shown in Figure 3. In its current form, the

model uses only contour information from post-implant CT imaging and does not require applicator geometry, effectively providing an external reference for expected T&O dosimetry. The value of this is that any institution could input their HRCTV and OAR contours and receive a prediction for the dose they could expect for a given T&O implant. In addition, because of the non-reliance on applicator geometry, this approach could be extended to create decision support tools that could identify cases where T&O applicators alone could/could not meet dosimetric constraints. Since interstitial needle implantation is challenging, increases treatment time and can reduce patient comfort, it would be valuable to identify cases that do not need needles up front.

The analysis of relative dose resulted in model performance metrics that were similar to those in absolute dose, using an average per fraction prescription of ~6-7 Gy to convert between the two. For example, the standard deviations of ΔD_{2cc} ranged from 7.1 to 9.4% (0.46 to 0.61 Gy in absolute dose), and mean ΔD_{2cc}ranged from -1.14 to 0.36% (-0.07 to 0.02 Gy). In addition, model accuracy did not significantly differ between patients treated with different total numbers of fractions (and by extension, cases with different prescriptions). Based on these results, we feel it is valid to train and apply these models on cases with different prescriptions.

Notably, the predictive accuracy of ΔD_{2cc} to within 0.46-0.61 Gy standard deviation is comparable to the range of inter-practitioner differences in ΔD_{2cc} means, e.g. [-0.69 Gy, 0.1 Gy] for rectum. There are numerous unaccounted for inter-practitioner variations that could influence dose, including differences in applicator loading, updates in target and OAR constraints over time, variability in contouring and vaginal packing, differences in patient groups, and variations in internal optimization stopping criteria, which is further complicated by the disconnect between isodose tuning and DVH dose evaluation. The fact that our model already predicts estimation error to within inter-practitioner variability in spite of these complexities is promising, and

suggests that the model can be used for plan quality control. In fact, the model was able to identify a case (see * in Figure 6A) that featured a non-ideal implant due to difficulties with tandem insertion through a stenotic os into an anteverted uterus. In later fractions, where implants were improved (e.g. Figure 6B), the actual doses were in better agreement with the model-predicted doses (ΔD_{2cc} for bladder ranged from 1.07-2.05 Gy for fractions 2-4, while ΔD_{2cc} for fraction 1 was 3.52 Gy). This suggests that the model could help physicians decide whether a re-implant is warranted for challenging cases. It should be noted that although variability was observed between practitioners, nearly all clinical plans met our institutional dose constraints.

The HRCTV and OAR contours used in this study were primarily CT-based. It stands to question whether the current models would be directly applicable for MR-based contouring and planning. MR-based tumor volumes for cervical cancer have been shown to be smaller than those drawn on CT (25). Since the model relies only on the extracted contours, it should still provide accurate predictions for any given HRCTV.

There are limitations to our presented approach. The study was restricted to a single institution and, presently, only considers standard T&O cases. Like similar EBRT KBP methods (11,14), the target coverage is taken as a set variable, and therefore the model cannot be used to predict target coverage, or predict optimal target and OAR dose tradeoffs. The identification of any significant geometric predictors of ΔD_{2cc} hints that more sophisticated accounting of HRCTV and applicator geometry could yield more accurate predictions. Figure 6 shows examples of a few geometric features that are not well captured by the simple HRCTV boundary-distance model, such as asymmetry of the HRCTV with respect to the tandem and dose from the ovoids. However, it is evident that a linear correction for these features is

insufficient; although model error was reduced in the training dataset by 0-4 cGy after the correction, there was no improvement in predictive accuracy in the validation set, leading to concerns of over-fitting. Regardless, despite the model's reliance on a simplistic assumption of equivalent dose fall-off for points within an organ that are equidistant from the target (which is arguably more applicable to EBRT treatment plans), it seems to perform quite well. Future work will explore voxel-based dose prediction that accounts for relative positioning of targets, OARs and all applicator components.

Despite the current model's limitations, the ~0.5 Gy prediction accuracy demonstrates that the model could function as a multi-institutional quality control tool for T&O BT planning, since it can compute BT predictions from contours alone. Predictions are produced fully within the MIM environment and require only the structure set of an RT DICOM, so any institution with MIM could upload a structure set for a patient and receive dose predictions with the click of a button. As seen in the context of multi-institutional clinical trials of EBRT (9,26), objective measures of plan quality can highlight previously uncontrolled quality variability across multiple institutions. Patient-specific dose predictions can not only quantify unknown quality variations in BT practice, but also provide a means to reduce inter-practitioner variability. Future work will extend these models to other applicators and intracavitary/interstitial hybrid cases and examine whether predictions could guide further plan optimization and improve plan quality by re-planning cases with large discrepancies between predicted and actual dose. Finally, we will deploy this tool in the multi-institutional context and utilize dose predictions for fully-automated BT planning.

In summary, we have adapted knowledge-based dose prediction methods to predict OAR DVHs and, in particular, the critical OAR D_{2cc} quality metric for GYN brachytherapy. To our

knowledge, this is the first such application of knowledge-based methods to GYN brachytherapy and could form the basis for treatment plan quality control and automated brachytherapy planning.

Disclosure

Dr. Meyers, Moore and Mayadev report grants from Padres Pedal the Cause, during the conduct of the study. Dr. Moore acknowledges funding support from AHRQ (R01 HS025440-01), has a patent Developing Predictive Dose-Volume Relationships for a Radiotherapy Treatment licensed to Varian Medical Systems, and a patent Knowledge-based prediction of three-dimensional dose distributions pending. In addition, Drs. Moore, Brown and Scanderbeg acknowledge research funding, travel support, and honoraria from Varian Medical Systems, outside the submitted work. Dr. Mayadev reports personal fees from AstraZeneca, grants from NRG Oncology, grants from GOG Foundation, personal fees from Varian Medical Systems, outside the submitted work. Dr. Simon reports personal fees from Courage Health, Inc., outside the submitted work.

Acknowledgments:

We thank James Murphy for helpful discussions.

Funding: This work was supported by a Translational Cancer Research Award from Padres Pedal the Cause.

References

1. Hanks GE, Herring DF, Kramer S. Patterns of care outcome studies. Results of the national practice in cancer of the cervix. Cancer. 1983 Mar 1;51(5):959–67.

2. Eifel PJ, Moughan J, Erickson B, Iarocci T, Grant D, Owen J. Patterns of radiotherapy practice for patients with carcinoma of the uterine cervix: a patterns of care study. Int J Radiat Oncol Biol Phys. 2004 Nov 15;60(4):1144–53.

3. Mayadev J, Qi L, Lentz S, Benedict S, Courquin J, Dieterich S, et al. Implant time and process efficiency for CT-guided high-dose-rate brachytherapy for cervical cancer. Brachytherapy. 2014 Jun;13(3):233–9.

4. Ma TM, Harkenrider MM, Yashar CM, Viswanathan AN, Mayadev JS. Understanding the underutilization of cervical brachytherapy for locally advanced cervical cancer. Brachytherapy. 2019 Jun;18(3):361–9.

5. Mayadev J, Klapheke A, Yashar C, Hsu I-C, Kamrava M, Mundt AJ, et al. Underutilization of brachytherapy and disparities in survival for patients with cervical cancer in California. Gynecol Oncol. 2018;150(1):73–8.

6. Albuquerque K, Tumati V, Lea J, Ahn C, Richardson D, Miller D, et al. A Phase II Trial of Stereotactic Ablative Radiation Therapy as a Boost for Locally Advanced Cervical Cancer. Int J Radiat Oncol Biol Phys. 2020 01;106(3):464–71.

7. Pötter R, Tanderup K, Kirisits C, de Leeuw A, Kirchheiner K, Nout R, et al. The EMBRACE II study: The outcome and prospect of two decades of evolution within the GEC-ESTRO GYN working group and the EMBRACE studies. Clin Transl Radiat Oncol. 2018 Feb;9:48–60.

8. Nelms BE, Chan MF, Jarry G, Lemire M, Lowden J, Hampton C, et al. Evaluating IMRT and VMAT dose accuracy: practical examples of failure to detect systematic errors when applying a commonly used metric and action levels. Med Phys. 2013 Nov;40(11):111722.

9. Moore KL, Schmidt R, Moiseenko V, Olsen LA, Tan J, Xiao Y, et al. Quantifying Unnecessary Normal Tissue Complication Risks due to Suboptimal Planning: A Secondary Study of RTOG 0126. Int J Radiat Oncol Biol Phys. 2015 Jun 1;92(2):228–35.

10. Moore KL, Brame RS, Low DA, Mutic S. Experience-based quality control of clinical intensity-modulated radiotherapy planning. Int J Radiat Oncol Biol Phys. 2011 Oct 1;81(2):545–51.

11. Appenzoller LM, Michalski JM, Thorstad WL, Mutic S, Moore KL. Predicting dose-volume histograms for organs-at-risk in IMRT planning. Med Phys. 2012 Dec;39(12):7446–61.

12. Shiraishi S, Tan J, Olsen LA, Moore KL. Knowledge-based prediction of plan quality metrics in intracranial stereotactic radiosurgery. Med Phys. 2015 Feb;42(2):908.

13. Shiraishi S, Moore KL. Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy. Med Phys. 2016 Jan;43(1):378.

14. Ge Y, Wu QJ. Knowledge-based planning for intensity-modulated radiation therapy: A review of data-driven approaches. Med Phys. 2019 Jun;46(6):2760–75.

15. Viswanathan AN, Beriwal S, De Los Santos J, Demanes DJ, Gaffney D, Hansen J, et al. The American Brachytherapy Society Treatment Recommendations for Locally Advanced Carcinoma of the Cervix Part II: High Dose-Rate Brachytherapy. Brachytherapy. 2012;11(1):47–52.

16. Cornell M, Kaderka R, Hild SJ, Ray XJ, Murphy JD, Atwood TF, et al. Noninferiority Study of Automated Knowledge-Based Planning Versus Human-Driven Optimization Across Multiple Disease Sites. International Journal of Radiation Oncology*Biology*Physics [Internet]. 2019 Oct 31 [cited 2020 Jan 6]; Available from: http://www.sciencedirect.com/science/article/pii/S0360301619339586

17. Nicolae A, Morton G, Chung H, Loblaw A, Jain S, Mitchell D, et al. Evaluation of a Machine-Learning Algorithm for Treatment Planning in Prostate Low-Dose-Rate Brachytherapy. Int J Radiat Oncol Biol Phys. 2017 15;97(4):822–9.

18. Bélanger C, Cui S, Ma Y, Després P, Adam M Cunha J, Beaulieu L. A GPU-based multi-criteria optimization algorithm for HDR brachytherapy. Phys Med Biol. 2019 May 8;64(10):105005.

19. Guthier CV, Aschenbrenner KP, Müller R, Polster L, Cormack RA, Hesser JW. Real-time inverse high-dose-rate brachytherapy planning with catheter optimization by compressed sensing-inspired optimization strategies. Phys Med Biol. 2016 21;61(16):5956–72.

20. Luong NH, Alderliesten T, Pieters BR, Bel A, Niatsetski Y, Bosman PAN. Fast and insightful bi-objective optimization for prostate cancer treatment planning with high-dose-rate brachytherapy. Applied Soft Computing. 2019 Nov 1;84:105681.

21. Zhou Y, Klages P, Tan J, Chi Y, Stojadinovic S, Yang M, et al. Automated high-dose rate brachytherapy treatment planning for a single-channel vaginal cylinder applicator. Phys Med Biol. 2017 07;62(11):4361–74.

22. Deist TM, Gorissen BL. High-dose-rate prostate brachytherapy inverse planning on dose-volume criteria by simulated annealing. Phys Med Biol. 2016 Feb 7;61(3):1155–70.

23. Shen C, Gonzalez Y, Klages P, Qin N, Jung H, Chen L, et al. Intelligent inverse treatment planning via deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer. Phys Med Biol. 2019 May 29;64(11):115013.

24. Damato AL, Viswanathan AN, Cormack RA. Validation of mathematical models for the prediction of organs-at-risk dosimetric metrics in high-dose-rate gynecologic interstitial brachytherapy. Med Phys. 2013 Oct;40(10):101711.

25. Viswanathan AN, Erickson B, Gaffney DK, Beriwal S, Bhatia SK, Lee Burnett O, et al. Comparison and consensus guidelines for delineation of clinical target volume for CT- and MR-based brachytherapy in locally advanced cervical cancer. Int J Radiat Oncol Biol Phys. 2014 Oct 1;90(2):320–8.

26. Li N, Carmona R, Sirak I, Kasaova L, Followill D, Michalski J, et al. Highly Efficient Training, Refinement, and Validation of a Knowledge-based Planning Quality-Control System for Radiation Therapy Clinical Trials. Int J Radiat Oncol Biol Phys. 2017 01;97(1):164–72.

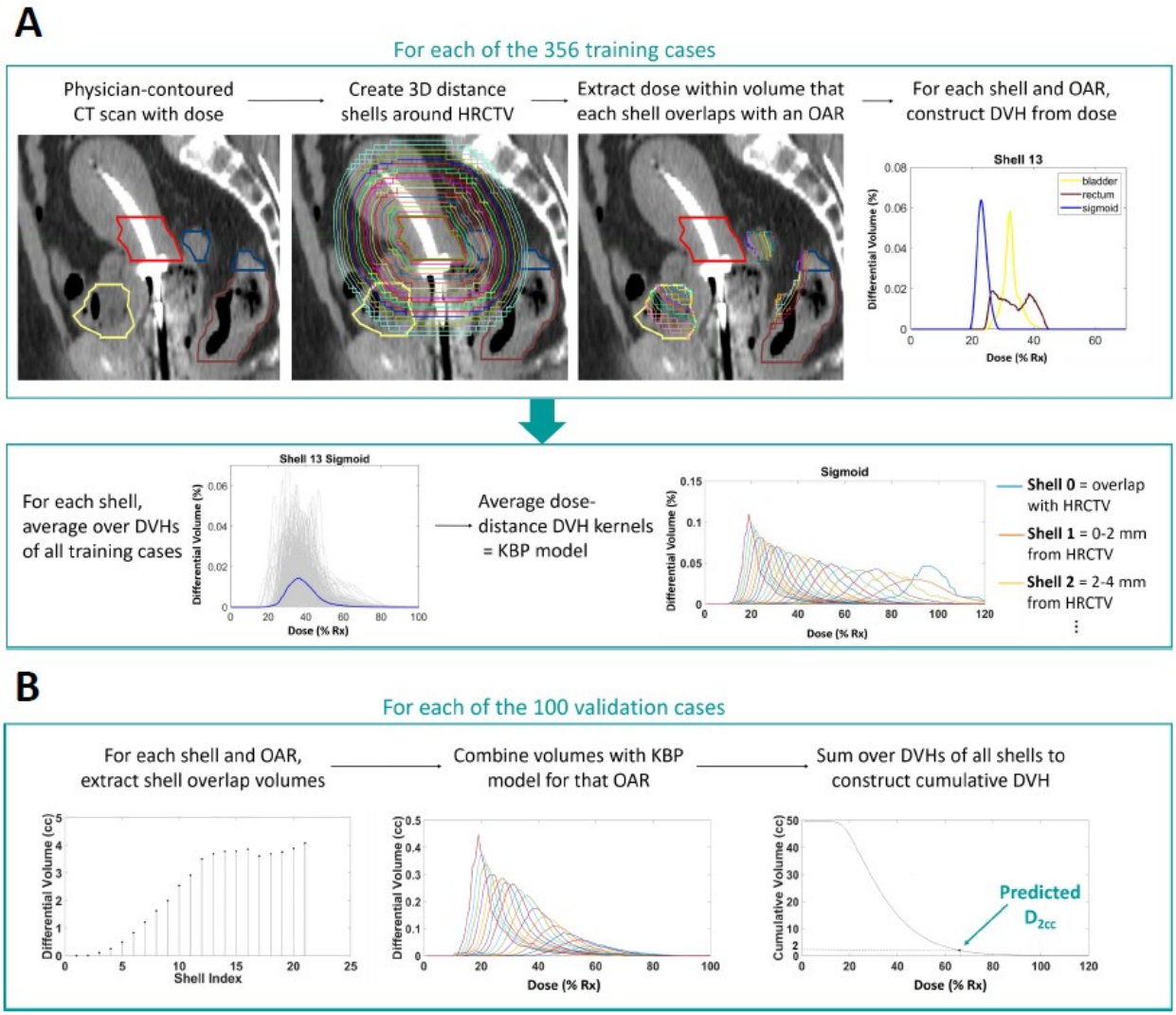


Figure 1. Methodology and workflow applied to A) train the model, and B) validate on a separate dataset.

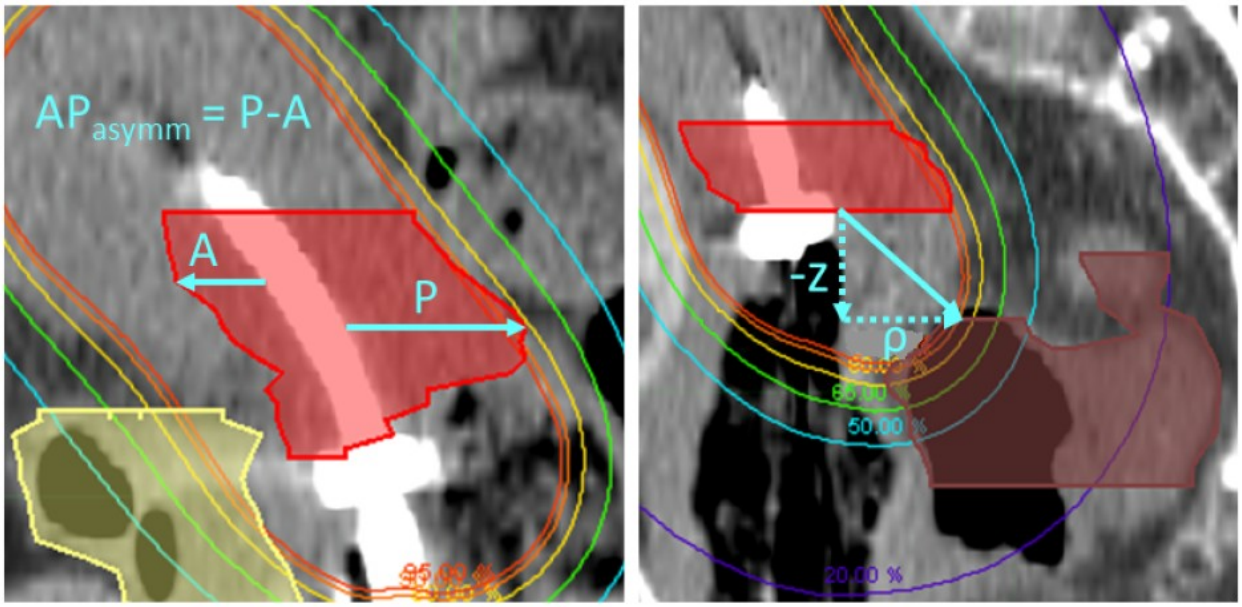


Figure 2. Illustration of some of the geometric features and how they were defined, including the anterior-posterior asymmetry metric, $AP_{asymmetry}$ (left), which quantifies how well the tandem was centered in the HRCTV, superior-inferior distance, z, and radial distance ρ (right), which identify the relative positioning of the closest 2cc of an OAR to the base of the HRCTV.

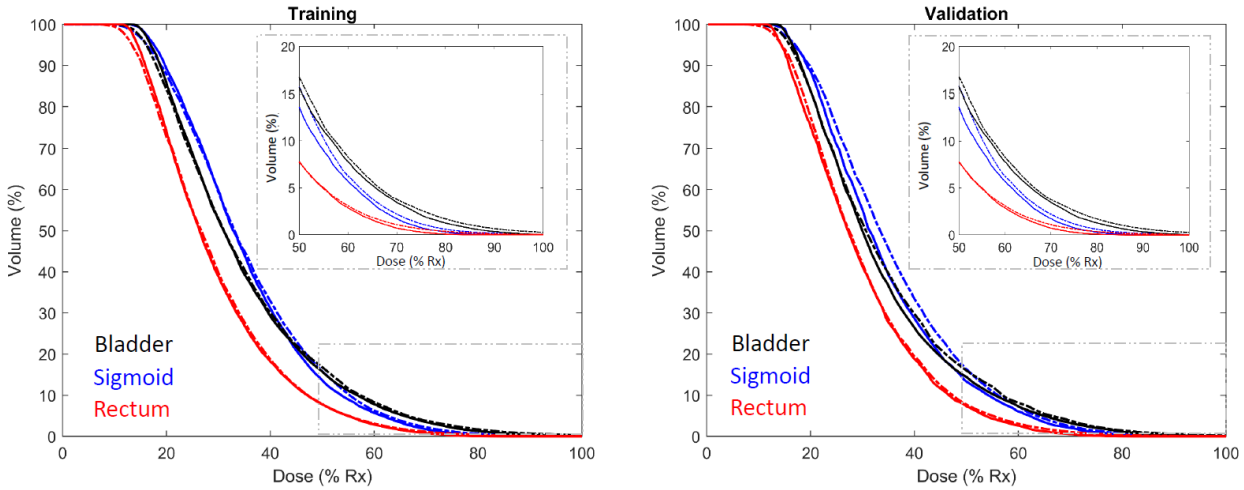


Figure 3. Actual (solid line) and predicted (dotted line) OAR DVHs, averaged over all cases in training (A) and validation (B) datasets.

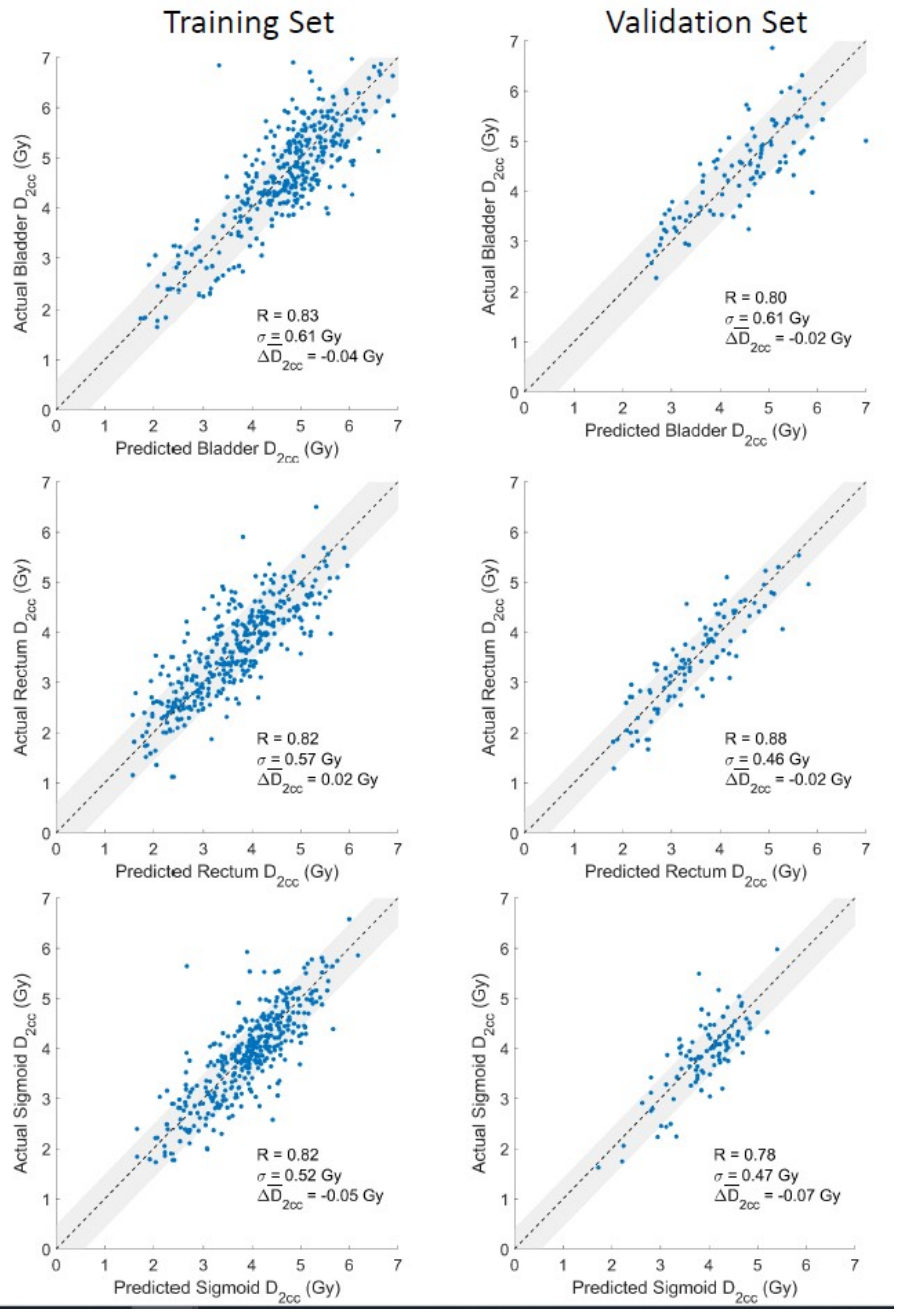
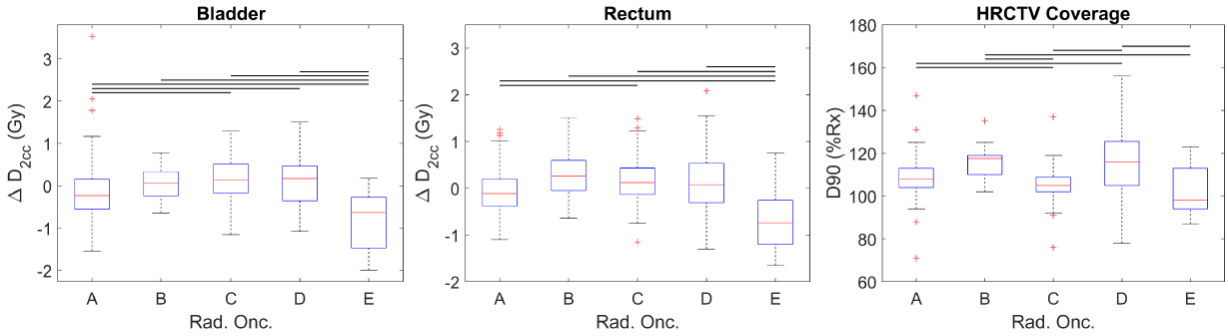


Figure 4. Actual vs. predicted D_{2cc} for each organ for training and validation datasets, along with Pearson correlation coefficients (R), standard deviation (indicated by σ as well as grey colorwash) and mean of ΔD_{2cc}'s (equal to zero for the training set due to bias subtraction). Black lines indicate hypothetical perfect model predictions.



443

444 **Figure 5.** Box-and-whisker plots showing physician-dependent variation in ΔD_{2cc} for bladder and

445 rectum, along with corresponding variation in the HRCTV coverage metric D90. Black line

446 segments connect pairs of physicians whose average ΔD_{2cc} or D90 values significantly differed at

447 the confidence level $p<0.01$, after accounting for multiple-comparison corrections.

448

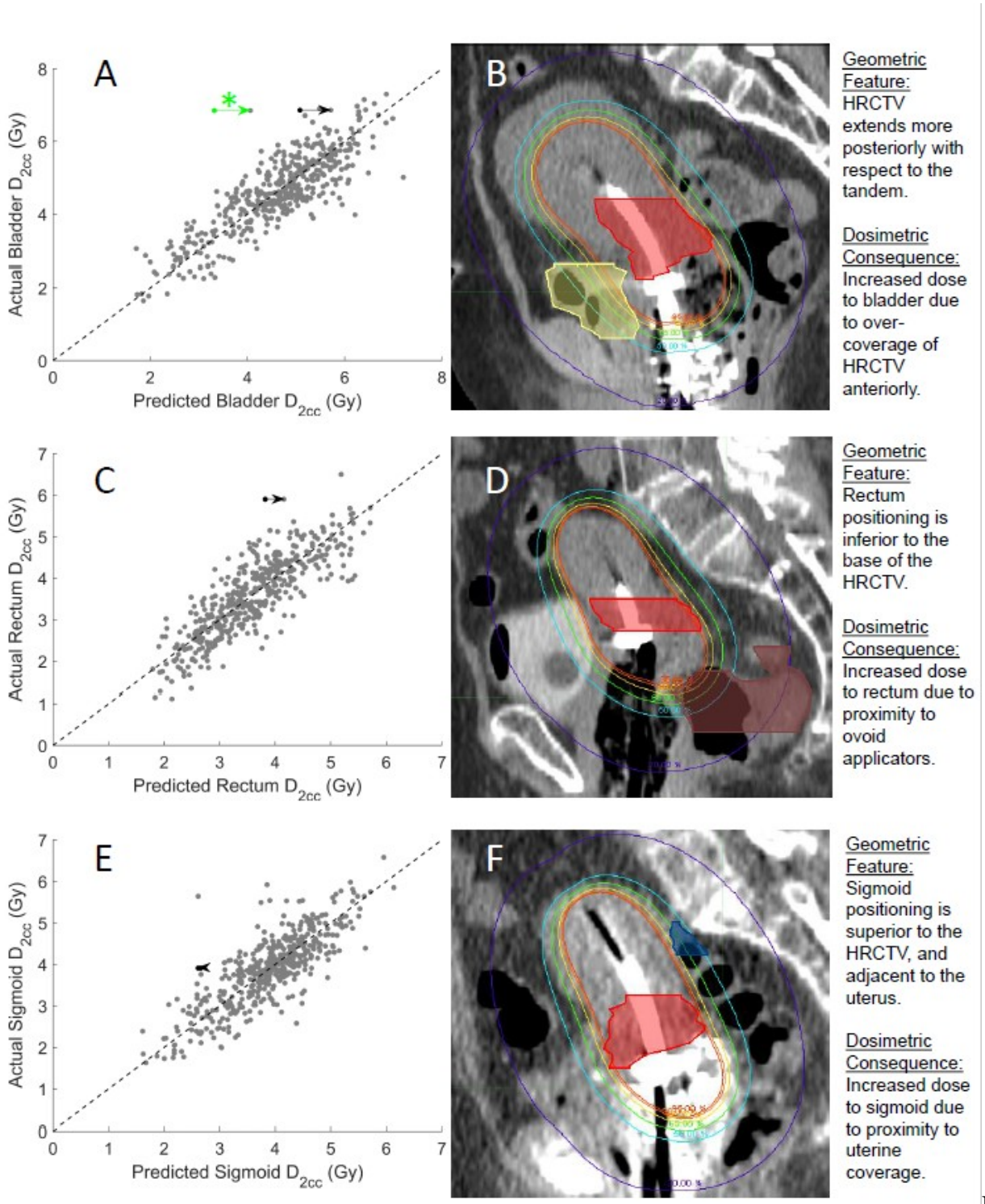


Figure 6. (A, C, E) show pooled plots of both the training and validation dataset predictions after geometric correction, compared to the actual values. (B, D, E) display example cases that

452 featured a large deviation between actual and predicted D_{2cc}, along with explanations of the
453 underlying geometric features that likely led to these discrepancies. The black arrow in each plot
454 shows how each of these cases changed with the geometric correction. The most extreme outlier
455 in (A) (indicated by the green arrow and *) corresponded to a first-fraction case that featured a
456 non-ideal implant due to difficulties with tandem insertion through challenging anatomy. The
457 implants improved for later fractions (e.g. (B), which was the second fraction of the same
458 patient), and as a result the difference between actual and predicted dose was much smaller.

459

Table 1. Breakdown of training and validation data by various categorical features, including the stage, attending radiation oncologist, total number of fractions and fraction number of the case. Some patients received T&O for all fractions of their treatment ("All T&O"), while other patients ("Component T&O") had at least one fraction that was treated with a different applicator (tandem and cylinder (T&C) or tandem and ring (T&R)), or needles. Although only T&O fractions were included in training and validation datasets, the "Brachytype" provides a breakdown of the number of cases that corresponded to patients that fell into each of these categories.

Number of patients	Training (P = 114)	Validation (P= 52)	Total (P = 126)
Stage			
T1	35	11	40
T2	51	26	57
T3	25	15	26
T4	3	0	3
Prescribed total number of fractions			
2	1	0	1
3	5	4	6
4	54	14	54
5	54	34	64
Number of cases	**Training (N = 356)**	**Validation (N = 100)**	**Total (N = 456)**
Physician			
A	154 (43 %)	44 (44 %)	198 (44 %)
B	16 (5 %)	8 (8 %)	24 (5 %)
C	101 (28 %)	18 (18 %)	119 (26 %)
D	69 (19 %)	27 (27 %)	96 (21 %)
E	16 (5 %)	3 (3 %)	19 (4 %)
Brachytype			
All T&O	304 (85 %)	96 (96 %)	400 (88 %)
Component T&O	52 (15 %)	4 (4%)	56 (12 %)
- T&C	- 13 (4 %)	- 2 (2%)	- 15 (3 %)
- T&R	- 16 (6 %)	- 1 (1%)	- 17 (5 %)
- Needles	- 23 (5 %)	- 1 (1%)	- 24 (4 %)

Fraction Number			
1	87 (24 %)	18 (18%)	105 (23 %)
2	84 (24 %)	19 (19%)	103 (23 %)
3	73 (21 %)	28 (28%)	101 (22 %)
4	72 (20 %)	19 (19%)	91 (20 %)
5	40 (11 %)	16 (16%)	56 (12 %)

Table 2. The results of categorical stratification analysis indicate that, after Bonferroni corrections, only stratification by radiation oncologist for bladder and rectum yields significant differences in ΔD_{2cc} (as indicated by bold text).

	ΔD_{2cc} p-Value		
	Bladder	Rectum	Sigmoid
Physician	**$1.6*10^{-12}$**	**$6.6*10^{-11}$**	$1.9*10^{-3}$
Brachytype	0.30	0.08	0.05
Tumor Stage	0.98	0.46	0.34
Fraction Number	0.60	0.50	0.33
Total Number of Fractions	0.77	0.80	0.71

Table 3. A comparison of model performance metrics on both the training and validation sets, both before and after including geometric corrections.

OAR	Model Performance	Training Set (N=356)		Validation Set (N=100)	
		Pre-correction	Post-correction	Pre-correction	Post-correction
Bladder	$<D_{2cc}>$	-0.04 Gy	0.00 Gy	-0.02 Gy	0.02 Gy
	R	0.83	0.85	0.80	0.80
	σ	0.61 Gy	0.57 Gy	0.61 Gy	0.61 Gy
Rectum	$<D_{2cc}>$	0.02 Gy	0.00 Gy	-0.01 Gy	-0.02 Gy
	R	0.82	0.83	0.89	0.86
	σ	0.57 Gy	0.53 Gy	0.43 Gy	0.49 Gy
Sigmoid	$< D_{2cc}>$	-0.05 Gy	0.00 Gy	-0.07 Gy	-0.02 Gy
	R	0.82	0.82	0.79	0.78
	σ	0.52 Gy	0.52 Gy	0.47 Gy	0.47 Gy

Table S1. Displayed are the best-fit linear regression models of ΔD_{2cc} to selected geometric features for each of the three OARs. Mean estimates of the dominant coefficients are listed, along with the corresponding standard errors (S.E.). The adjusted R-squared, which measures the goodness-of-fit adjusted for the number of fitting parameters, is also listed.

OAR	Regression Equation (Coefficients listed as Mean ±S.E.)	Adj. R-Squared
Bladder	$\Delta D_{2cc} = (-0.11 \pm 0.03) + (0.05 \pm 0.01)\ AP_{asymmetry}$	0.13
Rectum	$\Delta D_{2cc} = (-0.00 \pm 0.03) - (-0.19 \pm 0.02)\ z_{rectum}$	0.14
Sigmoid	$\Delta D_{2cc} = (-0.05 \pm 0.03)$	0.00