Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Daniel E. Spratt, Siyi Tang, Yilun Sun, et al. Artificial Intelligence Predictive Model for Hormone Therapy Use in Prostate Cancer. NEJM Evid.
DOI: 10.1056/EVIDoa2300023.

Supplementary Appendix

Supplementary Appendix for: Artificial Intelligence Predictive Model for Hormone Therapy Use in Prostate Cancer

Table Of Contents

Section 1: NRG Prostate Cancer AI Consortium Members

The following authors were critical to the accrual of the trials used in this study:

Michael Kucharczyk, Luis Souhami, Leslie Ballas, Christopher A. Peters, Sandy Liu, Alexander G. Balogh, Pamela D. Randolph-Jackson, David L. Schwartz, Michael R. Girvigian, Naoyuki G. Saito, Adam Raben, Rachel A. Rabinovitch & Khalil Katato

Section 2: Clinical Risk Group Defined by National Comprehensive Cancer Network (NCCN) Guidelines Prostate Cancer V.1.2022

Patients were considered to be in the high- risk group if they had any of the following: a prostate-specific antigen (PSA) > 20ng/mL, Gleason score of 8 to 10, clinical stage T3 or higher; in the intermediate-risk group if they had Gleason score of 7 or a Gleason score of 6 or less with a PSA 10-20 ng/mL or a clinical stage T2b and not high-risk; and in the low-risk group if they had a Gleason score of 6 or less, a PSA < 10 ng/mL, and a clinical T-stage of T2a or lower.

Section 3: Test Availability

The test can be accessed and run through a CLIA-certified lab that is available online (https://artera.ai/our-test/). This test has not been cleared nor approved by the FDA, and is offered as a single-site Laboratory-Developed Test (LDT) in a dedicated CLIA-certified laboratory using the approach for the AI assay described in our manuscript and that is available as noted. Samples for the test are shipped to the Artera laboratory, where they are digitized, and the AI algorithm is run there. Then, a report is returned to the ordering clinician. This approach of offering testing as a single-site LDT through a CLIA-licensed laboratory is used by other commercial risk-stratification tests such as Decipher Prostate® (https://decipherbio.com/) and Prolaris® (https://myriad.com/genetic-tests/prolaris-prostate-tumor-test/).

Section 4: Methods for Multimodal Deep Learning Model Development

Clinical Data Preprocessing

Categorical clinical variables (T-stage, Gleason score and primary/secondary Gleason pattern) and binary treatment type (0 for radiotherapy alone, 1 for radiotherapy with short-term androgen-deprivation therapy [ADT]) were fed through neural network embedding layers to generate continuous vector embeddings. Groupings are as follows: Gleason total (≤6, 7, 8, and ≥9), both primary and secondary Gleason patterns (≤3, 4, and 5), T-stage (Tx, T0, T1a, T1b, T1c or T1, T2a, T2b, T2c or T2, T3a, T3b, T3c or T3, T4a, T4b, T4). Continuous clinical variables (age, baseline PSA) were standardized based on the mean and standard deviation of the training data. In the development set, missing continuous clinical variables (age and PSA) were imputed using sklearn SimpleImputer[1] with the default mean strategy; missing categorical clinical variables (T stage and Gleason information) were treated as separate "N/A" category by the encoder. In the validation set, NRG/RTOG 9408, missing clinical variables (age, Gleason total, T stage, and PSA) were considered numeric and imputed using sklearn KNNImputer with the default parameters (average of 5 nearest neighbors)[1,2]. Missing Gleason primary or secondary patterns were imputed based on the most frequent patterns of their non-missing Gleason combination.

Image Feature Extraction Model Development

For each patch from a patient's histopathology images, a 128-dimensional feature vector was extracted using the self-supervised pre-trained Resnet-50 image feature extraction model and was standardized based on the mean and standard deviation of the training data. All the patch-level feature vectors from the same patient were stacked to form an image feature tensor, which was fed to the downstream predictive model.

Inverse Probability Treatment Weighting

As the development set comprised two phase III randomized trials (NRG/RTOG 9910 and 0126), inverse probability of treatment weighting (IPTW) was used to ensure that patients in two treatment types had comparable clinical baseline characteristics[3]. Propensity score was calculated using a logistic regression model with elastic net penalty, where treatment types were regressed against patients' age, baseline PSA, Gleason score, Gleason primary/secondary patterns, and T-stage variables. To mitigate the high variability introduced by large weights, IPTW weights were trimmed based on the 1st and 99th percentiles[4,5].

Downstream Predictive Model Development

The downstream predictive model took the image feature tensor, preprocessed clinical data, and treatment type (rx) as input for each patient. An attention multiple instance learning network was employed to learn a weight for each patch from the patient[6]. A single 128-dimensional image vector was generated from the image feature tensor for each patient by taking the weighted sum of the image vectors of all patches from the same patient, where the weights were learned by the attention mechanism. A concatenation of this single 128-dimensional image vector, preprocessed clinical data, and treatment type was further processed through the joint fusion pipeline to effectively learn predictive feature encodings of differential treatment benefit from the addition of short-term ADT to radiotherapy.

The multimodal predictive model was trained in a multitask manner. The first task was to predict the relative risk of distant metastasis using the factual rx ("Task 1" in Supplementary Figure 1A). The image, clinical, and factual rx vectors were concatenated and fed through a few layers of fully connected neural networks to produce a continuous score for each patient that estimates the relative risk of distant metastasis (referred to as "factual model prediction score" hereafter). The negative log-partial likelihood was used as the training objective for the first task and the factual model prediction scores were the estimated log relative hazards[7].

The negative log-partial likelihood loss was parameterized by the model weights θ and formulated as follows:

$$loss(\theta) := -\frac{1}{N_{E=1}} \sum_{i:E_i=1} \left(f_\theta(x_i) - log \sum_{j \in \Re(T_i)} e^{f_\theta(x_j)} \right),$$

where the values T_i, E_i, and x_i are the respective event time or time of last follow-up, an indicator variable for whether the event is observed, and the model input for the i^{th} observation. The function f_θ represents the factual branch of the multimodal model, and $f_\theta(x)$ is the estimated log relative hazard given an input x. The value $N_{E=1}$ represents the number of patients with an observable event. The set of patients with an observable event is represented as $E_i = 1$. The risk set $\Re(t) = \{i : T_i \geq t\}$ is the set of patients still at risk of failure at time t. We used Breslow's approximation for handling tied event times[8].

Based on the estimated relative risk on the first task, the second task was to predict the delta score, defined as the difference in factual model prediction score and counterfactual model prediction score ("Task 2" in Supplementary Figure 1A). To this end, a counterfactual rx variable was created by toggling the patient's factual rx (radiotherapy for patients who received radiotherapy with short-term ADT, and vice versa). The counterfactual rx variable was fed through the same rx embedding layer and concatenated with the image and clinical vectors. Then, the concatenated vectors were fed through the same fully connected neural network layers yielding another continuous score (referred to as "counterfactual model prediction score" hereafter). For patients who received radiotherapy alone, delta would be the factual model prediction score minus the counterfactual prediction score; whereas for patients with radiotherapy and short-term ADT, delta would be the counterfactual prediction score minus the factual model prediction score. The delta indicates the magnitude of therapeutic benefit for each patient, where a larger delta suggests a larger benefit from additional short-term ADT, and vice versa.

For this predictive task, the delta loss was designed and used as the training objective. Essentially, the delta loss was the deviation between the predicted delta scores and the expected delta scores. The expected delta scores were dependent on patients falling into one of four subgroups based on their treatment types and distant metastasis outcomes as follows: (a) Subgroup A consisted of metastasis-free patients who received radiotherapy alone; (b) Subgroup B consisted of patients who received radiotherapy alone and had distant metastasis; (c) Subgroup C comprised of metastasis-free patients who received radiotherapy with short-term ADT; (d) Subgroup D consisted of patients who received radiotherapy with short-term ADT and had distant metastasis. For Subgroup A, the delta should be close to 0 as the patients had no distant metastasis when receiving radiotherapy alone treatment and additional short-term ADT would not affect their risk of distant metastasis; for Subgroup B, the delta should be greater than or equal to 0 since the patients may benefit from additional short-term ADT treatment; for Subgroup C, the delta should be greater than 0 since the patients were free of distant metastasis when receiving additional short-term ADT; finally, Subgroup D's delta should be close to 0 as the patients had distant metastasis even if they received additional short-term ADT treatment. During training, the model was penalized when the delta scores did not fall in the expected range described above. The training objective for the predictive task was defined using the softplus function[9].

During training, we approximated the weighted sum of both losses from the prognostic task and the predictive task, and each data point was weighted by its IPTW weight. Once the model was trained, a cutoff was selected at the 67th percentile of the delta scores in the development set such that all patients in the validation set with a delta score greater than the cutoff were considered to be predictive model positive, with predicted benefit of additional short-term ADT, and those with a delta score less than the cutoff were considered to be predictive model negative. The final model was chosen based on the lowest ratio of IPTW-weighted hazard ratios of predictive model positive and negative subgroups on the tuning set.

Section 5: Supplementary Figures and Tables

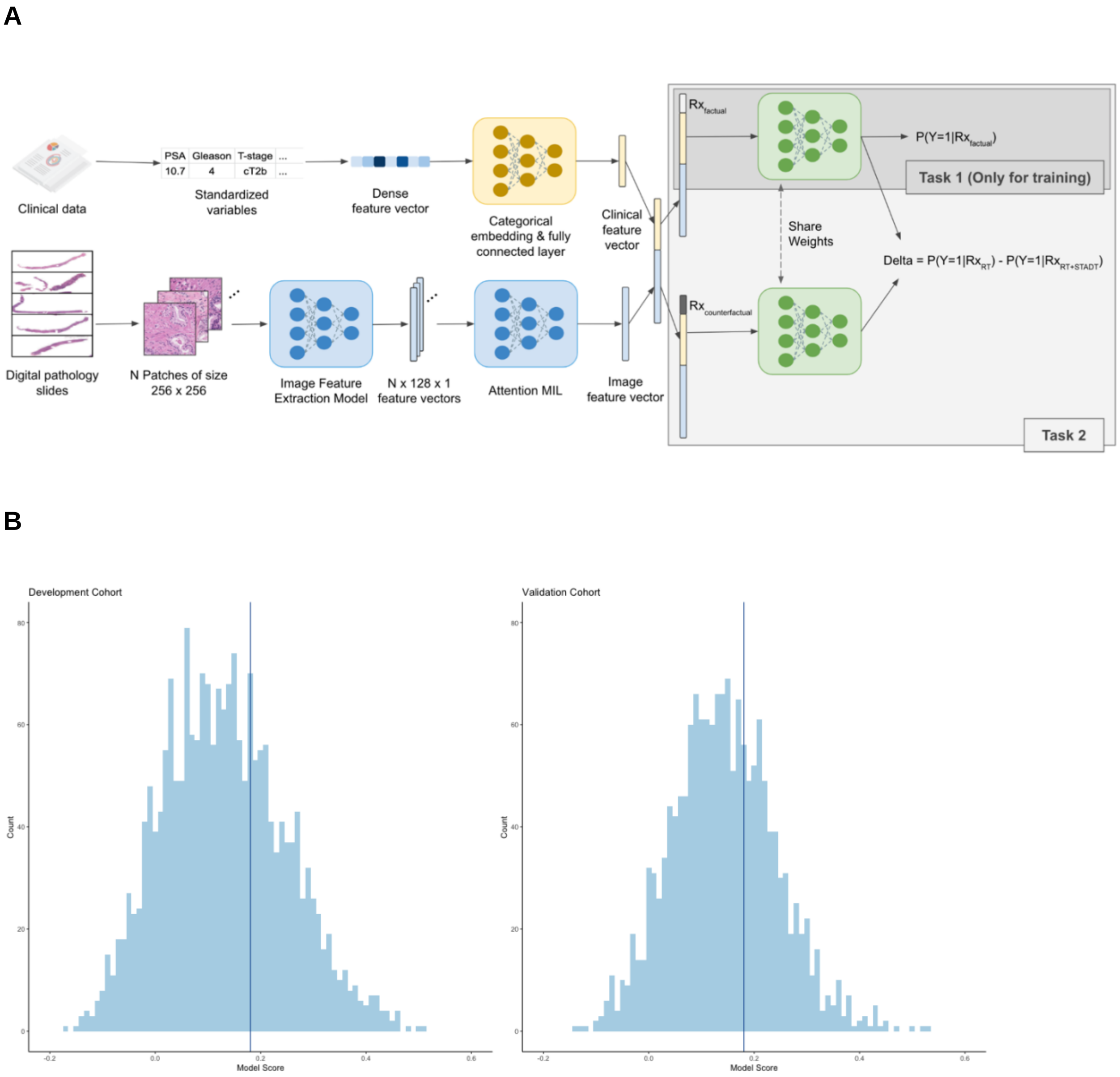


Figure S1. Multimodal deep learning architecture and distribution of delta for development and validation set.

(A) The multimodal architecture accepts both clinical data and histopathology image data and outputs a delta score that captures the magnitude of therapeutic benefit. (B) The 67th percentile of the delta scores in the development set was selected as the cutoff threshold such that predictive positive patients had a delta greater than the cutoff, and predictive negative patients had a delta less than the cutoff.

PSA = prostate-specific antigen; Rx = treatment type; P = probability; Y = outcome; RT = radiotherapy; ST-ADT = short-term androgen-deprivation therapy; N = number; MIL = multiple instance learning.

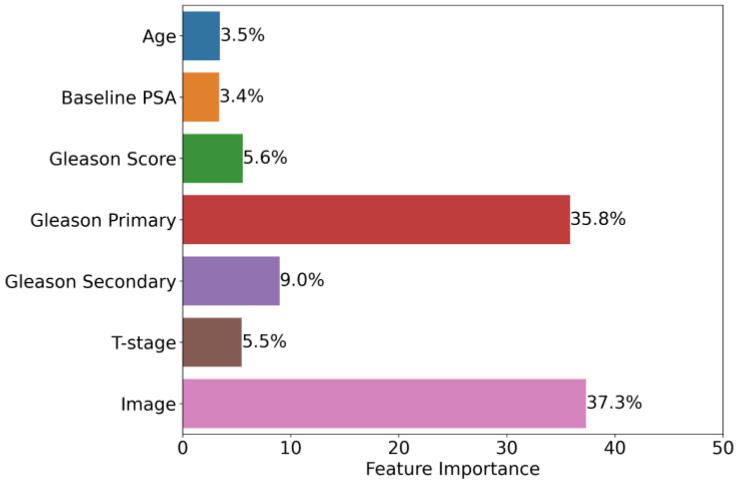


Figure S2. Model feature weights and their associated feature importance.

The greater the feature importance, the more contribution the individual clinical (e.g. baseline PSA) or histopathology-derived feature has to the downstream predictive model prediction. The feature importance is calculated based on the absolute Shapley value for each variable, averaged over the patients in the validation set (NRG/RTOG 9408) to obtain a global measure of feature importance, and is normalized across the features for visualization purposes[10]. Histopathology-derived features–including Gleason score and imaging features–contributed the most to the artificial intelligence-derived predictive model. Of note, the downstream predictive model has a capacity to jointly learn complex interactions between image and clinical features, and the Shapley value provides a combined contribution of main and interaction effects. Note that some percentages may not add up to a hundred percent due to rounding.

PSA = prostate-specific antigen.

Endpoint	NCCN	Predictive Model	RT+ST-ADT	RT	15-yr Absolute Benefit of	15-yr RMST Benefit of	s/Hazard Ratio
	Risk	Group	Incidence/N	Incidence/N	ADT (%, CI 95%)	ADT (Years, CI 95%)	(95% CI)
DM	Low-Intermediate	Positive	11/237	31/232	10.4 (5.2, 15.6)	0.8 (0.3, 1.2)	0.34 (0.17, 0.67)
		Negative	31/461	30/476	-0.5 (-3.7, 2.8)	0.0 (-0.2, 0.3)	1.05 (0.63, 1.73)
PCSM	Low-Intermediate	Positive	8/237	27/232	10.1 (5.2, 14.9)	0.6 (0.3, 1.0)	0.28 (0.13, 0.62)
		Negative	22/461	26/476	0.1 (-2.8, 3.0)	0.0 (-0.2, 0.3)	0.86 (0.49, 1.51)
MFS	Low-Intermediate	Positive	162/237	151/232	-1.2 (-10.7, 8.3)	0.2 (-0.6, 1.1)	1.12 (0.90, 1.40)
		Negative	310/461	319/476	-0.7 (-7.4, 6.1)	0.8 (0.2, 1.4)	0.91 (0.78, 1.06)
OS	Low-Intermediate	Positive	159/237	147/232	-1.5 (-11.1, 8.0)	0.1 (-0.7, 0.9)	1.13 (0.90, 1.42)
		Negative	301/461	315/476	-0.2 (-7.0, 6.6)	0.8 (0.2, 1.4)	0.90 (0.76, 1.05)

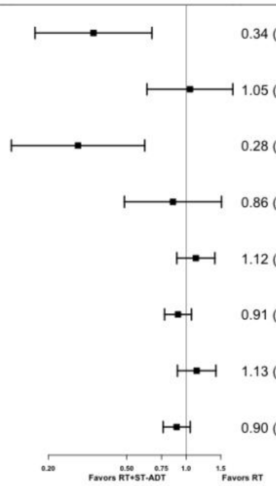


Figure S3. Forest plots for all endpoints in positive and negative predictive model groups of NRG/RTOG 9408 (validation set) for the subgroup of NCCN low-intermediate-risk patients.

NCCN = National Comprehensive Cancer Network; RT = radiotherapy; ST-ADT = short-term androgen-deprivation therapy; yr = year; RMST = restricted mean survival time; s/HR = subdistribution hazard ratio or hazard ratio; CI = confidence interval; N = number of patients; DM = distant metastasis; PCSM = prostate cancer-specific mortality; MFS = distant metastasis-free survival; OS = overall survival.

Endpoint	NCCN Risk	Predictive Model Group	RT+ST-ADT Incidence/N	RT Incidence/N	15-yr Absolute Benefit of ADT (%, CI 95%)	15-yr RMST Benefit of ADT (Years, CI 95%)	HR (95% CI)
MFS	All	Positive	186/273	181/270	0.6 (-8.2, 9.4)	0.4 (-0.4, 1.2)	1.04 (0.85, 1.27)
		Negative	350/515	367/536	0.0 (-6.3, 6.3)	0.8 (0.2, 1.3)	0.91 (0.79, 1.06)
OS	All	Positive	182/273	176/270	0.2 (-8.7, 9.1)	0.3 (-0.5, 1.1)	1.05 (0.85, 1.29)
		Negative	340/515	363/536	0.5 (-5.8, 6.9)	0.8 (0.2, 1.3)	0.90 (0.77, 1.04)

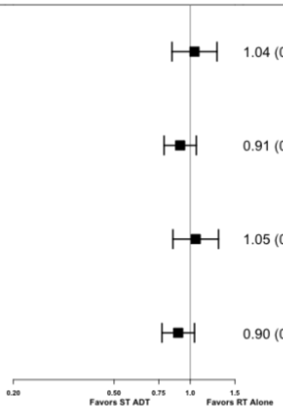


Figure S4. Forest plots for exploratory endpoints in positive and negative predictive model groups of NRG/RTOG 9408 (validation set) for all patients.

NCCN = National Comprehensive Cancer Network; RT = radiation therapy; ST-ADT = short-term androgen-deprivation therapy; yr = year; RMST = restricted mean survival time; HR = hazard ratio; CI = confidence interval; N = number of patients; MFS = distant metastasis-free survival; OS = overall survival.

Table S1. Trials included in meta-analyses.

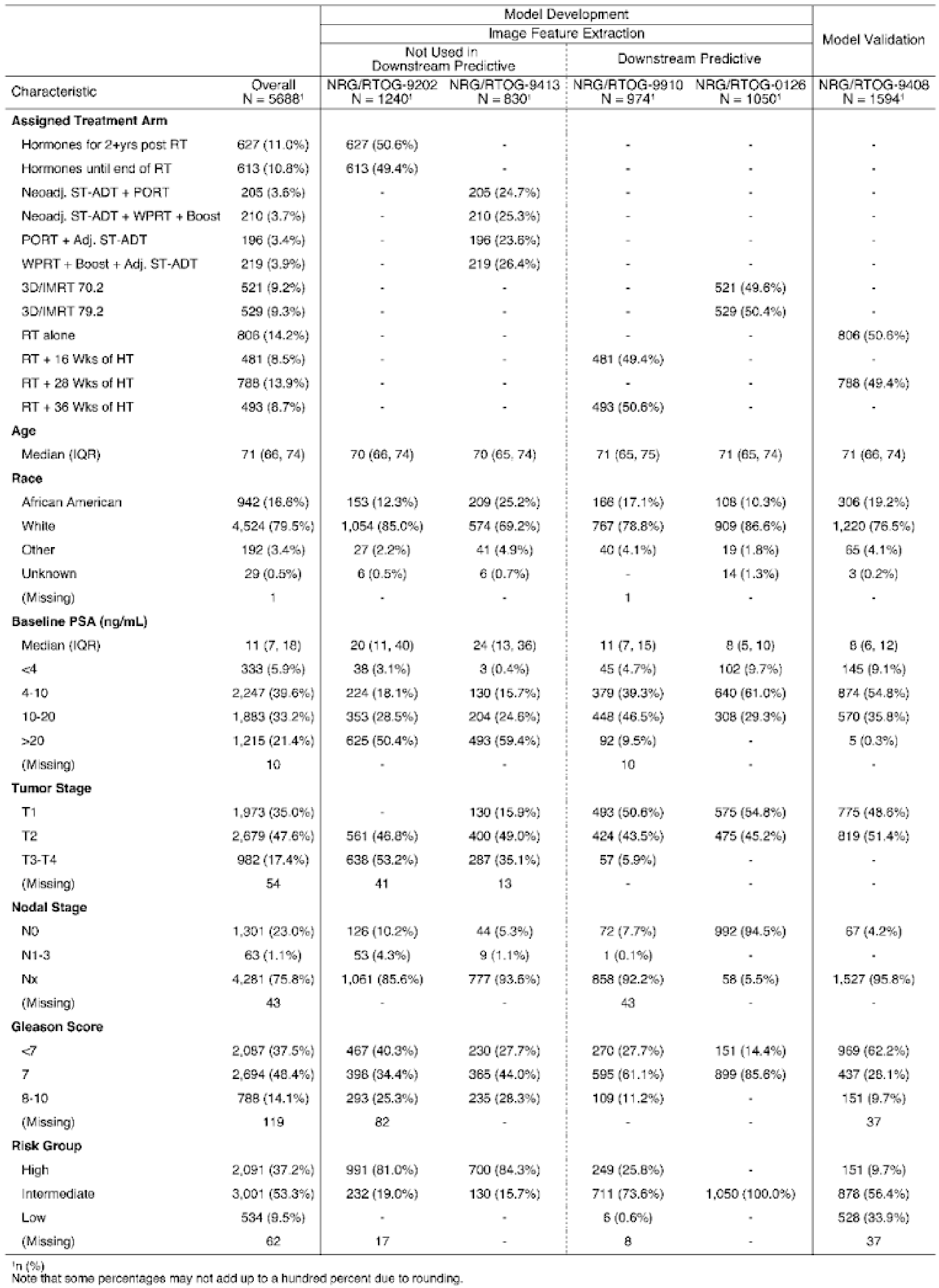
	Population	Dates	Experimental Arm	Control Arm	Patients in Experimental Arm	Patients in Control Arm	Primary Endpoint	Primary Endpoint Result	Secondary Efficacy Endpoints
NRG/RTOG 9408	T1b-T2b and PSA ≤20	1994-2001	RT + 4 mo ADT	RT alone	987	992	OS	Negative[1]	DSM BF DM
NRG/RTOG 9413	>15% LN risk by Roach AND PSA <100; T2c-T4 tumors and Gleason >=6 were allowed regardless of LN risk	1995-1999	2x2 factorial: 4 mo neoadjuvant/concurrent vs adjuvant ADT Prostate only RT vs whole pelvis RT		1323 total		PFS	Positive[2]	BF OS LP DM RF PCSF
NRG/RTOG 9910	T1b-T4 GS2-6 PSA>10 ≤100 T1b-T4 GS 7 PSA<20 T1b-c GS 8-10 PSA <20	2000-2004	RT + 36 weeks ADT	RT + 4 mo ADT	737	752	DSS	Negative[3]	OS DFS LRP DM BF
NRG/RTOG 9202	T2c-T4 AND PSA<150	1992-1995	RT + 28 mo ADT	RT + 4 mo ADT	758	762	DFS	Positive[4]	LP DM BF DSS OS
NRG/RTOG 0126	cT1b-T2b, GS 2-6, and PSA 10-20; or GS 7 and PSA <15	2002-2008	79.2 Gy RT alone	70.2 Gy RT alone	748	751	OS	Negative[5]	BF DSS LP DM

[1] Jones et al., IJROBP 2021; [2] Roach et al., Lancet Oncol 2018; [3] Pisansky et al., JCO 2015;

[4]Lawton et al., IJROBP 2017; [5]Michalski et al., JAMA Oncol 2018

PSA = prostate-specific antigen; RT = radiation therapy; ADT = androgen-deprivation therapy; OS = overall survival; DSM = disease-specific mortality; BF = biochemical failure; DM = distant metastasis; PFS = progression-free survival; LP = local progression; RF = regional failure; PCSF = prostate cancer-specific failure; DSS = disease-specific survival; DFS = disease-free survival; LRP = locoregional progression

Table S2. Study summary and patient baseline characteristics for all NRG/RTOG trials used for model development and validation.



		Model Development				Model Validation
		Image Feature Extraction				
		Not Used in Downstream Predictive		Downstream Predictive		
Characteristic	Overall N = 5688[1]	NRG/RTOG-9202 N = 1240[1]	NRG/RTOG-9413 N = 830[1]	NRG/RTOG-9910 N = 974[1]	NRG/RTOG-0126 N = 1050[1]	NRG/RTOG-9408 N = 1594[1]
Assigned Treatment Arm						
Hormones for 2+yrs post RT	627 (11.0%)	627 (50.6%)	-	-	-	-
Hormones until end of RT	613 (10.8%)	613 (49.4%)	-	-	-	-
Neoadj. ST-ADT + PORT	205 (3.6%)	-	205 (24.7%)	-	-	-
Neoadj. ST-ADT + WPRT + Boost	210 (3.7%)	-	210 (25.3%)	-	-	-
PORT + Adj. ST-ADT	196 (3.4%)	-	196 (23.6%)	-	-	-
WPRT + Boost + Adj. ST-ADT	219 (3.9%)	-	219 (26.4%)	-	-	-
3D/IMRT 70.2	521 (9.2%)	-	-	-	521 (49.6%)	-
3D/IMRT 79.2	529 (9.3%)	-	-	-	529 (50.4%)	-
RT alone	806 (14.2%)	-	-	-	-	806 (50.6%)
RT + 16 Wks of HT	481 (8.5%)	-	-	481 (49.4%)	-	-
RT + 28 Wks of HT	788 (13.9%)	-	-	-	-	788 (49.4%)
RT + 36 Wks of HT	493 (8.7%)	-	-	493 (50.6%)	-	-
Age						
Median (IQR)	71 (66, 74)	70 (66, 74)	70 (65, 74)	71 (65, 75)	71 (65, 74)	71 (66, 74)
Race						
African American	942 (16.8%)	153 (12.3%)	209 (25.2%)	168 (17.1%)	108 (10.3%)	306 (19.2%)
White	4,524 (79.5%)	1,054 (85.0%)	574 (69.2%)	767 (78.8%)	909 (86.6%)	1,220 (76.5%)
Other	192 (3.4%)	27 (2.2%)	41 (4.9%)	40 (4.1%)	19 (1.8%)	65 (4.1%)
Unknown	29 (0.5%)	6 (0.5%)	6 (0.7%)	-	14 (1.3%)	3 (0.2%)
(Missing)	1	-	-	1	-	-
Baseline PSA (ng/mL)						
Median (IQR)	11 (7, 18)	20 (11, 40)	24 (13, 36)	11 (7, 15)	8 (5, 10)	8 (6, 12)
<4	333 (5.9%)	38 (3.1%)	3 (0.4%)	45 (4.7%)	102 (9.7%)	145 (9.1%)
4-10	2,247 (39.6%)	224 (18.1%)	130 (15.7%)	379 (39.3%)	640 (61.0%)	874 (54.8%)
10-20	1,883 (33.2%)	353 (28.5%)	204 (24.6%)	448 (46.5%)	308 (29.3%)	570 (35.8%)
>20	1,215 (21.4%)	625 (50.4%)	493 (59.4%)	92 (9.5%)	-	5 (0.3%)
(Missing)	10	-	-	10	-	-
Tumor Stage						
T1	1,973 (35.0%)	-	130 (15.9%)	493 (50.6%)	575 (54.8%)	775 (48.6%)
T2	2,679 (47.6%)	561 (46.8%)	400 (49.0%)	424 (43.5%)	475 (45.2%)	819 (51.4%)
T3-T4	982 (17.4%)	638 (53.2%)	287 (35.1%)	57 (5.9%)	-	-
(Missing)	54	41	13	-	-	-
Nodal Stage						
N0	1,301 (23.0%)	126 (10.2%)	44 (5.3%)	72 (7.7%)	992 (94.5%)	67 (4.2%)
N1-3	63 (1.1%)	53 (4.3%)	9 (1.1%)	1 (0.1%)	-	-
Nx	4,281 (75.8%)	1,061 (85.6%)	777 (93.6%)	858 (92.2%)	58 (5.5%)	1,527 (95.8%)
(Missing)	43	-	-	43	-	-
Gleason Score						
<7	2,087 (37.5%)	467 (40.3%)	230 (27.7%)	270 (27.7%)	151 (14.4%)	969 (62.2%)
7	2,694 (48.4%)	398 (34.4%)	365 (44.0%)	595 (61.1%)	899 (85.6%)	437 (28.1%)
8-10	788 (14.1%)	293 (25.3%)	235 (28.3%)	109 (11.2%)	-	151 (9.7%)
(Missing)	119	82	-	-	-	37
Risk Group						
High	2,091 (37.2%)	991 (81.0%)	700 (84.3%)	249 (25.8%)	-	151 (9.7%)
Intermediate	3,001 (53.3%)	232 (19.0%)	130 (15.7%)	711 (73.6%)	1,050 (100.0%)	878 (56.4%)
Low	534 (9.5%)	-	-	6 (0.6%)	-	528 (33.9%)
(Missing)	62	17	-	8	-	37

[1] n (%)

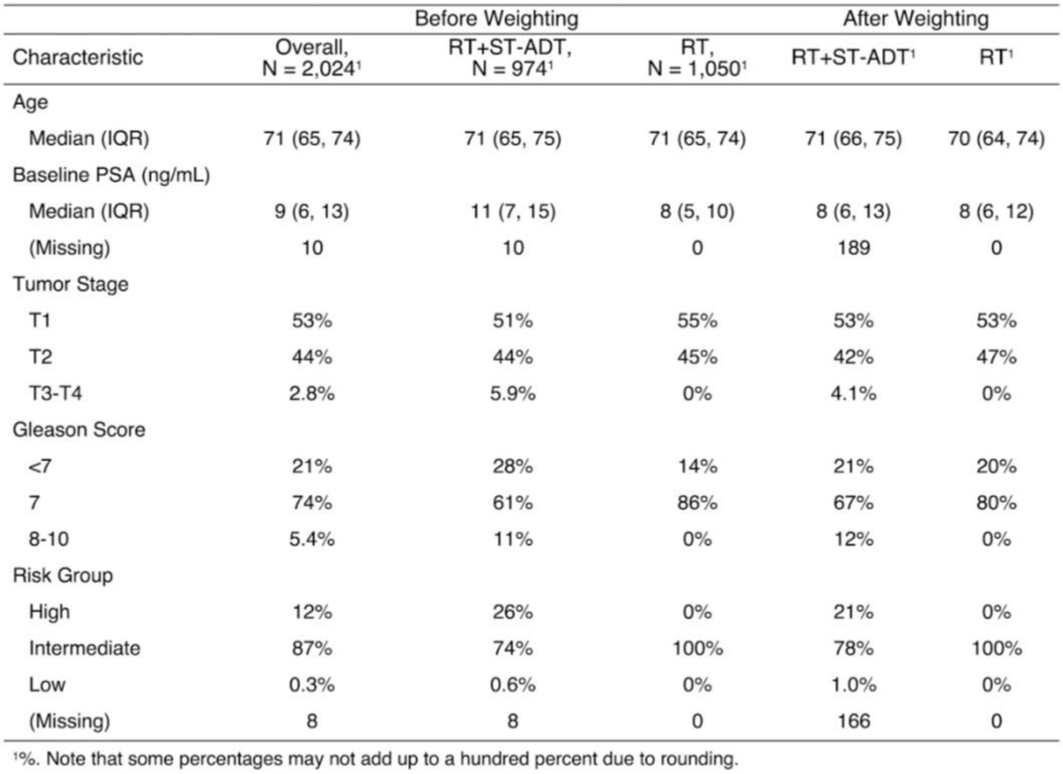
Note that some percentages may not add up to a hundred percent due to rounding.

3D/IMRT = intensity-modulated radiotherapy; Wks = weeks; RT = radiotherapy; HT = hormone therapy; IQR = interquartile range; PSA = prostate-specific antigen; ng/mL = nanograms per milliliter.

Table S3. Representativeness of patient cohort

Disease under investigation	Localized Prostate cancer
Race or ethnic group	Prostate cancer disproportionately affects African American men in the United States
Age	Incidence and mortality rates are strongly related to age with the highest incidence being seen in elderly men (> 65 years of age)
Geography	Incidence and mortality rates for prostate cancer vary worldwide, with the greatest incidence in North America and highest mortality rate in Asia.
Overall representativeness of study	The study population included patients from North America. Prostate cancer patients are younger outside North America thus, the age distribution in the current study population differs from that in some countries. In the RTOG-9408 validation set of 1,594 patients, there were 306 (19.2%) Black, 1,220 (76.5%) White, 68 (4.3%) Other/Unknown men. In the US, 488,375 men were diagnosed with localized prostate cancer from 2015-2019[11]. Of these men, there were 76,374 Black (15.6%) and 253,697 White (52%) men.

Table S4. Patient baseline characteristics for the development cohort.

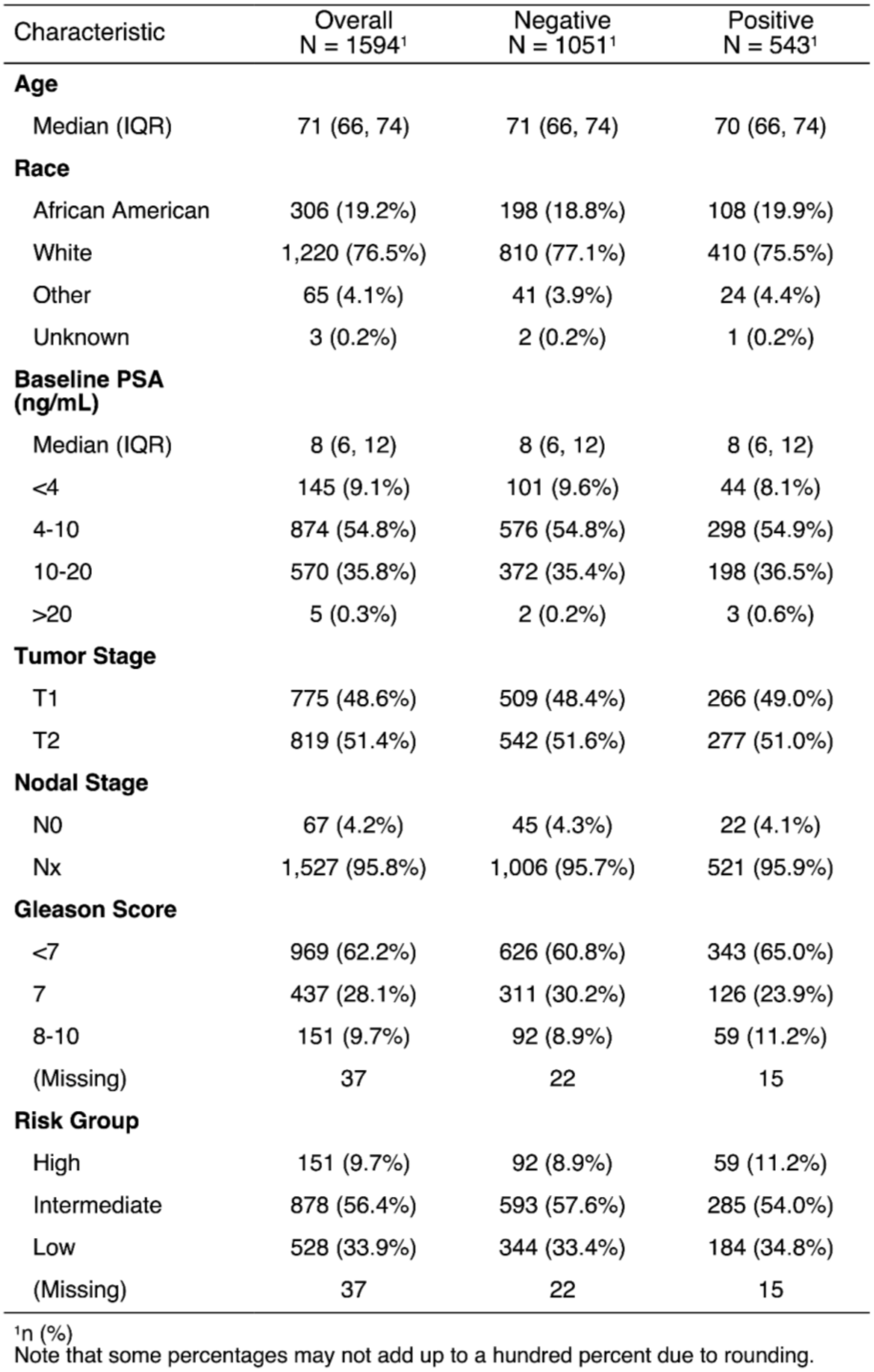


	Before Weighting			After Weighting	
Characteristic	Overall, N = 2,024[1]	RT+ST-ADT, N = 974[1]	RT, N = 1,050[1]	RT+ST-ADT[1]	RT[1]
Age					
Median (IQR)	71 (65, 74)	71 (65, 75)	71 (65, 74)	71 (66, 75)	70 (64, 74)
Baseline PSA (ng/mL)					
Median (IQR)	9 (6, 13)	11 (7, 15)	8 (5, 10)	8 (6, 13)	8 (6, 12)
(Missing)	10	10	0	189	0
Tumor Stage					
T1	53%	51%	55%	53%	53%
T2	44%	44%	45%	42%	47%
T3-T4	2.8%	5.9%	0%	4.1%	0%
Gleason Score					
<7	21%	28%	14%	21%	20%
7	74%	61%	86%	67%	80%
8-10	5.4%	11%	0%	12%	0%
Risk Group					
High	12%	26%	0%	21%	0%
Intermediate	87%	74%	100%	78%	100%
Low	0.3%	0.6%	0%	1.0%	0%
(Missing)	8	8	0	166	0

[1] %. Note that some percentages may not add up to a hundred percent due to rounding.

RT = radiation therapy; ADT = androgen-deprivation therapy; IQR = interquartile range; PSA = prostate-specific antigen; ng/mL = nanograms per milliliter.

Table S5. Patient baseline characteristics by predictive model group in NRG/RTOG 9408.



Characteristic	Overall N = 1594[1]	Negative N = 1051[1]	Positive N = 543[1]
Age			
Median (IQR)	71 (66, 74)	71 (66, 74)	70 (66, 74)
Race			
African American	306 (19.2%)	198 (18.8%)	108 (19.9%)
White	1,220 (76.5%)	810 (77.1%)	410 (75.5%)
Other	65 (4.1%)	41 (3.9%)	24 (4.4%)
Unknown	3 (0.2%)	2 (0.2%)	1 (0.2%)
Baseline PSA (ng/mL)			
Median (IQR)	8 (6, 12)	8 (6, 12)	8 (6, 12)
<4	145 (9.1%)	101 (9.6%)	44 (8.1%)
4-10	874 (54.8%)	576 (54.8%)	298 (54.9%)
10-20	570 (35.8%)	372 (35.4%)	198 (36.5%)
>20	5 (0.3%)	2 (0.2%)	3 (0.6%)
Tumor Stage			
T1	775 (48.6%)	509 (48.4%)	266 (49.0%)
T2	819 (51.4%)	542 (51.6%)	277 (51.0%)
Nodal Stage			
N0	67 (4.2%)	45 (4.3%)	22 (4.1%)
Nx	1,527 (95.8%)	1,006 (95.7%)	521 (95.9%)
Gleason Score			
<7	969 (62.2%)	626 (60.8%)	343 (65.0%)
7	437 (28.1%)	311 (30.2%)	126 (23.9%)
8-10	151 (9.7%)	92 (8.9%)	59 (11.2%)
(Missing)	37	22	15
Risk Group			
High	151 (9.7%)	92 (8.9%)	59 (11.2%)
Intermediate	878 (56.4%)	593 (57.6%)	285 (54.0%)
Low	528 (33.9%)	344 (33.4%)	184 (34.8%)
(Missing)	37	22	15

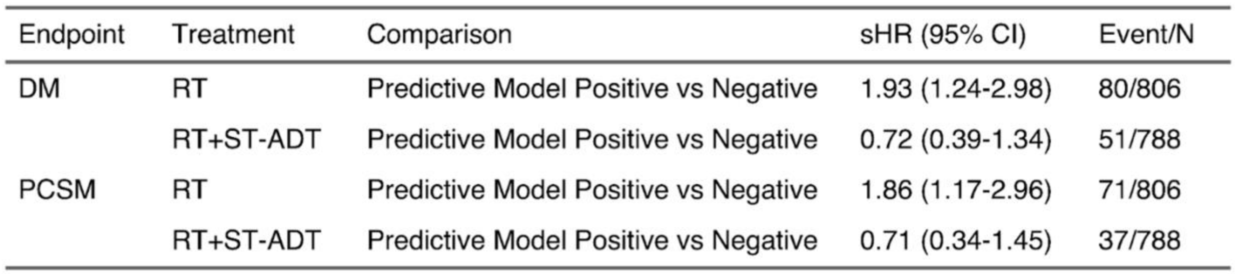
[1] n (%)

Note that some percentages may not add up to a hundred percent due to rounding.

IQR = interquartile range; PSA = prostate-specific antigen; ng/mL = nanograms per milliliter.

DM = distant metastasis; PCSM = prostate cancer-specific mortality; RT = radiation therapy; ST-ADT = short-term androgen-deprivation therapy; sHR = subdistribution hazard ratio; CI = confidence interval.

Table S6. Prognostic evaluation of the predictive model in NRG/RTOG 9408



Endpoint	Treatment	Comparison	sHR (95% CI)	Event/N
DM	RT	Predictive Model Positive vs Negative	1.93 (1.24-2.98)	80/806
	RT+ST-ADT	Predictive Model Positive vs Negative	0.72 (0.39-1.34)	51/788
PCSM	RT	Predictive Model Positive vs Negative	1.86 (1.17-2.96)	71/806
	RT+ST-ADT	Predictive Model Positive vs Negative	0.71 (0.34-1.45)	37/788

DM = distant metastasis; PCSM = prostate cancer-specific mortality; RT = radiation therapy; ST-ADT = short-term androgen-deprivation therapy; sHR = subdistribution hazard ratio; CI = confidence interval; N = number of patients.

References

1. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.

2. Mucherino A, Papajorgji P, Pardalos PM. Data Mining in Agriculture. Springer Science & Business Media; 2009.

3. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015;34: 3661–3679.

4. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol. 2008;168: 656–664.

5. Lee BK, Lessler J, Stuart EA. Weight Trimming and Propensity Score Weighting. PLoS ONE. 2011. p. e18174. doi:10.1371/journal.pone.0018174

6. Ilse M, Tomczak J, Welling M. Attention-based Deep Multiple Instance Learning. In: Dy J, Krause A, editors. Proceedings of the 35th International Conference on Machine Learning. PMLR; 10--15 Jul 2018. pp. 2127–2136.

7. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018;18: 24.

8. Breslow N. Covariance analysis of censored survival data. Biometrics. 1974;30: 89–99.

9. Leen TK, Dietterich TG, Tresp V. Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference. MIT Press; 2001.

10. Shapley LS. 17. A Value for n-Person Games. In: Kuhn HW, Tucker AW, editors. Contributions to the Theory of Games (AM-28), Volume II. Princeton: Princeton University Press; 1953. pp. 307–318.

11. Cancer of the Prostate - Cancer Stat Facts. In: SEER [Internet]. [cited 10 Apr 2023]. Available: https://seer.cancer.gov/statfacts/html/prost.html