Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials

Andre Esteva (andre@artera.ai)
Jean Feng
Douwe van der Wal
Shih-Cheng Huang
Jeffry P Simko
Sandy DeVries
Emmalyn Chen
Edward M Schaeffer
Todd M Morgan
Yilun Sun
Amirata Ghorbani
Nikhil Naik
Dhruv Nathawani
Richard Socher
Jeff M. Michalski
Mack Roach III
Thomas M. Pisansky
Jedidiah M. Monson
Farah Naz
James Wallace
Michelle J. Ferguson
Jean-Paul Bahary
James Zou
Matthew Lungren
Serena Yeung
Ashley E. Ross
NRG Prostate Cancer AI Consortium
Howard M. Sandler
Phuoc T. Tran
Daniel E. Spratt
Stephanie Pugh
Felix Y. Feng
Osama Mohamad

Research Article

Keywords: Prostate cancer, biomarker, digital pathology, AI, deep learning, clinical trials

Posted Date: April 20th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1573559/v1

Version of Record: A version of this preprint was published at npj Digital Medicine on June 8th, 2022. See the published version at https://doi.org/10.1038/s41746-022-00613-w.

EDITORIAL NOTE:

April 26, 2022: The author has provided information from the executive editor of the journal *npj Digital Medicine* indicating that this manuscript has been peer-reviewed and formally accepted for publication at *npj Digital Medicine* with an acceptance date of 13 April 2022.

Abstract

Prostate cancer is the most frequent cancer in men and a leading cause of cancer death. Determining a patient's optimal therapy is a challenge, where oncologists must select a therapy with the highest likelihood of success and the lowest likelihood of toxicity. International standards for prognostication rely on non-specific and semi-quantitative tools, commonly leading to over- and under-treatment. Tissue-based molecular biomarkers have attempted to address this, but most have limited validation in prospective randomized trials and expensive processing costs, posing substantial barriers to widespread adoption. There remains a significant need for accurate and scalable tools to support therapy personalization. Here we demonstrate prostate cancer therapy personalization by predicting long-term, clinically relevant outcomes using a multimodal deep learning architecture and train models using clinical data and digital histopathology from prostate biopsies. We train and validate models using five phase III randomized trials conducted across hundreds of clinical centers. Histopathological data was available for 5,654 of 7,764 randomized patients (71%) with a median follow-up of 11.4 years. Compared to the most common risk stratification tool—risk groups developed by the National Cancer Center Network (NCCN)—our models have superior discriminatory performance across all endpoints, ranging from 9.2–14.6% relative improvement in a held-out validation set. This artificial intelligence-based tool improves prognostication over standard tools and allows oncologists to computationally predict the likeliest outcomes of specific patients to determine optimal treatment. Outfitted with digital scanners and internet access, any clinic could offer such capabilities, enabling global access to therapy personalization.

Introduction

In 2020, 1,414,259 new cases and 375,304 deaths from prostate cancer occurred worldwide[1]. While prostate cancer is often indolent and treatment can be curative, prostate cancer represents the leading global cause of cancer-associated disability due to the negative effects of over- and under-treatment, and is a leading cause of cancer death in men[2,3]. Determining the optimal course of therapy for an individual patient is difficult, and involves considering their overall health, the characteristics of their cancer, the side effect profiles of many possible treatments, outcomes data from clinical trials involving patient groups with similar diagnoses, and the prognostication of their expected future outcomes. This challenge is compounded by the lack of readily accessible prognostic tools to better risk stratify patients.

One of the most common systems used to risk stratify patients worldwide is the National Comprehensive Cancer Network (NCCN), or D'Amico, risk groups developed in the late 1990s. This system is based on digital rectal examination of the prostate, serum prostate-specific antigen (PSA) level, and tumor biopsy grade assessed by histopathology. This three-tier system forms the basis of treatment recommendations used for localized prostate cancer throughout the world[4], but has repeatedly been shown to have suboptimal prognostic and discriminatory performance[5]. This in part is due to the subjective and non-specific nature of the core variables in these models. For instance, Gleason grading[6] was developed in the 1960s and has suboptimal interobserver reproducibility even amongst expert urologic pathologists[7,8].

Although newer clinicopathologic risk stratification systems have been created, three variables remain at their core—Gleason score, T-stage, and PSA[9].

More recently, tissue-based genomic biomarkers[10] have demonstrated superior prognostic performance. However, nearly all of these tests lack validation in prospective randomized clinical trials in the intended use population, and there has been little to no adoption outside of the United States due to costs, laboratory requirements, and processing time[11]. Importantly, prognostic models developed on cohort study data and not on randomized clinical trials are subject to selection biases from treatment decisions made in the clinic, and often have less accurate clinical and long-term outcome data. As such, there remains a serious unmet clinical need for improved and more accessible tools to personalize therapy for prostate cancer[12].

Artificial intelligence (AI) has demonstrated remarkable capabilities across a number of use-cases in medicine, ranging from physician-level diagnostics[13] to workflow optimization[14], and has the potential to support cancer therapy.[15,16] As clinical adoption of digital histopathology continues[17], AI can be implemented more broadly in the care of cancer patients. Advances and progress in the use of AI for histopathology-based prognostics have already begun, for instance by predicting short-term patient outcomes[18] or by improving the accuracy of Gleason-based cancer grading on postoperative surgical samples[19]. Whereas standard risk-stratification tools are fixed and based on few variables, AI can learn from large amounts of minimally processed data across various modalities. In contrast to genomic biomarkers, AI systems leveraging digitized images are lower-cost and massively scalable. In addition, these tools can incrementally improve over time through continued learning to optimize test performance and health care value.

In this study, we demonstrate that a multi-modal AI (MMAI) system can be used to address an unmet need for accessible and scalable prognostication in localized prostate cancer. This MMAI system has the potential to be a generalizable digital AI biomarker for global adoption. Herein, we train and validate prognostic biomarkers in localized prostate cancer using five NRG Oncology phase III randomized clinical trials by leveraging multi-modal deep learning on digital histopathology and clinical data[20–24]. By utilizing data from large clinical trials with long-term follow-up and treatment information that is standardized and less subject to bias, our model learns from and is trained on some of the most accurate clinical and outcome data available.

Results

We created a unique MMAI architecture that ingests both tabular clinical and image data, and trains with self-supervised learning to leverage the substantial amount of data available. We trained and validated six distinct models on a dataset of 16,204 histopathology slides (~ 16 TB of image data) and clinical data from 5,654 patients to predict six binary outcomes varying by endpoints and timeframes (5- and 10-year distant metastasis, 5- and 10-year biochemical failure, 10-year prostate cancer-specific survival, and

10-year overall survival). Notably, accurate prediction of distant metastasis at 5- and 10-years is particularly important for identifying patients who may have more aggressive disease and require additional treatment. We measured the performance of these models with the area under the time-dependent receiver operator characteristic curve (AUC) of sensitivity and specificity, based on censored events accounting for competing risks, and the NCCN risk groups served as our baseline comparator. Prior to model development, data from all five clinical trials were split into training (80%) and validation (20%). The MMAI model consistently outperformed the NCCN risk groups across all tested outcomes when comparing the performance results for the validation set.

Developing the MMAI architecture

For each patient, the MMAI model took as input clinical variables—including the NCCN variables (combined Gleason score, clinical T-stage, baseline PSA), as well as age, Gleason primary, and Gleason secondary—and digitized histopathology slides (median of 2 slides). Joint learning across both data streams is complex and involves building two separate machine learning pipelines—one for learning feature embeddings from the pathologic image data (Image pipeline) and the other to jointly learn from both clinical and image data to output risk scores for an outcome of interest (Fusion pipeline, see Fig. 1a). We standardized the image features across the trials for consistency.

Effective learning of relevant features from a variable number of digitized histopathology slides involves both image standardization and self-supervised pre-training. For each patient, we segmented out all the tissue sections in their biopsy slides, and combined them into a single large image, called an image quilt, of a fixed width and height across all patients (Supplementary Fig. 1). We then overlaid a grid over the image quilt which cut it into patches of size 256 x 256 pixels across its RGB channels. These patches were then used to train a self-supervised learning model[25] to learn histomorphological features useful for downstream AI tasks (Fig. 1b). Once trained, the self-supervised learning model took the patches of an image quilt and output a 128-dimensional vector representation for each patch. Concatenating all vectors in the same spatial orientation as the original patches yielded a feature tensor, which we called a feature-quilt, that effectively compressed the initially massive image quilt into a compact representation useful for further downstream learning. Before concatenation, this feature-quilt was averaged to further compress the representation into a 128-dimensional feature vector for each patient. The tabular clinical data was concatenated with the output of the image pipeline. The concatenated vector was then fed to a CatBoost classifier[26] and the model output a risk score for the task at hand.

Assembling NRG/RTOG clinical trials data

With approval from NRG Oncology, a National Clinical Trials Network (NCTN) group funded by the National Cancer Institute (NCI), we assembled a unique dataset from five large multinational randomized phase III clinical trials of men with localized prostate cancer (NRG/RTOG 9202, 9408, 9413, 9910, and 0126)[20–24]. All patients received definitive external radiotherapy (RT), with or without pre-specified use of androgen-deprivation therapy (ADT). Combined RT with short-term ADT was of four-month duration, with

medium-term ADT of 36-week duration, and with long-term ADT of 28-month duration (Table 1). Of the 7,764 eligible patients randomized in these five trials, there were 5,654 with high quality digital histopathology image data. This represented 16.1 TB of histopathology image data from 16,204 histopathology slides of pre-treatment biopsy samples.

Identifying human-interpretable self-supervised learning image features

The internal data representations of the self-supervised learning model are shown in Fig. 2. We fed the entire dataset's image patches through the self-supervised learning model and extracted model features—a 128-dimensional vector outputted by the model—for each patch. The Uniform Manifold Approximation and Projection algorithm (UMAP)[27] was applied to these features, projecting them from 128 dimensions down to two, and each patch was plotted as an individual point. Neighboring data points represent image patches that the model considered similar. UMAP grouped the feature vectors into 25 clusters, some of which are shown in various colors, and a pathologist was asked to interpret the 20 nearest-neighbor image patches of the cluster centroids and try to identify trends observed for each cluster. Insets in Fig. 2 show example image patches (and pathologist descriptions) that are close in feature space to the cluster centroids, and the full interpretation for all 25 clusters is shown in Supplementary Fig. 2.

Evaluating performance of the MMAI models on a validation set

The MMAI prognostic models developed using pathology images, NCCN variables (combined Gleason score, T-stage, baseline PSA), age, Gleason primary, and Gleason secondary, had superior discriminatory performance on the entire held-out test set across all clinical endpoints and timeframes when compared to the most commonly used NCCN risk stratification tool.

The model performance results for the validation sets are shown in Fig. 3. In Fig. 3a and Fig. 3d-h, the blue bars represent the performance of the MMAI models, each trained on a specific endpoint timeframe, and the gray bars represent the performance of the corresponding NCCN model. Figure 3b shows the relative improvement of the MMAI over NCCN across the outcomes, and across the subsets of the validation set that come from the five individual trials. As can be seen, our model consistently outperforms the NCCN model across all tested outcomes, with a substantial relative improvement in AUC varying from 9.2–14.6%. Further, the trial subsets unanimously see a relative improvement over NCCN except for prediction of 10-year biochemical failure in RTOG 9910. This trial had one of the lower event rates and shortest follow-up times compared to the remaining trials, and all patients received hormone therapy. With a short follow-up time, patients were less likely to recover their testosterone and prostate-specific antigen levels to be able to experience biochemical failure.

To evaluate the incremental benefit of various data components, we ran an ablation study in which we sequentially removed data features to understand their effect on the overall performance of the system[28]. We trained additional MMAI models using the following data setups: pathology images only, pathology images + the NCCN variables (combined Gleason score, T-stage, baseline PSA), and pathology images + NCCN variables + three additional variables (age, Gleason primary, Gleason secondary). Each additional

data component improved performance, with the full setup (pathology images and six clinical variables) yielding the best results (Fig. 3c). The MMAI prognostic model had superior discrimination compared to the NCCN model for all outcomes, including 5-year distant metastasis (AUC of 0.84 vs 0.74, p-value < 0.001), 5-year biochemical failure (AUC of 0.67 vs 0.59, p-value < 0.001), 10-year prostate cancer-specific survival (AUC of 0.77 vs 0.68, p-value < 0.001), and 10-year overall survival (AUC of 0.65 vs 0.59, p-value < 0.001).

Discussion

Prior work prognosticating outcomes from histopathology slides typically leverages extensive region-level pathologist annotations on slides (e.g., to train AI models to predict Gleason grading[19,29,30], or to prognosticate short-term outcomes[18]). In contrast, our technique learns from patient-level clinical data and unannotated histopathology slides and can prognosticate long-term outcomes. Moreover, the self-supervised learning of the image model allows it to learn from new image data without the need for additional annotations. When deploying machine learning (ML) models in a new domain (e.g., a new scanner or a new clinic) including data from that domain during training can help improve generalization and performance. This system also lowers the barrier for physicians to begin using this tool and for clinics to easily continue sharing histopathology image data to obtain prognostic information and aid their treatment decisions. This lower barrier for usage is a valuable advantage when deploying the system in new locations and attempting to adapt to their inherent biases—a challenge previously observed in medical AI deployment[31]. In addition, this tool focuses on supporting the oncologist in making treatment decisions and provides complementary information to the diagnostic and histopathology information identified by a pathologist.

Self-supervised learning[25] is a method recently popularized in the ML community for learning from datasets without annotations. Typical ML setups leverage supervised learning, in which datasets are composed of data points (e.g., images) and data labels (e.g., object classes). In contrast, during self-supervised learning, *synthetic* data labels are *extracted* from the original data, and used to train generic feature representations which can be used for downstream tasks. Here we find that momentum contrast[32]—a technique that takes the set of image patches, generates augmented copies of each patch, then trains a model to predict whether any two augmented copies come from the same original patch—is effective at learning features from digital pathology slides. The structural setup is shown in Figure 1b, with further details in the Methods. One challenge with real-world medical datasets is the sheer volume of image data available and potential class imbalance (tissue vs. no tissue) that is a result of how histopathology slides are created. To overcome this and guide the self-supervised learning process towards patch regions that are likely to be more clinically useful, we only use images from patients with a Gleason primary ≥4. Future work could investigate further whether other training techniques such as

transfer learning are effective, and whether training or fine-tuning models end-to-end using patient-level information improves final performance.

When new models are introduced, an understanding of how to use them in routine clinical care is critical to the adoption of such tools. In this study, we focus on the unique AI architecture used to develop the prognostic models trained on data from thousands of patients with accurate, long-term follow-up data and clinically relevant outcomes. The MMAI model consistently outperforms the NCCN model across all tested outcomes. However, further work will be required to evaluate the clinical utility of this model, including identifying actionable information from defined patient risk groups and calibration with contemporary data. Importantly, interpretability of the features used by the model to predict prognosis should be investigated, and will be the subject of future work.

By creating a deep learning architecture that simultaneously ingests multiple data types, including histopathology image data of variable sizes as well as clinical data, we built a deep learning system capable of inferring long-term patient outcomes that substantially outperforms established clinical models. This study leverages robust and large-scale clinical data from five prospective, randomized, multinational phase III trials with up to 20 years of patient follow-up for 5,654 patients across a varied population, enrolled at hundreds of different diverse medical centers. Validation of these prognostic classifiers on a large amount of clinical trials data—in the intended use population—uniquely positions these tools as aids to therapeutic decision-making. Barriers to the adoption of urine-, blood-, and tissue-based molecular assays include their invariably high costs, the collection and consumption of biospecimen samples, and long turnaround times. In contrast, AI tools lack these limitations, substantially lowering their barrier to large-scale adoption. Moreover, the growing adoption of digital histopathology will support the global distribution of AI-based prognostic and predictive testing. This will enable broad access to therapy personalization and enable AI algorithms to continue improving by learning from diverse multi-national data.

Methods

Dataset preparation. In collaboration with NRG Oncology, we obtained access to full patient-level baseline clinical data, digitized histopathology slides of pretreatment prostate biopsies, and longitudinal outcomes from five landmark, large-scale, prospective, randomized, multinational clinical trials containing 5,654 patients, 16,204 histopathology slides, and longer than 10 years of median follow-up: NRG/RTOG 9202, 9408, 9413, 9910, and 0126 (Table 1). Patients in these trials were randomized across various combinations of external radiotherapy (RT) with or without different durations of androgen deprivation therapy (ADT). The slides were digitized over a period of one year by NRG Oncology using a Leica Biosystems Aperio AT2 digital pathology scanner at a resolution of 20x. The histopathology images were manually reviewed for quality and clarity. Six baseline clinical variables that were collected across all

trials (combined Gleason score, Gleason primary, Gleason secondary, T-stage, baseline PSA, age), along with the digital histopathology images, were used for model training and validation. The patients from five trials were split into training (80%) and validation (20%) datasets, and there was no patient overlap among splits. To ensure that the test set captured a clinically relevant and representative subset of patients, the final test set was selected such that the NCCN risk group's 5-yr distant metastasis AUC performance was between 0.7 and 0.75, as observed in the literature[34,35]. Institutional Review Board approval was obtained from NRG Oncology (IRB00000781) and informed consent was waived because this study was performed with anonymized data.

Image pipeline. All tissue from digitized slides were segmented into a single image quilt of size 200 by 200 patches for each patient prior to model training. A simple grid was then laid over the image quilt to obtain contiguous and adjacent patches of size 256 x 256 pixels. We used a ResNet-50 model[36], together with the MoCo-v2 training protocol[37] (parameters: learning rate = 0.03 with a cosine learning rate schedule for 200 epochs, moco-t = 0.2, multilayer perceptron head, batch size of 256, the default MoCo-v2 parameters for augmentation), to train the self-supervised learning model used in the system architecture of Figure 1b. For the validation results shown in Figure 3a, we used images of patients with a Gleason primary ≥4 to pre-train a corresponding self-supervised learning model to effectively learn relevant histomorphologic features. Once self-supervised pre-training was complete, we fed in all patches with usable tissue (See tissue segmentation section) to the self-supervised pretrained ResNet-50 model to generate an image feature vector for each patch. These image feature vectors were averaged to produce a 128-dimensional image feature for each patient.

Fusion pipeline. To leverage information from both modalities (image and clinical features), we used a joint fusion approach. The tabular clinical data were all considered as numerical variables, and a CatBoost[26] model that took in a concatenation of numerical clinical variables and image features as input was used for model prediction and to output a risk score (parameters: learning rate 0.003, depth 5, L2 leaf regularization 10, 2000 iterations).

Tissue segmentation. After the slides were cut into 256 x 256 pixel patches, we developed an artifact classifier by training a ResNet-18 to classify whether a patch showed usable tissue, or whether it showed whitespace or artifacts. The artifact classifier was trained for 25 epochs, optimized using stochastic gradient descent with a learning rate of 0.001. The learning rate was reduced by 10% every 7 epochs. We manually annotated 3661 patches (tissue vs not tissue) and trained this classifier on 3366 of them, achieving a validation accuracy of 97.6% on the remaining patches. This artifact classifier was then used to segment tissue sections and filter out low quality images during image feature generation.

Model performance metrics (AUC). For each model and each outcome, we estimated the time-dependent receiver operating characteristic curve, accounting for competing events and censoring, using the R-package timeROC[38]. The area under this curve defines the model's performance. Each time-dependent curve was constructed by evaluating the sensitivities and specificities based on the disease statuses fixed at time *t* and the model predictions determined by sweeping through a threshold *c*. Methods detailed in Blanche et al. were used to compute pointwise 95% confidence intervals (1.96 x standard error) for AUCs and two-sided p-values for comparing AUCs of two models (e.g., MMAI vs. NCCN)[38].

NCCN risk groups. Three variables—clinical T-stage, Gleason score, and baseline PSA—were used to group patients into low-, intermediate-, and high-risk groups. The risk groups were defined as follows: low risk (cT1–cT2a, Gleason score ≤6, and PSA <10 ng/mL), intermediate risk (cT2b–cT2c, Gleason score 7, and/or PSA 10-20 ng/mL), and high risk (≥cT3a or Gleason score 8-10 or PSA >20 ng/mL)[9].

Declarations

Data Availability

The data that support the findings of this study included pathology slides, clinicopathologic variables, and outcomes information from NRG Oncology. Data may be made available for noncommercial academic use from the authors with permission from NRG Oncology. For access to the clinicopathology variables and outcomes information, please contact AP@nrgoncology.org. For the digitized pathology slides, please contact A.E. (aesteva@artera.ai).

Code Availability

The multi-modal AI architecture was developed using PyTorch Python library (https://pytorch.org/). In addition, scikit-learn, NumPy, statsmodels, pandas, Matplotlib, and MoCo-v2 have been used for computation and plotting (available under: https://scikit-learn.org/stable/, https://numpy.org/, https://www.statsmodels.org/, https://pandas.pydata.org/, https://matplotlib.org/, and https://github.com/facebookresearch/moco). The trained model used in this study has not yet undergone regulatory review and cannot be made available at this time. Interested researchers can contact A.E. (aesteva@artera.ai) for questions on its status and access.

Acknowledgements

This project was supported by grants U10CA180868 (NRG Oncology Operations), U10CA180822 (NRG Oncology SDMC), UG1CA189867 (NCORP), U24CA196067 (NRG Specimen Bank) from the National Cancer Institute (NCI). This work was also funded by Artera, Inc. The authors would like to thank Florence

Lo, Asad Ullah, and Jen Chieh Lee for digitizing the histopathology slides, as well as the accrual authors listed under the NRG Prostate Cancer AI Consortium.

Author Contributions

A.E. and F.Y.F. conceptualized the study, assembled the team, and advised this work. A.E. conceptualized the AI system, prepared the data, and trained the models. A.E, SC.H., and D.vdW. developed the multi-modal AI architecture. A.G., N.N., D.N., and E.C. supported data preprocessing and contributed to model development. R.S., J.Z., M.L., and S.Y. advised the multi-modal AI architecture. J.F. and Y.S. performed statistical analyses and model validation. J.P.S. directed histology slide preparation for scanning and annotated the image clusters. S.D. digitized the histopathology slides. F.Y.F., E.M.S., T.M.M., A.E.R., and P.T.T. provided strategic guidance and urology domain expertise. F.Y.F., H.M.S., T.M.P., J.M.M., M.R., F.N., J.W., M.J.F., and JP.B. supported patient accrual for all clinical trials in this study. S.P. and the NRG Prostate Cancer AI Consortium prepared and collected the data for all clinical trials in this study for over 20 years. D.E.S., A.E., F.Y.F., and E.C. prepared the manuscript with input from all authors. O.M. and F.Y.F. assembled the clinical team, obtained the dataset, and managed the scanning operations for slide digitization. O.M. and F.Y.F. contributed equally as co-last authors.

Competing Interests

A.E., D.vdW., and E.C. are employees at Artera. A.E., D.vdW., D.N., R.S. and N.N are or were employees of Salesforce.com, Inc. F.Y.F. is an advisor to and holds equity in Artera and is a consultant for Janssen, Roivant, Myovant, Bayer, Novartis, Varian, Blue Earth Diagnostics and Exact Sciences. L.S. received travel support and honorarium from Varian Medical Systems and is on the advisory board for AbbVie. M.K. received funding from Limbus AI, is a consultant for Palette Life Sciences, and is on the advisory board for AbbVie, Ferring, Janssen, and TerSera. A.E.R. is a consultant for Astellas, Bayer, Blue Earth, Janssen, Myovant, Pfizer, Progenics, and Veracyte. H.S. is a member of the ASTRO Board and a member of the clinical trials steering committee for Janssen. P.T.T. is a consultant for Johnson & Johnson, RefleXion Medical, Myovant and AstraZeneca. D.E.S. is a consultant for AstraZeneca, Blue Earth, Bayer, Boston Scientific, Gammatile, Janssen, Novartis, and Varian. The remaining authors declare no competing interests.

NRG Prostate Cancer AI Consortium

Felix Y. Feng[4], Howard M. Sandler[18], Michael Kucharczyk[22], Luis Souhami[23], Leslie Ballas[24], Christopher A. Peters[25], Sandy Liu[26], Alexander G. Balogh[27], Pamela D. Randolph-Jackson[28], David L. Schwartz[29], Michael R. Girvigian[30], Naoyuki G. Saito[31], Adam Raben[32], Rachel A. Rabinovitch[33], Khalil Katato[34]

[22]Department of Radiation Oncology, Accrual-Nova Scotia Cancer Centre, Halifax, Nova Scotia, Canada, B3H 1V7

[23]Department of Oncology, The Research Institute of the McGill University Health Centre (MUHC), Montreal, QC H4A 3J1 CA

[24]Department of Radiation Oncology, University of Southern California, Los Angeles, CA

[25]Department of Radiation Oncology, Northeast Radiation Oncology Centers (NROC), Dunmore, PA

[26]Department of Medical Oncology, UCLA Medical Center, Los Angeles, CA

[27]Department of Radiation Oncology, Tom Baker Cancer Centre, Calgary, AB T2N 4N2 CA

[28]Department of Radiation Oncology, Washington Cancer Institute, Washington, DC

[29]Department of Radiation Oncology, Veteran Affairs New York Harbor Healthcare System-Brooklyn Campus, Brooklyn, NY

[30]Department of Radiation Oncology, Kaiser Permanente Los Angeles Medical Center, Los Angeles, CA

[31]Department of Radiation Oncology, Indiana University School of Medicine, Indianapolis, IN

[32]Department of Radiation Oncology, Christiana Care Health System-Christiana Hospital, Newark, DE

[33]Department of Radiation Oncology, University of Colorado Hospital, Aurora, CO

[34]Department of Medical Oncology, Hurley Medical Center, Flint, MI

References

1. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J. Clin. **71**, 209–249 (2021).
2. Carroll, P. H. & Mohler, J. L. NCCN Guidelines Updates: Prostate Cancer and Prostate Cancer Early Detection. J. Natl. Compr. Canc. Netw. **16**, 620–623 (2018).
3. Ward, E. M. *et al.* Annual Report to the Nation on the Status of Cancer, Featuring Cancer in Men and Women Age 20–49 Years. *J. Natl. Cancer Inst.* **111**, 1279–1297 (2019).
4. Houshdar Tehrani, M. H., Gholibeikian, M., Bamoniri, A. & Mirjalili, B. B. F. Cancer Treatment by Caryophyllaceae-Type Cyclopeptides. Front. Endocrinol. **11**, 600856 (2020).
5. Daskivich, T. J., Wood, L. N., Skarecky, D., Ahlering, T. & Freedland, S. Limitations of the national comprehensive cancer network (NCCN®) guidelines for prediction of limited life expectancy in men with prostate cancer. J. Urol. **197**, 356–362 (2017).

6. Chen, N. & Zhou, Q. The evolving Gleason grading system. Chin. J. Cancer Res. **28**, 58–64 (2016).
7. Allsbrook, W. C., Jr *et al.* Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. Hum. Pathol. **32**, 74–80 (2001).
8. Allsbrook, W. C., Jr *et al.* Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. Hum. Pathol. **32**, 81–88 (2001).
9. Schaeffer, E. *et al.* NCCN Guidelines Insights: Prostate Cancer, Version 1.2021: Featured Updates to the NCCN Guidelines. J. Natl. Compr. Canc. Netw. **19**, 134–143 (2021).
10. Kornberg, Z., Cooperberg, M. R., Spratt, D. E. & Feng, F. Y. Genomic biomarkers in prostate cancer. Transl. Androl. Urol. **7**, 459–471 (2018).
11. Gaudreau, P.-O., Stagg, J., Soulières, D. & Saad, F. The Present and Future of Biomarkers in Prostate Cancer: Proteomics, Genomics, and Immunology Advancements. Biomark. Cancer **8**, 15–33 (2016).
12. Eggener, S. E., Bryan Rumble, R. & Beltran, H. Molecular Biomarkers in Localized Prostate Cancer: ASCO Guideline Summary. JCO Oncology Practice vol. 16 340–343 (2020).
13. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115–118 (2017).
14. Naik, N. *et al.* Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. Nat. Commun. **11**, 5727 (2020).
15. Ho, D. Artificial intelligence in cancer therapy. Science **367**, 982–983 (2020).
16. Kann, B. H., Hosny, A. & Aerts, H. J. W. L. Artificial intelligence for clinical oncology. Cancer Cell **39**, 916–927 (2021).
17. van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. Nat. Med. **27**, 775–784 (2021).
18. Wulczyn, E. *et al.* Interpretable survival prediction for colorectal cancer using deep learning. NPJ Digit Med **4**, 71 (2021).
19. Wulczyn, E. *et al.* Predicting prostate cancer specific-mortality with artificial intelligence-based Gleason grading. Communications Medicine **1**, 1–8 (2021).
20. Jones, C. U. *et al.* Radiotherapy and short-term androgen deprivation for localized prostate cancer. N. Engl. J. Med. **365**, 107–118 (2011).
21. Michalski, J. M. *et al.* Effect of Standard vs Dose-Escalated Radiation Therapy for Patients With Intermediate-Risk Prostate Cancer: The NRG Oncology RTOG 0126 Randomized Clinical Trial. JAMA Oncol **4**, e180039 (2018).
22. Pisansky, T. M. *et al.* Duration of androgen suppression before radiotherapy for localized prostate cancer: radiation therapy oncology group randomized clinical trial 9910. J. Clin. Oncol. **33**, 332–339 (2015).
23. Horwitz, E. M. *et al.* Ten-year follow-up of radiation therapy oncology group protocol 92 – 02: a phase III trial of the duration of elective androgen deprivation in locally advanced prostate cancer. J. Clin. Oncol. **26**, 2497–2504 (2008).

24. Roach, M., 3rd *et al.* Phase III trial comparing whole-pelvic versus prostate-only radiotherapy and neoadjuvant versus adjuvant combined androgen suppression: Radiation Therapy Oncology Group 9413. J. Clin. Oncol. **21**, 1904–1911 (2003).
25. Jing, L. & Tian, Y. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. IEEE Trans. Pattern Anal. Mach. Intell. **PP**, (2020).
26. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. *arXiv:1706.09516 [cs.LG]* Preprint at https://doi.org/10.48550/arXiv.1706.09516 (2017).
27. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [stat.ML]* Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2018).
28. Fawcett, C. & Hoos, H. H. Analysing differences between algorithm configurations through ablation. J. Heuristics **22**, 431–458 (2016).
29. Nagpal, K. *et al.* Development and Validation of a Deep Learning Algorithm for Gleason Grading of Prostate Cancer From Biopsy Specimens. JAMA Oncol **6**, 1372–1380 (2020).
30. Bulten, W. *et al.* Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol. **21**, 233–241 (2020).
31. Beede, E. *et al.* A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. in *Proceedings of the* 2020 *CHI Conference on Human Factors in Computing Systems* 1–12 (Association for Computing Machinery, 2020).
32. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9729–9738 (2020).
33. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 779–788 (2016).
34. Klein, E. A. *et al.* Decipher Genomic Classifier Measured on Prostate Biopsy Predicts Metastasis Risk. Urology **90**, 148–152 (2016).
35. Spratt, D. E. *et al.* Development and Validation of a Novel Integrated Clinical-Genomic Risk Group Classification for Localized Prostate Cancer. J. Clin. Oncol. **36**, 581–590 (2018).
36. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (openaccess.thecvf.com, 2016).
37. Chen, X., Fan, H., Girshick, R. & He, K. Improved Baselines with Momentum Contrastive Learning. *arXiv:2003.04297 [cs.CV]* Preprint at https://doi.org/10.48550/arXiv.2003.04297 (2020).
38. Blanche, P., Dartigues, J.-F. & Jacqmin-Gadda, H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat. Med. **32**, 5381–5397 (2013).

Tables

Table 1 is in the supplementary files section.

Figures

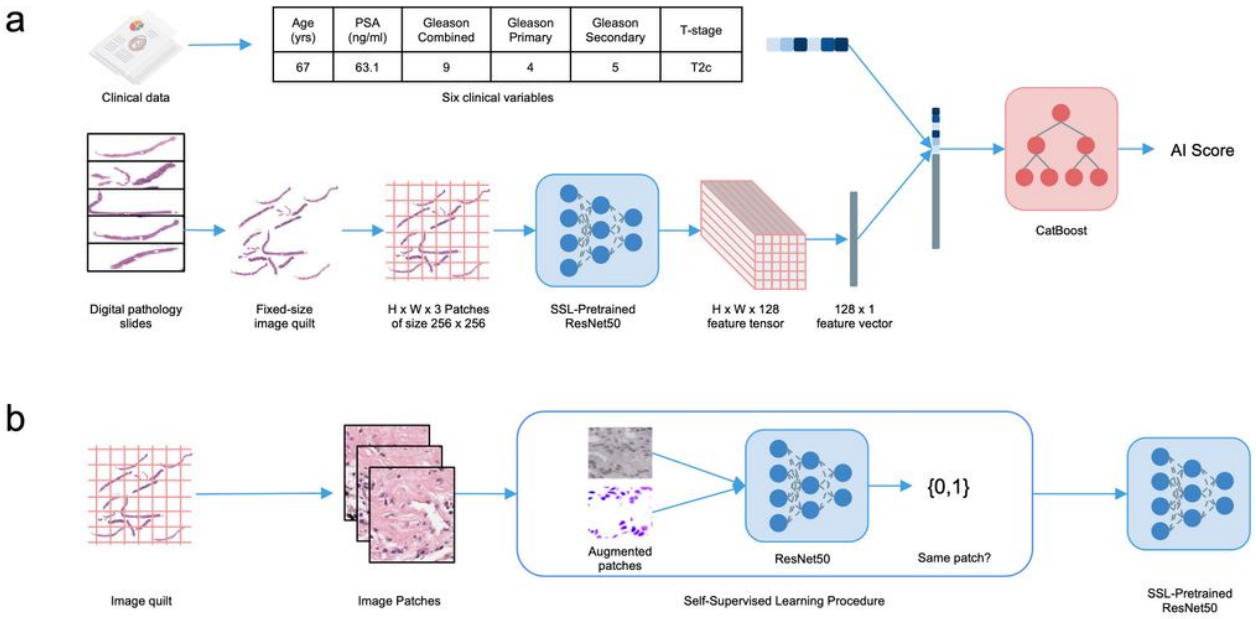


Figure 1

Multi-modal deep learning system and dataset. (a) The multi-modal architecture is composed of two parts: a tower stack to parse a variable number of digital histopathology slides and another tower stack to merge the resultant features and predict binary outcomes. **(b)** The training of the self-supervised model of the image tower stack.

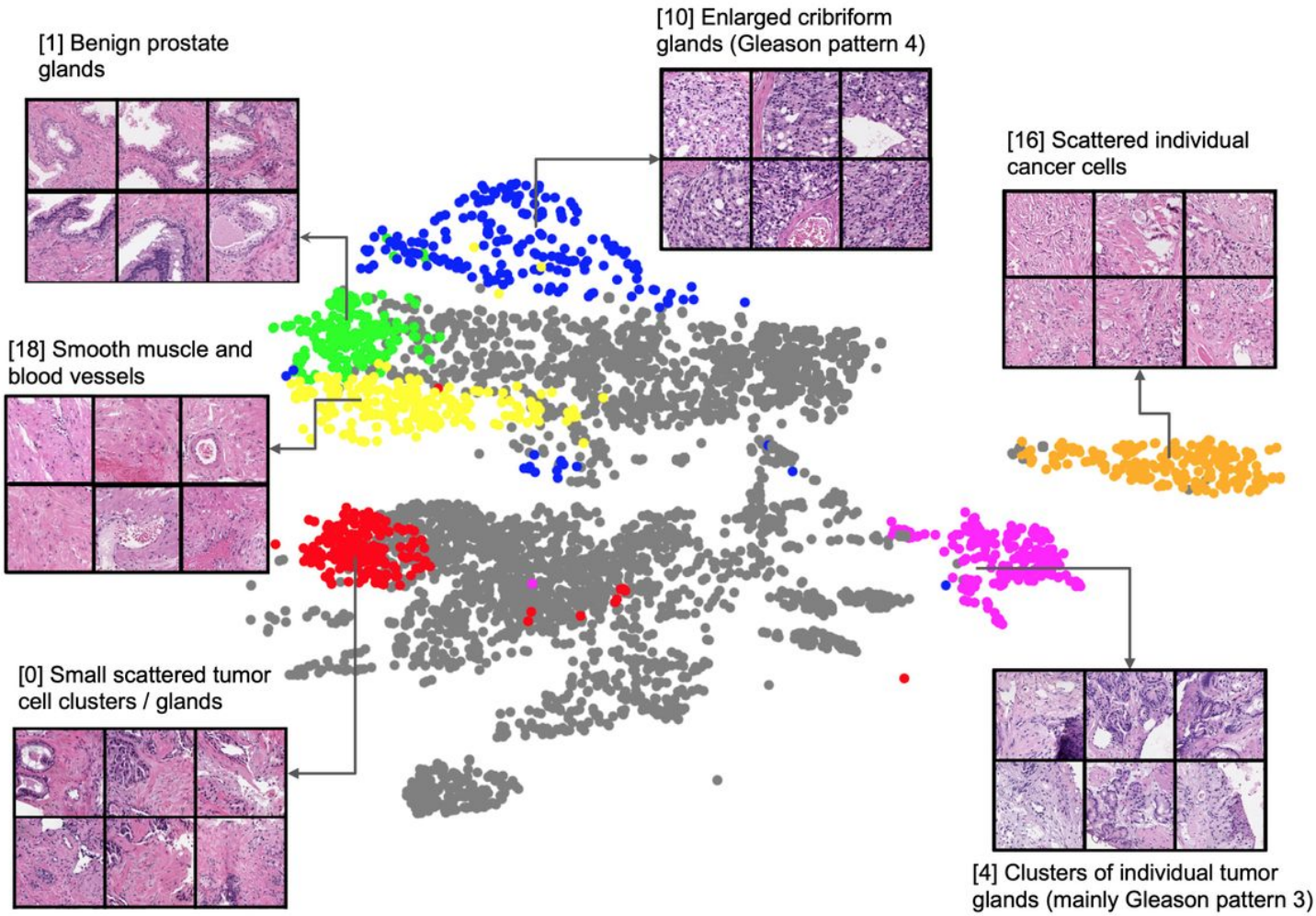
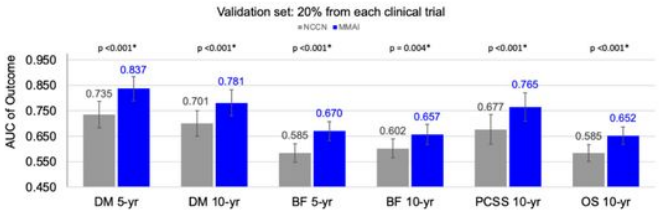


Figure 2

Pathologist interpretation of self-supervised model tissue clusters. The self-supervised model in the multi-modal model was trained to identify whether or not augmented versions of small patches of tissue came from the same original patch, without ever seeing clinical data labels. After training, each image patch in the dataset of 10.05M image patches was fed through this model to extract a 128-dimensional feature vector, and the UMAP algorithm[27] was used to cluster and visualize the resultant vectors. A pathologist was then asked to interpret the 20 image patches closest to each of the 25 cluster centroids—the descriptions are shown next to the insets. For clarity, we only highlight 6 clusters (colored), and show the remaining clusters in gray. See Supplementary Figure 2 for full pathologist annotation.

a



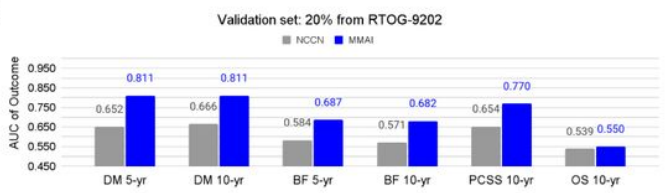
b

Relative Improvement						
20% Testset	**DM 5-yr**	**DM 10-yr**	**BF 5-yr**	**BF 10-yr**	**PCSS 10-yr**	**OS 10-yr**
All Trials	13.85%	11.43%	14.57%	9.19%	13.06%	11.46%
RTOG-9202	24.39%	21.77%	17.64%	19.44%	17.74%	2.04%
RTOG-9408	16.69%	13.66%	18.33%	16.21%	16.40%	14.07%
RTOG-9413	24.39%	11.97%	8.19%	0.18%	14.48%	20.19%
RTOG-9910	22.26%	36.56%	19.59%	-3.62%	10.19%	8.18%
RTOG-0126	32.91%	27.19%	18.95%	11.55%	31.53%	33.60%

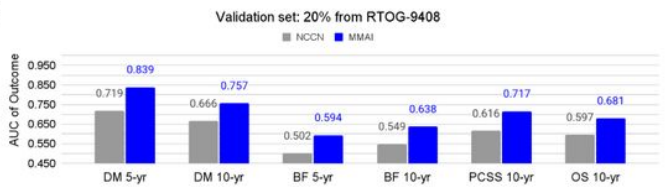
c

Ablation Study						
Model	**DM 5-yr**	**DM 10-yr**	**BF 5-yr**	**BF 10-yr**	**PCSS 10-yr**	**OS 10-yr**
Path+NCCN+3	0.837	0.781	0.670	0.657	0.765	0.652
Path+NCCN	0.827	0.782	0.666	0.660	0.761	0.615
Path	0.779	0.728	0.646	0.612	0.766	0.587

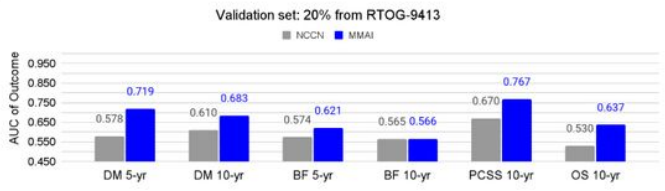
d



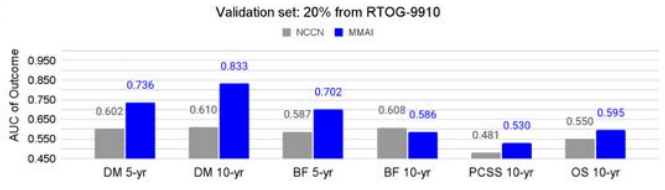
e



f



g



h

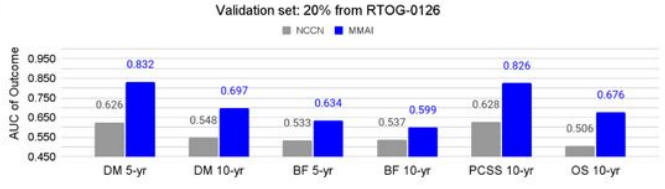


Figure 3

Comparison of the multi-modal deep learning system to NCCN risk groups across trials and outcomes. **(a)** Performance results reporting on the area under the curve (AUC) of time-dependent receiver operator characteristics of the MMAI (blue bars) vs NCCN (gray bars) models, include 95% confidence intervals and two-sided p-values. Comparison is made across 5-year and 10-year timepoints on the following binary outcomes: distant metastasis (DM), biochemical failure (BF), prostate cancer-specific survival (PCSS), and overall survival (OS). **(b)** Summary table of the relative improvement of the MMAI model over the NCCN model across the various outcomes and broken down by performance on the data from each trial in the validation set. Relative improvement is given by (AUC_{MMAI} - AUC_{NCCN}) / AUC_{NCCN}. **(c)** Ablation study showing model performance when trained on a sequentially decreasing set of data inputs, including the pathology images only (path), pathology images + NCCN variables (path+NCCN), and pathology images + NCCN variables + age + Gleason primary + Gleason secondary (path+NCCN+3). **(d-h)** Performance comparison on the individual clinical trial subsets of the validation set—together, these five comprise the entire validation set shown in (a).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ArteraNPJDIGITALMED03305R1SupplementaryInformation.pdf
- ArteraNPJDIGITALMED03305R1Table1NoLegend.jpg