AI in Precision Oncology
Volume 1, Number 2, 2024

DOI: 10.1089/aipo.2024.0004

AI IN PRECISION ONCOLOGY

Open camera or QR reader and scan code to access this article and other resources online.

ORIGINAL RESEARCH

Analytical Validation of a Clinical Grade Prognostic and Classification Artificial Intelligence Laboratory Test for Men with Prostate Cancer

Paul Gerrard,[1,*] Jingbin Zhang,[1] Rikiya Yamashita,[1] Huei-Chung Huang,[1] Sanghita Nag,[1] Sokha Nhek,[1] Joshua Kish,[1] Adam Cole,[2] Nathan Silberman,[1] Trevor J. Royce,[1] and Tim Showalter[1]

Abstract

Introduction: This is the first study of which we are aware to describe the analytical validation (AV) of clinical grade artificial intelligence (AI) algorithms for a commercially available prostate cancer test performed on hematoxylin and eosin stained specimens that is not dependent on *a priori* established molecules or *a priori* semantically meaningful morphology.

Methods: We adapted AV methods used in molecular diagnostics and clinical pathology to two AI biomarkers used in a clinical test for prostate cancer biopsy specimens. The two algorithms included one algorithm with prognostic performance and a second algorithm predictive for treatment benefit from short-term androgen deprivation therapy (ST-ADT). We assessed analytical accuracy, intra-operator reliability, and inter-operator reliability, and biopsy set completeness reliability on two AI algorithms deployed into a clinical laboratory setting. Analytical accuracy was measured using intraclass correlation coefficient (ICC). Reliability studies were assessed using ICC for the prognostic algorithm and percent agreement for the ST-ADT classification algorithm.

Results: Analytical accuracy ICC was 0.991 and 0.934 for the prognostic and ST-ADT algorithms, respectively. Intra-operator reliability was 0.981 (ICC) and 100% (percent agreement) for the prognostic and ST-ADT algorithms, respectively. Inter-operator reliability was 0.994 (ICC) and 93.3% (percent agreement) for the prognostic and ST-ADT algorithms, respectively. Biopsy-completeness reliability for one versus three prostate biopsy cores was 0.894 (ICC) and 91.67% (percent agreement) for the prognostic and ST-ADT algorithms respectively. For one versus six cores, reliability was 0.857 (ICC) and 95.00% (percent agreement) for the prognostic and ST-ADT algorithms respectively.

Conclusion: This study describes a novel approach to AV of AI algorithms in prostate cancer and applies this approach to two algorithms translated for use as a clinical grade AI-based laboratory test, supporting analytical validity of the test.

Keywords: precision medicine, machine learning, artificial intelligence, prostatic neoplasms, clinical laboratory techniques

[1]ArteraAI, Los Altos, California, USA.
[2]PathNet, Little Rock, Arkansas, USA.

*Address correspondence to: Paul Gerrard, MD, ArteraAI, 108 First Street, Los Altos, CA 94022, USA, E-mail: paul@artera.ai

Introduction

There is growing evidence supporting the use of artificial intelligence (AI) applied to hematoxylin and eosin (H&E) stained histopathology images to create AI biomarkers that predict adverse outcomes in prostate cancer.[1–3] These studies demonstrate clinical validity (CV) of novel AI biomarkers (e.g., the ArteraAI Prostate biomarker) in a research setting. We sought to implement AI in a clinical laboratory, which calls for protocols for establishing analytical performance and ensuring consistent high-quality testing for clinical specimens.

While prior guidelines or studies have described analytical validation (AV) of stains for specific proteins and AI interpretation of these specific stains,[4,5] to date we are unaware of AV methods that can be generalized to AI tests on H&E stained images that provide patient-level rather than slide-level results. Therefore, we aimed to develop AV approaches to such AI while establishing AV of the ArteraAI Prostate Test.

Given the dearth of published literature for translating AI algorithms into laboratory tests and the mounting interest in similar AI tests in the future, it was essential to develop an AV approach that would both meet specific projects needs as well as establish a framework that can broadly be used for other similar AI tests used on non-specific stains that prognosticate patient level events (e.g., distant metastasis [DM]).

ArteraAI Prostate Test moves some of the heavy lifting in the laboratory from a physical to a software workflow. This abstraction from physical processes using specialized equipment and reagents to a workflow based on more general use devices relying on non-specific stains has important implications for AV and the definition of the ''biomarker.'' Fundamentally, AV of an assay is based on establishing the ability of the assay to accurately detect the biomarker of interest.

Therefore, a critical question for the ArteraAI Prostate Test is ''What are the biomarkers of interest?'' For an AI algorithmic test reliant on measuring IHC stains for *a priori* specified epitopes, the epitopes themselves are the biomarkers of interest, and AV of epitope detection may largely be a matter of probe performance rather than the software, which is now detecting specific probes rather than more general tissue morphology.

Alternatively, for an AI algorithmic test that utilizes non-specific probes (e.g., H&E), the output of the algorithm rather than the algorithm measured input is the only meaningful biomarker of interest. This is indeed how Paige Prostate showed AV, using algorithm output rather than input. However, there remains a fundamental difference between Paige Prostate and ArteraAI Prostate Test, which is that Paige Prostate is used to point out a slide-level finding to a pathologist, specifically regions of likely cancer within a slide with associated geometric coordinates on that slide. ArteraAI Prostate Test uses algorithms that yield patient-level findings rather than slide-level findings, specifically outputs associated with risks of oncologic endpoints for the patient.

Fundamentally, these challenges reflect the underlying issue of ''What are the analytes or biomarkers measured within the ArteraAI Prostate Test?'' Unlike most laboratory tests, for ArteraAI Prostate Test the ''biomarker'' is not only algorithm output rather than input, but also output directly predictive of a patient–level clinical outcome as opposed to a slide-level finding with specific geometric coordinates. This informs how we assessed AV.

Here, we briefly summarize the operation and CV of the current ArteraAI Prostate Test followed by an explanation of AV.

Materials and Methods

Descriptions of AI algorithms

ArteraAI Prostate Test utilizes both image data and clinical data in two distinct algorithms, which are described here along with their CV. These algorithms operate by breaking a whole slide image into patches and analyzing the patches as illustrated in Figure 1. More detailed descriptions of the AI architecture are contained in the Supplementary Data. The first algorithm provides AI scores associated with DM and prostate cancer-specific mortality (PCSM) (prostate prognostic algorithm or model).

For clinically meaningful interpretation, continuous probabilistic estimates of 10 year DM risk, 5 year DM risk, and 10 year PCSM are reported based on monotonic transformations of the AI scores. The second AI algorithm reports a binary classification with one of the two following outcomes: (1) Likely to benefit from short-term androgen deprivation therapy (ST-ADT) or (2) Unlikely to benefit from ST-ADT as described in Spratt et al. (the ST-ADT predictive algorithm or model).[2]

Validations of prior versions of these algorithms have been previously published,[1–3] but descriptions of clinically implemented algorithms are described here.

The test is intended to accept prostate core needle biopsy tissue and can be performed on one to six prostate cores, including the core with the highest Gleason Grade. The test also accepts the following clinical variables, which are provided by the ordering clinician and/or originating lab for the specimen:

- Age
- Baseline prostate specific antigen (PSA)
- T-stage

Pathologist-assessed Primary, Secondary, and Overall Gleason Scores were previously captured as clinical variables, but they have been removed to reduce the test's dependence on an operator dependent input variable.[6–8]

The ArteraAI Prostate Test is currently performed as a laboratory-developed test (LDT) in Jacksonville, FL on a 3DHistech P1000 scanner.

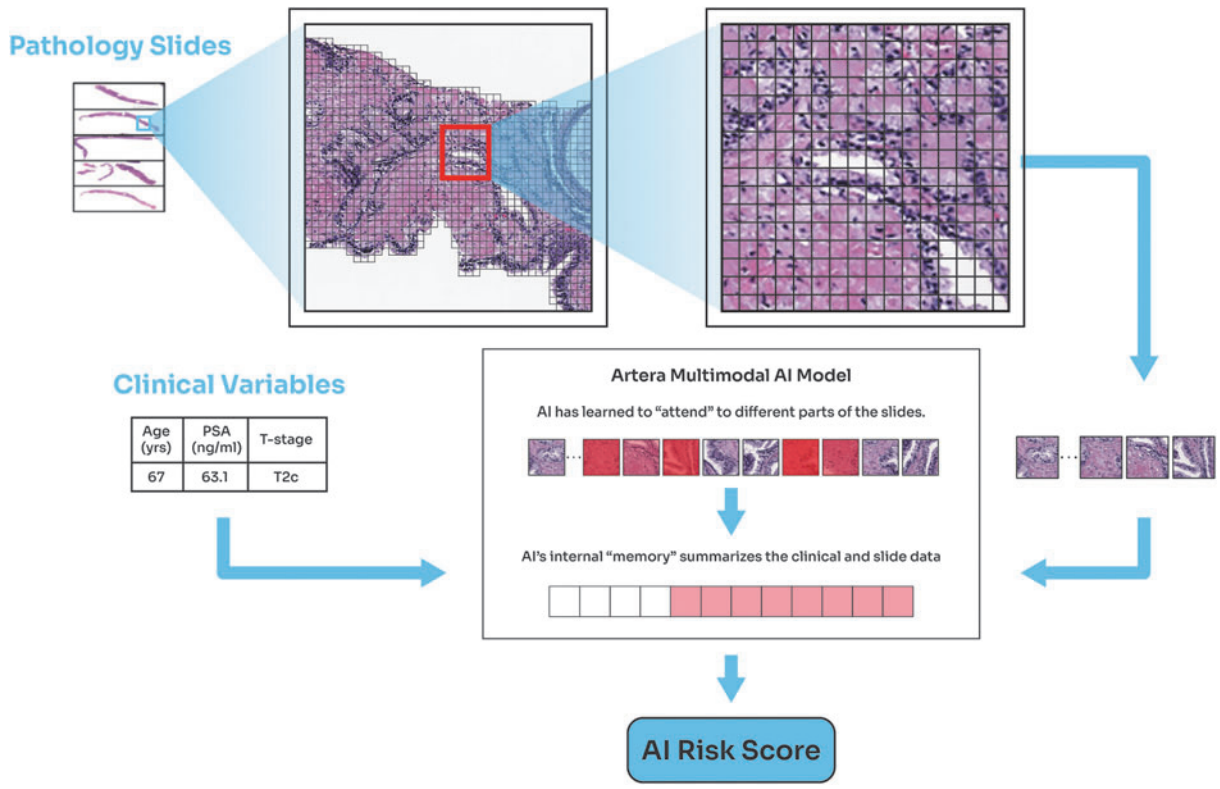

Fig. 1. ArteraAI Prostate Tissue Classification and AI. A tissue specimen on a slide is digitized into a single large image. This image is "patchified." Each patch is classified as containing tissue or not, and artificial intelligence is applied to tissue-containing patches. AI, artificial intelligence; PSA, prostate specific antigen.

Summary of clinical validation

CV on algorithm with prognostic performance. We obtained data from 8 randomized controlled clinical trials (NRG/RTOG protocols 9202, 9408, 9413, 9910, 0126, 0415, 0521, and 9902). These clinical trials enrolled over 10,000 patients. Cases from these clinical trials were used for which clinical data and H&E stained slides were available yielding a total of 7026 cases. Cases with missing age, baseline PSA, Gleason grade, T-stage, and outcome data were excluded. Cases were split into downstream model training and selection ($n = 5259$) and validation ($n = 1767$) cohorts.

No data from the training cohort was used in the validation cohort. H&E stained specimens were digitized on a Leica AT2. The algorithm was clinically validated for DM and PCSM using Fine-Gray models (sub-distribution hazard ratio [sHR] 2.41, 95% confidence interval [CI] 2.05–2.82, p-value <0.001 for DM and sHR 2.59, 95% CI 2.17–3.10, p-value <0.001 for PCSM) (Table 1).

CV on algorithm with predictive performance. Data from seven trials (NRG/RTOG 9202, 9413, 9902, 9910, 0126, 0415, and 0521) were used for development, and the NRG/RTOG 9408 trial was used for validation. NRG/RTOG 9408 enrolled 2028 patients. Cases from clinical trials were used for which clinical data and H&E stained slides were available yielding a total of 3977 cases used for development and 1509 cases from NRG/RTOG 9408 with 851 NCCN intermediate-risk cases used for validation. Cases with missing age, baseline PSA, Gleason grade, clinical T stage, and DM follow-up were excluded. H&E stained specimens were digitized on a Leica AT2.

Table 1. Fine and Gray Regression Results of the Algorithm Associated with Endpoints: Distant Metastasis and Prostate Cancer-Specific Mortality in a Multi-Trial Validation Cohort ($n = 1767$)

Endpoint	*sHR (95% CI),* p
DM	2.41 (2.05–2.82), <0.001
PCSM	2.59 (2.17–3.10), <0.001

sHRs reported out per 1 standard deviation increase in the algorithm score.

CI, confidence interval; DM, distant metastasis; PCSM, prostate cancer-specific mortality; sHR, sub-distribution hazard ratio.

We assessed CV for predictiveness using the significance of ST-ADT treatment interaction of the Fine and Gray regression model, with a two-sided p-value <0.05 indicative of a statistically significant difference in DM outcomes due to ST-ADT use between biomarker positive and biomarker negative patients when comparing hazard ratios. A total of 276 patients (32%) were classified as biomarker positive, where additional ST-ADT significantly reduced the risk of DM compared with radiation therapy alone (sHR 0.33, 95% CI [0.15–0.72], $p=0.006$).

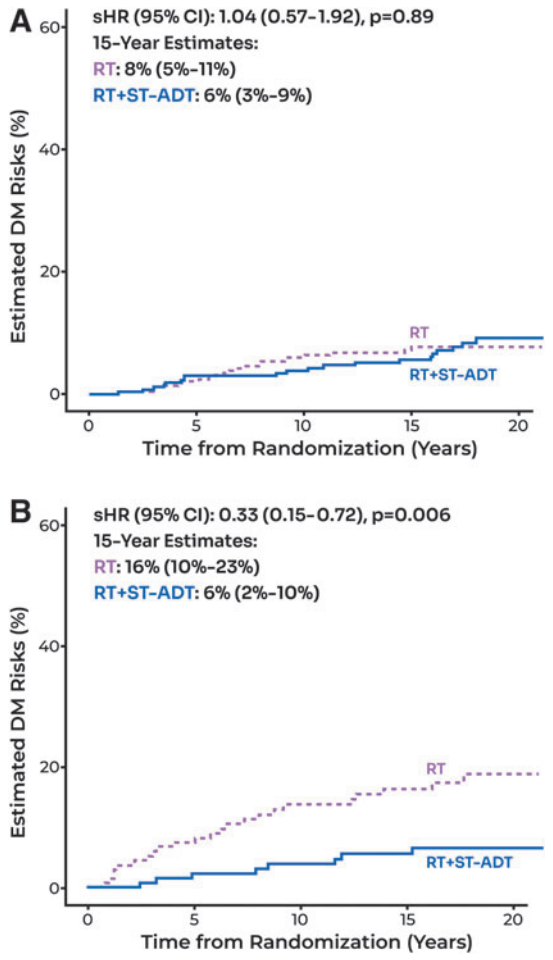

Fig. 2. Cumulative incidence estimates of treatment arms by biomarker subgroups in the validation cohort of RTOG 9408 NCCN intermediate patients ($n=851$). **(A)** The biomarker negative group. **(B)** sThe biomarker positive group. CI, confidence interval; DM, distant metastasis; Est., estimated; RT, radiation therapy; RT + ST-ADT, radiation therapy + short-term adjuvant deprivation therapy; sHR, sub-distribution hazard ratio.

In contrast, there was no significant difference between treatment for biomarker negative patients ($n=575$, sHR 1.04, 95% CI [0.57–1.92], $p=0.89$). The treatment-by-algorithm interaction for DM was observed ($p=0.02$). Cumulative incidence estimates of DMs stratified by biomarker are shown in Figure 2.

Analytical Validity Methods

Before AV of the AI test, we separately validated the whole slide imaging system as per College of American Pathology guidelines for digital pathology[9] for diagnostic pathology. We did this via a comparison of a pathologist's interpretation of glass slides and digitized images of those slides with a 2-week washout period in between. These guidelines are intended for validation for human interpretation using digital pathology rather than AI interpretation. Therefore, while an important step for the laboratory, we deemed it insufficient to establish AV of the assay using AI analyses.

Before designing AV experiments, we searched the medical literature, guidelines, and publications from the Food and Drug Administration (FDA) on other devices. We found that guidelines exist for AV of numerous analytical methods, but were able to find no specific guidelines on similar tests. We therefore sought to adapt a hybrid approach to AV informed by AV study designs from two other devices/devices classes, Paige Prostate.

From the perspective of intended use to a clinician, other prognostic tests seemed most relevant. In the absence of FDA-cleared or -approved tests in prostate cancer, we examined the FDA's validation guidance for devices classed under the NYI device class,[10] ''Classifier, Prognostic, Recurrence Risk Assessment, RNA Gene Expression, Breast Cancer,'' which is more fully defined as

> A device which uses a gene expression profile of a breast cancer tumor, from patients stage i or stage ii lymph node negative, with a tumor size of <5.0 cm, to provide a risk assessment for distant recurrence of breast cancer. The result is indicated for use only as a prognostic marker by physicians along with a number of other factors to assess the risk of recurrence of breast cancer.

includes Mammaprint and Prosigna., developing analogous tests for the relevant elements of each. While the NYI device class does not include image-based AI tests, it includes tests such as Mammaprint and Prosigna that provide algorithmic results regarding prognosis, giving it a similar intended use in breast cancer to the ArteraAI Prostate Test in prostate cancer.

We sought to examine validation methods from tests that were methodologically similar to ArteraAI Prostate Test. In developing our approach, we considered that, although Paige Prostate is operates an AI test performed on H&E stained prostate cancer specimens, it does not

provide patient-level results. Therefore, this test's workflow and methodology were the most similar to our test that we could identify. However, the intended use of the test is quite distinct from Artera AI Prostate Test, and since the results of Paige Prostate are slide-level results rather than patient-level., Therefore, we also considered AV of tests that provide similar patient-level results as well.

There are numerous gene expression classifiers that apply algorithmic analysis to provide statistical results regarding patient outcomes (e.g., prognostic tests) in a number of cancers, most notably in prostate and breast cancer. Of these tests, we found only two that had FDA clearance, both of which are breast cancer tests[11,12] classed under the NYI device class. While FDA clearance does not necessarily indicate a superior validation as compared with LDTs, FDA clearance is one of the most widely accepted standards for indicating a device performs its claimed indications. Furthermore, both Congress and the FDA have clearly been considering FDA oversight of LDTs as evidenced by proposed VALID Act legislation[13] and an FDA-proposed rule[14] regarding the regulation of LDTs. While neither of these has yet become law, this has indicated the importance of being ready for FDA oversight in laboratory test development as a means of mitigating evolving regulatory risks.

Analytical accuracy

Analytical accuracy refers to the ability of an assay to detect the analyte of interest. The analytes of interest for ArteraAI Prostate Test are the outputs of AI algorithms trained and validated on a Leica AT2.

ArteraAI prostate test was developed and validated on cases with typically only a small number of cores. In the validation data set for the prognostic model, having 1 to 2 cores was the most common, and 71% of specimens had no more than 6 cores. In the validation data set for the ST-ADT algorithm, having 3 to 4 cores was the most common, and 69% of specimens had no more than 6 cores. In combination with the biopsy completeness study results, tests were done using 1 core.

To establish analytical accuracy, we set AI biomarker results from cases digitized on a Leica AT2 as the gold standard against which results in the lab using the 3DHistech P1000 could be compared. We tested the same 60 specimens on a Leica AT2 and the 3DHistech P1000 used in Artera's CLIA lab. We used the intraclass correlation coefficient (ICC) for establishing concordance.

Reliability Studies

Intra-operator, inter-day study

An experiment was designed to assess consistency of results when the test was performed on the same specimen on three distinct days. A single core with the highest grade cancer from 30 cases was used for these studies. Specimens were scanned at time 0, after 24 h, and after 48 h for a total of 3 scans on 3 different days of each case by the same operator. Reliability was evaluated using ICC for the prognostic model. The associated 95% CI is calculated using approximate formulas for the standard error of the ICC estimate.[15] Percent agreement was used to assess reliability for the ST-ADT algorithm.

Inter-operator study

We designed this experiment to assess the consistency of results when the test was performed by two different operators to ensure consistency between operators. A single core with the highest grade cancer from 30 cases was used for this study. Two different operators scanned the images for analysis. Reliability was evaluated using ICC for the prognostic model. Reliability was evaluated using ICC for the prognostic model between the two operators. The associated 95% CI is calculated using approximate formulas for the standard error of the ICC estimate.[15] Percent agreement was used to assess reliability for the ST-ADT algorithm.

Biopsy completeness study

We sought to determine consistency of test results for different numbers of prostate cancer cores, which is analogous to limits of detection and quantification. As noted above, use of a single core was considered most representative of the validation data set, but multiple cores, commonly 12, are typically obtained in prostate cancer biopsies. Therefore, we sought to ensure that the test could perform consistently even with a higher number of cores as sometimes more than one core may be placed on a slide, or it may be challenging to identify which core has the highest grade tumor and be most appropriate for testing.

To evaluate this, we evaluated 60 cases. For each of the 60 cases, we ran the test on a single core with the highest grade tumor (the control), 3 cores including the core with the highest grade tumor, and 6 cores including the core with the highest grade tumor. We compared the results for 1 core versus 3 cores and 1 core versus 6 cores. Reliability was evaluated using ICC for the prognostic model. The associated 95% CI is calculated using approximate formulas for the standard error of the ICC estimate.[15] Percent agreement was used to assess reliability for the ST-ADT algorithm.

AV Results

The same overall experiments were used for establishing AV of both the prognostic model and ST-ADT model, though statistical evaluation was different, since the prognostic model outputs a continuous score whereas the ST-ADT model results are reported as binary. Results

Table 2. Analytical Accuracy

Model	*ICC (95% CI)*
Prognostic	0.993 (0.986–0.996)
ST-ADT	0.934

ICC, intraclass correlation coefficient; ST-ADT, short-term androgen deprivation therapy.

are summarized in Tables 2 and 3 given next for the prognostic and ST-ADT models, respectively. For all AV experiments, the results satisfied pre-specified acceptance criteria.

Discussion

In this study we sought to establish AV of a clinical grade AI-based LDT that evaluates morphological features from non-specific stains to predict patient level outcomes. We see this as a critical advance for technical implementation for the use of AI in pathology, as analytical or technical validation is a critical element of routine adoption of AI in pathology.[16] To our knowledge, AV of a similar test has never before been published, and methodology has not been described for doing this, requiring that we first establish appropriate experiments.

While at first glance it may seem that we could readily adopt approaches used in the AV of other AI products, we discovered that many of the experiments and questions addressed in other AI products were simply not applicable to our test.

There were two key issues that required us to develop novel approaches to determining AV. The first was that we were establishing AV of a test reliant on non-specific stains, so measures like primer and probe specificity were not relevant or readily adapted to a test reliant on H&E. The second key issue was that we were establishing AV of an AI biomarker that is trained to predict future events without relying on specific human-identifiable morphological features, so we could not use a pathologist's independent interpretation as a gold standard in the assessment.

While we have shown that a prior version of the AI model appeared to be associated with some human interpretable features,[1] the correspondence is not necessarily consistent, nor was the model designed for it to be so.

Table 3. Reliability Studies

	Prognostic model		*ST-ADT model*
Study	*ICC*	*95% CI*	*Percent agreement*
Intra-operator, inter-day study	0.981	0.965–0.990	100.000
Inter-operator study	0.994	0.988–0.997	93.333
Biopsy completeness: 1 vs. 3 cores	0.894	0.828–0.935	91.667
Biopsy completeness: 1 vs. 6 cores	0.857	0.772–0.912	95.000

Generally speaking, we have observed that most AI tests used in the laboratory do one of two things that make them meaningfully different from our own biomarker, and limit the generalizability of prior methods concerning laboratory implementation:

1. AI algorithms that quantify immunohistochemistry (IHC) or other specially colored stains. Examples include HalioDX (Now Veracyte) Immunoscore or Cernostics (now Castle Biosciences) TissueCypher algorithms.
2. AI algorithms that detect specific morphological features, a prime example of which is PaigeAI's Paige Prostate algorithm.

Both types of algorithms are designed to detect specific *a priori* established morphological features on a slide, which implies that AV of these other assays can be evaluated by examining the degree to which the assay detects the defined features of interest. For example, Immunoscore relies on IHC stains for CD3 and CD8,[5] making the sensitivity and specificity of these antibodies to the intended epitopes the critical components of the AV.

However, for an AI test that examines H&E, there is no targeted stain in use. H&E do have differential affinities for different structures within cells and tissues, making them useful stains, but they provide this color to cellular and tissue structures by binding to or complexing with a large number of different types of molecules with varying affinities, making each stain individually highly non-specific for any particular molecule, in contrast to IHC.[7,8] Therefore, studying analytical performance of the probes in a test, where the probes are H&E, is not helpful in establishing analytical performance of the test.

Alternatively, AI, such as Paige Prostate, is intended to examine H&E. However, this AI is intended to provide information about specific human-identifiable morphological or pathological features in the slide and provide a location. Therefore, studies regarding the ability of the AI to detect these established features provide information about analytical performance, but this has limited relevance to ArteraAI Prostate Test.

For example, the FDA Clearance Summary for Paige Prostate describes a localization study, which is important for AI that seeks to show pathologists where morphology concerning for cancer is located on the slide.[17] However, the ArteraAI Prostate Test is not intended to highlight pathological morphology on a glass slide, but rather to prognosticate outcomes and predict benefit from therapy based on the full specimen it is shown. Moreover, it does this without restricting itself to detection of *a priori* established features that can be used for investigation of analytical performance.

Notably, following the performance of the experiments in the current report (but before publication), AV of another H&E prognostic test was published,[18] but this prognostic

test is many ways more similar to Paige Prostate than to ArteraAI Prostate Test as this test uses AI to detect pathologist identified features, and algorithmically uses the pathologist identified features to prognosticate outcomes.

While we report on AV of a specific laboratory test, this study has broader implications as a first of its kind publication on AV of AI applied to H&E-stained tissue that makes inferences about the disease outcome rather than specific slide findings. While we have performed this AV on a test that uses H&E stained specimens, our AV methodology could generally be seen as applying to laboratory tests that use AI on non-specific stains more generally, or potentially even unstained tissue.

The studies comparing one versus multiple cores are interesting not only regarding the validation of this specific test, but also because of the implications that they have on prostate cancer biological heterogeneity. These results show that while the test performs reasonably well with one versus multiple cores, there is sensitivity to the selection of tissue used for testing.

However, this is a limitation that the AI test shares with gene expression testing, which appears quite sensitive to the choice of tumor focus for testing.[19] This is not surprising given that heterogeneity within tumors is a well-known phenomenon that is thought to contribute significantly to treatment resistance and may predispose patients to worse clinical outcomes.[20] In general, little has been done in clinical testing to date to consider intratumoral heterogeneity, which tends to require specialized tools such as spatial genomics, that are limited largely to the research setting. Image analysis AI may offer an additional mode of research to study intratumoral heterogeneity, and may 1 day offer a feasible method for clinically assessing intratumoral heterogeneity in patients.

Therefore, this is a limitation that seems to be caused by prostate cancer rather than a limitation unique to image analysis AI. However, from the standpoint of clinical use, AI may have an operational advantage over genomics since AI does require any additional tissue beyond a diagnostic H&E stained slide. When there is insufficient tissue for reliable genomic testing of the highest grade core, either genomic testing cannot be done, or it must be done on a different core, which may yield a different result from the primary core.

However, image analysis AI can be performed on the original diagnostic H&E stained slides, making testing of the highest grade core virtually always feasible unless the slide containing this core is damaged.

There are a number of important limitations of this study. Perhaps the most important limitation concerns scanner generalizability. While the study shows that the AI model generalized well from the research setting using a Leica AT2 to the clinical setting using a 3DHistech P1000, this should not be understood to imply that either the AI models studied here or AI models more generally are scanner agnostic. In exploratory work, we have found substantial differences between scanners and deliberately designed models to be insensitive to scanner differences.

The detailed computational approach to this will be the topic of a future manuscript. In general, our approach corrected for color differences, which are due to the ways different scanners capture red, green, and blue channel data. A description of this work is outside of the scope of this manuscript. At this point, we believe that validation should be performed on a specific scanner model before the AI is trusted to work on that model of scanner. In our case, we used a concordance study between scanner models to establish generalizability.

An additional limitation is the use of prostate core count as the unit of measurement in quantifying the amount of tissue used. This measure makes sense in the case of prostate biopsies for which there are cores, but this approach would not generalize well to pathology cases not based on cores, such as surgical excisions.

Conclusion

We have developed an approach for AV of a prognostic algorithm and an ST-ADT response algorithm applied to H&E stained tissue specimens and applied this approach to the AV of novel prostate AI LDT. The analytical performance of the test as assessed by these techniques supports adequacy for clinical use.

Acknowledgments

The authors would like to acknowledge Wayne Lai for graphic design of included figures.

Authors' Contributions

P.G.: Conceptualization, methodology, and writing—original draft. J.Z. and H.-C.H.: Formal analysis, writing—original draft. R.Y.: Software. S.Nag.: Methodology, investigation. S.Nhek. and A.C.: Writing—review and editing. J.K.: Supervision. N.S.: Supervision, project administration. T.J.R.: Writing—original draft, supervision. T.S.: Conceptualization, supervision, project administration, and writing—review and editing.

Data Availability

The data that support the findings of this study have been originated by Artera Inc. Requests for data sharing by license or by permission for the specific purpose of replicating results in this manuscript can be submitted to support@artera.ai.

Author Disclosure Statement

P.G., J.Z., R.Y., H.-C.H., S.Nag., S.Nhek., N.S., T.J.R., and T.S. are current employees of Artera Inc., and have a personal financial interest in Artera Inc. J.K. receives compensation from Artera. A.C. receives compensation from PathNet.

Funding Information

This study was funded by Artera Inc.

Supplementary Material

Supplementary Data

References

1. Esteva A, Feng J, van der Wal D, et al. Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. NPJ Digit Med 2022;5(1):71.
2. Spratt DE, Tang S, Sun Y, et al. Artificial intelligence predictive model for hormone therapy use in prostate cancer. NEJMEvid 2023;2(8); doi: 10.1056/EVIDoa2300023
3. Ross AE, Zhang J, Huang HC, et al. External validation of a digital pathology-based multimodal artificial intelligence architecture in the NRG/RTOG 9902 Phase 3 Trial. Eur Urol Oncol 2024:S2588-9311(24)00029-4; doi: 10.1016/j.euo.2024.01.004
4. Goldsmith JD, Fitzgibbons PL, Swanson PE. Principles of analytic validation of clinical immunohistochemistry assays. Adv Anat Pathol 2015;22(6):384–387.
5. Marliot F, Lafontaine L, Galon J. Immunoscore assay for the immune classification of solid tumors: Technical aspects, improvements and clinical perspectives. Methods Enzymol 2020;636:109–128.
6. Sooriakumaran P, Lovell DP, Henderson A, et al. Gleason scoring varies among pathologists and this affects clinical risk in patients with prostate cancer. Clin Oncol R Coll Radiol G B 2005;17(8):655–658.
7. Allsbrook WC, Mangold KA, Johnson MH, et al. Inter observer reproducibility of Gleason grading of prostatic carcinoma: Urologic pathologists. Hum Pathol 2001;32(1):74–80.
8. Ozkan TA, Eruyar AT, Cebeci OO, et al. Interobserver variability in Gleason histological grading of prostate cancer. Scand J Urol 2016;50(6):420–424.
9. Evans AJ, Brown RW, Bui MM, et al. Validating whole slide imaging systems for diagnostic purposes in pathology. Arch Pathol Lab Med 2022;146(4):440–450.
10. Food and Drug Administration. Gene Expression Profiling Test System for Breast Cancer Prognosis-Class II Special Controls Guidance for Industry and FDA Staff. FDA GuidDoc 2021. Available from: https://www.fda.gov/medical-devices/guidance-documents-medical-devices-and-radiation-emitting-products/gene-expression-profiling-test-system-breast-cancer-prognosis-class-ii-special-controls-guidance [Last accessed: March 28, 2024].
11. Food and Drug Administration. Substantial Equivalence Decision Summary. Report No.:510(k) Number: k062694. Available from: https://www.accessdata.fda.gov/cdrh_docs/reviews/k062694.pdf [Last accessed: April 9, 2024].
12. Food and Drug Administration. Substantial Equivalence Decision Summary. Report No.:510(k) Number: K130010. Available from: https://www.accessdata.fda.gov/cdrh_docs/reviews/K130010.pdf [Last accessed: April 9, 2024].
13. Buschon, Larry. VALID Act of 2023. H.R.2369 April 7, 2023. Available from: https://www.congress.gov/bill/118th-congress/house-bill/2369/text [Last accessed: April 9, 2024].
14. Federal Register. Medical Devices; Laboratory Developed Tests. 2023. Available from: https://www.federalregister.gov/documents/2023/10/03/2023-21662/medical-devices-laboratory-developed-tests [Last accessed: February 6, 2024].
15. Smith C. On the estimation of the intraclass correlation. Ann Hum Genet 1956;21:363–373.
16. Colling R, Pitman H, Oien K, et al. Artificial intelligence in digital pathology: A roadmap to routine use in clinical practice. J Pathol 2019;249(2):143–150.
17. Food and Drug Administration. Evaluation of Automatic Class III Designation for Paige Prostate Decision Summary. Report No.: DEN200080. Available from: https://accessdata.fda.gov/cdrh_docs/reviews/DEN200080.pdf [Last accessed: February 16, 2024].
18. Fernandez G, Zeineh J, Prastawa M, et al. Analytical validation of the PreciseDx digital prognostic breast cancer test in early-stage breast cancer. Clin Breast Cancer 2024;24(2):93.e6–102.e6.
19. Wei L, Wang J, Lampert E, et al. Intratumoral and intertumoral genomic heterogeneity of multifocal localized prostate cancer impacts molecular classifications and genomic prognosticators. Eur Urol 2017;71(2):183–192.
20. Proietto M, Crippa M, Damiani C, et al. Tumor heterogeneity: Preclinical models, emerging technologies, and future applications. Front Oncol 2023;13:1164535.