Prostate Cancer and Prostatic Diseases (2018) 21:594–603
https://doi.org/10.1038/s41391-018-0067-4

ARTICLE

Development and validation of a novel automated Gleason grade and molecular profile that define a highly predictive prostate cancer progression algorithm-based test

Michael J. Donovan[1] · Gerardo Fernandez[2] · Richard Scott[2] · Faisal M. Khan[2] · Jack Zeineh[2] · Giovanni Koll[2] · Nataliya Gladoun[2] · Elizabeth Charytonowicz[2] · Ash Tewari[3] · Carlos Cordon-Cardo[1]

Received: 26 January 2018 / Revised: 1 May 2018 / Accepted: 16 May 2018 / Published online: 7 August 2018

Abstract

Background Postoperative risk assessment remains an important variable in the effective treatment of prostate cancer. There is an unmet clinical need for a test with the potential to enhance the Gleason grading system with novel features that more accurately reflect a personalized prediction of clinical failure.

Methods A prospectively designed retrospective study utilizing 892 patients, post radical prostatectomy, followed for a median of 8 years. In training, using digital image analysis to combine microscopic pattern analysis/machine learning with biomarkers, we evaluated Precise Post-op model results to predict clinical failure in 446 patients. The derived prognostic score was validated in 446 patients. Eligible subjects required complete clinical-pathologic variables and were excluded if they had received neoadjuvant treatment including androgen deprivation, radiation or chemotherapy prior to surgery. No patients were enrolled with metastatic disease prior to surgery. Evaluate the assay using time to event concordance index (C-index), Kaplan–Meier, and hazards ratio.

Results In the training cohort ($n = 306$), the Precise Post-op test predicted significant clinical failure with a C-index of 0.82, [95% CI: 0.76–0.86], HR:6.7, [95% CI: 3.59–12.45], $p < 0.00001$. Results were confirmed in validation ($n = 284$) with a C-index 0.77 [95% CI: 0.72–0.81], HR = 5.4, [95% CI: 2.74–10.52], $p < 0.00001$. By comparison, a clinical feature base model had a C-index of 0.70 with a HR = 3.7. The Post-Op test also re-classified 58% of CAPRA-S intermediate risk patients as low risk for clinical failure.

Conclusions Precise Post-op tissue-based test discriminates low from intermediate high risk prostate cancer disease progression in the postoperative setting. Guided by machine learning, the test enhances traditional Gleason grading with novel features that accurately reflect the biology of personalized risk assignment.

Introduction

In the United States, prostate cancer (PCa) remains the most commonly diagnosed solid tumor among men with an estimated 161,360 new cases and 27,000 deaths in 2017 [1]. Noteworthy, the world-wide burden is expected to increase to 1.7 million new cases and 499,000 deaths by 2030 [2]. Surgery remains a selected treatment option with overall good prognosis; however, a subset of ~25–30% of men will have a biochemical (PSA) recurrence, with 4–25% developing metastatic disease within 15 years [3, 4]. Accurate risk stratification in the post-surgical setting is essential to identify patients at high risk of PCa-specific mortality who would potentially benefit from aggressive multimodal therapy from the majority who are cured by surgery alone [5]. Although surgical pathology clinical features such as pre-op PSA ≥ 20

These authors contributed equally: Michael J. Donovan, Gerardo Fernandez

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41391-018-0067-4) contains supplementary material, which is available to authorized users.

✉ Michael J. Donovan
Michael.Donovan@mssm.edu

[1] Department of Pathology, Icahn School of Medicine at Mt. Sinai, 1468 Madison Avenue, New York City, NY 10029, USA

[2] Department of Pathology, Icahn School of Medicine at Mt. Sinai, 1111 Amsterdam Ave, New York City, NY 10025, USA

[3] Department of Urology, Sinai Hospital, 1470 Madison Avenue, New York City, NY 10029, USA

ng/mL, Gleason pattern ≥ 4, extracapsular extension, and seminal vesicle invasion all represent potential triggers for high-risk disease, clinicians are reluctant to recommend secondary treatment due to morbidity, cost, and questions on clinical efficacy [6, 7]. Additional approaches are clearly necessary to assign objective risk in the post-surgical setting.

Major advances have been made in cataloging the genomic alterations in PCA and understanding the molecular mechanisms underlying clinically significant disease; however, there has been little development on the phenotypic characterization of the intact tissue specimen. Currently, surgical pathology features, most notably the Gleason grade and stage, are the primary means to identify men at highest risk of prostate cancer specific mortality. Nevertheless, despite the ability to stratify clinical significant risk, there remains considerable variability in outcomes when applied to individual men with PCa.

A significant element of this unpredictability resides in Gleason grading, specifically the presence and volume of Gleason pattern 4 disease. Tumor grading was reported using Grade Groups endorsed by the 2014 International Society of Urological Pathology Consensus Conference, whereby GG1 (Grade Group 1) = Gleason score ≤ 6, GG2 = Gleason score 3 + 4 = 7, GG3 = Gleason score 4 + 3 = 7, GG4 = Gleason score 8, and GG5 = Gleason score 9–10 [8]. Several molecular assays, notably *Decipher* (Genome Dx, Vancouver, Canada) and *Prolaris* (Myriad Genetics, Salt Lake City, UT) have provided a more tangible assessment of tumor biology and although promising, extended validation studies have demonstrated that clinical calculators such as "Cancer of the Prostate Risk Assessment Post-Surgical Score (CAPRA-S)" remain important tools for predicting risk [9].

We have previously reported on the significance of precision (systems) pathology and the role of quantitative tissue-based biomarkers in prostate cancer outcome models [10]. In the current study we confirm these original observations with the development and validation of a post-surgical risk model for prostate cancer. By applying novel artificial intelligence image analysis feature selection tools, such as machine vision and machine learning, as well as quantitative biomarker multiplexing centering on proliferation and androgen signaling, we aimed to enhance traditional Gleason grading in patient-focused prognostic modeling. Thereby providing risk assignment that is unbiased, broadly applicable, and independent of interpretive histology.

Methods

Study design and participants

This prospectively designed retrospective study included patients post-radical prostatectomy from 1993 to 2005 at the Henry Ford Hospital ($n = 373$; HFH, Detroit, MI) and the Roswell Park Cancer Center ($n = 519$; RPCC, Roswell Park, NY). The Institutional Review Boards at the participating institutions approved the use of human patient specimens and their clinical data for this study. Patients were selected with complete clinicopathologic variables and excluded if they had received any neoadjuvant treatment including androgen deprivation, radiation or chemotherapy prior to surgery. No patients were enrolled with metastatic disease prior to surgery. All samples were obtained as deidentified.

We constructed a pooled sample set from both HFH and RPCC to produce 50:50 training and validation cohorts which were further sub-divided into four individual groups. Absolute PSA values were utilized; surgical margin status, extracapsular extension, seminal vesicle invasion and lymph node status were treated as binary variables and defined by each institution. We did not have information on prostate size, tumor volume and length of a positive surgical margin. By report, patient follow-up after RP was not standardized between the two institutions but differences were minor. The average time for radiotherapy was 42D for Roswell and 52D for Henry Ford with ADT dependent upon clinicians at each institution, e.g., average at Henry Ford was 751D. Two "time to event" outcomes were evaluated: (1) clinical failure (CF) endpoint, defined as the first PSA rise post-adjuvant therapy of a sustained series of elevations above nadir (>0.1 ng/mL) or a single rise of ≥0.5 ng/mL, metastatic disease or death from prostate cancer; and (2) post-surgical PSA recurrence (PSAR) after nadir of either a single 0.4 ng/mL or two consecutive 0.2 ng/mL PSA values. Censored patients who received secondary therapy were considered to have had a PSAR [10]. Metastasis end point was defined as bone, viscera, or lymph node metastasis documented radiographically by computed tomography or bone scan. Forty-six patients did not reach PSA nadir post surgery and were removed from the PSAR cohort.

Procedures

Tissue microarrays containing three to eight 0.6 mm cores per patient representing the highest clinical Gleason grade and score were prepared from formalin-fixed paraffin-embedded (FFPE) tissue specimens. For RPCC, patient's cores were distributed over three arrays; while HFF patients were grouped together. All personnel involved in data collection were blinded to patient information including demographics and event status. TMA tissue sections from RPCC and HFF were all stained over a 10-week period by two operators in a single laboratory in the Department of Pathology at the Icahn School of Medicine with a multiplex immunofluorescence (MIF) assay [11–13] of five biomarkers: Androgen receptor (AR), Ki67, Cytokeratin 18 (CK18), Cytokeratin 5/6 (CK5/6), and Alpha-methylacyl-CoA Racemase (AMACR). The MIF assay (see

Supplemental Materials, S1, eTable S1 and S2) was performed on a Leica Bond RX automated stainer.

Image feature construction

Digital image analysis

We leveraged novel image analysis and machine learning software to generate a comprehensive suite of morphometric and biologic features that represent the complete Gleason grading system and known drivers of prostate cancer progression including proliferation and androgen signaling (see Supplemental methods for a complete description of the feature generation methodology). In brief, the analytic process begins with machine vision segmentation of MIF images to create distinct glandular and stromal histologic compartments (see Supplemental Materials, S3). All images were screened and edited for tumor content and staining artifacts by genitourinary (GU) pathologists (GF, MJD).The algorithm pipeline then incorporates a graph theory-based approach for characterization of fusion (i.e., cribriform patterns) and fragmentation of tumor architecture (e.g., ring structure adjacency features) and biomarker (i.e., AR and Ki67) quantitation. The pipeline generates combination features that quantify relative biomarker levels in stroma vs. glandular epithelium, producing discrete biomarker values per gland ring [i.e., relative rise (dynamic range) AR and Ki67 positive ratio features]. In the final phase, Ki67 levels are differentially combined with the ring and AR features to discriminate low vs. higher Gleason grades. This sequential, multi-layered approach interrogates distinct histologic compartments and biological capacity to mathematically characterize prostate cancer progression [14, 15].

Statistical analysis

The primary endpoint of the study was to validate the ability of the Precise Post-op test to accurately predict post-surgical CF. Exploratory analyses were performed as secondary endpoints including PSAR, and test performance in National Comprehensive Cancer Network Guidelines (2017) NCCN intermediate and high risk patients. All study participants were blinded to outcome and clinical data before data generation and analysis. Cohort-specific clinical base models and the published post-surgical risk assessment tool, CAPRA-S, which utilizes a point structure associated with six clinicopathologic variables: pre-operative PSA, prostatectomy Gleason Score, surgical margin status, extra-prostatic extension, seminal vesicle invasion and lymph node invasion were employed to assess added significance of the Precise Post-op test [9]. Noteworthy, the CAPRA-S score is based on biochemical recurrence and not clinical significance as defined in the current study.

Image derived features were correlated with CF and PSAR using the Concordance index (C-index) which is similar to the ROC AUC. Features with a C-index of <0.5 or >0.5 reflect an increased or reduced risk of CF, respectively. Association of the Precise Post-op assay with individual clinical-pathologic variables and outcome was assessed using support vector regression multivariate logistic regression models for censored data (SVRc) [16]. Individual feature weights in the final model are covariates within the SVRc algorithm employing a linear kernel. Analysis of individual patient data used C-index univariate and multivariate analysis proportional hazards models along with cumulative Kaplan–Meier incidence curves to evaluate the association of Precise Post-op with time to CF or PSAR. Analyses were conducted in Matlab using standard packages. All C-indices and hazard ratios (HR) are reported as 95% CIs. Significance was set as a two-tailed p value of <0.05.

The Precise Post-op model score is on a continuous scale of 0–100, with an identified threshold based on an optimized sensitivity and specificity for discriminating low vs. high risk for either CF or PSAR. For the CF model, a score above the identified threshold predicts high risk of CF at 96 months (8 years) post surgery. For the PSAR model, a score above the threshold predicts high risk of PSAR at 60 months (5 years) post surgery. The generation of the score and its derivation have been previously published (see Donovan et al. [16]).

Results

Patient characteristics

Combined patient groups were generated from both HFH ($n = 373$) and RPCC ($n = 519$) and divided 50:50 into balanced training and validation cohorts of 446 patients, further subdivided for model development. (see Supplemental Materials, S4, and eTable S3 for complete demographics). Although a breakdown of pathology, pT2 stage was provided only a single pT2 was employed in modeling. Digitized prostatectomy TMA cores were evaluated for histology and grid integrity. A total of 22% of patients ($n = 255$) were removed for minimal tumor content (<3 tumor glands) and/or poor biomarker staining; and 7% ($n = 47$) were removed for missing clinical data; yielding a total of 590 patients for train/test development and validation.

Automated Gleason grading integrated with a molecular phenotype

Morphometric and biomarker feature development was restricted to training sub-groups with step-wise performance

Table 1 Demographics and clinical characteristics for patients in both training and validation cohorts

CF models	Training		Validation	
	Number	%	Number	%
Total	306	100	284	100
Race				
Caucasian	243	79	222	78
African American	60	20	61	21
Native American	2	1	1	0
Other	1	0	0	0
Average age, yrs	61		60	
Pathologic stage				
T2	247	81	212	75
T3	55	18	71	25
T4	4	1	1	0
Average Pre-Op PSA (ng/mL)	7		7	
RP-dominant Gleason grade				
3	232	76	216	76
4	69	23	60	21
5	5	2	8	3
RP-secondary Gleason grade				
3	149	49	143	50
4	141	46	126	44
5	16	5	15	5
RP Gleason score				
Group grade 1	103	33	110	39
Group grade 2	122	40	102	36
Group grade 3	45	15	32	11
Group grade 4	24	8	23	8
Group grade 5	12	4	17	6
Lymph node invasion				
No	296	97	282	99
Yes	10	3	2	1
Positive surgical margins				
No	224	73	196	69
Yes	82	27	88	31
ECE				
No	188	61	160	56
Yes	118	39	124	44
SVI				
No	280	92	266	94
Yes	26	8	18	6
Adjuvant therapy				
No	235	77	197	69%
Yes	71	23	87	31
PSA rise post adjuvant Tx				
No	262	86	245	86
Yes	44	14	39	14

using the C-index for CF event outcome as a measure of risk accuracy. Features were prioritized on C-index <0.5 or >0.5; i.e., for features <0.5, a higher/greater the feature value (i.e., Gleason grade 4, 5) equates to a shortened time to event/clinical failure. From an original set of 10,000 image features, we applied an outcome-driven, C-index CF feature selection process to identify 19 features representing quantitative attributes of the Gleason grade, proliferation and androgen signaling. In addition to performance, each feature was evaluated for cohort-site stability. Noteworthy, the number of Ki67 positive nuclei was a rare event with "0" in 70% of cases, further emphasizing the necessity for intact tissue assessment coupled with both AR and Ki67 in discrete gland assignment. The curated feature set was combined with 10 clinical variables, including age, pre-op PSA, Gleason dominant grade, Gleason score, Grade Group categorization, pathology stage, positive margins, positive seminal vesicle invasion, and positive lymph node, for training model generation. As previously reported, we confirmed the association of a positive lymph node with clinical failure [10, 11] and, although important, the feature remains a generally rare event in clinical practice, including the current cohort (12 of 590 patients, 2%). To maximize potential for earlier medical intervention while enriching the model feature selection process, we excluded the lymph node feature but retained patients with this characteristic (see Supplemental Materials, eTable S4 for univariate feature performance). Of note, all Capra-S models included lymph node status.

Primary objective

Precise Post-op CF training

A final training cohort of 306 patients, median follow-up 7.6 years and a 15% CF event rate was used for SVR model development, Table 1. Twenty-four percent of patients ($n = 105$) had received adjuvant therapy (ADT or RT) as per standard regimens at each institution. Importantly 94% events were a PSA rise post adjuvant therapy, 36% metastases, and 23% subjects exhibited both endpoints. From the original 29 (clinical + imaging) features, 4 compound imaging features, and 1 clinical variable were selected (Table 2, greater values equate to priority in the model) with a performance C-index of 0.82 [95% CI: 0.76–0.86], Hazards Ratio of HR: 6.7 [95% CI: 3.59–12.45], $p < 0.00001$ (Table 3). When patients were stratified by model score below vs. above the optimized cut-point of 37 (range 0–100), corresponding to a 27% model predicted probability of CF, the HR was 6.7, sensitivity 72% and specificity 78% for correctly predicting CF within 8 years. The HR is measuring high vs. low risk according to the cut-point of 37. Increasing Precise Post-op scores

Table 1 (continued)

CF models	Training		Validation	
	Number	%	Number	%
Metastasis				
No	289	94	272	96
Yes	17	6	12	4
Deceased				
No	258	84	243	86
Yes	48	16	41	14
CF status				
Censored	259	85	242	85
Event	47	15	42	15
Median target time	7 years 7 months		7 years 11 months	

Table 2 MPIF and clinical features selected in the CF model with respective weights

Feature	Weight in final model
1. Pathology stage	−38.2201878297942
2. Ring + AR relative rise,weighted (RAR1)	−19.7034000943070
3. Ring x AR Relative Rise, unweighted (RAR2)	−16.1010748269531
4. RAR1 + RARK1 weighted (RARK2)	−11.1768929453938
5. RAR1 x Ki67, unweighted (RARK1)	−10.1984254719578

Higher negative feature value equates to a shorter time to disease progression

Table 3 Clinical failure model training and test results

CF training CI [95% CI]	0.8200 [0.76–0.87]
Training Cut-Point	37.1344
Train sensitivity	0.7179 [0.58-0.86]
Train specificity	0.7793 [0.71-0.85]
Train hazard ratio	6.6837 [3.59–12.45] $p < 0.00001$
CF validation CI [95% CI]	0.7702 [0.70–0.83]
Test sensitivity	0.7879 [0.64–0.93]
Test specificity	0.7021 [0.63–0.78]
Test hazard ratio	5.3684 [2.74–10.52] $p < 0.00001$

indicate greater risk of PSA rise post adjuvant therapy and metastasis. (Fig. 1a). Noteworthy, 15 of the 19 patients (79%) in the training cohort with either lymph node positive disease or metastasis were correctly classified as high risk by the model. The probability of risk for CF with increasing Precise Post-op scores as a continuous distribution is illustrated in Supplementary Materials, Figure S1A and B. In addition to the clinical + IF feature models we also employed base clinical multivariate regression models and the on-line risk calculator, CAPRA-S, to further assess the significance of IF features for predicting outcome.

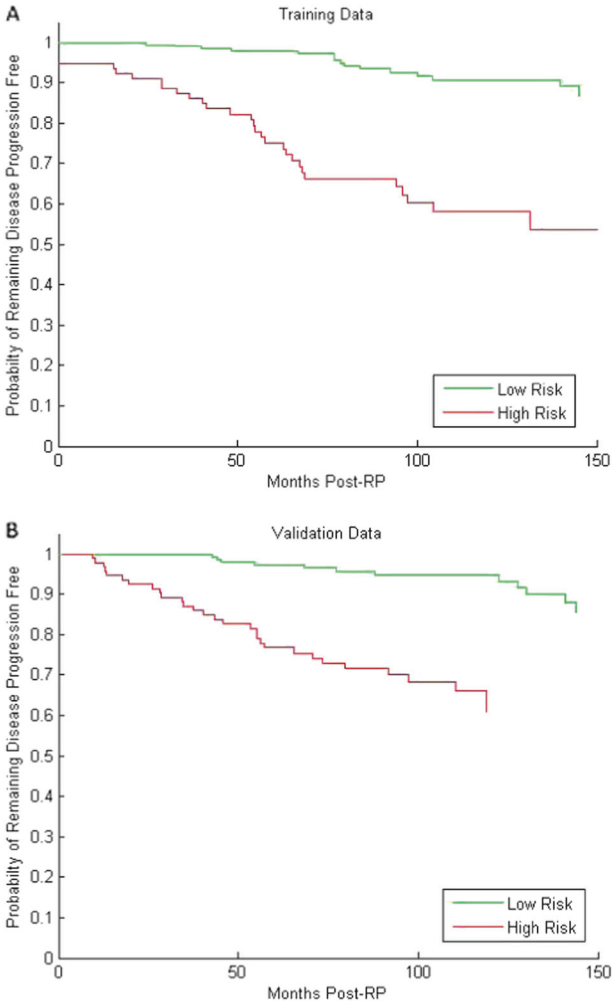

Fig. 1 **a** Kaplan–Meier curves show freedom from CF in training (**a**) and validation (**b**) with the Precise Post-op model. Training: HR 6.7, [95% CI, 3.59–12.45], $p < 0.00001$; validation: HR 5.36 [95% CI, 2.74–10.52], $p < 0.00001$

A modified clinical, non-CAPRA-S model (i.e., without lymph node status) achieved a C-index of 0.78, HR of 5.8 (two features selected: pathology stage and pre-op PSA), while a combined Precise Post-op model that included CAPRA-S had a C-index of 0.85. Importantly, neither the dominant Gleason grade, nor Grade Group (GG) (univariate C-index of 0.31, respectively) were selected by the Precise Post-op model. Furthermore, the three most significant compound ring plus AR or Ki67 features in the model all had highly significant univariate CF C-indices of 0.24 ($p < 0.0001$).

Precise Post-op CF validation

The model was validated in 284 patients (Table 1) and produced a C-index of 0.77. Twenty-nine percent of patients ($n = 128$) had received adjuvant therapy (ADT or

RT) as per standard regimens at each institution. As in training, the majority of the events (93%) were PSA rise post adjuvant therapy with 29% metastases and 18% of subjects experiencing both endpoints. When patients were stratified by model score below (low risk) vs. above (high risk) a cut-point of 37 (scale 0–100), the HR was 5.4 [95% CI (2.74, 10.52), $p < 0.00001$], sensitivity 79% and specificity of 70% (Fig. 1b and Table 3). By comparison, a clinical (non-CAPRA-S) base model had a C-index of 0.70, and HR of 3.7. The probability of risk for CF with increasing Precise Post-op scores as a continuous distribution is illustrated in Supplementary Materials, Figures S1A, B.

Of the five features, pathologic stage was most important followed by four imaging features which characterize Gleason in the context of Ki67 and AR levels (Table 2). These four imaging features enhanced the clinical Gleason grading through morphometry, Ki67 quantitation and differences between AR levels in stromal vs. epithelial nuclei (Fig. 2a–i). Of note, the top two compound morphometry and AR imaging features selected by the model were independently correlated with the dominant Gleason grade (correlation coefficient range, $r = 0.28$–0.31) and Gleason score (range, $r = 0.31$–0.33), but not with the secondary Gleason grade (range, $r = 0.18$–0.19).

To further assess independent value of the Precise Post-op test over CAPRA-S and individual clinical/pathological values we constructed two multi-variable time-dependent cox models. Each had the Precise Post-op score as either a continuous or categorical variable. In the first model with a continuous score, the Precise Post-op had a p-value of 0.0057. In the second model with a categorical score, the Precise Post-op had a p-value of 0.0414. Both results suggest independent value of the Precise Post-op test; however, improved significance was observed when the score is assessed as a continuous variable.

Secondary objectives

PSA recurrence training and validation models

A training cohort of 285 patients, median follow-up time of 8 years with a 20% PSAR rate, identified two clinical features, pathology stage and positive margins, and one imaging feature, reflective of Gleason grading combined with the dynamic range of AR levels between stroma and epithelial cells. The difference in cohort size between CF and PSAR was due to the removal of 46 patients that did not nadir post surgery. As in the CF model, the one imaging feature was positively correlated with both the dominant Gleason grade (0.25) and Gleason sum (0.24). The Precise Post-op PSAR training model produced a C-index of 0.79 (95% CI: 0.74–0.83), HR = 7 (95% CI: 3.85–12.71), $p < 0.00001$ (see Supplement 1, S8 and S9). Of this group of patients, 81% (230 out of 285) had not received any adjuvant therapy, of which there were nine PSAR events. The Precise Post-op model successfully classified five of these patients as high risk for PSAR. The probability of risk for PSAR with increasing Precise Post-op scores as a continuous distribution is illustrated in Supplementary Materials, Figures S1C and D.

In a 259 patient validation cohort, the Precise Post-op PSAR test yielded a C-index of 0.71 (95% CI: 0.68–0.75), HR = 3.8 (95% CI: 2.3–6.12), $p < 0.00001$, sensitivity 78% and specificity of 59% (see Supplemental Materials 1, S6 and S7); including CAPRA-S improved both the specificity (59% vs. 69%) and HR (3.8 vs. 4.2). By comparison, the clinical only model had a C-index of 0.70 with a HR 3.4, sensitivity 57%, and specificity 75%.

Subpopulation analyses

To explore subpopulations, we combined intermediate and high-risk patients based on the following characteristics: pT2 with positive margins, pT3, pre-op PSA > 20 ng/mL, lymph node invasion, SVI, GS 4 + 3 = 7 (GG3), PSAR and GS≥8 (≥GG4), from training and validation groups. A total of 345 from 590 evaluable patients (58%) were identified with a 24% CF event rate. The Precise Post-op model produced a C-index of 0.73 compared to a CAPRA-S of 0.70. Of note, combining the Precise Post-op model with CAPRA-S, further improved the AUC to 0.83. When patients were stratified by model score below vs. above a cut-point of 37, the HR was 3.5 [95% CI (2.20–5.46), $p < 0.00001$], sensitivity 75% and specificity of 63% (Fig. 3). Additionally, in the remaining 245 low risk patients there were no CF events, sensitivity of 0% and specificity 88.5%, demonstrating the good performance of the Precise Post-op assay across multiple risk groups.

We also compared the distribution of CAPRA-S scores according to their low (<3), intermediate [3–5], and high (≥6) risk categorization in the validation cohorts. Analysis of the data revealed that the Precise Post-op CF model re-classified 58% of the intermediate CAPRA-S 3-to-5 category as low-risk for CF, and 42% as high risk. The opposite was observed with the PSAR model in that 73% of CAPRA-S intermediate risk patients were re-classified as high risk and 27% as low risk. The results support the necessity of providing additional discrimination for intermediate risk patients.

Discussion

Postoperative risk assessment remains an important variable in the treatment of prostate cancer. We aimed to validate a

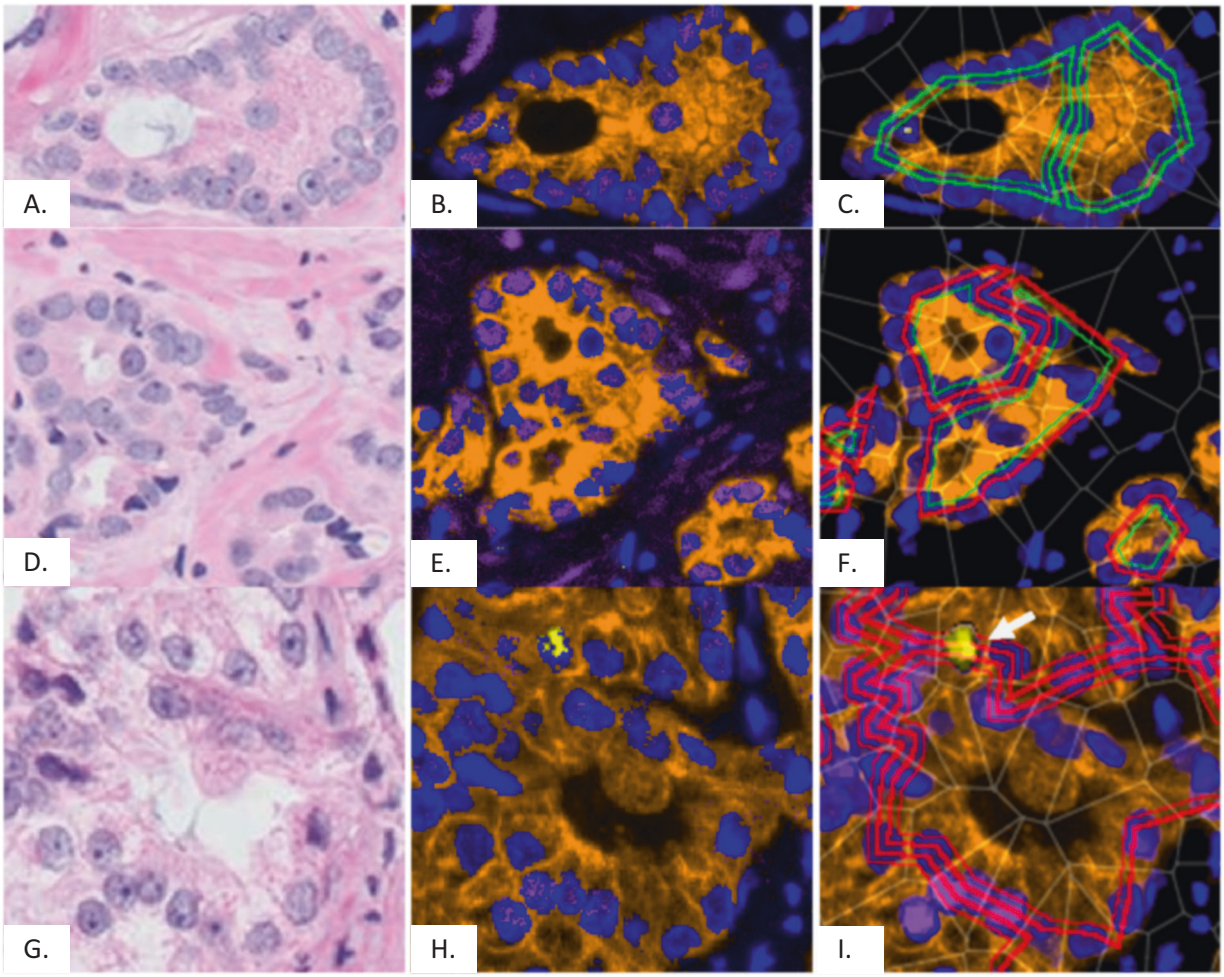

Fig. 2 **a**–**i** Ring feature visualization. Row 1, **a**–**c** Gleason pattern 3, low AR, risk score 17.6. Row 2, **d**–**f** pattern 4, low AR, risk score 15.0. Row 3, **g**–**i** pattern 4, high AR, high Ki67, risk score 85.8. **a**, **d**, **e**: **h**, **e** images for reference. **b**, **e**, **h** Input IF images—CK18 gold, DAPI blue, AR purple, Ki67 yellow, showing max strength biomarker per pixel. **c**, **f**, **i** Resultant combination features—outer rings (double width): gland morphology (stroma-.lumen adjacency loss feature, SLP). High/low-grade rings shown in red/green, respectively. Inner rings: AR gland/stromal contrast. Nuclear AR is normalized by stromal contrast (75th–25th percentile): high/low-grade nuclei are shown on a purple to blue color scale. Non-nuclear AR is masked. Note: normalization reduces gland AR signal (blue) in (**c**) and **f** with presence of strong stromal AR; increased in **i** (purple). Ki67-positive nuclei, yellow (**i**, arrow)

tissue-based on-slide risk model (Precise Post-op) to identify which patients would benefit most from early intervention. Here, we report on a test which enhances clinical Gleason grading to predict post-surgical outcomes. The Precise Post-op test was trained and independently validated on two well-characterized patient populations to provide postoperative prognostic outcome risk.

More specifically, this prospectively designed retrospective study used advanced image analysis tools to develop and validate prognostic post-surgical outcome models. Accurately understanding prostate cancer recurrence risk after initial therapy is necessary for more effective patient management, defining therapeutic recommendations, and subsequent future trial enrollment [17]. The dilemma facing many urologists is that the clinical course for the majority of men post-surgery is quite favorable, and even men with high risk features by National Comprehensive Cancer Network (NCCN 2017) guidelines, have an estimated median time to metastatic disease of 5–8 years [18].

Unfortunately, the subset of men who will eventually have clinical progression continues to remain poorly understood. A recent approach to address this issue has been the introduction of RNA signature assays, notably *Decipher* (Genome Dx Biosciences, Vancouver, Canada), and *Prolaris* (Myriad Genetics, Salt Lake City, UT). Both tests, although with non-overlapping gene sets and different intended use patient characteristics, provide risk scores associated with early metastasis and prostate cancer-specific mortality. These genomic assays produce improved

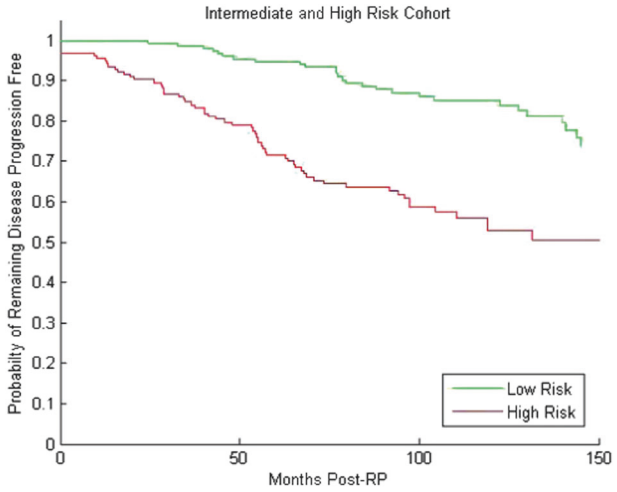

Fig. 3 Kaplan–Meier of intermediate and high risk patients stratified by Precise Post-op model. HR 3.5, p-value < 0.00001

accuracy both independently and when combined with clinical algorithms such as CAPRA-S, shifting the C-indices overall by 0.1 [17–19].

Clinical features have always been important risk predictors in prostate cancer risk calculators; however, the only feature that is both highly significant yet subjective is the Gleason grade/score. This represents a limitation when applying population risk nomograms on individual patients, but an advantage for applications such as the Precise Post-op test which utilizes the patient's intact tissue specimen to predict outcome. The validated Precise Post-op CF model, independent of the clinical Gleason, produced a C-index of 0.77 vs. a clinical only model C-index of 0.70, demonstrating significant patient-specific value for risk assignment. Furthermore, when the Precise platform was applied to the intermediate and high risk group, identified as the most to potentially benefit from more effective risk stratification, the Precise model had a C-index of 0.73 (vs. a C-index of 0.70 for CAPRA-S) including a HR of 3.5, and a significant p-value ($p < 0.0001$). From a clinical use perspective, patients with scores above the validated cut-point of 37 would be classified as higher risk for CF. Given the observed model prediction of a poor durable response to adjuvant androgen therapy in the high risk group, adjustments in patient management post-surgery including frequency of PSA testing and or consideration of selected adjuvant treatment may be warranted. It is important to point out that 79% of men with either lymph node positive disease or evidence of metastasis were correctly classified as high risk by the Precise model.

We have previously reported on a post-operative clinical and image-analysis biomarker-based outcome model that was more accurate for determining risk of progression post surgery than other standard risk group models or a 10-year postoperative nomogram [12, 19, 20]. Some of the limitations for this previous assay included reliance on the H&E-stained section and inclusion of the clinical Gleason. Notably, the current validated Precise Post-op model no longer requires the H&E image and more importantly contains quantitative features which serve to automate the Gleason grading process. Moreover, displaying a HR of 5 and an AUC of 0.77, the model appears to be superior to currently available tests to define risk, including genome-based assays. Finally, in a sub-cohort of intermediate and high-risk patients the Precise Post-op test when combined with the CAPRA-S algorithm resulted in a further improved C-index of 0.83. Of importance, the CAPRA-S was chosen as the clinical standard for predicting post-prostatectomy PSA recurrence and prostate cancer specific death [21].

Of note, the experimental design demonstrated that from the available six clinical variables for multivariate selection, only pathologic tumor stage was selected for the CF model while both stage and positive margin status were selected for the PSAR model. The absence of a clinical Gleason in the model would potentially address current practice requirements for diagnostic uniformity while retaining the innate heterogeneity within the intact tissue specimen. The quantitative and standardized assessment of cell-specific biomarkers, such as AR and Ki67, utilizing quantitative multiplex immunofluorescence (MIF), image analysis and morphometry, have proven effective in stratifying patients and guiding treatment decisions, including enrollment in AS programs, brachytherapy ± androgen ablation, salvage radiotherapy and surgical approach (e.g., incorporation of a lymph node dissection and extent of surgical margin) [22]. As identified, the restriction of AR levels to PSAR and the combined AR-Ki67 in CF, reflect biological principles underpinning the predicted endpoint.

A series of recent publications have continued to demonstrate the importance of elevated proliferative indices as defined by a positive Ki67 phenotype as it relates to overall survival [23]. Recently, high Ki67 expression was shown to be predictive of reduced survival and increased risk of metastasis, independent of PSA, Gleason score and D'Amico risk category [24]. Finally, the prognostic value of Ki67 was further assessed in a multi-institutional study utilizing over 1000 prostatectomy specimens and found that a high proliferative index of Ki67 was associated with worse recurrence free (HR = 1.47, $p = 0.02$–0.0008) and worse survival (HR = 2.03, $p = 0.03$) [25].

Of significance, a Ki67 phenotype reflects the proliferative state of the tumor, while AR expression levels signify androgen signaling and ultimately mediate tumor progression and response to anti-androgens. We have previously reported the role of elevated AR expression and its association with reduced durable response to androgen deprivation therapy resulting in PSAR and CF [12, 13]. In the current study, we have validated AR features to

accurately detect the >90% post-adjuvant PSA rise event state independent of proliferation, which provides additional mechanistic rationale for such a working hypothesis. We have expanded this interrogation to refine AR expression as a ratio of stromal and epithelial AR levels, thereby incorporating processes of AR stromal loss with prostate cancer invasion and outcome. The progressive loss of stromal AR has been associated with increasing Gleason grade, remodeling of the cancer microenvironment to promote invasion, and prostate cancer specific death [26], illustrating the importance for maintaining cellular compartment analysis in patient specific phenotyping. Such cellular attributes will undoubtedly be lost in all traditional genomic studies which abolish tissue architecture.

Most recently, the deregulation of AR expression has been shown to be a driver of chromatin relaxation, specifically at histone acetylation sites, such as bromodomain-containing protein-4 and prostate cancer progression [27]. We hypothesize that our ability to quantify AR composition at the micro-anatomical cellular level is providing the necessary tools to begin to dissect these types of processes more effectively than standard immunohistochemistry and subjective interpretation.

There are limitations to the Precise Post-op test, including the retrospective nature of the study design and the inability to assess clinical utility in a randomized prospective environment. This is true for all post-operative prognostic models where 'level 1' evidence is not feasible. It is for this reason that one of our primary goals was to develop a model that is relatively easy to interpret and implement in general practice given easy access to fixed tissue specimens, binary with respect to risk, cost-effective compared to genomic analyses and utilizes the intact specimen to derive significant features which have biological implications. We do acknowledge that such testing platforms are not routinely available in the majority of pathology laboratories and as is true of many novel genomic tests, speciality reference laboratories are necessary. We also chose to combine and construct training and test patients using both hospital cohorts to address variability in tissue processing and ultimately reduce generation of unstable features. Ultimately such models could be incorporated into the risk assessment NCCN guidelines with an opportunity for inclusion in various therapeutic clinical trials including neoadjuvant studies. We also acknowledge that the majority of events in the CF cohort were due to a post-adjuvant PSA rise and not documented metastases. Of note, 19 of these patients had either lymph node invasion or metastatic disease and the Precise Post-op test classified 79% as high risk. Additional limitations include the use of tissue microarrays and the percentage of unevaluable patients during modeling. We believe that tissue microarrays, when constructed appropriately, are quite effective in capturing patient specific tumoral heterogeneity and also allow for consistency when analyzing biomarkers across multiple tissue samples. Finally, the majority of excluded patients was due to limited tumor content in the available cores which will not be a limiting factor in clinical practice and was not detrimental with respect to event loss or statistical power of the analyses. Furthermore, additional extended validation studies are underway utilizing whole sections to further understand performance of the assay

In sum, this study introduces an innovative platform to assess prostate cancer risk, revealing that patients with high Precise Post-op scores have a higher likelihood of having clinical failure within 8 years. The Precise Post-op test guided by machine learning competes with Gleason grading utilizing novel image features that combine morphometry with biological attributes that appear to more accurately reflect disease potential. The clinical tangible is a robust, easy to interpret postoperative risk model that helps to further define clinical failure. The automated enhancement of the Gleason grade in a multivariate modeling approach represents an objective interrogation of prostate cancer heterogeneity which can be easily incorporated into clinical pathology practice. Accurately understanding prostate cancer recurrence risk after initial therapy is critical for patient management, defining therapeutic recommendations and subsequent future clinical trials. Further studies are now in progress to extend the validation studies, including the application to the biopsy specimen at diagnosis.

Acknowledgements We would like to thank members of the Biorepository and Pathology Core and all support personnel in the Department of Pathology at the Icahn School of Medicine at Mt. Sinai. We would also like to acknowledge both Roswell Park Cancer Center and the Henry Ford Hospital for access to their respective prostatectomy tissue cohorts.

Funding The study was funded by the Icahn School of Medicine at Mt. Sinai but Mt. Sinai was not directly involved in the design and conduct of the study, the collection, management, analysis and interpretation of the data; preparation, review, or approval of the manuscript nor the decision to submit the manuscript for publication.

Author contributions MJD, GF, RS, JZ, FMK, AT, and CCC contributed to study concept and design. MJD, GF, RS, GK, NG, EC, and FMK contributed to study development and methods, and GF, RS, NG, and EC collected the data. All authors analyzed and interpreted the data. FMK provided statistical and model development support. GK, NG, EC, and JZ provided administrative, technical, and material support. GF, JZ, RS, and MJD supervised the study. MJD, GF, and CCC wrote the manuscript. All authors have approved the final version of the manuscript.

Compliance with ethical standards

Conflict of interest MJD, GF, RS, FMK, JZ, and CCC have patents in varying aspects of the methods, technology and modeling platform utilized in the study.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. CA Cancer J Clin. 2017;67:7–30.
2. Center MM, Jemal A, Lortet-Tieulent J, Ward E, Ferlay J, Brawley O, et al. International variation in prostate cancer incidence and mortality rates. Eur Urol. 2012;61:1079–92.
3. Wilt TJ, Brawer MK, Jones KM, Barry MJ, Aronson WJ, Fox S, et al. Radical prostatectomy versus observation for localized prostate cancer. N Engl J Med. 2012;367:203–13.
4. Bill-Axelson A, Holmberg L, Garmo H, Rider JR, Taari K, Busch C, et al. Radical prostatectomy or watchful waiting in early prostate cancer. N Engl J Med. 2014;370:932–42.
5. Cooperberg MR, Davicioni E, Crisan A, Jenkins RB, Ghadessi M, Karnes RJ. Combined value of validated clinical and genomic risk stratification tools for predicting prostate cancer mortality in a high-risk prostatectomy cohort. Eur Urol. 2015;67:326–33.
6. Boorjian SA, Thompson RH, Tollefson MK, Rangel LJ, Bergstralh EJ, Blute ML, et al. Long-term risk of clinical progression after biochemical recurrence following radical prostatectomy: the impact of time from surgery to recurrence. Eur Urol. 2011;59:893.
7. Hoffman KE, Nguyen PL, Chen MH, Chen RC, Choueiri TK, Hu JC, et al. Recommendations for post-prostatectomy radiation therapy in the United States before and after the presentation of randomized trials. J Urol. 2011;185:116.
8. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA, et al. The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. Am J Surg Pathol. 2016;40:244–52.
9. Cooperberg MR, Hilton JF, Carroll PR. The Capra-S Score. A straightforward tool for improved prediction of outcomes after radical prostatectomy. Cancer. 2011;117:5039–46.
10. Donovan MJ, Cordon-Cardo C. Implementation of a precision pathology program focused on oncology-based prognostic and predictive outcomes. Mol Diagn Ther. 2017;21:115–23.
11. Cordon-Cardo C, Kotsianti A, Verbel DA, Teverovskiy M, Capodieci P, Hamann S, et al. Improved prediction of prostate cancer recurrence through systems pathology. J Clin Invest. 2007;117:1876–83.
12. Donovan MJ, Hamann S, Clayton M. A systems pathology approach for the prediction of prostate cancer progression after radical prostatectomy. J Clin Oncol. 2008;26:3923–9.
13. Donovan MJ, Khan FM, Fernandez G, Mesa-tejada R, Sapir M, Zubek VB, et al. Personalized prediction of tumor response and cancer progression on prostate needle biopsy. J Urol. 2009;182:123–30.
14. Scott R, Khan FM, Zeineh J, Donovan M, Fernandez G. Gland ring morphometry for prostate cancer prognosis in multispectral immunofluorescence images. Medical Image Computing and Computer Assisted Intervention, MICCAI 2014. Lect Notes Comput Sci. 2014;8673:585–92.
15. Khan FM, Scott R, Donovan MJ, Fernandez G. Predicting and replacing the pathological Gleason grade with automated gland ring morphometric features from immunofluorescent prostate cancer images. J Med Imag. 2017;4:021103.
16. Donovan MJ, Khan FM, Fernandez G, Mesa-Tejada R, Sapir M, Zubek VB, et al. Personalized prediction of tumor response and cancer progression on prostate needle biopsy. J Urol. 2009;182:125–32.
17. Cuzick J, Stone S, Fisher G, Yang ZH, North BV, Berney DM, et al. Validation of an RNA cell cycle progression score for predicting death from prostate cancer in a conservatively managed needle biopsy cohort. Br J Cancer. 2015;11:382–9.
18. Ross AE, Johnson MH, Yousefi K, Davicioni E, Netto GJ, Marchionni L, et al. Tissue-based genomics augments post-prostatectomy risk stratification in a natural history cohort of intermediate and high risk men. Eur Urol. 2016;69:157–65.
19. Spratt DE, Yousefi K, Deheshi S, Ross AE, Den RB, Schaeffer EM, et al. Individual patient-level met-analysis of the performance of the Decipher Genomic Classifier in high-risk men after prostatectomy to predict development of metastatic disease. J Clin Oncol. 2017;35:1991–8.
20. Donovan MJ, Khan FM, Powell D, Bayer-Zubek V, Cordon-Cardo C, Costa J, et al. Postoperative systems models more accurately predict risk of significant disease progression than standard risk groups and a 10-year postoperative nomogram: potential impact on the receipt of adjuvant therapy after surgery. BJUI. 2012;109:40–5.
21. Punnen S, Freedland SJ, Presti JC Jr, Aronson WJ, Terris MK, Kane CJ, et al. Multi-institutional validation of the CAPRA-S score to predict disease recurrence and mortality after radical prostatectomy. Eur Urol. 2015;65:1171–7.
22. Donovan MJ, Cordon-Cardo C. Genomic analysis in active surveillance: predicting high-risk disease using tissue biomarkers. Curr Opin Urol. 2014;24:303–10.
23. Pascale M, Aversa C, Barbazza R, Maronqiu B, Siracusano S, Stoffel F, et al. The proliferation marker Ki67, but not neuroendocrine expression, is an independent factor in the prediction of prognosis of primary prostate cancer patients. Radiol Oncol. 2016;50:313–20.
24. Green WJ, Ball G, Hulman G, Johnson C, Van Schalwyk G, Ratan HL, et al. Ki67 and DLX2 predict increased risk of metastasis formation in prostate cancer-a targeted molecular approach. Br J Cancer. 2016;115:236–42.
25. Tretiakova MS, Wei W, Boyer HD, Newcomb LF, Hawley S, Auman H, et al. Prognostic value of Ki67 in localized prostate carcinoma: a multi-institutional study of > 1000 prostatectomies. Prostate Cancer Prostatic Dis. 2016;19:264–70.
26. Leach D, Need E, Toivanen R, Trotta AP, Palethorpe HM, Tamblyn DJ, et al. Stromal androgen receptor regulates the composition of the microenvironment to influence prostate cancer outcome. Oncotarget. 2015;6:16135–50.
27. Urbanucci A, Barfield S, Kytola V, Itkonen HM, Coleman IM, Vodak D, et al. Androgen receptor deregulation drives bromodomain-mediated chromatin alterations in prostate cancer. Cell Rep. 2017;19:2045–59.