Clinical Cancer Research

Development and Clinical Validation of an *In Situ* Biopsy-Based Multimarker Assay for Risk Stratification in Prostate Cancer

Peter Blume-Jensen[1], David M. Berman[2], David L. Rimm[3], Michail Shipitsin[1], Mathew Putzi[4], Thomas P. Nifong[1], Clayton Small[1], Sibgat Choudhury[1], Teresa Capela[1], Louis Coupal[5], Christina Ernst[1], Aeron Hurley[1], Alex Kaprelyants[1], Hua Chang[1], Eldar Giladi[1], Julie Nardone[1], James Dunyak[1], Massimo Loda[6], Eric A. Klein[7], Cristina Magi-Galluzzi[8], Mathieu Latour[9], Jonathan I. Epstein[10], Philip Kantoff[6], and Fred Saad[9]

Abstract

Purpose: Prostate cancer aggressiveness and appropriate therapy are routinely determined following biopsy sampling. Current clinical and pathologic parameters are insufficient for accurate risk prediction leading primarily to overtreatment and also missed opportunities for curative therapy.

Experimental Design: An 8-biomarker proteomic assay for intact tissue biopsies predictive of prostate pathology was defined in a study of 381 patient biopsies with matched prostatectomy specimens. A second blinded study of 276 cases validated this assay's ability to distinguish "favorable" versus "nonfavorable" pathology independently and relative to current risk classification systems National Comprehensive Cancer Network (NCCN and D'Amico).

Results: A favorable biomarker risk score of ≤ 0.33, and a nonfavorable risk score of >0.80 (possible range between 0 and 1) were defined on "false-negative" and "false-positive" rates of 10% and 5%, respectively. At a risk score ≤ 0.33, predictive values for favorable pathology in very low-risk and low-risk NCCN and low-risk D'Amico groups were 95%, 81.5%, and 87.2%, respectively, higher than for these current risk classification groups themselves (80.3%, 63.8%, and 70.6%, respectively). The predictive value for nonfavorable pathology was 76.9% at biomarker risk scores >0.8 across all risk groups. Increased biomarker risk scores correlated with decreased frequency of favorable cases across all risk groups. The validation study met its two coprimary endpoints, separating favorable from nonfavorable pathology (AUC, 0.68; $P < 0.0001$; OR, 20.9) and GS-6 versus non–GS-6 pathology (AUC, 0.65; $P < 0.0001$; OR, 12.95).

Conclusions: The 8-biomarker assay provided individualized, independent prognostic information relative to current risk stratification systems, and may improve the precision of clinical decision making following prostate biopsy. *Clin Cancer Res; 21(11); 2591–600. ©2015 AACR.*

Introduction

In 2014, there will be an estimated 233,000 new diagnoses of prostate cancer in the United States (1). The majority of these patients will have early-stage, clinically localized disease (1–3). Given the marked genomic and pathologic heterogeneity of prostate cancer and concerns regarding its overtreatment (4, 5), significant effort has been devoted to the discovery and validation of biomarkers capable of distinguishing indolent cases with good prognosis from more aggressive cases with poor survival (6). Pathologic evaluation of needle biopsy tissue establishes a prostate cancer diagnosis and provides data to define a patient's risk category that will guide treatment decisions (6). Although a number of useful classification systems have been developed that combine available clinical and pathologic parameters (6, 7), it is widely recognized that improved classification methods are needed to ascribe an individualized risk score and help clinical management (8–10).

A significant proportion of patients undergoing biopsy ultimately will have their tumor pathology (as defined by Gleason score) upgraded or downgraded following assignment of a more accurate "surgical" or pathologic Gleason score after prostatectomy (11). For example, 25% to 30% of patients with low-risk disease at biopsy are subsequently upgraded upon pathologic review of their radical prostatectomy (10–12). These revisions may reflect initial biopsy sampling error (13), or pathologist discordance in tumor grading (14), both of which can contribute either to overtreatment or delayed treatment of disease (4). In particular, there are concerns around overcalling

[1]Metamark Genetics Inc., Cambridge, Massachusetts. [2]Department of Pathology and Molecular Medicine, Queen's University, Kingston, Ontario, Canada. [3]Department of Pathology, Yale University Medical School, New Haven, Connecticut. [4]Urology Austin, Austin, Texas. [5]Impacts Inc., Montreal, Quebec, Canada. [6]Dana-Farber Cancer Institute, Boston, Massachusetts. [7]Glickman Urological and Kidney Institute, Cleveland Clinic, Cleveland, Ohio. [8]Department of Anatomic Pathology, Cleveland Clinic, Cleveland, Ohio. [9]Centre Hospitalier de l'Université de Montréal, Montreal, Quebec, Canada. [10]Department of Pathology, Urology and Oncology, Johns Hopkins Hospital, Baltimore, Maryland.

Corresponding Author: Fred Saad, Centre Hospitalier de l'Université de Montréal, 900 rue Saint Denis, Montreal H2X0A9, QC, Canada. Phone: 514-890-8000, ext. 27466; Fax: 514-412-7620; E-mail: fred.saad@umontreal.ca

AACR **2591**

Translational Relevance

Early-stage prostate cancers exhibit significant heterogeneity on the genomic, pathologic and clinical levels, motivating efforts to identify prognostic determinants capable of guiding patient management at the time of biopsy. Current pathologic and clinical parameters are insufficient for accurate prediction of the disease state from prostate cancer biopsies, particularly those with Gleason grades 3+3 or 3+4. Here, we sought to establish a novel assay for automated quantitative measurements of protein biomarkers in defined areas of intact formalin-fixed paraffin-embedded biopsy-derived tissue. The assay provides independent prognostic information about the disease state at the time of biopsy and could aid in stratification of patients for active surveillance versus therapeutic intervention.

or undercalling Gleason pattern 4 in needle biopsy samples, because these cases usually fall in a gray zone between clear active surveillance and surgical intervention (11, 15–17). In addition, there is an ongoing need to determine whether a cancer is organ confined, or non-organ confined with ultimate metastatic potential in patients with low- to intermediate-grade disease.

Advances have been made in identifying genetic markers that may inform clinical risk prognostication in prostate cancer (18, 19). There has been focus on identifying *in situ* protein biomarkers on the cellular level that, under circumstances of intratumor cellular heterogeneity, would enable more accurate classification of the most aggressive tumor areas that may consist of only few cancer cells (20, 21). We recently reported the development and use of a robust quantitative multiplex proteomics imaging (QMPI) approach to identify and measure protein biomarker candidates resistant to such sampling error (see Supplementary Appendix; ref. 22). Starting with a large pool of potential candidates, this biopsy simulation study identified 12 biomarker candidates that predict both prostate pathology aggressiveness and lethal outcome despite sampling error (22) based on prostatectomy tissue samples (Supplementary Appendix). The 12 biomarkers are ACTN1, CUL2, DERL1, FUS, HSPA9, PDSS2, PLAG1, pS6, SMAD2, SMAD4, VDAC1, and YBX1.

In this study, we report further development and subsequent blinded validation of an 8-biomarker assay derived from 11 of these markers in two independent clinical biopsy studies (Figure 1). Specifically, we first designed a biomarker development study to define an optimal biomarker subset and develop a predictive mathematical model. This effort resulted in an 8-biomarker prognostic model that provided "risk scores" predictive of final disease pathology. Next, benchmarked against Gleason scores and TNM staging data of matched post-surgical prostatectomy tissue, the performance of this 8-biomarker biopsy assay to predict the clinically relevant endpoint of favorable versus nonfavorable pathology was evaluated in an independent clinical validation study. The risk score generated by the 8-biomarker assay for each patient was compared with two current risk stratification systems, the National Comprehensive Cancer Network (NCCN) guideline categories and D'Amico system (6, 7), and its potential to provide Supple-

mentary Data of high-predictive value that could aid in treatment decision for individual patients was considered. The intended use of this biomarker assay is to supplement current biopsy-based prostate cancer risk assessment methods in cases where a clinical decision regarding active surveillance versus active treatment is not straightforward.

Materials and Methods

Clinical samples

Biomarker assay development. To develop a robust assay, multiple institutions were recruited representing typical North American patient cohorts (Table 1): Urology Austin, Chesapeake Urology Associates, Cleveland Clinic, Michigan Urology, and Folio Biosciences. Biopsy sample inclusion/exclusion criteria matched those that would be in place during routine clinical use of the assay. Because they are usually not candidates for active surveillance, patients with biopsy Gleason ≥4+3 were excluded, except for a limited number of biopsies that had been discordantly graded as both 3+4 and 4+3 by two expert pathologists. Annotations, including information on matched biopsy and prostatectomy pathology reports, were obtained. Researchers were blinded to sample identity during laboratory processing. The final number of cases included in the biomarker assay development cohort was 381.

Clinical validation. The validation study cohort (N = 276) was separate and independent from the assay development cohort, and comprised biopsy samples with matched prostatectomy annotation from patients managed at the University of Montreal Hospital Center, Canada (Table 1). Consent criteria and Institutional Review Board (IRB) approval steps were as for the assay development phase of the study, above. Inclusion criteria were biopsies with a centralized Gleason score 3+3 or 3+4 (biopsies with discordant grading by two expert pathologists of 3+4 and 4+3 were included as well), and matched prostatectomy with pathologic TNM staging, PSA level, and resulting surgical Gleason score. Central review was performed on biopsies, whereas for prostatectomies we had to rely on the original pathology annotation, due to practical challenges in accessing all the different locations where these had been conducted; however, all prostatectomy pathology annotation was done according to new ISUP classification. Sample sizes were statistically designed, with power greater than 95% for both cohorts.

Quantitative multiplex immunofluorescence format

The QMPI approach for protein *in situ* measurements was as described (Supplementary Appendix; ref. 23). Briefly, formalin-fixed, paraffin-embedded prostate cancer biopsy tissue slides were analyzed using a quantitative multiplex proteomics imaging (QMPI) platform for intact tissue that integrates morphologic object recognition and molecular biomarker measurements from tumor epithelium at the individual slide level. Four combinations of three (triplex) biomarkers each were used: (i) PLAG1, SMAD2, ACTN1; (ii) VDAC1, FUS, SMAD4; (iii) pS6, YBX1, DERL1; (iv) PDSS2, CUL2, HSPA9. When each of the primary antibodies used was validated for specificity, it was found that PLAG1 was insufficiently specific; it was, thus, excluded from the potential signature. Prostate tissue slides that passed initial quality control were subjected to the multiplex staining procedure with subsequent image acquisition, Definiens analysis, and bioinformatics

Table 1. Summary of the clinical patient cohorts in both biomarker development and clinical validation studies, including tumors scored both at biopsy (Biopsy Gleason) and after radical prostatectomy (Surgical Gleason)

	Biomarker assay development study	Clinical validation study
Total (N)	381	276
Age, mean (SD), y	60.6 (7.0)	59.9 (5.7)
PSA, mean (SD), ng/mL	7.7[a] (9.6)	6.7 (5.7)
Biopsy Gleason score	N (%)	N (%)
3+3	162 (42.5)	191 (69.2)
3+4	191 (50.1)	68 (24.6)
3+4[b]	28 (7.3)	17 (6.2)
Surgical Gleason score	N (%)	N (%)
3+3 = 6	138 (36.2)	80 (29)
3+4	97 (25.5)	150 (54.3)
4+3	119 (31.2)	39 (14.1)
8	14 (3.7)	4 (1.5)
9	13 (3.4)	3 (1.1)
Pathologic stage	N (%)	N (%)
T1	2 (0.6)	
T2	269 (70.5)	168 (60.8)
T3a	72[c] (18.9)	91 (33)
T3b-c	37 (9.7)	15 (5.5)
T4	1 (0.3)	0 (0)
Missing	0 (0)	2 (0.7)
Source Institution	N (%)	N (%)
University of Montreal	—	276 (100)
Urology Austin	147 (38.6)	—
Chesapeake Urology Associates	8 (2.1)	—
Cleveland Clinic	71 (18.6)	—
Folio Biosciences	36 (9.4)	—
Michigan Urology	119 (31.2)	

NOTE: In the clinical validation study, preoperative PSA was missing for 8 patients, and clinical staging was missing for 12 patients.
[a]Excludes two patients with PSA reported at diagnosis of 791 and 600 ng/mL, which is atypical of newly diagnosed patients.
[b]Patients with discordant Gleason (3+4/4+3) based on centralized Gleason review were included; see Materials and Methods.
[c]Includes four samples annotated only as T3.

Figure 1.
A simplified overview of biomarker selection, biomarker assay model development, and validation of biomarker assay model.
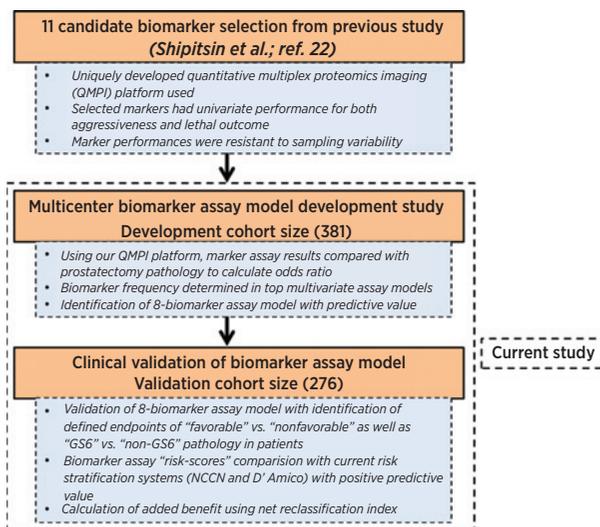
analysis. (See Supplementary Appendix for further details of antibody validation, staining protocols, image acquisition, image analysis, and interexperimental controls.)

Biomarker assay development

A noninterventional, retrospective clinical assay development study using biopsy case tissue samples was devised to define the best marker subset from those candidates previously shown to correlate with both prostate pathology aggressiveness and lethal outcome (22). The study goal was to define a model able to distinguish between prostate pathology usually recommended for active surveillance ("GS 6"): surgical Gleason 3+3 and ≤T3a versus those more likely to require prostatectomy ("non–GS 6"): Surgical Gleason ≥3+4 or nonlocalized >T3a or N or M. This GS 6 versus non–GS 6 definition was based on studies showing that tumors with surgical Gleason 3+3 at prostatectomy do not metastasize (15, 16). IRBs approved the biomarker development protocol, and patient consent was obtained or waived accordingly.

Train-test to define biomarker set

To rank our previously identified biomarkers in terms of their predictive importance, we evaluated each potential combination

of biomarkers using 500 rounds of computation. In each round, a training set (65% of cases) was obtained by bootstrap and remaining samples (35% of cases) were used as a testing set. A logistic regression model was trained on the training set, and risk scores were obtained for samples in both the training and testing sets. The final marker coefficients were used in a logistic regression model for calculation of the risk score, provided as a continuous scale from 0 to 1. After computation of these statistics, we sorted the biomarker sets in three different ways. Biomarker sets were characterized by the area under the receiver operating characteristic (ROC) curve (AUC), and sorted by increasing value of Akaike information criterion, decreasing value of the AUC on the training set, and decreasing value of the AUC on the testing set. To finalize the set of biomarkers for our assay, we determined the frequency of usage of each biomarker in the 10% of most highly ranked sets. Optimal biomarker set size was likewise determined through examination of test-set AUC performance during the exhaustive model search. We defined a resulting risk score, a continuous number between 0 and 1, which estimated the likelihood of "non–GS 6" pathology. Sensitivity analyses confirm the defined, locked-down biomarker assay set. We also explored a number of sensitivity analyses to confirm performance robustness. These analyses identified an optimal set of 8 biomarkers with the best performance in distinguishing GS 6 from non–GS 6 cases (Fig. 2).

Clinical validation study

A noninterventional, blinded, prospectively designed, retrospectively collected clinical study was conducted to validate the performance of the 8-biomarker assay to predict prostate pathology on its own and relative to current systems for patient risk categorization. This study used biopsy samples from a patient cohort independent of the one used for assay development (Table 1), and generated a "risk score" for each sample. Matched prostatectomy samples with annotated surgical

Blume-Jensen et al.
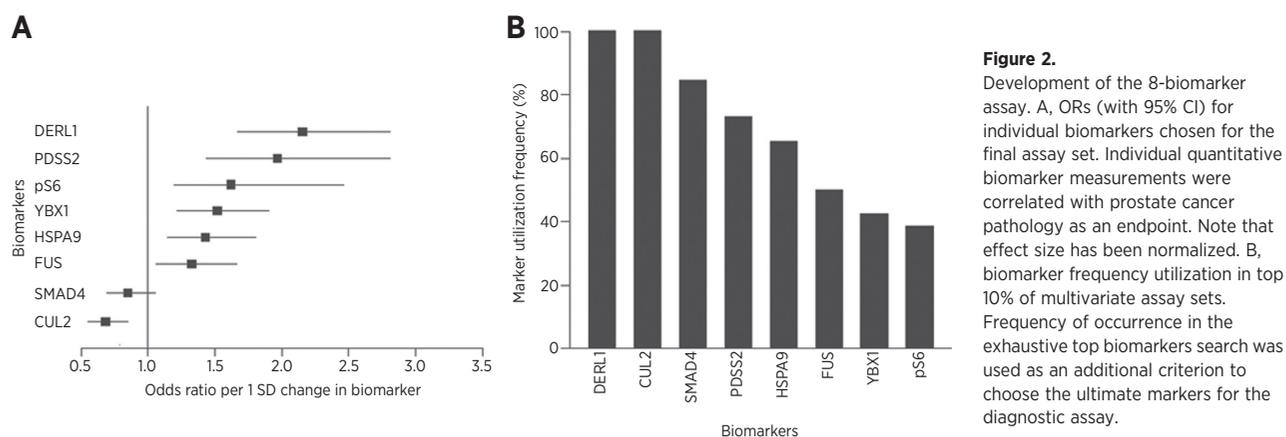
Figure 2.
Development of the 8-biomarker assay. A, ORs (with 95% CI) for individual biomarkers chosen for the final assay set. Individual quantitative biomarker measurements were correlated with prostate cancer pathology as an endpoint. Note that effect size has been normalized. B, biomarker frequency utilization in top 10% of multivariate assay sets. Frequency of occurrence in the exhaustive top biomarkers search was used as an additional criterion to choose the ultimate markers for the diagnostic assay.

Gleason scores ultimately classified the tumor as "favorable" or "nonfavorable" for the purposes of evaluating assay results. ROCs and corresponding AUCs determined for the 8-biomarker risk score quantitatively evaluated performance of the assay.

Two coprimary endpoints for validation study

Favorable versus nonfavorable pathology was chosen for final patient categorization throughout the validation study. It reflects the increasing awareness that organ-confined disease with minimal Gleason 4 pattern is likely to remain indolent with a significantly better long-term prognosis than higher-grade (dominant Gleason 4 pattern) or non–organ-confined disease (16, 24, 25). Thus, one of our two coprimary endpoints was defined as "favorable" pathology: Surgical Gleason ≤3+4 and organ-confined disease (≤T2) versus "nonfavorable" pathology: Surgical Gleason ≥4+3 or non–organ-confined disease (T3a, T3b, N, or M). In addition, in recognition of clinical variability, a second coprimary endpoint was defined as "GS 6" pathology: Surgical Gleason of 3+3 and localized disease (≤T3a) versus "non–GS 6" pathology: Surgical Gleason ≥3+4 or nonlocalized disease (T3b, N, or M).

Statistical analysis

Using biopsy samples, the 8-biomarker assay–derived risk score validated these two coprimary endpoints for prostate pathology, as assessed by AUC. Secondary analyses included ORs for the highest quartile versus lowest quartile of risk score, and OR (point estimate) for the continuous scale. We compared the risk outcomes from our diagnostic test with current risk classification categories as defined by the NCCN and D'Amico (6, 7), using positive predictive values (PPV). Definition and statistical analysis of the Net Reclassification Improvement (NRI) was done as described by Pencina and colleagues (26).

See the Supplementary Appendix for the statistical plan for both clinical studies.

Results

Biomarker assay development

We first sought to identify a protein biomarker set and predictive model capable of correlating biopsy-based prognosis with final post-prostatectomy pathology. Biomarker assay development used prostate biopsies and post-prostatectomy anno-

tation from 381 patients recruited from several U.S. institutions (Table 1). Eleven previously identified biomarker candidates were evaluated using a 5-step optimization process to determine the top-performing biomarker set from among all possible combinations (Supplementary Appendix, Supplementary Fig. S1). Individual quantitative biomarker measurements were correlated with prostate cancer pathology as an endpoint, and univariate ORs were calculated (Fig. 2A). The performance of all possible biomarker combinations were assessed and several high-performing sets were identified; the biomarker set with the highest performance had a test AUC of 0.79 [95% confidence interval (CI), 0.72–0.84]. Given that many biomarker sets had similar performance in the bootstrapped test AUC, both univariate performance and frequency of occurrence in top-performing models (Fig. 2B) were used to choose the final biomarker set for validation. Model coefficients for each marker were derived from logistic regression on the full dataset. The resulting locked-down biomarker assay set consisted of eight biomarkers (CUL2, DERL1, FUS, HSPA9, PDSS2, pS6, SMAD4, and YBX1) as shown in Fig. 2.

Clinical validation study

Clinical validation used biopsy specimens, and post-prostatectomy annotation from a second independent cohort of 276 patients. As shown in Table 2, the study met its two coprimary endpoints and validated the 8-biomarker assay for both endpoints: "favorable" versus "nonfavorable" and "GS 6" versus "non–GS 6" (see Materials and Methods for endpoint details). The analysis for "favorable" pathology yielded an AUC of 0.68 with 95% CI, 0.61–0.74. The associated P value was <0.0001, with an OR for risk score of 20.9 per unit change. "GS 6" pathology yielded an AUC of 0.65 with 95% CI, 0.58–0.72. The associated P value was <0.0001, with an OR for risk score of 12.6 per unit change. Further details are shown in Supplementary Figs. S2 and S3 in the Supplementary Appendix.

We had sufficient prostatectomy annotation to classify 256 cases within our validation cohort according to NCCN and D'Amico criteria, two current systems for risk classification. The performance for "favorable" pathology of the 8-biomarker assay on associated biopsy samples from this cohort is shown in Supplementary Fig. S4 and was similar to the full cohort of 276 cases (AUC, 0.69; 95% CI, 0.63–0.76; P < 0.0001; OR for risk score, 26.2/unit change).

Table 2. Clinical validation study: prognostic test performance of the 8-biomarker assay against the two coprimary endpoints

Population (N)	Endpoint definition	AUC (95% CI)	P (Bonferroni-adjusted)	OR lowest- to highest-risk score quartile (95% CI)	OR as point estimate for continuous range of risk scores (95% CI)
Coprimary endpoints					
(N = 274)[a]	Favorable pathology—Surgical Gleason ≤3+4 and organ confined (≤T2) vs. nonfavorable—surgical Gleason ≥ 4+3 or non-organ confined (T3a, T3b, N, or M)	0.68 (0.61–0.74)	<0.0001	3.3 (1.8–6.1)	20.9 (6.4–68.2)
(N = 276)	"GS 6"—Surgical Gleason = 3+3 and localized ≤T3a vs. "non–GS 6"—surgical Gleason ≥3+4 or nonlocalized (T3b, N, or M)	0.65 (0.58–0.72)	<0.0001	4.2 (1.9–9.3)	12.6 (3.5–47.2)

[a]Two patient samples (of 276) lacked sufficient annotation for this calculation.

The AUC for identification of the "favorable" group based on NCCN risk stratification alone was 0.69 (95% CI, 0.62–0.75) and 0.65 (95% CI, 0.59–0.71) based on D'Amico alone. The AUCs for the combined models of the 8-biomarker assay with NCCN and D'Amico were 0.75 (95% CI, 0.69–0.81) and 0.75 (95% CI, 0.69–0.81), respectively.

Multivariate analysis was also performed during development of the combined models. The OR for aggressive disease per 0.25 change in the 8-biomarker assay risk score, with NCCN held as a constant, was 3.5 (pv $< 10^{-3}$; 95% CI, 2.6–4.7). The OR for aggressive disease per 0.25 change in 8-biomarker assay risk score, when the D'Amico status was held constant, was also 3.5 (pv $< 10^{-3}$, 95% CI, 2.6–4.7). These results are consistent with the notion that the 8-biomarker assay, the results of which are entirely based on integrated molecular and morphologic information, is capturing information that is partly independent of the prognostic features represented in the NCCN and D'Amico models, both of which are entirely based on clinical and pathologic parameters.

The 8-marker proteomic assay was also considered in the context of standard clinical parameters, including PSA, the percentage of cores positive as a surrogate for tumor volume, and biopsy Gleason pattern. These parameters were available to the investigators for the clinical development ('train-test') cohort and held as blinded data by the independent validating statistician for the validation cohort. A logistic regression model was built for the "favorable" endpoint on the train–test cohort using the clinical nomogram parameters alone as well as for the integrated model combining the clinical and 8-marker assays. The clinical parameters alone resulted in an AUC of 0.67 (95% CI, 0.61–0.74), and the combined model resulted in an AUC of 0.71 (95% CI, 0.64–0.77). Multivariate analysis of the combined model showed that the OR for aggressive disease per 0.25 change in 8-biomarker assay risk score, when clinical parameters were held constant, was 3.1 (P $< 10^{-3}$; 95% CI, 2.2–4.4). Again, these results are in accordance with the molecular information captured by the 8-biomarker assay is partly independent of the prognostic features based on the clinical and pathologic parameters in the clinical nomogram.

Next, to define the risk score ranges (cutoff points) for favorable versus nonfavorable pathology, we generated sensitivity and specificity curves for the 8-biomarker assay (Fig. 3A and B, respectively). Figure 3A shows an example of a favorable risk score category identified in this study population on the basis of the 8-biomarker assay. Here, a sensitivity of 90% (95% CI, 82%–

94%) corresponds to a biomarker assay risk score of 0.33 [sensitivity: P(risk score > 0.33 | nonfavorable pathology)]. Below this threshold of risk score, the false-negative rate among patients with nonfavorable pathology is limited to 10% (95% CI, 6%–18%). Stated differently, among patients with risk score ≤0.33, the chance of having nonfavorable pathology is less than or equal to 1 in 10. Conversely, in Fig. 3B, a nonfavorable category was defined by a biomarker assay risk score of 0.80, which in this study population corresponded to a Specificity of 95% (95% CI, 90%–98%). Above this threshold of risk score, the false-positive rate among patients with favorable pathology is limited to 5% (95% CI, 2%–10%), where Specificity = P(risk score ≤ 0.80| favorable pathology). Stated differently, among patients with risk score greater than 0.80, the likelihood of having a favorable pathology is less than 5 in 100.

Clinical utility of the 8-biomarker assay

Similar to prognostic assay such as OncotypeDx, which stratifies an individual patient to a particular risk level through cutoff points that define risk categories (27), the intended use of this 8-biomarker assay is to categorize an individual patient to favorable versus nonfavorable disease pathology based on where his risk score is on the specificity curves of Fig. 3. To this end, we first calculated the predictive value of the risk scores. The PPV for identifying favorable disease with a risk score of ≤0.33 was 83.6% (specificity, 90%; Table 3). Conversely, at a risk score of >0.80, 76.9% had nonfavorable disease (i.e., only 23.1% of patients in the high-risk score category had favorable disease). This translates to 39% of patients in this study population having risk scores ≤0.33 or >0.8, of which 81% were correctly identified by our 8-biomarker assay (Table 3). Of patients with intermediate risk scores (0.33 < risk score ≤ 0.8), 58.3% had favorable disease.

Next, we compared the PPV of this 8-biomarker assay with those of the NCCN and D'Amico risk categories. Here, within each NCCN category, we analyzed the distribution of patients categorized as having favorable disease by our biomarker risk score. The PPV for favorable pathology, at a risk score of ≤0.33, was 75% for patients in the NCCN intermediate-risk category; 81.5% for NCCN low-risk patients; and 95% for NCCN very low-risk, with an overall PPV for favorable pathology of 85% (Table 4, in green, and Fig. 3C). This contrasts favorably with the PPVs obtained for the NCCN risk categories alone, which were 40.9% for intermediate-risk, 63.8% for low-risk, and 80.3% for very low-risk (Table 4, blue). In other words, when combined with the current

Blume-Jensen et al.
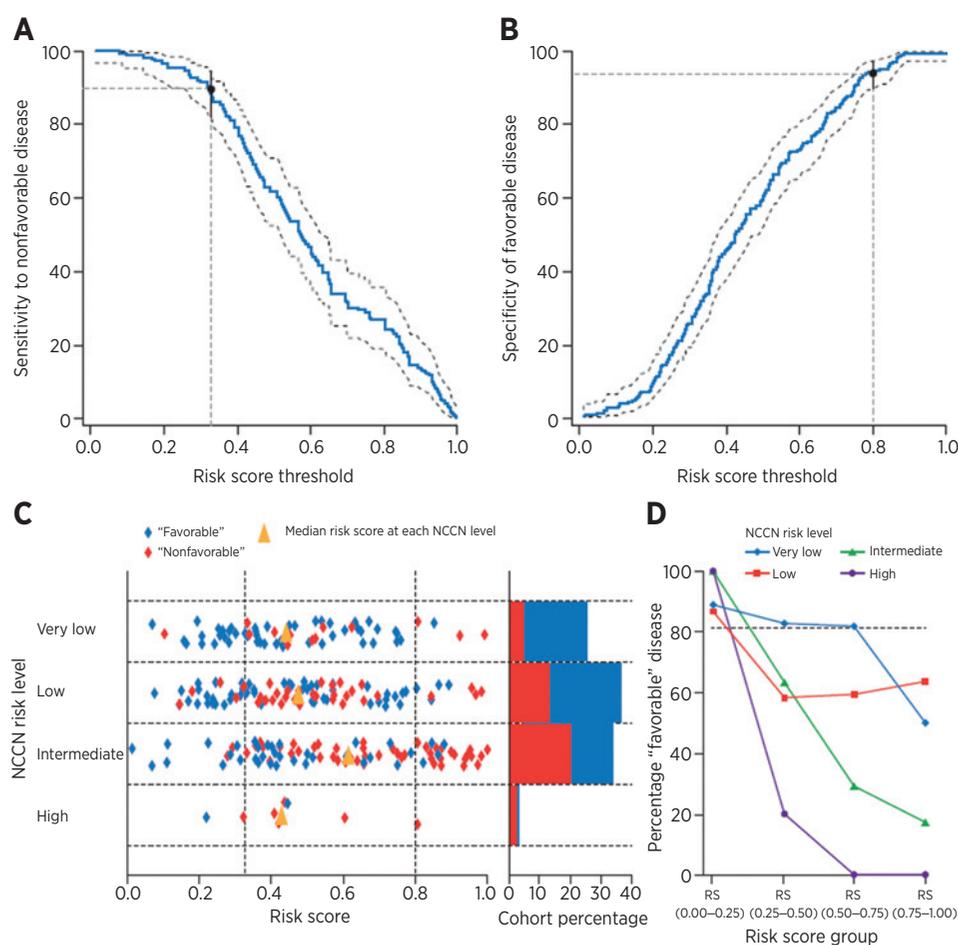
Figure 3.
Clinical validation study: performance for prediction of favorable pathology. Sensitivity and specificity curves (A and B, respectively) may be used to identify appropriate risk classification groups. Risk score distribution relative to NCCN risk classification groups (C and D) showing that the biomarker assay adds significant additional risk information within each NCCN classification level. A, sensitivity as a function of biomarker risk score. A sensitivity of 90% (95% CI, 82%–94%) corresponded to a biomarker assay risk score of 0.33, that is, sensitivity = P[risk score > 0.33 | nonfavorable pathology]. Using this risk score, a patient with nonfavorable pathology would have a 10% chance of incorrectly receiving a favorable classification (95% CI, 6%–18%). This false negative might lead to delayed treatment. B, specificity as a function of biomarker risk score. A specificity of 95% (95% CI, 90%–98%) in this cohort corresponded to a biomarker assay risk score of 0.8, that is, specificity = P[risk score ≤ 0.80| favorable pathology]. Using this risk score threshold, a patient with favorable pathology would have a 5% chance of incorrectly receiving a nonfavorable classification (95% CI, 2%–10%). This false positive might lead to overtreatment. C, biomarker risk score further stratifies tumor pathology within the existing NCCN classification system. Each rhomboid represents an individual patient case, and favorable (blue) and nonfavorable (red) categories refer to final classification of patient cases based on prostatectomy pathology. Note, the relative concentration of favorable (blue) cases at biomarker risk score <0.33, and the relative concentration of nonfavorable (red) cases at biomarker risk score >0.8. The predictive value (+PV) for favorable pathology was 85% at risk score cutoff <0.33. The predictive value (−PV) for nonfavorable pathology was 100% at risk score cutoff >0.9, and 76.9% at risk score >0.8. D, observed frequency of favorable cases as a function of the biomarker risk score quartile. Increased risk score quartile largely correlates with decreased observed frequency of favorable cases in each NCCN category. Observed frequency of patients with favorable pathology identified by the 8-biomarker assay versus the NCCN stratification alone increases from 0% to 23.8% at a confidence level of 81%.

prognostic standard, this 8-biomarker assay can increase the confidence that a patient categorized to the NCCN very low-risk group indeed has a favorable pathology based on his biopsy from 80.3% to 95%. Conversely, the predictive value for nonfavorable pathology was 76.9% at a risk score >0.8, but reached 100% at a risk score cutoff point of >0.9. Correspondingly, a higher biomarker risk score correlated with decreased frequency of favorable cases within each NCCN category (Fig. 3D). Taken together, we conclude that the risk score generated by our 8-biomarker assay offered additional prognostic value for individual patients relative to NCCN risk categories alone. Similar results were obtained

when comparing with the D'Amico categories (Supplementary Fig. S5).

To further confirm the benefit of having the 8-biomarker risk score in the context of current risk stratification systems, we performed an NRI analysis for the favorable and nonfavorable categories determined by our 8-biomarker assay, relative to NCCN and D'Amico categories (Supplementary Fig. S6). Using the underlying data shown in Tables 3 and 4, we found an NRI for NCCN of 0.34 (P < 0.00001; 95% CI, 0.20–0.48) and for D'Amico of 0.24 (P = 0.0001; 95% CI, 0.12–0.35). This provides strong evidence that this 8-biomarker assay provides additional

Table 3. Clinical validation study

Biomarker assay risk score range	Number of patients according to biomarker assay scores			Percentage of PPV (95% CI)
	Total	Favorable	Nonfavorable	
≤0.33	61	51	10	83.6% (71.9%–91.8%)
0.33–0.80	156	91	65	58.3% (50.2%–66.2%)
>0.80	39	9	30	23.1% (11.1%–39.3%)
All	256	151	105	59% (52.7%–65.1%)

NOTE: Comparison of predictive value of the 8-biomarker assay for favorable pathology. Green, predictive value for favorable disease based on test risk score less than 0.33 in the indicated risk category.

discriminatory capability to that offered by current risk stratification systems alone.

We also performed decision curve analysis, as this provides another method for characterizing performance of different risk systems and risk cutoff points (28). Supplementary Fig. S7 provides an example of such an analysis and describes the net benefit for a number of treatment regimes based on the 8-marker model, NCCN, and the combined model NCCN and 8-marker model. Using the illustrated cutoff points, we see that the combined model provides benefit through the span of risk thresholds. The range of incremental net benefit improves from small (0.005 at risk threshold of 20%) to significant (0.06 at a risk threshold of 50%) across the scale of risk thresholds. This further demonstrates the added prognostic information achieved by combining the 8-marker model with a standard-of-care risk stratification system.

Discussion

Recent studies indicate that long-term survival for patients with organ-confined Gleason 3+4 disease is significantly better than for patients with non–organ-confined disease or for tumor with dominant Gleason pattern 4 or higher (17, 24, 25), and that deferred therapy for the former group does not significantly change long-term outcome (29–31). Most risk stratification systems do not discriminate between Gleason 7 biopsies, so typically patients considered candidates for active surveillance belong to "very low-risk" or "low-risk" groups, which only contain biopsy Gleason score ≤6 (2, 6). However, around 25% of Gleason grade

3+4 biopsies are "downgraded" and a similar percentage of Gleason grade 3+3 biopsies are "upgraded" when comparing with the surgical Gleason scores, primarily owing to biopsy sampling error and pathologist discordance (11, 15). Therefore, there is a clear unmet clinical need for a molecular evidence-based prognostic assay that can complement the current pathology-based stratification system(s) to more accurately predict disease aggressiveness of patients with Gleason grade 3+3 and 3+4 biopsies (16).

To address this unmet need, we have embarked on an assay development effort, building on prior work that had led to the identification of 11 biomarker candidates shown to be predictive of both prostate cancer pathology and lethal outcome despite tissue sampling error (22). Of importance is that, we have used a multiplex proteomics imaging platform (23) for quantitative measurements of eight independent biomarkers in intact tissue to generate the risk score (Supplementary Appendix). This *in situ* approach has the advantage compared with other tests using homogenized tissue (19, 32), as it requires relatively few cancer cells (which is common of the biopsy specimens) and is not sensitive to variations in the ratios of benign tissue relative to tumor tissue.

Here, we report the results demonstrating that we have successfully developed an 8-biomarker biopsy-based prognostic assay that was validated in an independent blinded study and shown to complement current risk stratification systems. Specifically, our study shows that, at a risk score of ≤0.33, our biomarker assay can identify patients with favorable pathology in the very low-risk and low-risk NCCN and low-risk D'Amico

Table 4. Clinical validation study

NCCN	Biomarker assay risk score range	Number of patients according to biomarker assay scores and current classification systems categories			Percentage of PPV (95% CI)
		Total	Favorable	Nonfavorable	
Very low by NCCN	—	66	53	13	80.3% (68.7%–89.1%)
Very low	≤0.33	20	19	1	95% (75.1%–99.9%)
Very low	0.33–0.80	42	33	9	78.6% (63.2%–89.7%)
Very low	>0.80	4	1	3	25% (0.6%–80.6%)
Low by NCCN	—	94	60	34	63.8% (53.3%–73.5%)
Low	≤0.33	27	22	5	81.5% (61.9%–93.7%)
Low	0.33–0.80	59	34	25	57.6% (44.1%–70.4%)
Low	>0.80	8	4	4	50% (15.7%–84.3%)
Intermediate by NCCN	—	88	36	52	40.9% (30.5%–51.9%)
Intermediate	≤0.33	12	9	3	75% (42.8%–94.5%)
Intermediate	0.33–0.80	50	23	27	46% (31.8%–60.7%)
Intermediate	>0.80	26	4	22	15.4% (4.4%–34.9%)
High by NCCN	—	8	2	6	25% (3.2%–65.1%)
High	≤0.33	2	1	1	50% (1.3%–98.7%)
High	0.33–0.80	5	1	4	20% (0.5%–71.6%)
High	>0.80	1	0	1	0% (0%–97.5%)

NOTE: Comparison of predictive value of the 8-biomarker assay for favorable pathology with NCCN risk categories. Green, predictive value for favorable disease based on test risk score less than 0.33 in the indicated risk category; blue, predictive value for favorable disease by the standard risk category alone.

groups with PPVs of 95%, 81.5%, and 87.2%, respectively, that represents a significant improvement over NCCN or D'Amico stratification system alone. We suggest that information provided by the biomarker assay in addition to existing pathologic parameters can help both the physician and patient to consider active surveillance option with added confidence.

Furthermore, our 8-biomarker assay is also able to identify with high-confidence patients with nonfavorable pathology who are arguably unsuitable for active surveillance. Although a patient with a risk score >0.8 has a PPV of 76.9% for nonfavorable pathology, the PPV reaches 100% if the risk score is >0.9 across all risk groups for both risk stratification systems tested. This shows how this biomarker assay can be used in combination with current stratification system(s) to personalize the risk level for each individual patient, enabling rational and evidence-based decision by both physician and patient. Note, that the settings used for clinical decision-making thresholds are a matter of risk tolerance in the patient and physician. Because our assay provides continuous specificity and sensitivity curves, and individualized risk for each patient, other thresholds that reflect different individual risk tolerance could be chosen. For example, if 80% specificity is sufficient for a particular patient, then a risk score of 0.6 or higher would favor treatment over active surveillance.

Because of the heterogeneous and multifocal nature of prostate cancer, prostate biopsies frequently contain only lower-grade morphologic features with higher-grade molecular features that are not reflected in morphology. Our 8 biomarkers were specifically selected and evaluated to equally well-predict pathology outcome regardless whether they were measured in low-grade or high-grade regions from the same patient (22). Accordingly, by measuring proteins only from intact tumor regions, our test can accurately and sensitively assess such high-grade molecular features *in situ*, even in tissue samples with variable amounts of tumor versus benign components. This is a significant advantage over previously reported gene expression-based technologies that rely on tissue homogenization (18, 32–34), and where molecular (genetic) information from higher-grade regions is variably diluted by the values of the individual markers in intermixed benign tissue regions that inadvertently are part of the tissue homogenate. Genomic Health's "Oncotype DX prostate cancer test" is a 17-gene PCR-based assay, similar to the biopsy-based test described in this study, that provides genomic prostate score (GPS) for predicting cancer aggressiveness in prostate cancer patients based on the same endpoints we have developed (32, 33). Although a stand-alone AUC for the "Oncotype Dx" test has not been reported, the combined AUC for GPS and CAPRA is reported to be 0.67, whereas the AUC for CAPRA alone was reported as 0.63 (for favorable pathology). Obviously, a direct comparison between these studies is difficult, as a head-to-head study would be required to fully establish relative performance. Myriad's "Prolaris" test is another gene expression–based assay that measures activity of a panel of cell-cycle progression (CCP) genes and provides risk of prostate cancer–specific progression and mortality (18, 34). Although this test was originally developed on prostatectomy samples, it was assessed on needle biopsy material for its ability to predict lethal outcome from prostate cancer. Unfortunately, in this study (19) patients with biopsy Gleason scores of 4+3, 8, 9, and 10 were included and made up approximately 26% of the

study population. Arguably, these patients are in need of aggressive therapy, and therefore not included in our intended use, clinical need population. The "Prolaris" CCP score showed an HR of 1.65 in multivariate analysis, and although the majority of the performance of the test was derived from the patients with advanced disease, no specific measure of the performance on patients with biopsy scores of 3+3 or 3+4 was reported. Finally, the "Prolaris" test provides a prostate cancer risk score for patients by combining the CCP score with standard clinicopathologic parameters, it is not a stand-alone test (18, 32–34).

In this study, Gleason scores and TNM staging of matched post-prostatectomy annotations were used as benchmark to evaluate performance of the 8-biomarker assay, instead of metastatic disease or death as endpoints. This is due to relatively low grade and early stage of the target population of our test, rendering it necessary to have very long follow-up of very large and conservatively managed cohorts to be sufficiently powered for such study. Indeed, only few biopsy studies have had metastatic disease or lethal outcome as endpoints. As expected, to compensate for the power, these studies have included a significant proportion of high-risk cases (biopsies with Gleason score $\geq$4+3) in which time to death is much shorter; however, in such cases, the clinical utility of a complementary molecular test is arguably more limited (19).

Finally, our methodology was based on a binary endpoint derived from prostate pathology that was shown in a previous clinical study to correlate with lethal outcome (22). The binary endpoint was chosen to match clinical practice, in which patients are considered for active surveillance or aggressive treatment. For future investigations, assays could be developed on the basis of either an ordered categorical (e.g., low-, medium-, high-risk prostate pathology), or a continuous endpoint by combining inherently categorical prostate pathology assessment with other clinical variables. Although development of such models is straightforward, they would inherently be based on subjective weighting and combination of parameters that would require a new assessment of correlation with long-term outcome before general acceptance.

In conclusion, this study supports further evaluation of this biopsy-based prognostic biomarker assay for personalized prognostication of prostate cancer and its impact on therapeutic choice. The ability to provide differential information for the individual patient relative to current risk stratification systems, in which prognostic values are more limited, makes it a potentially useful addition in practice to improve accuracy of clinical decision.

Disclosure of Potential Conflicts of Interest

Authors' Contributions

Conception and design: P. Blume-Jensen, D.M. Berman, D.L. Rimm, T.P. Nifong, A. Hurley, E.A. Klein, J.I. Epstein, P. Kantoff, F. Saad

References

1. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. CA Cancer J Clin 2014;64:9–29.
2. Epstein JI. An update of the Gleason grading system. J Urol 2010;183:433–40.
3. Barocas DA, Cowan JE, Smith JA Jr, Carroll PR, Ca PI. What percentage of patients with newly diagnosed carcinoma of the prostate are candidates for surveillance? An analysis of the CaPSURE database. J Urol 2008;180:1330–4.
4. Loeb S, Bjurlin MA, Nicholson J, Tammela TL, Penson DF, Carter HB, et al. Overdiagnosis and overtreatment of prostate cancer. Eur Urol 2014;65:1046–55.
5. Sandhu GS, Andriole GL. Overdiagnosis of prostate cancer. J Natl Cancer Inst Monogr 2012;2012:146–51.
6. NCCN. NCCN clinical practice guidelines in oncology: prostate cancer. Version 3.2012. Available from: http://www.nccn.org/professionals/physician_gls/f_guidelines.asp#prostate_detection. Accessed 18 Feb 2014.
7. D'Amico AV, Whittington R, Malkowicz SB, Schultz D, Blank K, Broderick GA, et al. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. JAMA 1998;280:969–74.
8. Bangma CH, Roobol MJ. Defining and predicting indolent and low risk prostate cancer. Crit Rev Oncol Hematol 2012;83:235–41.
9. Brimo F, Montironi R, Egevad L, Erbersdobler A, Lin DW, Nelson JB, et al. Contemporary grading for prostate cancer: implications for patient care. Eur Urol 2013;63:892–901.
10. Truong M, Slezak JA, Lin CP, Iremashvili V, Sado M, Razmaria AA, et al. Development and multi-institutional validation of an upgrading risk tool for Gleason 6 prostate cancer. Cancer 2013;119:3992–4002.
11. Epstein JI, Feng Z, Trock BJ, Pierorazio PM. Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: incidence and predictive factors using the modified Gleason grading system and factoring in tertiary grades. Eur Urol 2012;61:1019–24.
12. Pinthus JH, Witkos M, Fleshner NE, Sweet J, Evans A, Jewett MA, et al. Prostate cancers scored as Gleason 6 on prostate biopsy are frequently Gleason 7 tumors at radical prostatectomy: implication on outcome. J Urol 2006;176:979–84.
13. Porten SP, Whitson JM, Cowan JE, Cooperberg MR, Shinohara K, Perez N, et al. Changes in prostate cancer grade on serial biopsy in men undergoing active surveillance. J Clin Oncol 2011;29:2795–800.
14. Goodman M, Ward KC, Osunkoya AO, Datta MW, Luthringer D, Young AN, et al. Frequency and determinants of disagreement and error in gleason scores: a population-based study of prostate cancer. Prostate 2012;72:1389–98.
15. Eggener SE, Scardino PT, Walsh PC, Han M, Partin AW, Trock BJ, et al. Predicting 15-year prostate cancer specific mortality after radical prostatectomy. J Urol 2011;185:869–75.
16. Ross HM, Kryvenko ON, Cowan JE, Simko JP, Wheeler TM, Epstein JI. Do adenocarcinomas of the prostate with Gleason score (GS) </ = 6 have the potential to metastasize to lymph nodes? Am J Surg Pathol 2012;36:1346–52.
17. Stark JR, Perner S, Stampfer MJ, Sinnott JA, Finn S, Eisenstein AS, et al. Gleason score and lethal prostate cancer: does 3 + 4 = 4 + 3? J Clin Oncol 2009;27:3459–64.
18. Cooperberg MR, Simko JP, Cowan JE, Reid JE, Djalilvand A, Bhatnagar S, et al. Validation of a cell-cycle progression gene panel to improve risk stratification in a contemporary prostatectomy cohort. J Clin Oncol 2013;31:1428–34.
19. Cuzick J, Berney DM, Fisher G, Mesher D, Moller H, Reid JE, et al. Prognostic value of a cell cycle progression signature for prostate cancer death in a conservatively managed needle biopsy cohort. Br J Cancer 2012;106:1095–9.
20. Donovan MJ, Hamann S, Clayton M, Khan FM, Sapir M, Bayer-Zubek V, et al. Systems pathology approach for the prediction of prostate cancer progression after radical prostatectomy. J Clin Oncol 2008;26:3923–9.
21. Ding Z, Wu CJ, Chu GC, Xiao Y, Ho D, Zhang J, et al. SMAD4-dependent barrier constrains prostate cancer growth and metastatic progression. Nature 2011;470:269–73.
22. Shipitsin M, Small C, Choudhury S, Giladi E, Friedlander S, Nardone J, et al. Identification of proteomic biomarkers predicting prostate cancer aggressiveness and lethality despite biopsy-sampling error. Br J Cancer 2014;111:1201–12.
23. Shipitsin M, Small C, Giladi E, Siddiqui S, Choudhury S, Hussain S, et al. Automated quantitative multiplex immunofluorescence in situ imaging identifies phospho-S6 and phospho-PRAS40 as predictive protein biomarkers for prostate cancer lethality. Proteome Sci 2014;12:40.
24. Pierorazio PM, Walsh PC, Partin AW, Epstein JI. Prognostic Gleason grade grouping: data based on the modified Gleason scoring system. BJU Int 2013;111:753–60.
25. Mullins JK, Feng Z, Trock BJ, Epstein JI, Walsh PC, Loeb S. The impact of anatomical radical retropubic prostatectomy on cancer control: the 30-year anniversary. J Urol 2012;188:2219–24.
26. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27:157–72.
27. Dowsett M, Cuzick J, Wale C, Forbes J, Mallon EA, Salter J, et al. Prediction of risk of distant recurrence using the 21-gene recurrence score in node-negative and node-positive postmenopausal patients with breast cancer treated with anastrozole or tamoxifen: a TransATAC study. J Clin Oncol 2010;28:1829–34.
28. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74.
29. Graefen M, Walz J, Chun KH, Schlomm T, Haese A, Huland H. Reasonable delay of surgical treatment in men with localized prostate cancer—impact on prognosis? Eur Urol 2005;47:756–60.
30. Holmstrom B, Holmberg E, Egevad L, Adolfsson J, Johansson JE, Hugosson J, et al. Outcome of primary versus deferred radical prostatectomy in the

National Prostate Cancer Register of Sweden Follow-Up Study. J Urol 2010;184:1322–7.

31. Vickers AJ, Bianco FJ Jr, Boorjian S, Scardino PT, Eastham JA. Does a delay between diagnosis and radical prostatectomy increase the risk of disease recurrence? Cancer 2006;106:576–80.

32. Klein EA, Cooperberg MR, Magi-Galluzzi C, Simko JP, Falzarano SM, Maddala T, et al. A 17-gene assay to predict prostate cancer aggressiveness in the context of gleason grade heterogeneity, tumor multifocality, and biopsy undersampling. Eur Urol 2014;66:550–60.

33. Cullen J, Rosner IL, Brand TC, Zhang N, Tsiatis AC, Moncur J, et al. A biopsy-based 17-gene genomic prostate score predicts recurrence after radical prostatectomy and adverse surgical pathology in a racially diverse population of men with clinically low- and intermediate-risk prostate cancer. Eur Urol. 2014 Nov 29. [Epub ahead of print].

34. Crawford ED, Scholz MC, Kar AJ, Fegan JE, Haregewoin A, Kaldate RR, et al. Cell cycle progression score and treatment decisions in prostate cancer: results from an ongoing registry. Curr Med Res Opin 2014; 30:1025–31.

Clinical Cancer Research

Development and Clinical Validation of an *In Situ* Biopsy-Based Multimarker Assay for Risk Stratification in Prostate Cancer

Peter Blume-Jensen, David M. Berman, David L. Rimm, et al.

Updated version	Access the most recent version of this article at: doi:10.1158/1078-0432.CCR-14-2603
Supplementary Material	Access the most recent supplemental material at: http://clincancerres.aacrjournals.org/content/suppl/2015/03/04/1078-0432.CCR-14-2603.DC1

Cited articles	This article cites 32 articles, 5 of which you can access for free at: http://clincancerres.aacrjournals.org/content/21/11/2591.full#ref-list-1
Citing articles	This article has been cited by 11 HighWire-hosted articles. Access the articles at: http://clincancerres.aacrjournals.org/content/21/11/2591.full#related-urls

E-mail alerts	Sign up to receive free email-alerts related to this article or journal.
Reprints and Subscriptions	To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.
Permissions	To request permission to re-use all or part of this article, use this link http://clincancerres.aacrjournals.org/content/21/11/2591. Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.