Intended for healthcare professionals

Get the support you need to get published Sage Author Services

Sage Journals
Search this journal Enter search terms... Advanced search
Access/Profile Cart
Browse by discipline Information for

Cancer Biomarkers

IOS Press Impact Factor: 1.9 / 5-Year Impact Factor: 2.0 Journal Homepage Submission Guidelines

Open access | Research article | First published online March 20, 2025

A simplified MyProstateScore2.0 for high-grade prostate cancer

Tiffany M Tang, Yuping Zhang, [...] for the EDRN-PCA3 Study Group (+12) View all authors and affiliations

All Articles | https://doi.org/10.1177/18758592241308755 | View article versions

Abstract

Background:

The limited diagnostic accuracy of prostate-specific antigen screening for prostate cancer (PCa) has prompted innovative solutions, such as the state-of-the-art 18-gene urine test for clinically-significant PCa (MyProstateScore2.0 (MPS2)). **Objective:** We aim to develop a non-invasive biomarker test, the simplified MPS2 (sMPS2), which achieves similar state-of-the-art accuracy as MPS2 for predicting high-grade PCa but requires substantially fewer genes than the 18-gene MPS2 to improve its accessibility for routine clinical care. **Methods:** We grounded the development of sMPS2 in the Predictability, Computability, and Stability (PCS) framework for veridical data science. Under this framework, we stress-tested the development of sMPS2 across various data preprocessing and modeling choices and developed a stability-driven PCS ranking procedure for selecting the most predictive and robust genes for use in sMPS2. **Results:** The final sMPS2 model consisted of 7 genes and achieved a 0.784 AUROC (95% confidence interval, 0.742–0.825) for predicting high-grade PCa on a blinded external validation cohort. This is only 2.3% lower than the 18-gene MPS2, which is similar in magnitude to the 1–2% in uncertainty induced by different data preprocessing choices. **Conclusions:** The 7-gene sMPS2 provides a unique opportunity to expand the reach and adoption of non-invasive PCa screening.

Introduction

Prostate specific antigen (PSA) is the most widely used biomarker for early screening of prostate cancer, contributing to significant decline in prostate cancer (PCa) mortality.[1] Despite the benefit, the low specificity of PSA due to its organ-specific rather than cancer-specific nature, can lead to overdiagnosis and overtreatment. To combat these challenges, tremendous effort has been made to develop non-invasive and more specific biomarker tests, especially to distinguish clinically-significant or high-grade PCa (i.e., Grade Group 2 PCa and higher) from dormant or low-grade PCa. In particular, the discovery of recurrent gene fusion *TMPRSS2:ERG* (*T2:ERG*) in prostate cancer[2] led to the development of MyProstateScore test (MPS), which measures expression of prostate cancer antigen 3 (*PCA3*) and *T2:ERG* in clinical urine specimens[3] and has been endorsed by the National Comprehensive Cancer Network for consideration prior to biopsy in men with elevated PSA.[4]

Since the development of MPS, high-throughput gene expression profiling has become readily available through RNA sequencing from large-scale cancer studies, such as The Cancer Genome Atlas (TCGA). Leveraging public and in-house RNA-seq data from PCa tumor and normal prostate, we previously nominated 54 biomarkers and established the second generation of MyProstateScore named MPS2, an 18-gene urine assay to predict high-grade PCa.[5] The gene panel was validated on an independent or external cohort, achieving equivalent performance as in the development cohort (area under the receiver operating characteristic [AUROC] 0.81) and surpassing the performance of MPS (AUROC 0.74).

Given the previous concerns regarding PSA and possible overdiagnosis of indolent, low-grade prostate cancer, we stress the importance of demonstrating the robustness and generalizability of non-invasive biomarker tests such as MPS and MPS2. Notably, in any statistical or data science process, including those leading to the development of aforementioned biomarkers, there are numerous human judgment calls that must be made. These decisions include but are not limited to data preprocessing (e.g., how to impute missing values, how to handle outliers or erroneous measurements, how to filter of genes or handle correlated variables) and modeling choices (e.g., which models, which tuning parameters). When left unchecked, these necessary but often arbitrary choices introduce sources of uncertainty that are ignored in the traditional statistical uncertainty quantification, unknowingly alter downstream conclusions, and lead to poor generalizability in new patient cohorts.

Recently, the Predictability, Computability, and Stability (PCS) framework[6,7] was developed to provide an overarching philosophical and practical framework for both stress-testing these choices and mitigating their unwanted impacts so as to bolster more reliable, trustworthy scientific conclusions. In short, the PCS framework advocates the need for transparent documentation of human judgment calls throughout the analysis pipeline and revolves around three core principles – *predictability* as a quantifiable assessment of whether the model sufficiently captures reality, *computability* as a necessary consideration in algorithmic design and data collection, and *stability* across data and model perturbations as a minimum requirement for veridical (trustworthy) science. These principles were originally motivated by extensive interdisciplinary research[8,9] and have since led to numerous novel and extensively-validated scientific findings including epistatic mechanisms,[10,11] stable biomarker discoveries,[12,13] and interpretable clinical decision-making.[14,15]

In this work, our primary aim is to ease the translation of MPS2 into routine clinical care by substantially reducing the number of required genes in the test while preserving its high accuracy for predicting high-grade PCa. To this end, we leverage the PCS framework to perform a reliable, stability-driven selection of genes and distill the 18-gene MPS2 test into a simplified 7-gene MPS2 test (s^7MPS2) while maintaining its high predictive power. Through this investigation, we demonstrate the accuracy and robustness of s^7MPS2 as well as the original MPS2 test to various data preprocessing and modeling choices, thereby bolstering their trustworthiness and generalizability.

Materials and methods

Development cohort

The same development cohort used to build the original MPS2 models was used.[5] Briefly, prebiopsy urine samples (first-catch urine following digital rectal examination) were prospectively collected at the University of Michigan from patients presenting for 12-core or greater prostate biopsy due to elevated PSA levels (3–10 ng/mL) from 2008 to 2020. A total of 761 samples were included in the final development cohort. We defer additional details to Tosoian et al. (2024).[5]

External validation cohort

The external validation cohort was the same one used in the original MPS2 study.[5] The cohort consisted of 743 patients in the prospective NCI EDRN PCA3 Evaluation Trial.[16] This trial enrolled consecutive patients presenting for biopsy across 11 academic centers, primarily due to elevated PSA levels or abnormal digital rectal examination findings.

Data preprocessing of gene expression data

Using qPCR profiling from OpenArray™, we measured gene expression in each urine sample via the cycle threshold (Ct), or the number of amplification cycles required for sample fluorescence to exceed the background level. Lower Ct values suggest higher gene expression. In this work, we focus our analysis on the 54 genes nominated in the MPS2 study.[5] Then, because of the inevitably many different but equally-reasonable choices that could be made during the data preprocessing, we chose to preprocess this gene expression data using four different data preprocessing pipelines (detailed in Appendix A) and proceeded to evaluate the robustness of our model and conclusions across each of these data preprocessing pipelines in accordance with the PCS framework for veridical data science.[6,7] This helps to ensure that the downstream scientific findings are not solely reliant on a particular data preprocessing decision.

Predictor and outcome variables

Throughout this study, the set of predictor variables included the 54 nominated genes (from Section Data Preprocessing of Gene Expression Data) as well as available clinical variables that are generally associated with high-grade PCa (age, race, family history of prostate cancer, abnormal DRE, prior negative biopsy, and PSA).[17] These variables were used to predict the primary outcome of interest – namely, high-grade PCa, defined as a dichotomous variable indicating whether the PCa is Grade Group 2 or higher. This is considered clinically significant PCa in this patient group (i.e., patients who should then undergo MRI or biopsy).

Modeling choices

We considered many different statistical and/or machine learning models for predicting high-grade PCa. Namely, we considered common statistical or machine learning models such as ordinary logistic regression, logistic regression with L_1 (LASSO) regularization,[18] L_2 (ridge) regularization,[19] and combined L_1+L_2 (elastic net) regularization,[20] random forests (RF),[21] gradient boosting decision trees,[22] and RuleFit.[23] We also investigated recently developed tree-based machine learning methods including random forest+ (RF+),[24] a PCS-guided generalization of random forests which combines the strength of both linear models and nonlinear trees, and fast interpretable greedy-tree sums (FIGS),[25] which grows a flexible but controllable number of shallow decision trees in summation. We focused primarily on these interpretable linear and tree-based models given our goal of identifying important genes for reliable biomarker development. Moreover, we note that tree-based machine learning models are often well-suited for biological tasks such as this, in part due to the resemblance between the thresholding behavior of decision trees and the on-off switch-like behavior commonly thought to govern genetic processes.[26] Specific implementations of each model and their hyperparameters are detailed in Table 1.

Table 1. Prediction methods, software implementations, and hyperparameters under study.

Prediction Method	Implementation	Hyperparameters*
Logistic regression	`LogisticRegression()` in sklearn python package	No hyperparameters
Logistic regression with L_1 (Lasso) regularization	`LogisticRegression(penalty="l1")` in sklearn python package	C = 10^i for i = -3, -2.5, 2, …, 2, 2.5, 3
Logistic regression with L_2 (ridge) regularization	`LogisticRegression(penalty="l2")` in sklearn python package	C = 10^i for i = -3, -2.5, 2, …, 2, 2.5, 3
Logistic regression with combined L_1+L_2 (elastic net) regularization	`LogisticRegression(penalty="elasticnet")` in sklearn python package	C = 10^i for i = -3, -2.5, 2, …, 2, 2.5, 3 l1_ratio = 0.1, 0.25, 0.5, 0.75, 0.9
Random forest	`RandomForestClassifier()` in sklearn python package	min_samples_leaf = 1, 3, 5, 10 n_estimators = 500
Gradient boosting decision trees	`GradientBoostingClassifier()` in sklearn python package	learning_rate = 0.05, 0.1, 0.15, min_samples_leaf = 1, 5, 10 max_depth = 3, 5 n_estimators = 500
RuleFit	`RuleFitClassifier()` in imodels python package	max_rules = 5, 10, 30, 50
Random forest+	`RandomForestPlusClassifier()` in imodels python package	Default hyperparameter grid used
Fast interpretable greedy-tree sums	`FIGSClassifier()` in imodels python package	max_rules = 5, 10, 12, 15, 20, 30, 50

* Hyperparameters were tuned using 5-fold cross-validation. Unless specified, the default hyperparameters for each method were used.

Model development: PCS ranking

Leveraging expression data from the 54 nominated genes and available clinical variables (age, race, family history of prostate cancer, abnormal DRE, prior negative biopsy, and PSA),[17] we built the simplified MPS2 model (sMPS2) to predict high-grade PCa defined as grade group 2 or higher PCa. More specifically, using the Development Cohort data, we developed and applied the PCS ranking procedure, which consists of three main stages (Figure 1) – (1) a prediction check stage, where we evaluated model prediction performance, (2) a stability-driven gene ranking stage, where we ranked the importance of each gene using the models that have passed the prediction check in stage 1, and (3) a selection of stable genes stage, where we identified the most stably important genes for use in the final sMPS2 model. Each of these stages is heavily rooted in the PCS framework for veridical data science.[6,7] This work builds upon similar ideas from the RF literature[9,10,11] and extends them to different classes of methods. Further details regarding the PCS ranking can be found in Appendix B alongside a supplementary PCS documentation on GitHub with extensive justification for the many human judgment calls made throughout the data analysis pipeline (https://github.com/Yu-Group/sMPS2). We note that proper data splitting is a crucial ingredient to enable the generalizability of our developed model. We thus outline the data splitting procedure in the supplementary materials in Figure S1.

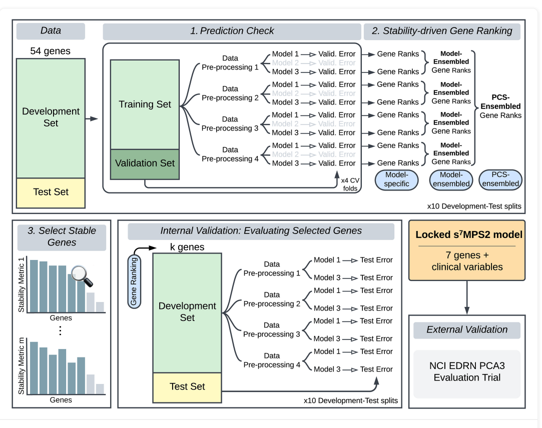

Figure 1. To develop the s^7MPS2 model, we first split the Development cohort data into a Development and Test Set and leveraged the Development Set to (1) conduct a prediction check, where we evaluated the prediction performance for various machine learning models across four different data preprocessing pipelines, filtering out models with poor prediction performance, and (2) subsequently rank the most important genes across all data preprocessing pipelines and prediction-checked models. Only the most important and stable genes, measured according to multiple stability metrics (stage 3), were selected to be used in the final locked s^7MPS2 model. To validate the quality and size of the gene panel used in s^7MPS2, we performed an internal validation, examining the test prediction performance using alternative gene rankings and panel sizes (k). We lastly validated the final, locked s^7MPS2 model on a blinded, external validation set using the NCI EDRN PCA3 Evaluation Trial.

Internal validation for evaluating selected genes

To evaluate the gene rankings from Stage 2 as well as the choice of the number of selected genes, we performed an internal validation using the test set from each of our 10 Development-Test splits (the same splits used in Stage 1). That is, for each Development-Test split, gene ranking from that given Development-Test split, and choice of k (k = 1, 2, 3, 4, 5, 6, 7, 10, 12, 15, 17, 20, 25, 30, 40, 54), we (a) took the top k-ranked genes and the available clinical features as covariates, (b) trained each prediction-checked model on each preprocessed Development set split, and (c) evaluated the prediction error (i.e., AUROC, AUPRC, and classification accuracy) on the test set. These errors were then averaged across the 10 Development-Test splits. We examined these averaged prediction errors across different choices of the gene panel size k and prediction models. Note that we also evaluated the different ways of obtaining the gene rankings (i.e., model-specific, model-ensembled, and PCS-ensembled).

Final simplified MPS2 model (sMPS2)

The final covariate gene set in the simplified MPS2 model consisted of the 6 topmost important stable genes: *T2:ERG*, *SCHLAP1*, *OR51E2*, *TFF3*, *PCAT14*, and *PCA3*. Using these 6 genes and 6 clinical features (age, race, family history of prostate cancer, abnormal DRE, prior negative biopsy, and PSA) as covariates, we trained a logistic regression with L_2 (ridge) regularization to predict high-grade PCa using the full Development Cohort dataset. This final trained model is referred to as the s^7MPS2 model as it requires the measurement of 7 genes (i.e., the 6 genes used as covariates and the reference gene *KLK3* which is necessary for data preprocessing). Here, we used a logistic regression with L_2 (ridge) regularization due to its strong performance in the internal validation assessment (Section Internal Validation for Evaluating Selected Genes). Since the inclusion of prostate volume in clinical models is well-known to improve the prediction of high-grade PCa,[27,28] we also trained an analogous model, termed s^7MPS2+, which includes all of the covariates in s^7MPS2 plus prostate volume, for use when a patient's prostate volume is readily available. For comparison, we also investigated and provided the external validation results for the s^8MPS2 (and s^8MPS2+) model, which is analogous to s^7MPS2 (and s^7MPS2+) but includes the top 7 most stable genes as covariates (*T2:ERG*, *SCHLAP1*, *OR51E2*, *TFF3*, *PCAT14*, *PCA3*, and *APOC1*). Finally, we calibrated the models to account for differences in outcome prevalence between the development and the validation cohorts following a logistic recalibration method[29] as in Tosoian et al. (2024).[5] These calibrated sMPS2 models were locked prior to external validation.

Model validation on blinded, external cohort

We evaluated the final locked sMPS2 models on a blinded, external cohort (n = 743) from the NCI EDRN PCA3 Evaluation Trial[16] described in Section External Validation Cohort. We importantly note that this external validation cohort data was only accessible by two investigators (C.X. and Y. Zheng), who performed the validation. The AUROC from the locked sMPS2 models were compared against MPS[3] and MPS2.[5] These AUROCs were then used to approximate the minimum test tradeoff (MTT) for an additional predictor.[30] Specifically, the test tradeoff is defined as the minimum number of data collections per true positive to yield a positive net benefit.[30] The MTT helps to formally quantify the improvement in risk prediction while accounting for the additional data collection cost of the 18-gene MPS2, compared to the smaller sMPS2 models. Moreover, we also computed a range of MTT corresponding to the range of mean AUROCs from different data preprocessing choices in order to shed light on how this important source of uncertainty might affect the MTT.

Results

Development of the simplified MyProstateScore 2.0 (sMPS2) model

Grounded by the PCS framework for veridical data science,[6] we developed a stability-driven machine learning pipeline, termed PCS ranking, to build a robust and accurate risk score model of high-grade PCa using substantially fewer genes than MPS2. In this PCS-guided development pipeline, we rigorously assessed the accuracy and stability of modeling results across both data preprocessing and modeling pipelines in order to account for the inherent uncertainty arising from such human judgment calls and thus more faithfully capture the uncertainty and generalizability when deployed in reality.[6] Briefly, the PCS ranking pipeline for sMPS2 consists of three main stages: (1) A prediction check stage, where we evaluated the prediction performance for a variety of machine learning models across four different data preprocessing pipelines and filtered out models with poor prediction performance; (2) A stability-driven gene ranking stage, where we ranked the importance of each gene according to both its magnitude of importance and the stability of its importances across model fits, data preprocessing pipelines, and data splits; and (3) A selection of stable genes stage, where we selected the set of most stable important genes for use in the final, locked sMPS2 model. Details regarding each step are provided in Section Model Development: PCS Ranking and Appendix B. Here, by identifying and focusing on the most important genes that were highly stable across both data preprocessing and modeling choices, we ensure that the final locked sMPS2 model is not solely a random artifact resulting from human analysis decisions, but rather a robust clinical risk model, which we will show to have highly predictive and generalizable performance (Sections Internal assessment and validation of the sMPS2 models and External Cohort Validation).

Prediction check of machine learning models via cross validation

As part of the prediction check stage in the development of sMPS2 (Figure 1), we assessed the cross-validation prediction performance for models trained with all 54 genes and available clinical variables across nine different machine learning models and four different preprocessing pipelines. We summarize the AUROC results from 4-fold CV repeated over 10 Development-Test splits in Figure 2 and provide the analogous AUPRC and classification accuracy results in Figure S2. As shown in Figure 2(A), the linear-based models (i.e., both regularized and unregularized logistic regression) tended to outperform the non-linear tree-based models (i.e., RF, GBDT, RuleFit, and FIGS), suggesting some smooth underlying structure in the data which can be more easily captured via linearity as opposed to trees (which are non-smooth piecewise constant functions). This is further supported by the observation that RF (AUROC 0.769) yielded a lower AUROC than RF+ (AUROC 0.781), a generalization of RF which models both smooth linear structure and nonlinear tree structure.

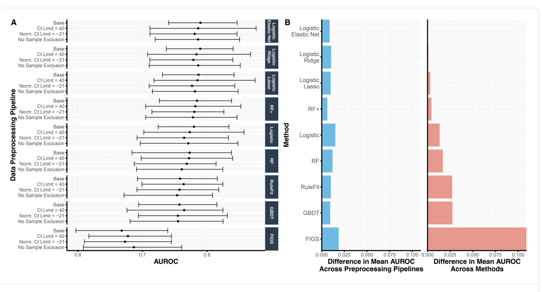

Figure 2. (A) For each choice of data preprocessing and prediction model, the validation AUROC, averaged across 4 CV folds and 10 repeated development-test splits, is shown. The error bars represent the inner 95% quantile range of the distribution of AUROCs. That is, across the 40 AUROCs, generated from 4-fold CV over 10 repeated Development-Test splits, the error bars correspond to the lower 2.5th and upper 97.5th quantiles of the distribution. (B) We compare the variation in AUROC across data preprocessing pipelines (left) and methods (right). In the left subplot, we show the range of mean AUROCs across the four data preprocessing pipelines for each method. In the right subplot, we show the difference between the mean AUROC from each method and the best performing method (i.e., logistic regression with the elastic net penalty) across all data preprocessing pipelines. The difference in AUROC across data preprocessing pipelines is substantially smaller than that across prediction methods, suggesting that the development pipeline and downstream findings are robust to data preprocessing choices.

Remarkably, the variation in prediction accuracy across data preprocessing pipelines was substantially smaller than the variation in prediction accuracy across models. Figure 2(B) shows the range of mean AUROCs across the four data preprocessing pipelines for each method (Figure 2(B), left), compared to the difference between each method's mean AUROC and that of the best-performing method (i.e., logistic regression with elastic net regularization) (Figure 2(B), right). Across all methods, the range of mean AUROCs across data preprocessing pipelines never exceeded 0.020 whereas the difference between logistic regression with elastic net regularization (the best-performing model) and RuleFit, GBDT, and FIGS were larger, exhibiting differences of 0.027, 0.027, and 0.109, respectively. These observations provide evidence that the trained models are not highly dependent on human choices made during the data preprocessing pipeline – a crucial stability check for fostering trust in our model development process.

Before proceeding to stage 2, we ultimately used this prediction check to filter out models with poor prediction performance, a possible indicator that the model does not accurately reflect reality and would generate unreliable interpretations.[6] Here, we chose to use ordinary logistic regression as the "reference" model given its simplicity yet decent cross-validated prediction performance (AUROC 0.772) in this problem, and we dropped all models with worse prediction performance than logistic regression. Specifically, this prediction check excludes RuleFit, GBDT, and FIGS from the remainder of the analysis. Note that though RF (AUROC 0.769) has slightly lower prediction performance than logistic regression on average across the different data preprocessing pipelines, we did not exclude RF since at least one of its data preprocessing pipelines led to higher accuracy than that for logistic regression. In other words, the uncertainty due to data preprocessing is larger than the modeling difference between RF and logistic regression. We hence deemed that RF passed the prediction check and included RF in the remainder of the analysis.

Stability-driven genes associated with high-grade prostate cancer

Having filtered out poor-performing prediction models and established that the prediction-checked models are indeed robust to different data preprocessing choices, we identified genes, which were both ranked highly important and highly stable across the four data preprocessing pipelines, six prediction-checked models, and ten development-test splits (i.e., 4 × 6 × 10 = 240 combinations), for use in the sMPS2 model.

Top-ranked genes across all data preprocessing pipelines and models

In Figure 3(A), we show the mean ranking of each gene across the 240 preprocessing-model-split combinations alongside numerous stability metrics, including the standard deviation of each gene's ranking (Figure 3(B)) and the proportion of times (out of 240) that the gene ranked in the top 5, 10, or 17 (out of 54) (Figure 3(C)–(E)). The top six-ranked genes *T2:ERG*, *SCHLAP1*, *OR51E2*, *TFF3*, *PCAT14*, and *PCA3* were all highly stable, each appearing in the top 10 ranked genes in more than 70% of the preprocessing-model-split combinations. The seventh-ranked gene, *APOC1*, also appears to be stably important; however, its stability declines when using only two of the four logistic-based regression models in the PCS-ensembled gene rankings (Figure S3). Notably, these top seven-ranked genes, selected from the 54 candidate genes under consideration, are a subset of the genes used in the validated MPS2 model.[5]

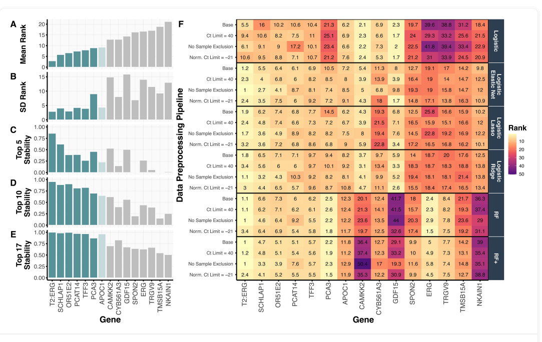

Figure 3. We summarize the top 15 ranked genes, according to **(A)** their mean gene ranking across four data preprocessing pipelines and six prediction-checked models, alongside **(B)** the variability of their gene rankings as measured by the standard deviation (SD) of this distribution and **(C-E)** the proportion of times that the gene appeared in the top 5, 10, and 17 genes. The top six genes were selected for the s^7MPS2 model as they all achieved a low (i.e., highly important) mean rank (< 10) and good stability across the various metrics. The *APOC1* gene was also considered and used in the s^8MPS2 model. **(F)** The heatmap shows a more granular view of the gene rankings, displaying the mean gene ranking per data preprocessing and model choice, averaged across 10 Development-Test splits. While the various data preprocessing and model choices generally yield similar gene rankings for the genes used in the simplified MPS models, logistic regression yields the most different gene rankings compared to the other models. In particular, it is the driving factor for *PCA3*'s large SD in (B).

When examining the gene rankings per data preprocessing pipeline and method in Figure 3(F), we confirmed the robustness and stability of these trained models across data preprocessing choices, not only in terms of their resulting prediction accuracy as discussed in Section Prediction Check of Machine Learning Models via Cross Validation, but also in terms of their most important genes and model architecture. Figure 3(F) further reveals that the two genes *T2:ERG* and *PCA3* comprising the original MPS model were stably ranked as the top two most important genes according to RF and RF+. Moreover, the top 6 genes were particularly stable across the regularized logistic regression models, RF, and RF+ fits while the ordinary logistic regression model generally produced a different set of gene rankings. Interestingly, the ordinary logistic regression model did not rank *PCA3* highly and is the main source of instability seen in the high SD for *PCA3* in Figure 3(B).

Top-ranked genes from specific models

Beyond these top genes, Figure 3(F) illuminates several additional interesting gene ranking patterns across the different models and data preprocessing pipelines. First, there are genes, such as *CAMKK2* and *GDF15*, that tend to be more highly ranked when considering only linear structure (i.e., in the logistic-based models) while other genes, such as *ERG* and *TRGV9*, tend to be more highly ranked when allowing for nonlinear structures (i.e., in the tree-based models, RF and RF+). However, given that these prediction models yielded similar prediction accuracies, it's unclear which model's interpretation to trust over another. We hence selected the top 6 genes (*T2:ERG*, *SCHLAP1*, *OR51E2*, *TFF3*, *PCAT14*, and *PCA3*), together with the reference gene *KLK3*, for use in the final simplified 7-gene MPS2 model (s^7MPS2), as these genes were highly stable across the various data-preprocessing, modeling, and Development-Test split combinations. Given the borderline stability status of *APOC1*, we also developed a simplified 8-gene MPS2 model (s^8MPS2), which includes those genes in s^7MPS2 along with *APOC1*.

Internal assessment and validation of the sMPS2 models

Thus far, we have been primarily guided by the stability of the feature rankings when selecting the number of top-ranked genes in our final simplified MPS2 model. To assess the impact of this choice of the gene panel size (i.e., the number of top-ranked genes used in the model) on the prediction accuracy, we performed an internal validation assessment using the test set (details in Section Internal Validation for Evaluating Selected Genes). Figure 4 highlights the test prediction accuracies when using the logistic regression model with ridge regularization (other model results can be found in Figure S4). Taking the top 7 PCS-ensembled genes (as in s^8MPS2) yielded the highest test AUROCs in the base and the Ct Limit = 40 data preprocessing pipelines (0.811 and 0.807, respectively) while also demonstrating competitive performance in the remaining two data preprocessing pipelines. The top 6 PCS-ensembled genes (as in s^7MPS2) yielded similarly strong AUROCs, showing only a 0.01 drop in AUROC compared to taking the top 7 PCS-ensembled genes across all data preprocessing pipelines.

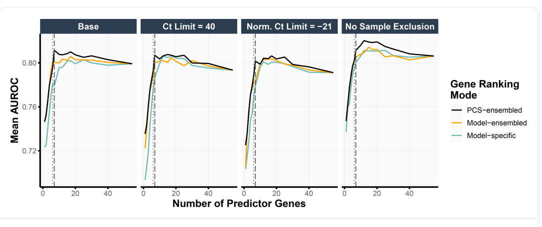

Figure 4. We show the mean AUROC, evaluated on the test set, when training the logistic regression model with ridge regularization using various choices of gene panel sizes (x-axis), data preprocessing pipelines (subplots), and gene rankings (color). The PCS-ensembled gene rankings yield the highest test AUROCs compared to other procedures for obtaining the gene rankings. Moreover, using 7 predictor genes (vertical dashed line) as in s^8MPS2 gave the highest test AUROC in the base and Ct Limit = 40 data preprocessing pipelines and very competitive prediction performance in the remaining data preprocessing pipelines. Taking 6 predictor genes (vertical dotted line) as in s^7MPS2 gave slightly lower but similarly competitive test prediction performance as 7 predictor genes.

Moreover, like the cross-validation prediction accuracies (Section Prediction Check of Machine Learning Models via Cross Validation), these test prediction accuracies were very stable across the different data preprocessing choices. In particular, the AUROCs when taking the top 6 and 7 PCS-ensembled genes ranged between [0.788, 0.801] (difference of 0.013) and [0.801, 0.811] (difference of 0.010), respectively, across the different data preprocessing pipelines. When taking the top 17 genes using logistic regression with elastic net regularization as done in the original MPS2 model, the test prediction performance was also highly stable across data preprocessing pipelines, giving an AUROC range of [0.800, 0.811] (difference of 0.012). This demonstrates the robustness of both the sMPS2 models and the original MPS2 model against alternative data preprocessing choices and indicates that these potentially different, but equally-reasonable choices do not solely drive downstream conclusions.

We further assessed the impact of our stability-driven PCS-ensembled ranking approaches, which averages the gene rankings across both data preprocessing pipelines and models, as compared to alternative approaches – namely, a model-ensembled approach, which averages the gene rankings across models only, and a model-specific approach, which produces a unique gene ranking per model and data preprocessing choice (details in Appendix B.2). Across all data preprocessing pipelines and gene panel sizes, the top genes according to the PCS-ensembled gene rankings led to higher prediction accuracies than that from the model-ensembled or the model-specific gene rankings (Figure 4). This pattern also holds across different choices of prediction models (Figure S4).

External cohort validation

When evaluated on the NCI EDRN PCA3 Evaluation trial,[16] the locked s^7MPS2 model yielded an AUROC of 0.784 (95% confidence interval [CI], 0.742–0.825) for predicting high-grade PCa (Figure 5). This was a 4.7% improvement over MPS, which gave an AUROC of 0.737 (95% CI, 0.694–0.780), and only a 2.3% drop relative to the more complex 18-gene MPS2 model, which gave an AUROC of 0.807 (95% CI, 0.769–0.846). In the case when prostate volume is available, the s^7MPS2 + model gave an AUROC of 0.806 (95% CI, 0.768–0.845), which was only a 1.2% drop relative to MPS2+ (AUROC: 0.818, 95% CI: 0.781–0.855). We also compared s^7MPS2 to s^8MPS2 in Figure 5, showing that the improvement when adding one additional gene (*APOC1*) is small. s^8MPS2 and s^8MPS2 + yielded AUROCs of 0.785 (95% CI, 0.744–0.826) and 0.809 (95% CI, 0.771–0.847), respectively. Importantly, the drops in AUROC (2.3%/2.2% for s^7MPS2/s^8MPS2 and 1.2%/0.9% for s^7MPS2+/s^8MPS2 + relative to MPS2 and MPS2+, respectively) are within the 1–2% uncertainty intervals induced by different data preprocessing choices. These AUROCs confirm the high diagnostic accuracy of the sMPS2 models as well as their similarly strong performance compared to the MPS2 models.

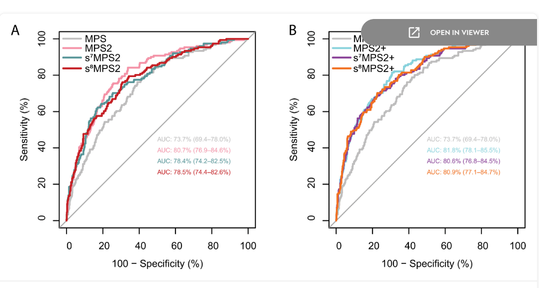

Figure 5. The AUROC curves from the blinded external validation cohort are shown for various urine-based biomarker tests without prostate volume (MPS, MPS2, s^7MPS2, and s^8MPS2, left) and with prostate volume (MPS2+, s^7MPS2+, and s^8MPS2+ right). The 7- and 8-gene simplified MPS2 models (s^7MPS2/s^8MPS2 and s^7MPS2+/s^8MPS2+) yield higher AUROCs compared to the 2-gene MPS model and competitive AUROCs compared to the 18-gene MPS2 model.

Get the support you need to get published Sage Author Services

Articles In the Same Collection(s):
Collection: EDRN 2023
Open Access
Polyclonal plasma cell (PolyPC) signature as a key indicator for predi...
Show Details
Open Access
Microbial biomarker development for detection and prognosis of early-s...
Show Details
Open Access
Development of a multiplexed lateral flow assay for the serologic dete...
Show Details

Sage Journal Reviewer Gateway Resources for peer reviewers Find out more >>

Similar articles:
Restricted access
Development and Validation of a UK-Specific Prostate Cancer Staging Predictive Model: UK Prostate Cancer Tables
Show Details
Open Access
Advancements in Biomarkers of Prostate Cancer: A Review
Show Details
Open Access
Biomarkers for prostate cancer detection and risk stratification
Show Details
View More

Sage recommends:
SAGE Knowledge
Entry
Recursive Partitioning
Show Details
SAGE Knowledge
Entry
Decision Curve Analysis
Show Details
SAGE Knowledge
Book chapter
Unlocking the Potential of Biomarkers in Targeted Oncology
Show Details
View More

However, from a clinical perspective, it is also important to evaluate the performance of a practical clinical testing approach using a specific decision threshold that yields high sensitivity for high-grade PCa. In particular, since the traditional clinical approach of biopsying all men with elevated PSA is highly sensitive for PCa detection but leads to excess unnecessary biopsies, secondary tests such as biomarkers have focused on improving the specificity of screening while preserving the high sensitivity attained from a "biopsy all" approach. This emphasis on high sensitivity is reflected in current clinical guidelines, which propose the use of "biomarkers that improve the specificity of screening".[31] Considering the relative harms of false-negative and false-positive testing results and the proposed role of biomarkers for rule out testing (i.e., for ruling out the possibility of prostate cancer), we thus assessed thresholds providing 95% sensitivity for high-grade cancer. At this 95% sensitivity, s^7MPS2/s^8MPS2 provided a specificity of 32%/30%, a negative predictive value (NPV) of 96%/96%, and positive predictive value (PPV) of 26%/26% (Table 2). This corresponds to an estimated reduction of 318/297 unnecessary biopsies avoided per 1000 patients based upon the s^7MPS2/s^8MPS2 models. More interestingly, both s^7MPS2 + and s^8MPS2 + achieve similar, if not higher, specificity (40.7%), NPV (96.8%), and PPV (28.9%) than MPS2+ (40.5% specificity, 96.8% NPV, 28.9% PPV) and leads to an estimated 407 unnecessary biopsies avoided per 1000 patients under this clinical testing approach at 95% sensitivity. These estimates of the number of unnecessary biopsies are also complemented by a low false negative rate across all of the sMPS2 and MPS2 models. For each of these models, an estimated 52 biopsies would have been missed per 1000 patients at the 95% sensitivity level, had no other screening been undertaken. Detailed performance of the sMPS2 models at a range of high sensitivities with and without clinical factors are provided in Supplementary Table 1 and 2.

Table 2. Performance of MPS2, MPS2+ and corresponding sMPS2 models (7- and 8-biomarkers) in the validation cohort at the 95% sensitivity level.

Model	Sensitivity (%)	Specificity (%) (TN/(TN + FP))	Negative Predictive Value (%) (TN/(TN + FN))	Positive Predictive Value (%) (TP/(TP + FP))	Estimated unnecessary biopsies avoided per 1000 patients
MPS	95.0	23.0 (136/592)	94.4 (134/144)	23.9 (143/599)	230
MPS2	95.0	37.0 (219/592)	96.5 (219/227)	27.7 (143/516)	370
s^7MPS2	95.0	31.8 (188/592)	95.9 (188/196)	26.1 (143/547)	318
s^8MPS2	95.0	29.7 (176/592)	95.7 (176/184)	25.6 (143/559)	297
MPS2+	95.0	40.5 (240/592)	96.8 (240/248)	28.9 (143/495)	405
s^7MPS2+	95.0	40.7 (241/592)	96.8 (241/249)	28.9 (143/494)	407
s^8MPS2+	95.0	40.7 (241/592)	96.8 (241/249)	28.9 (143/494)	407

TN = true negatives; FP = false positives; FN = false negatives; TP = true positives

Finally, we evaluated the MTT[30] to formally assess the tradeoff between the improvement in risk prediction and the additional cost of data collection between the 18-gene MPS2/MPS2 + and the smaller 7-gene s^7MPS2/s^7MPS2 +. At a 20% prevalence of high-grade prostate cancer (as seen in the validation cohort), the MTT is 229.5 and 122.0 for MPS2 + versus s^7MPS2 + and MPS2 versus s^7MPS2, respectively. Thus, at least 229 (or 122) sets of 18-gene MPS2+ (or MPS2) tests would be needed for every true positive to yield a positive net benefit relative to using s^7MPS2+ (or s^7MPS2). Furthermore, when incorporating the 1–2% uncertainty due to data preprocessing choices, the MTT ranges from [125.9, 1368.8] and [85.5, 214.7] for the s^7MPS2 + and s^7MPS2 tests, respectively. These large MTT ranges are a byproduct of the inverse decaying nature of MTT. Intuitively, as the prediction gap between the sMPS2 and MPS2 models decrease, the net benefit of using MPS2 over the sMPS2 models decreases rapidly and hyperbolically. Consequently, the small difference in AUROCs between the sMPS2 models and MPS2 after accounting for the uncertainty in data preprocessing translates to extremely large MTTs seen above. This tradeoff analysis suggests that MPS2 is not practically worthwhile from a data collection cost perspective, especially after accounting for the often-neglected uncertainty due to data cleaning, and further supports the accessibility and utility of the sMPS2 models for clinical care.

Discussion

From a practical standpoint, it is preferable to have fewer variables in clinical assays, as this simplifies interpretation, reduces the potential for variation in measurements, and decreases costs. However due to the heterogeneity of prostate cancer, it may be necessary to increase the number of biomarkers to capture different forms of aggressive prostate. For example, the recurrent *T2:ERG* fusion of prostate cancer occurs in just about half of the patients. With exhaustive screening previously done[5] and extensive stability testing across various perturbation situations in this study, we identified the most stable and important biomarkers that contribute to prediction of clinically significant prostate cancer. Unsurprisingly, all of the top 6 most stable genes identified by the PCS ranking (*T2:ERG*, *PCA3*, *SCHLAP1*, *OR51E2*, *TFF3* and *PCAT14*) have been reported to be overexpressed in prostate cancer and their roles as biomarkers have been explored by independent studies.[32–39] The additional gene included in the s^8MPS2, *APOC1* (apolipoprotein C1), is also frequently reported to play a role in various cancer types including breast, ovarian, pancreatic, gastric cancers and be associated with tumor infiltrated macrophage.[40–44] It is further worth noting that the genes used in the s^7MPS2/s^8MPS2 models were selected out of 54 candidate genes and yet are a subset of the 18 genes used in the validated MPS2 model,[5] reinforcing their potential biological functions and prognostic values. Moreover, the resultant simplified MPS2 models built upon the aforementioned genes (s^7MPS2/s^8MPS2) achieved comparable performance as the original models on the external validation cohort (AUROC 0.784/0.785 vs 0.807) (within the uncertainty size of 1–2% from making different but reasonable data preprocessing choices); for the models with prostate volume, the difference became even smaller (AUROC 0.806/0.809 vs 0.818) (Figure 5). At the clinically relevant sensitivity of 95%, s^7MPS2/s^8MPS2 was parallel to the original model with about 96% NPV though the specificity decreased from 37% to 32%/30% while s^7MPS2+/s^8MPS2 + was equivalent to the original model in both NPV and specificity (Table 2). These findings, combined with the MTT analysis, suggest that clinical use of the simplified MPS2 models would benefit patients to a similar extent as the 18-biomarker MPS2, avoiding unnecessary biopsy while maintaining high sensitivity for high-grade cancers from early detection, at a substantially lower data collection cost.

From a methodological perspective, we demonstrate the value of the PCS framework for stress-testing various human judgment calls throughout the model development process. To facilitate both transparency and reproducibility, we provide extensive documentation and justification of our data preprocessing and model development process in the supplementary PCS documentation alongside all code on GitHub (https://github.com/Yu-Group/sMPS2) to reproduce our analysis. This work not only sheds light on such decisions in an effort towards more transparent and truthful science, but ensures that these different but equally-reasonable choices do not drastically impact downstream conclusions, fostering greater trust in the model. Going beyond this rigorous stress-testing, the PCS framework can also enhance the model itself. Here, we exploit the stability of similarly predictive models in order to simplify an 18-gene model into a 7-gene (or 8-gene) model for predicting high-grade PCa without substantially reducing its accuracy. Our overarching PCS ranking pipeline is more generally applicable, and as exemplified by the strong and generalizable prediction performance of the sMPS2 models, opens the door to more rigorous and trustworthy development of robust, highly-interpretable biomarker tests.

This study has some limitations. First the patient cohorts in this study were not suitable for comparing with MRI screening or evaluating the clinical benefit of combining MRI with biomarkers for screening. Due to the proposed use of the sMPS2 assay as a pre-mpMRI, prebiopsy test to rule out the need for mpMRI or biopsy, we excluded men with a history of prostate mpMRI and targeted biopsy. Second, although the total sample size of the training and testing cohorts are adequate, larger sample sizes are needed to evaluate the performance of sMPS2 in specific subpopulations, for example, African American, repeat biopsy population, and patients of non ETS fusion molecular subtype. Additionally, the performance metrics reported in this study are based on 95% sensitivity and 20% high-grade prevalence in the validation cohort (which is within reported range from 17% to 31%), representing a single data point to illustrate the clinical applicability. However, given that predictive values are affected by prevalence, the high-grade prevalence of the target population may vary from institute to institute and change over time as more intervention strategies come into play, re-calibrating or updating the model is necessary to improve the performance. Lastly, given the complex correlation structure of the data, we cannot guarantee that the identified 7 (or 8) genes used in the sMPS2 models are the best set of 7 (or 8) genes that lead to the highest prediction accuracy. Nonetheless, the current sMPS2 models demonstrate excellent robustness and generalizability at a reduced data collection cost compared to existing models, underscoring its clinical relevance and applicability.

Acknowledgments

This work was funded by the Michigan-Vanderbilt EDRN Biomarker Characterization Center (U2C CA271854) and the EDRN DMCC (U24 CA086368). The EDRN DMCC carried out analyses on the blinded validation cohort. Other sources of funding not involved in the design and conduct of the study included the Michigan Prostate SPORE (P50 CA186786), NCI Outstanding Investigator Award (Chinnaiyan, R35 CA231996), Johns Hopkins University Biomarker Reference Laboratory (U24 CA115102), NCI EDRN CVC (U01 CA113913), Prostate Cancer Foundation Young Investigator Award (Tosoian and Xiao), Prostate Cancer Foundation, Howard Hughes Medical Institute (Chinnaiyan), the American Cancer Society (Chinnaiyan) and the University of Michigan Health System-Peking University Health Science Center (UMHS-PUHSC) Joint Institute (Chinnaiyan, Xiao, and YZhang). We (Tang, Kenney, and Yu) also gratefully acknowledge partial support from NSF grants DMS-2413265 and DMS-2209975, NSF grant 2023505 on Collaborative Research: Foundations of Data Science Institute (FODSI), the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and 814639, and NSF grant MC2378 to the Institute for Artificial CyberThreat Intelligence and OperatioN (ACTION). Tang would like to acknowledge that the majority of the work presented was conducted while at the University of Michigan. The authors would also like to thank Steve Skates, the anonymous reviewers, and editor for their insightful comments that have greatly improved this work.

Declaration of conflicting interests

Dr Tang has a patent for a novel multiplex urine test for high-grade prostate cancer pending. Dr Zhang reported consulting fees from LynxDx outside the submitted work and has a patent for a novel multiplex urine test for high-grade prostate cancer pending. Dr Xiao reported grants from Prostate Cancer Foundation as well as consulting fees from LynxDx during the conduct of the study; and has a patent for a novel multiplex urine test for high-grade prostate cancer pending. Dr Kenney has a patent for a novel multiplex urine test for high-grade prostate cancer pending. Dr Tosoian reported consulting fees from LynxDx and equity interest from LynxDx outside the submitted work; and has a patent for a novel multiplex urine test for high-grade prostate cancer pending. Dr Chinnaiyan reported grants from the National Institutes of Health/National Cancer Institute, Prostate Cancer Foundation, and Howard Hughes Medical Institute; nonfinancial support from the American Cancer Society during the conduct of the study; and equity interest from LynxDx outside the submitted work; and has a patent for a novel multiplex urine test for high-grade prostate cancer pending. Dr Yu has a patent for a novel multiplex urine test for high-grade prostate cancer pending. No other disclosures were reported.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Tiffany M Tang

Data availability statement

As the data represented in this publication are from a blinded NCI-EDRN validation cohort, unblinding of the data is not allowed. To maintain patient confidentiality, no individual level data will be provided, only patient summaries. All code to reproduce this analysis can be found on GitHub: https://github.com/Yu-Group/sMPS2.

References

1. Etzioni R, Tsodikov A, Mariotto A, et al. Quantifying the role of PSA screening in the US prostate cancer mortality decline. *Cancer Causes Control* 2008; 19: 175–181.

2. Tomlins SA, Rhodes DR, Perner S, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 2005; 310: 644–648.

3. Tomlins SA, Aubin SMJ, Siddiqui J, et al. Urine TMPRSS2:ERG fusion transcript stratifies prostate cancer risk in men with elevated serum PSA. *Sci Transl Med* 2011; 3: 94ra72.

4. Moses KA, Sprenkle PC, Bahler C, et al. NCCN Guidelines® insights: prostate cancer early detection, Version 1.2023. *J Natl Compr Canc Netw* 2023; 21: 236–246.

5. Tosoian JJ, Zhang Y, Xiao L, et al. Development and validation of an 18-Gene urine test for high-grade prostate cancer. *JAMA Oncol* 2024; 10(6): 726–736.

6. Yu B, Kumbier K. Veridical data science. *Proc Natl Acad Sci U S A* 2020; 117: 3920–3929.

7. Yu B, Barter RL. *Veridical data science: The practice of responsible data analysis and decision making*. Cambridge, MA, USA: MIT Press, 2024.

8. Wu S, Joseph A, Hammonds AS, et al. Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proc Natl Acad Sci* 2016; 113: 4290–4295.

9. Basu S, Kumbier K, Brown JB,. et al. Iterative random forests to discover predictive and stable high-order interactions. *Proc Natl Acad Sci U S A* 2018; 115: 1943–1948.

10. Behr M, Kumbier K, Cordova-Palomera A, et al. Learning epistatic polygenic phenotypes with Boolean interactions. *PLoS One* 2024; 19: e0298906.

11. Wang Q, Tang TM, Youlton N, et al. Epistasis regulates genetic control of cardiac hypertrophy. *Res Sq [Preprint]* 2023. https://doi.org/10.21203/rs.3.rs-3509208/v1. 38045390; PMCID: PMC10690313.

12. Li X, Tang TM, Wang X, et al. A stability-driven protocol for drug response interpretable prediction (staDRIP), *arXiv [q-bioQM]*. 2020.

13. Irajizad E, Kenney A, Tang T, et al. A blood-based metabolomic signature predictive of risk for pancreatic cancer. *Cell Rep Med* 2023; 4: 101194.

14. Kornblith AE, Singh C, Devlin G, et al. Predictability and stability testing to assess clinical decision instrument performance for children after blunt torso trauma. *PLOS Digit Health* 2022; 1: e0000076.

15. Dwivedi R, Tan YS, Park B, et al. Stable discovery of interpretable subgroups via calibration in causal studies. *Int Stat Rev* 2020; 88: S135–S178.

16. Wei JT, Feng Z, Partin AW, et al. Can urinary PCA3 supplement PSA in the early detection of prostate cancer? *J Clin Oncol* 2014; 32: 4066–4072.

17. Thompson IM, Ankerst DP, Chi C, et al. Assessing prostate cancer risk: results from the prostate cancer prevention trial. *J Natl Cancer Inst* 2006; 98: 529–534.

18. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 1996; 58: 267–288.

19. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; 12: 55–67.

20. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 2005; 67: 301–320.

21. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.

22. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001; 29: 1189–1232.

23. Friedman JH, Popescu BE. Predictive learning via rule ensembles. *Aoas* 2008; 2: 916–954.

24. Agarwal A, Kenney AM, Tan YS, et al. MDI+: A Flexible Random Forest-Based Feature Importance Framework, *arXiv [statME]*. 2023.

25. Tan YS, Singh C, Nasseri K, et al. Fast Interpretable Greedy-Tree Sums, *arXiv [csLG]*. 2022.

26. Nelson DL, Lehninger AL, Cox MM. *Lehninger principles of biochemistry*. New York City, NY, USA: Macmillan, 2008.

27. Roobol MJ, Schröder FH, Hugosson J, et al. Importance of prostate volume in the European Randomised Study of Screening for Prostate Cancer (ERSPC) risk calculators: results from the prostate biopsy collaborative group. *World J Urol* 2012; 30: 149–155.

28. Ankerst DP, Till C, Boeck A, et al. The impact of prostate volume, number of biopsy cores and American urological association symptom score on the sensitivity of cancer detection using the prostate cancer prevention trial risk calculator. *J Urol* 2013; 190: 70–76.

29. Vergouwe Y, Nieboer D, Oostenbrink R, et al. A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med* 2017; 36: 4529–4539.

30. Baker SG, Kramer BS. Simple methods for evaluating 4 types of biomarkers: surrogate endpoint, prognostic, predictive, and cancer screening. *Biomark Insights* 2020; 15: 1177271920946715.

31. National Comprehensive Cancer Network, Prostate cancer early detection (version 2.2024), NCCN. https://www.nccn.org/guidelines/guidelines-detail?category=2&id=1460 (n.d., accessed September 20, 2024).

32. Tomlins SA, Day JR, Lonigro RJ, et al. Urine TMPRSS2:ERG Plus PCA3 for individualized prostate cancer risk assessment. *Eur Urol* 2016; 70: 45–53.

33. Prensner JR, Iyer MK, Sahu A, et al. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat Genet* 2013; 45: 1392–1398.

34. Xu LL, Stackhouse BG, Florence K, et al. PSGR, a novel prostate-specific gene with homology to a G protein-coupled receptor, is overexpressed in prostate cancer. *Cancer Res* 2000; 60: 6568–6572.

35. Garraway IP, Seligson D, Said J, et al. Trefoil factor 3 is overexpressed in human prostate cancer. *Prostate* 2004; 61: 209–214.

36. Shukla S, Zhang X, Niknafs YS, et al. Identification and validation of PCAT14 as prognostic biomarker in prostate cancer. *Neoplasia* 2016; 18: 489–499.

37. Rigau M, Morote J, Mir MC, et al. PSGR And PCA3 as biomarkers for the detection of prostate cancer in urine. *Prostate* 2010; 70: 1760–1767.

38. Vestergaard EM, Borre M, Poulsen SS, et al. Plasma levels of trefoil factors are increased in patients with advanced prostate cancer. *Clin Cancer Res* 2006; 12: 807–812.

39. Feng FY-C, Zhao S, Prensner J, et al. Investigating the long noncoding RNA SChLAP1 as a prognostic tissue and urine biomarker in prostate cancer. *J Clin Orthod* 2015; 33: 7–7.

40. Yang S, Du J, Wang W, et al. APOC1 Is a prognostic biomarker associated with M2 macrophages in ovarian cancer. *BMC Cancer* 2024; 24: 364.

41. Gu Q, Zhan T, Guan X, et al. Apolipoprotein C1 promotes tumor progression in gastric cancer. *Oncol Res* 2023; 31: 287–297.

42. Takano S, Yoshitomi H, Togawa A, et al. Apolipoprotein C-1 maintains cell survival by preventing from apoptosis in pancreatic cancer cells. *Oncogene* 2008; 27: 2810–2822.

43. Zhang H, Wang Y, Liu C, et al. The apolipoprotein C1 is involved in breast cancer progression via EMT and MAPK/JNK pathway. *Pathol Res Pract* 2022; 229: 153746.

44. Su W-P, Sun L-N, Yang S-L, et al. Apolipoprotein C1 promotes prostate cancer cell proliferation in vitro. *J Biochem Mol Toxicol* 2018; 32: e22158.

45. McCall MN, McMurray HR, Land H,. et al. On non-detects in qPCR data. *Bioinformatics* 2014; 30: 2310–2316.

Appendix A: Data Preprocessing of Gene Expression Data Details

We first preprocessed the gene expression data by following a similar data preprocessing procedure as taken in the original development of MPS2.[5] In this procedure, we first set the upper Ct value limit to 35. Specifically, Ct values greater than this limit were considered undetected and set to 35. Ct values from OpenArray™ that were "Undetermined" or "Inconclusive/No Amp" were also considered to be undetected and set to the upper Ct value limit of 35. We chose this value of 35 as opposed to the commonly used 40 since previous work showed that setting the Ct value limit to 40 can often introduce unwanted biases and that setting this limit to 35 can effectively reduce this bias for non-detects in qPCR.[45] Next, we computed the standard deviation (SD) across 3 technical replicates. If SD >= 1, the replicate farthest from the mean was removed; otherwise, all 3 replicates were kept. The average Ct value across these kept technical replicates was then calculated. As a filter for poor quality samples, all samples with an average Ct value of the reference gene *KLK3* above the 95th percentile were excluded. We next normalized the average Ct values for each target gene by *KLK3* using the formula -[average Ct of gene X - average Ct of *KLK3*]. Finally, z-score scaling was applied to the normalized average Ct before downstream model development and feature selection. We refer to this data preprocessing pipeline as the *base preprocessing pipeline*.

While the base preprocessing pipeline serves as a reasonable starting point, it is important to note that human judgment calls or choices such as the procedure for handling undetectable Ct values and poor quality samples were inevitably made. Hence, to improve the robustness of our model and conclusions against these preprocessing choices, we examined three alternative data preprocessing pipelines, each equally-reasonable but a slight modification of the base preprocessing pipeline.

1. **Ct limit = 40 preprocessing pipeline:** Rather than setting the upper Ct value limit to 35 for undetected replicates, we instead set the upper Ct value limit to 40. All other preprocessing steps remain unchanged from the base preprocessing pipeline.
2. **Normalized Ct limit = -21 preprocessing pipeline:** In the aforementioned data preprocessing pipelines, the Ct values for undetected replicates were set prior to the normalization of the Ct values. Thus, the normalized Ct value for undetected replicates differs between genes. For comparison, in this preprocessing pipeline, we instead replace the Ct values for all undetected replicates after Ct normalization to have a constant value of –21 (which was the lowest Ct value post-normalization). All other preprocessing steps remain unchanged from the base preprocessing pipeline.
3. **No sample exclusion preprocessing pipeline:** Rather than excluding all samples with an average Ct value of the reference gene *KLK3* above the 95th percentile, this preprocessing pipeline does not exclude any samples based upon their Ct value for the reference gene *KLK3*. All other data preprocessing steps remain unchanged from the base preprocessing pipeline.

While we focus our attention on the four described data preprocessing pipelines (i.e., the base preprocessing pipeline and three alternative preprocessing pipelines above) due to computational constraints, we acknowledge that these preprocessing pipelines encompass only a small fraction of all possible data preprocessing choices. Still, we believe that assessing model robustness and stability across these different preprocessing pipelines provides a valuable step towards scientific findings that are not solely reliant on particular (reasonable) data preprocessing decisions. In the spirit of transparency, we provide extensive documentation of our data preprocessing decisions and justify them when possible in the supplementary PCS documentation on GitHub (https://github.com/Yu-Group/sMPS2).

Appendix B: Model Development Details

B1. Prediction Check

In the first stage of our model development process, we conducted a prediction check, where we assessed model prediction performance in order to filter out models which may not accurately represent the scientific phenomena under study as suggested by their poor prediction performance.[6] Specifically, we first randomly split the Development Cohort (Section Development Cohort) into a Development (80%) and Test Set (20%). Then using the Development Set, we performed 4-fold cross-validation (CV). For each CV fold, we trained each model (described in Section Modeling Choices) on each of the four preprocessed datasets (described in Section Data Preprocessing of Gene Expression Data and Appendix A) using the training fold and evaluated the prediction error on the validation fold. The covariates used included the 54 nominated genes from the original MPS2 study and available clinical variables that are generally thought to be associated with high-grade PCa (age, race, family history of prostate cancer, abnormal DRE, prior negative biopsy, and PSA).[17] The cross-validation (CV) error was computed using three different evaluation metrics - area under the receiver operating characteristic curve (AUROC), area under the precision recall curve (AUPRC), and classification accuracy - and averaged across the four CV folds. This process was repeated for 10 different Development-Test splits. Models that did not outperform ordinary logistic regression in any of the data preprocessing pipelines across the three evaluation metrics were excluded from further analysis and failed the prediction check.

B2. Stability-driven Gene Ranking

For those models that passed the prediction check, we proceeded to compute a ranking of gene (or feature) importances for each preprocessing-model combination. More precisely, for each data preprocessing pipeline and model specification, we computed a measure of feature importance from the model fit on each CV training fold, averaged the feature importance measures across the 4 folds, and ranked the features according to this averaged feature importance for the given method. For random forest (RF) and random forest+ (RF+), we respectively used mean decrease in impurity (MDI)[21] and mean decrease in impurity+ (MDI+)[24] as the feature importance measures. For logistic regression and logistic regression with L_2 (ridge) regularization, we measured feature importance using the magnitude of the estimated regression coefficients. For logistic regression with L_1 (Lasso) or the combined $L_1 + L_2$ (elastic net) regularization, we measured feature importance by the number of times each feature had a non-zero coefficient across the 4 CV folds; if there are ties, we broke these ties based upon the magnitude of the estimated regression coefficients. Note that the covariate data was centered and scaled to have mean 0 and variance 1 prior to fitting these models. Additionally, while clinical variables were used as covariates in the trained models (and included in the feature importance computation), these clinical variables were dropped when computing the feature importance rankings since our aim is to identify the most important genes. Thus far, the aggregation of feature importances and resulting feature importance ranking is *model-specific*, i.e., each model or method yields its own feature importance ranking.

However, among the models that passed the prediction check and yielded fairly similar prediction performance, it is unclear whether or why one should trust the feature importance rankings from one method over another. Following the stability principle of the PCS framework,[6] we hence examined the features that were most stably important across these similarly-performing prediction models. To do so, we averaged the model-specific feature ranks across models, yielding a *model-ensembled* feature importance ranking, and across both data preprocessing pipelines and models, yielding a *PCS-ensembled* feature importance ranking. Note that this stability-driven feature ranking is performed per Development-Test split, using the same splits as that in the prediction check (Figure S1).

B3. Selection of Stable Genes

After obtaining the various gene rankings for each Development-Test split, we examined the stability of the top-ranked genes across the different Development-Test splits via various metrics (applying the stability principle in PCS to metric choices) in order to identify the sparse set of stable genes used in the final simplified MPS2 model. The first stability metric was the proportion of times each gene was ranked in the top k across all preprocessing-model specifications and Development-Test splits (4 data preprocessing pipelines × 6 models that passed the prediction check × 10 splits = 240 total configurations). Here, we tried several different choices of k: $k = 5$ and 10 to identify genes that were almost always the most important (~top 10% or 20%) out of the genes under consideration, and $k = 17$ for comparison with the original MPS2 study.[5] The second stability metric was the average PCS-ensembled feature importance ranking across all Development-Test splits. Note that this is equivalent to the average feature importance ranking across all data preprocessing pipelines, models, and Development-Test splits. The third stability metric is the standard deviation (SD) of the gene's importance rankings across all data preprocessing pipelines, models, and Development-Test splits. Using these various stability metrics, we identified the genes which were frequently ranked in the top 5, 10, and 17 features, had a high average feature ranking (i.e., ranked close to 1), and whose feature rank was highly stable (or had low SD or variability) across the different preprocessing pipelines, models, and Development-Test splits. We acknowledge that a heuristic approach, combined with expert domain knowledge, guided our final selection of genes and that no formal pre-specified thresholds were used for selection. However, these selected genes were locked prior to internal and external validation, and thus, these validation studies (detailed in Sections Internal Validation for Evaluating Selected Genes and Model Validation on Blinded, External Cohort) provide a proper, unbiased assessment of our model's generalizability.

Supplementary Material

Please find the following supplemental material available below.

For Open Access articles published under a Creative Commons License, all supplemental material carries the same license as the article it is associated with.

For non-Open Access articles published, all supplemental material carries a non-exclusive license, and permission requests for re-use of supplemental material or any part of supplemental material shall be sent directly to the copyright owner as specified in the copyright notice associated with the article.

sj-xlsx-1-cbm-10.1177_18758592241308755.xlsx | 19.59 KB

sj-xlsx-2-cbm-10.1177_18758592241308755.xlsx | 19.75 KB

Also from Sage

CQ Library – Elevating debate
Sage Data – Uncovering insight
Sage Business Cases – Shaping futures
Sage Campus – Unleashing potential
Sage Knowledge – Multimedia learning resources
Sage Research Methods – Supercharging research
Sage Video – Streaming knowledge
Technology from Sage – Library digital services

About
About Sage Journals
Accessibility guide
Historical content
Advertising disclaimer
Permissions
Terms of use
Sage discipline hubs
Sage microsites

Information for
Authors
Editors
Librarians
Promoters / Advertisers
Researchers
Reviewers
Societies
Frequently asked questions

Cancer Biomarkers
ISSN: 1574-0153
Online ISSN: 1875-8592

About Sage | Contact us | CCPA - Do not sell my personal information | CCPA | Privacy Policy

Copyright © 2025 by SAGE Publications