OXFORD

JNCI J Natl Cancer Inst (2020) 112(4): djz127

doi: 10.1093/jnci/djz127
First published online June 21, 2019
Article

ARTICLE

Overdiagnosis and Lives Saved by Reflex Testing Men With Intermediate Prostate-Specific Antigen Levels

Roman Gulati, Todd M. Morgan, Teresa A'mar, Sarah P. Psutka, Jeffrey J. Tosoian, Ruth Etzioni

See the Notes section for the full list of authors' affiliations.

Correspondence to: Roman Gulati, MS, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M2-B230, P.O. Box 19024, Seattle, WA 98109-1024 (e-mail: rgulati@fredhutch.org).

Abstract

Background: Several prostate cancer (PCa) early-detection biomarkers are available for reflex testing in men with intermediate prostate-specific antigen (PSA) levels. Studies of these biomarkers typically provide information about diagnostic performance but not about overdiagnosis and lives saved, the primary drivers of associated harm and benefit.
Methods: We projected overdiagnoses and lives saved using an established microsimulation model of PCa incidence and mortality with screening and treatment efficacy based on randomized trials. We used this framework to evaluate four urinary reflex biomarkers (measured in 1112 men presenting for prostate biopsy at 10 US academic or community clinics) and two hypothetical ideal biomarkers (with 100% sensitivity or specificity for any or for high-grade PCa) at one-time screening tests at ages 55 and 65 years.
Results: Compared with biopsying all men with elevated PSA, reflex testing reduced overdiagnoses (range across ages and biomarkers = 8.8–60.6%) but also reduced lives saved (by 7.3–64.9%), producing similar overdiagnoses per life saved. The ideal biomarker for high-grade disease improved this ratio (by 35.2% at age 55 years and 42.0% at age 65 years). Results were similar under continued screening for men not diagnosed at age 55 years, but the ideal biomarker for high-grade disease produced smaller incremental improvement.
Conclusions: Modeling is a useful tool for projecting the implications of using reflex biomarkers for long-term PCa outcomes. Under simplified conditions, reflex testing with urinary biomarkers is expected to reduce overdiagnoses but also produce commensurate reductions in lives saved. Reflex testing that accurately identifies high-grade PCa could improve the net benefit of screening.

The United States Preventive Services Task Force recommends offering prostate-specific antigen (PSA) screening for prostate cancer (PCa) to men age 55–69 years with a stated preference for screening after discussion with a provider (1). Available evidence indicates that screening reduces the risk of PCa death by 20% (2,3) but carries risks of false-positive tests, unnecessary biopsies, overdiagnosis, and overtreatment.

In practice, if a man elects screening, an intermediate PSA level (eg, PSA between 4.0 and 10.0 ng/mL) may prompt additional diagnostic (reflex) testing to further inform the choice to undergo prostate biopsy. Over the past decade, several blood, urine, and tissue biomarkers have been studied as potential reflex tests in this setting (4–9). For example, urine-based PCa biomarkers, such as quantitative TMPRSS2:ERG gene fusion (T2:ERG) and prostate cancer antigen 3 (PCA3) assays, and multiplex combinations, such as the Michigan Prostate Score (MiPS), have been shown to predict risk of PCa and high-grade PCa (10,11).

By identifying men at low risk for any or aggressive PCa, reflex testing seeks to reduce unnecessary biopsies and the diagnosis of "nonthreatening" disease while preserving survival benefit. Nonthreatening disease is often defined using clinicopathologic variables available at diagnosis (12–14). For example,

Received: March 31, 2019; **Revised:** May 30, 2019; **Accepted:** June 13, 2019

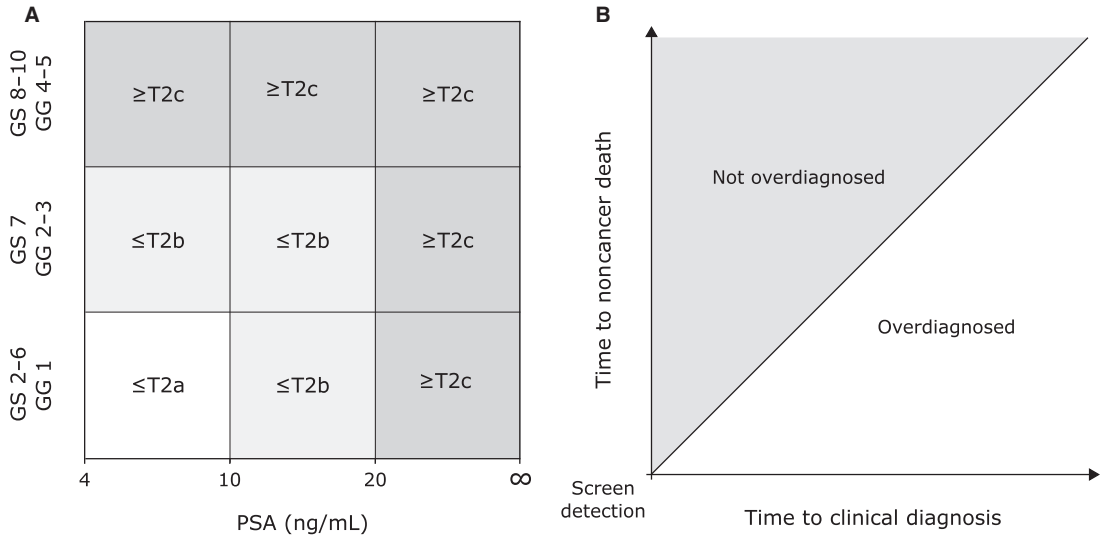

Figure 1. Two definitions of "nonthreatening" disease for a prostate cancer (PCa) diagnosed by prostate-specific antigen (PSA) screening. **A)** D'Amico classification of clinically localized PCa into low- (**white**), intermediate- (**light gray**), and high- (**dark gray**) risk groups. **B)** An overdiagnosed PCa (**white**) is any PCa that would not have been diagnosed during the patient's remaining lifetime in the absence of screening.

it might be defined as PCa with a low risk of biochemical recurrence according to the D'Amico system (12) (Figure 1A). Alternatively, nonthreatening disease can be defined as disease that would not have progressed to diagnosis in the absence of screening within a man's lifetime (Figure 1B). Such PCa will not cause symptoms or affect the quality or quantity of a man's life; in this article, these PCa are referred to as overdiagnosed (15–17). Most studies of reflex testing have focused on the clinicopathologic definition. However, overdiagnosis is the primary driver of harm in early-detection efforts and is the definition of nonthreatening disease used in this study.

Studies of reflex tests typically focus on their performance in terms of sensitivity, specificity, and reduction in unnecessary biopsies and do not examine impact on overdiagnosis or survival. Indeed, overdiagnosis is not directly observable, and empirical observation of survival impact requires lengthy follow-up. An alternative approach models disease natural history, progression, and survival to translate diagnostic performance of reflex tests into projections of long-term outcomes. In this article, we use this approach to quantify expected overdiagnosis and lives saved by reflex testing in men with serum PSA between 4.0 and 10.0 ng/mL. We use an established simulation model of PCa and intervention effects from randomized PCa screening and treatment trials to predict overdiagnosis and long-term survival benefit associated with four urinary biomarkers used as reflex tests (T2:ERG, PCA3, and MiPS for any and for high-grade PCa). For general insights, we also model hypothetical ideal biomarkers with 100% sensitivity and specificity for any and for high-grade PCa.

Methods

Simulated Life Histories

Our PCa microsimulation model (18–20) links a man's PSA levels with onset of preclinical PCa, progression from localized to distant- stage disease, and progression from preclinical PCa to clinical diagnosis (Figure 2). Baseline PCa survival without screening is based on the control group of the European Randomized Study of Screening for Prostate Cancer (ERSPC) (21). Early detection of PCa results in cure for a fraction of patients; the fraction of patients cured depends on how much earlier the cancer is detected (also known as "lead time") and declines to zero as the time of detection approaches the time of clinical diagnosis—when the cancer would have been diagnosed without screening. Curative treatment lowers the hazard of PCa death as estimated in the Scandinavian Prostate Cancer Group Study Number 4 (22) randomized trial of prostatectomy as primary treatment. Supplementary Figure 1 (available online) shows 13-year cumulative incidence of PCa death in the ERSPC and 20-year predictions from the model assuming these effects of early detection and treatment. The model simulates life histories, including ages at onset and at diagnosis and death with and without screening, and determines overdiagnoses and lives saved empirically from the simulated data (19).

To the extent that a more sensitive reflex test advances the timing of diagnosis (thereby lengthening the lead time), this increases the likelihood of overdiagnosis because there is a longer time window during which competing mortality can occur before clinical diagnosis. At the same time, according to our model of screening benefit, a more sensitive reflex test also increases the likelihood of cure (Figure 3). In this way the model provides a means of translating the diagnostic performance of a reflex test involving a more sensitive biomarker into changes in the downstream outcomes of interest.

Reflex Testing Using Empirical Biomarkers

Our empirical analysis used quantitative T2:ERG and PCA3 levels derived from whole urine after digital rectal exams collected from previously reported cross-sectional patient cohorts (7). These included a training cohort composed of 711 men presenting for biopsy at three US academic institutions and a validation cohort composed of 1225 men presenting for biopsy at seven US community clinics. After excluding men with missing age, on active surveillance, or with PSA outside the 4.0–10.0 ng/mL range, the combined patient cohort consisted of 1112 men (Supplementary Figure 2, available online) with age, PSA, T2:ERG, and PCA3 levels and biopsy results (benign or Gleason sum ≤6, 7, or ≥8). MiPS

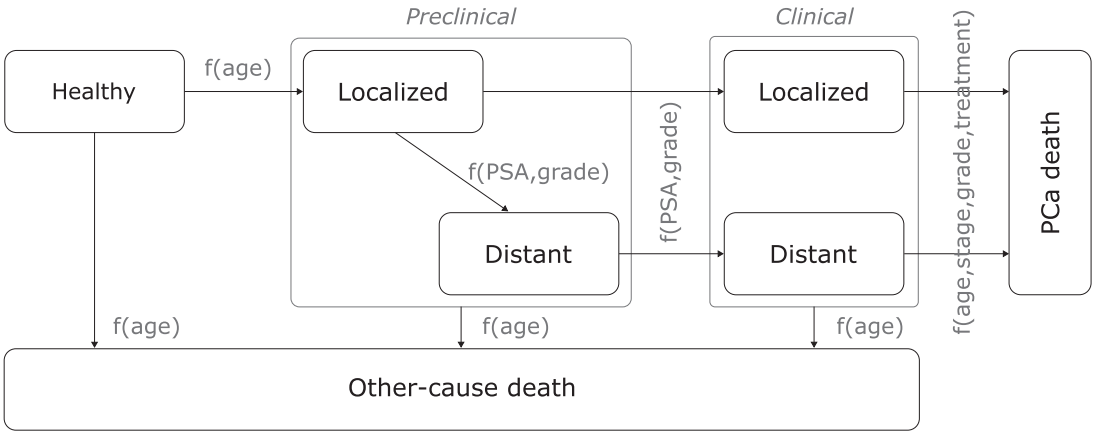

Figure 2. A microsimulation model of linked prostate-specific antigen (PSA) levels and prostate cancer (PCa) natural history and survival.

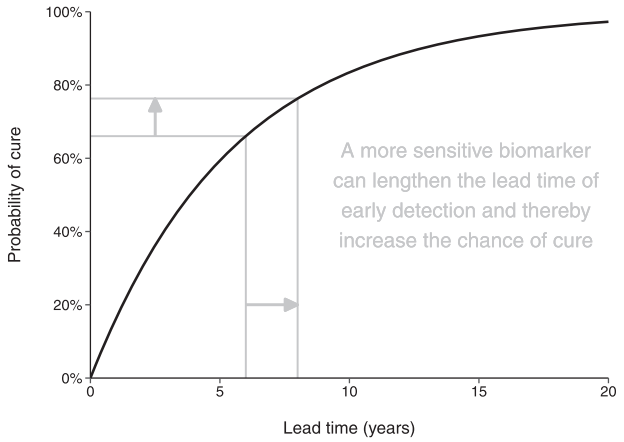

Figure 3. A model of the relationship between prostate cancer early detection and the probability of cure associated with definitive treatment.

levels were previously calculated from (log_2-transformed) PSA, T2:ERG, and PCA3 via logistic regression models fitted to the unrestricted training cohort to predict any PCa (MiPS) and high-grade (Gleason sum ≥7) PCa (MiPShg) (7). Biomarker levels for the combined cohort were used in this analysis.

Because only cross-sectional (one-time) test data for the reflex tests were available, our analysis focused on one-time screening tests. T2:ERG, PCA3, MiPS, and MiPShg levels were generated for simulated men age 55 or 65 years with PSA between 4.0 and 10.0 ng/mL at a screening test via bootstrap imputations. The bootstrapped values were selected to match age (±5 years), PSA level (±2.0 ng/mL), and cancer status (benign or Gleason sum ≤6, 7, or ≥8) of simulated individuals. Reflex test results were considered positive if generated biomarkers exceeded thresholds that have been used in practice (7,23,24): T2:ERG greater than 5, PCA3 greater than 35, MiPS greater than 30%, and MiPShg greater than 10%. Otherwise, no additional testing occurred.

Reflex Testing Using Ideal Biomarkers

A hypothetical analysis used perfect knowledge of PCa status to model reflex test results for simulated men age 55 or 65 years with PSA between 4.0 and 10.0 ng/mL at a screening test using ideal biomarkers with 100% sensitivity and specificity to detect any PCa and high-grade (Gleason sum ≥7) PCa.

Statistical Analysis

We assessed the validity of our biomarker model by comparing modeled diagnostic performance against empirical performance of the reflex biomarkers. We calculated sensitivity and specificity at the stated thresholds, receiver operating characteristic curves, and calibration plots for detecting any PCa and high-grade PCa, and we tested changes in the area under the receiver operating characteristic curve (AUC) when using each biomarker compared with PSA alone (25). DeLong test was used to compare AUCs, and two-sided *P* values less than .05 were considered statistically significant. Modeled biopsies performed or avoided and any PCa and high-grade PCa detected or missed for a range of MiPS thresholds were also tabulated for comparison with empirical outcomes.

We predicted the long-term impact of reflex testing by simulating screening-naïve men with PSA levels between 4.0 and 10.0 ng/mL at ages 55 and 65 years, generating reflex test results, determining impacts on ages at diagnosis and death, and calculating overdiagnoses and lives saved. We simulated sufficient numbers of men (>10 million for each age and biomarker) to eliminate Monte Carlo error. For simplicity, 100% of men with positive reflex tests received biopsy, and 100% of men with confirmed PCa received definitive treatment. For comparison, corresponding outcomes were predicted if all patients or no patients were biopsied. The 20-year cumulative incidence of PCa death was estimated using Kaplan-Meier methods.

As a sensitivity analysis, we also predicted long-term impacts under serial screening for men age 55 years with PSA between 4.0 and 10.0 ng/mL who were not diagnosed at the initial screening test either because they were not biopsied or because their biopsies were negative. These men continued biennial screening up to age 69 years, with biomarker values regenerated using the same imputation procedure described above whenever their PSA was between 4.0 and 10.0 ng/mL. Men with PSA above 10.0 ng/mL were always biopsied.

Results

Performance of Modeled Reflex Testing Using Empirical Biomarkers

The model preserved the performance advantages of the urinary biomarkers over PSA in terms of predicting relative and absolute probabilities of preclinical PCa. The model reproduced

ARTICLE

Table 1. Sensitivity and specificity of PCa-specific biomarkers for predicting any and high-grade PCa in empirical and simulated records of men with PSA level between 4.0 and 10.0 ng/mL*

			Any PCa		High-grade PCa	
Source	Biomarker	Threshold	Sensitivity, % (95% CI)	Specificity, % (95% CI)	Sensitivity, % (95% CI)	Specificity, % (95% CI)
Empirical	T2:ERG	5	65.5 (61.2 to 69.6)	58.4 (54.4 to 62.3)	66.1 (60.3 to 72.0)	51.5 (47.8 to 54.9)
	PCA3	35	55.6 (51.3 to 59.6)	76.0 (72.6 to 79.5)	63.4 (57.2 to 69.3)	68.9 (66.0 to 71.9)
	MiPS	30%	86.9 (84.0 to 89.9)	45.6 (41.4 to 49.4)	91.8 (87.9 to 95.3)	37.3 (34.2 to 40.7)
	MiPShg	10%	92.0 (89.5 to 94.3)	25.9 (22.2 to 29.2)	95.3 (92.6 to 97.7)	21.4 (18.6 to 24.2)
Model	T2:ERG	5	66.2 (65.0 to 67.4)	58.1 (57.3 to 59.0)	65.2 (62.9 to 67.3)	51.6 (50.8 to 52.4)
	PCA3	35	59.7 (58.5 to 61.0)	69.5 (68.7 to 70.4)	66.2 (64.0 to 68.4)	62.6 (61.9 to 63.4)
	MiPS	30%	90.8 (90.0 to 91.5)	38.6 (37.8 to 39.5)	93.2 (92.1 to 94.3)	31.2 (30.5 to 31.9)
	MiPShg	10%	94.5 (93.9 to 95.0)	22.4 (21.6 to 23.1)	96.5 (95.6 to 97.3)	18.2 (17.6 to 18.8)

*CI = confidence interval; T2:ERG = TMPRSS2:ERG gene fusion; PCA3 = prostate cancer antigen 3; MiPS = Michigan Prostate Score for any PCa; MiPShg = Michigan Prostate Score for high-grade PCa; PCa = prostate cancer; PSA = prostate-specific antigen.

statistically significant improvement in the AUCs of T2:ERG, PCA3, MiPS, and MiPShg compared with PSA with respect to the ability to discriminate between patients with and without PCa or high-grade PCa (Supplementary Table 1, available online). Discrimination and calibration of actual and simulated biopsy results are shown in Supplementary Figures 3 and 4 (available online). At the thresholds considered in this study, MiPS and MiPShg had the highest sensitivity and lowest specificity, with model estimates close to empirical estimates (Table 1).

Observed and simulated biopsy results for selected MiPS thresholds are presented in Table 2. Absolute numbers of any PCa and high-grade PCa diagnosed were higher in the observed results than in the simulated results; this is unsurprising because the observed results were for men who presented for biopsy whereas the simulated results were for unselected, age-matched individuals. In the observed results, a MiPS threshold of 40% would have missed 11.1% of any PCa and 4.0% of high-grade PCa, and using simulated results, corresponding predictions were 7.0% and 1.4%.

Predicted Impact of Reflex Testing Using Empirical Biomarkers

Predicted outcomes for men with PSA between 4.0 and 10.0 ng/mL under different biopsy strategies are shown in Table 3. Biopsying all men detected all cancers, with fewer overdiagnoses at age 55 years (67 per 1000) compared with age 65 years (121 per 1000), reflecting the fact that younger men have a lower chance of dying from other causes before the cancer would have been clinically diagnosed. As expected, universal biopsy also saved the most lives, with more lives saved at age 55 years (49 per 1000) compared with age 65 years (20 per 1000) because, as a result of secular increases in PSA levels with age, younger men with PSA between 4.0 and 10.0 ng/mL have higher baseline risk of PCa death than older men with PSA in this range.

Performing biopsy only when reflex biomarkers exceeded conventional thresholds reduced the number of biopsies, overdiagnoses, and lives saved relative to universal biopsy by varying amounts (Table 3). Summarized as ranges across screening ages, T2:ERG, PCA3, MiPS, and MiPShg avoided 50.9–51.4%, 66.2–73.9%, 33.1–43.1%, and 21.2–26.8% of biopsies and reduced overdiagnoses by 35.8–36.4%, 48.0–60.6%, 14.0–21.2%, and 8.8–15.0%, respectively. However, the predicted number of lives saved was concomitantly reduced by 36.0–38.5%, 46.6–64.9%, 11.4–19.6%, and 7.3–13.1%, respectively. In general, all biopsy strategies resulted in similar overdiagnoses per life saved at each age: on average 1.4 at age 55 years and 6.0 at age 65 years.

Predicted Impact of Reflex Testing Using Ideal Biomarkers

Biopsies performed based on the ideal biomarker for detecting any PCa would also reduce the number of biopsies relative to universal biopsy but with no change to overdiagnoses or lives saved (by definition; see Table 3). In contrast, biopsies performed based on the ideal biomarker for detecting high-grade PCa would not only reduce overdiagnoses as well as lives saved relative to universal biopsy but also improve their ratio (on average from 1.4 to 0.9 at age 55 years and from 6.0 to 3.5 at age 65 years).

Predicted 20-year cumulative incidence of PCa death is shown in Figure 4 for the same biopsy strategies. Reflex testing using T2:ERG, PCA3, MiPS, and MiPShg reduced mortality by varying amounts relative to a predicted maximum (biopsy no patients) of 6.5% at age 55 years and 5.5% at age 65 years and a predicted minimum (biopsy all patients or use the ideal biomarker for any PCa) of 2.5%. Similar to using the urinary biomarkers, biopsies performed based on the ideal biomarker for high-grade PCa produced mortality rates that fell midway between these extremes.

Sensitivity Analysis

The overall impact of using urinary reflex biomarkers at one-time tests was essentially unchanged if men with PSA between 4.0 and 10.0 ng/mL who were not diagnosed at the initial test continued screening (Supplementary Table 2, available online). Compared with results for one-time tests, the model predicted more biopsies, overdiagnoses, and lives saved under serial screening, yet overdiagnoses per life saved remained comparable across biopsy strategies. The improvement from using an ideal biomarker for detecting high-grade PCa, however, became smaller (decreased from 1.4 to 1.3).

Discussion

This study evaluated downstream harm and benefit resulting from reflex testing in asymptomatic men with serum PSA between 4.0 and 10.0 ng/mL. Our results indicated that, under simplified conditions, performing biopsy when urinary biomarkers

Table 2. Impact of MiPS-based risk prediction on biopsies avoided and any and high-grade PCa detected in empirical and simulated records of men with PSA level between 4.0 and 10.0 ng/mL*

		Biopsies		Any PCa		High-grade PCa	
Source	Threshold, %	Done, n	Avoided, No. (%)	Detected, No. (%)	Missed, No. (%)	Detected, No. (%)	Missed, No. (%)
Empirical	0	1112	0 (0.0)	513 (46.1)	0 (0.0)	257 (23.1)	0 (0.0)
	10	1077	35 (3.1)	508 (45.7)	5 (0.4)	254 (22.8)	3 (0.3)
	20	938	174 (15.6)	479 (43.1)	34 (3.1)	247 (22.2)	10 (0.9)
	30	772	340 (30.6)	446 (40.1)	67 (6.0)	236 (21.2)	21 (1.9)
	40	627	485 (43.6)	390 (35.1)	123 (11.1)	213 (19.2)	44 (4.0)
Model	0	1112	0 (0.0)	375 (33.8)	0 (0.0)	114 (10.2)	0 (0.0)
	10	1084	28 (2.6)	373 (33.6)	2 (0.2)	112 (10.1)	1 (0.1)
	20	944	168 (15.1)	359 (32.3)	17 (1.5)	111 (9.9)	3 (0.3)
	30	791	321 (28.9)	340 (30.6)	35 (3.2)	106 (9.5)	8 (0.7)
	40	636	476 (42.8)	297 (26.7)	78 (7.0)	98 (8.8)	16 (1.4)

*MiPS = Michigan Prostate Score for any PCa; PCa = prostate cancer; PSA = prostate-specific antigen.

Table 3. Predicted immediate and long-term PCa outcomes per 1000 men with PSA between 4.0 and 10.0 ng/mL by screening age and biopsy strategy

				PCa diagnoses, n				Deaths, n				
Screening age, y	Biopsy strategy*	Biopsies done, n	Biopsies, %Δ†	Low grade	High grade	Over-diagnoses, n	Overdiagnoses, %Δ†	PCa	Other	Lives saved, n	Lives saved,%Δ†	Overdiagnoses per life saved
55	No patients	0	100.0	0	0	0	100.0	88	912	0	100.0	—
	T2:ERG >5	491	50.9	156	112	43	36.4	57	943	31	36.0	1.4
	PCA3 >35	261	73.9	93	75	27	60.6	71	929	17	64.9	1.6
	MiPS >30%	569	43.1	188	151	53	21.2	49	951	39	19.6	1.4
	MiPShg >10%	732	26.8	203	160	57	15.0	45	955	42	13.1	1.4
	IDEAL	421	57.9	248	173	67	0.2	39	961	49	0.1	1.4
	IDEALhg	173	82.7	0	173	25	62.9	60	940	28	42.6	0.9
	All patients	1000	0.0	248	173	67	0.0	39	961	49	0.0	1.4
65	No patients	0	100.0	0	0	0	100.0	42	958	0	100.0	—
	T2:ERG >5	486	51.4	144	57	77	35.8	30	970	12	38.5	6.3
	PCA3 >35	338	66.2	105	59	63	48.0	31	969	11	46.6	5.8
	MiPS >30%	669	33.1	184	87	104	14.0	24	976	18	11.4	5.8
	MiPShg >10%	788	21.2	197	88	110	8.8	23	977	19	7.3	5.9
	IDEAL	311	68.9	218	93	121	0.0	22	978	20	0.2	6.0
	IDEALhg	93	90.7	0	93	34	72.0	32	968	10	51.8	3.5
	All patients	1000	0.0	218	93	121	0.0	22	978	20	0.0	6.0

*IDEAL = ideal biomarker with 100% sensitivity or specificity for any PCa; IDEALhg = ideal biomarker with 100% sensitivity or specificity for high-grade PCa; MiPS = Michigan Prostate Score for any PCa; MiPShg = Michigan Prostate Score for high-grade PCa; PCa = prostate cancer; PCA3 = prostate cancer antigen 3; PSA = prostate-specific antigen; T2:ERG = TMPRSS2:ERG gene fusion.
†"%Δ" shows the percent reduction under each biopsy strategy relative to biopsying all patients.

ARTICLE

exceeded conventional thresholds could reduce overdiagnoses by roughly 10–60%, depending on the biomarker used and patient age. However, this practice would reduce lives saved by a similar amount, leaving overdiagnoses per life saved virtually unchanged.

In general, translating the diagnostic performance of reflex biomarkers to their expected long-term impact requires a model that can incorporate the biomarkers in a manner concordant with disease histories and that can project the long-term effects of early detection involving more sensitive and specific tests. Our analysis imputed empirical biomarker measurements into a model with effects of early detection and treatment that mimic randomized trials of PSA screening and definitive treatment. Our results show that reflex testing involves tradeoffs; a decrease in overdiagnosis is inherently coupled to a decrease in lives saved except when the reflex biomarker is highly specific for aggressive disease. Similar tradeoffs likely apply to reflex biomarkers other than the ones examined here.

For simplicity, we modeled overdiagnoses and lives saved when the reflex biomarkers for any PCa exceeded thresholds that have been used in practice. Higher thresholds decreased both outcomes whereas lower thresholds increased both outcomes, yet overdiagnoses per life saved remained stable. Effects of varying thresholds for biomarkers for high-grade PCa are more complicated, however, because putative false positives can lead to detection of low-grade PCa.

Our predictions are consistent with previous modeling studies of reflex tests. A previous analysis using our model to study reflex PCA3 testing similarly estimated that reductions in lives saved tracked reductions in overdiagnoses (26). Another microsimulation model predicted little change to either overdiagnoses or lives saved when using the Prostate Health Index as a

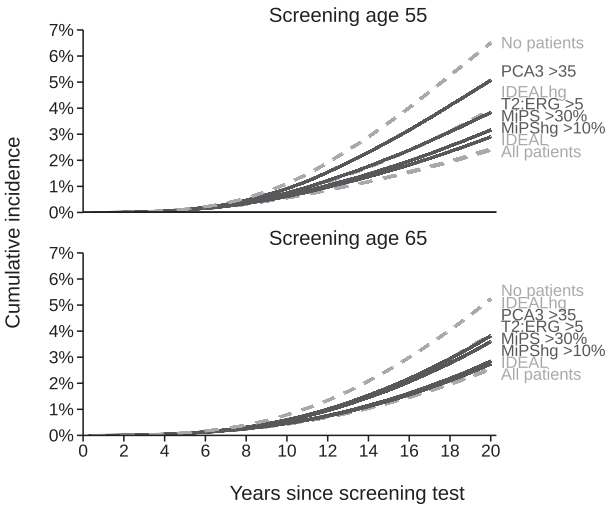

Figure 4. Predicted 20-year cumulative incidence of prostate cancer death for men with prostate-specific antigen between 4.0 and 10.0 ng/mL at screening age 55 or 65 years by biopsy strategy. *IDEAL = ideal biomarker with 100% sensitivity or specificity for any PCa; IDEALhg = ideal biomarker with 100% sensitivity or specificity for high-grade PCa; MiPS = Michigan Prostate Score for any PCa; MiPShg = Michigan Prostate Score for high-grade PCa; PCa = prostate cancer; PCA3 = prostate cancer antigen 3; T2:ERG = TMPRSS2:ERG gene fusion.

reflex test (27). The present study adds to previous work by validating diagnostic performance of modeled biomarkers, projecting long-term harm-benefit tradeoffs for several urinary biomarkers, and considering hypothetical ideal biomarkers for any PCa and for high-grade PCa.

Our study is not without limitations. The accuracy of our predictions may be limited by our model assumptions about the patterns of PCa development and progression and about how disease progression correlates with biomarker levels. Although the model approximates incidence trends in the United States (19,28) and in the ERSPC (21), it was originally developed and calibrated to stage- and grade-specific incidence data from the Surveillance, Epidemiology, and End Results program for the period 1975–2000. Consequently, it may not necessarily replicate contemporary low-grade, early-stage PCa, which is likely to have lower risk of progression than PCa previously classified as low grade. One implication is that the model-predicted risk of overdiagnosis for low-grade PCa may be lower than what would be expected based on contemporary grading practices. Conversely, also because of shifting grading schemes, the model may overstate the risk of PCa death for contemporary men with low-grade disease. Finally, the paucity of serial data on the reflex biomarkers motivated our focus on one-time screening tests. Our sensitivity analysis of serial screening supported our principal conclusions provided correlations in the cross-sectional data generalize to longitudinal relationships. It is also possible that active surveillance for low-risk PCa could modify the balance between overdiagnosis and lives saved, though our simplified assumptions about the biopsy decision process and definitive treatment for all confirmed disease provides a first approximation for plausible long-term outcomes.

Our findings have implications regarding how reflex tests might be used in practice and for future research. Our study shows that the performance of existing urinary reflex biomarkers, despite apparently favorable tradeoffs in terms of test diagnostics and biopsies avoided vs PCa missed, does not necessarily imply that their use will improve the long-term balance of harms and benefits. Indeed, an ideal biomarker for detecting any PCa could serve as a perfect surrogate for biopsy but would not reduce overdiagnosis by definition. This is because overdiagnosis is determined by a competition between cancer progression (or lack thereof) and other-cause mortality (Figure 1B), so even perfect knowledge about cancer status gives only an incomplete picture of a patient's risk for overdiagnosis. For this reason, the biopsy decision for a man with an intermediate PSA level should take into account his life expectancy. In contrast, an ideal biomarker for detecting high-grade PCa would reduce overdiagnoses per life saved at one-time screening tests (by 35.2% at age 55 years and 42.0% at age 65 years) although the incremental improvement is diminished under continued screening (by 6.1% at age 55 years). Our results therefore support efforts to develop more accurate reflex biomarkers for high-grade PCa, particularly for use in settings with less frequent screening.

In conclusion, previous studies have shown that reflex testing may substantially reduce unnecessary biopsies while missing few cancers with aggressive features. Investigations using a disease model like the one in this study can help to translate this improved diagnostic performance into impact on long-term clinical outcomes. Under simplified assumptions about the receipt of biopsy and treatment, our model predicts that reduced overdiagnoses due to existing urinary biomarkers are likely to be proportionally offset by reduced lives saved. This finding underscores the continued need for discovering biomarkers targeted to PCa destined to cause morbidity or mortality and for proven ways to control overdiagnosis and overtreatment. Disease modeling is certain to remain an essential partner in these efforts to improve the clinical outcomes that ultimately matter most.

Funding

This work was supported by the National Cancer Institute at the National Institutes of Health (grant numbers U01 CA199338 to RE, R50 CA221836 to RG, and P50 CA186786-05 to TMM).

Notes

Affiliations of authors: Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA (RG, TA, RE); Department of Urology, University of Michigan, Ann Arbor, MI (TMM, JJT); Department of Urology, University of Washington, Seattle, WA (SPP).

The contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Cancer Institute. The funding agency had no role in the design of the study; the collection, analysis, or interpretation of the data; the writing of the manuscript; or the decision to submit the manuscript for publication.

The authors have no conflicts of interest to disclose.

References

1. Grossman DC, Curry SJ, Owens DK, et al. Screening for prostate cancer: US Preventive Services Task Force recommendation statement. *JAMA*. 2018; 319(18):1901–1913.
2. Schröder FH, Hugosson J, Roobol MJ, et al. Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. *Lancet*. 2014;384(9959):2027–2035.

3. Hugosson J, Roobol MJ, Månsson M, et al. A 16-yr follow-up of the European Randomized Study of Screening for Prostate Cancer. *Eur Urol*. 2019;76(1):43–51. doi:10.1016/j.eururo.2019.02.009.
4. Catalona WJ, Bartsch G, Rittenhouse HG, et al. Serum pro-prostate specific antigen preferentially detects aggressive prostate cancers in men with 2 to 4 ng/ml prostate specific antigen. *J Urol*. 2004;171(6 Pt 1):2239–2244.
5. Vickers AJ, Cronin AM, Roobol MJ, et al. A four-kallikrein panel predicts prostate cancer in men with recent screening: data from the European Randomized Study of Screening for Prostate Cancer, Rotterdam. *Clin Cancer Res*. 2010;16(12):3232–3239.
6. Grönberg H, Adolfsson J, Aly M, et al. Prostate cancer screening in men aged 50–69 years (STHLM3): a prospective population-based diagnostic study. *Lancet Oncol*. 2015;16(16):1667–1676.
7. Tomlins SA, Day JR, Lonigro RJ, et al. Urine TMPRSS2:ERG plus PCA3 for individualized prostate cancer risk assessment. *Eur Urol*. 2016;70(1):45–53.
8. McKiernan J, Donovan MJ, O'Neill V, et al. A novel urine exosome gene expression assay to predict high-grade prostate cancer at initial biopsy. *JAMA Oncol*. 2016;2(7):882–889.
9. Klein EA, Chait A, Hafron JM, et al. The single-parameter, structure-based IsoPSA assay demonstrates improved diagnostic accuracy for detection of any prostate cancer and high-grade prostate cancer compared to a concentration-based assay of total prostate-specific antigen: a preliminary report. *Eur Urol*. 2017;72(6):942–949.
10. 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(94):94ra72.
11. Salagierski M, Schalken JA. Molecular diagnosis of prostate cancer: PCA3 and TMPRSS2:ERG gene fusion. *J Urol*. 2012;187(3):795–801.
12. D'Amico AV, Whittington R, Malkowicz SB, et al. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *JAMA*. 1998;280(11):969–974.
13. Kattan MW, Eastham JA, Stapleton AM, et al. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *J Natl Cancer Inst*. 1998;90(10):766–771.
14. Cooperberg MR, Broering JM, Carroll PR. Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis. *J Natl Cancer Inst*. 2009;101(12):878–887.
15. Etzioni R, Gulati R, Mallinger L, et al. Influence of study features and methods on overdiagnosis estimates in breast and prostate cancer screening. *Ann Intern Med*. 2013;158(11):831–838.
16. Loeb S, Bjurlin MA, Nicholson J, et al. Overdiagnosis and overtreatment of prostate cancer. *Eur Urol*. 2014;65(6):1046–1055.
17. Davies L, Petitti DB, Woo M, et al. Defining, estimating, and communicating overdiagnosis in cancer screening. *Ann Intern Med*. 2018;169(10):739.
18. Gulati R, Inoue L, Katcher J, et al. Calibrating disease progression models using population data: a critical precursor to policy development in cancer control. *Biostatistics*. 2010;11(4):707–719.
19. Gulati R, Gore JL, Etzioni R. Comparative effectiveness of alternative prostate-specific antigen-based prostate cancer screening strategies: model estimates of potential benefits and harms. *Ann Intern Med*. 2013;158(3):145–153.
20. Gulati R, Tsodikov A, Etzioni R, et al. Expected population impacts of discontinued prostate-specific antigen screening. *Cancer*. 2014;120(22):3519–3526.
21. De Koning HJ, Gulati R, Moss SM, et al. The efficacy of PSA screening: impact of key components in the ERSPC and PLCO trial. *Cancer*. 2018;124(6):1197–1206.
22. Bill-Axelson A, Holmberg L, Garmo H, et al. Radical prostatectomy or watchful waiting in prostate cancer—29-year follow-up. *N Engl J Med*. 2018;379(24):2319–2329.
23. de la Taille A, Irani J, Graefen M, et al. Clinical evaluation of the PCA3 assay in guiding initial biopsy decisions. *J Urol*. 2011;185(6):2119–2125.
24. Chevli KK, Duff M, Walter P, et al. Urinary PCA3 as a predictor of prostate cancer in a cohort of 3,073 men undergoing initial prostate biopsy. *J Urol*. 2014;191(6):1743–1748.
25. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
26. Birnbaum JK, Feng Z, Gulati R, et al. Projecting benefits and harms of novel cancer screening biomarkers: a study of PCA3 and prostate cancer. *Cancer Epidemiol Biomarkers Prev*. 2015;24(4):677–682.
27. Heijnsdijk EA, Denham D, de Koning HJ. The cost-effectiveness of prostate cancer detection with the use of prostate health index. *Value Health*. 2016;19(2):153–157.
28. Roth JA, Gulati R, Gore JL, et al. Economic analysis of prostate-specific antigen screening and selective treatment strategies. *JAMA Oncol*. 2016;2(7):890–898.