Optimal healthcare decision making under multiple mathematical models: application in prostate cancer screening

Noname manuscript No.
(will be inserted by the editor)

Optimal healthcare decision making under multiple mathematical models

Application in prostate cancer screening

Received: date / Accepted: date

Abstract Important decisions related to human health, such as screening strategies for cancer, need to be made without a satisfactory understanding of the underlying biological and other processes. Rather, they are often informed by mathematical models that approximate reality. Often multiple models have been made to study the same phenomenon, which may lead to conflicting decisions. It is natural to seek a decision making process that identifies decisions that all models find to be effective, and we propose such a framework in this work. We apply the framework in prostate cancer screening to identify prostate-specific antigen (PSA)-based strategies that perform well under all considered models. We use heuristic search to identify strategies that trade off between optimizing the average across all models' assessments and being "conservative" by optimizing the most pessimistic model assessment. We identified three recently published mathematical models that can estimate quality-adjusted life expectancy (QALE) of PSA-based screening strategies and identified 64 strategies that trade off between maximizing the average and the most pessimistic model assessments. All prescribe PSA thresholds that increase with age, and 57 involve biennial screening. Strategies with higher assessments with the pessimistic model start screening later, stop screening earlier, and use higher PSA thresholds at earlier ages. The 64 strategies outperform 22 previously published expert-generated strategies. The 41 most "conservative" ones remained better than no screening with all models in extensive sensitivity analyses. We augment current comparative modeling approaches by identifying strategies that perform well under all models, for various degrees of decision-makers' conservativeness.

Keywords comparative modeling · decision analysis · sensitivity analysis · model averaging · optimization · prostate cancer screening · simulation modeling

1 Introduction

Mathematical modeling has long been an important tool in medical decision making. Modeling is valuable in assessing and determining optimal cancer prevention and control strategies [15,52,17,38] because it is a principled way to estimate the consequences of the large number of plausible screening strategies (combinations of screening schedules, modalities and positivity thresholds) and to structure causally explicit analyses of screening trials when there are substantial protocol departures and extensive missing data [17,25].

All models are based on assumptions that are largely unverifiable; it is therefore good practice to perform comparative analyses of models, evaluating the same phenomenon using multiple models whose assumptions differ in important ways [50,16]. The National Cancer Institute (NCI)-sponsored Cancer Intervention and Surveillance Modeling Network (CISNET) consortium has employed comparative modeling to explore screening for lung, breast, and colorectal cancers, informing screening guideline recommendations for the United States Preventative Services Task Force (USPSTF) [34,39,51]. CISNET has also used comparative modeling to address other public health questions for prostate [18,12,25] and esophageal cancers [33].

Making sense of comparative model analyses is easy when the results from multiple models agree, because the conclusions from the multi-model exercise are the same as those drawn from any of the individual models. When models disagree, however, making sense of their results is more involved. As Habbema et al. comment in their reflections on a successful multiyear comparative modeling exercise on breast cancer, "The challenge for reporting multimodel results to policymakers is to keep it (nearly) as simple as reporting one-model results, but with an understanding that it is more informative and more credible. We have not yet met this challenge" [27].

To help address these challenges, we introduce a general framework in which we identify strategies that have desirable characteristics across the set of considered models. For cancer screening, it is often the case that a large number of screening strategies are practical, defined by combinations of clinically plausible screening schedules (i.e., at which ages to screen), screening modalities (tests or combinations thereof) and, for quantitative tests, thresholds beyond which the result is considered positive. To identify effective screening strategies from a large set (in the millions) of implementable decisions instead of a smaller number of hand-selected strategies, we propose to use mathematical optimization, a field with a rich history that provides techniques to make high quality decisions in the presence of models [5].

Consider the following fictitious comparative modeling example, which compares screening strategies with respect to patients' quality-adjusted life expectancy (QALE), using three equally plausible mathematical models. Table 1 shows the models' assessments of the difference in QALE between five screening strategies and no screening, ordered by the most pessimistic (minimum) model assessment. All models agree that strategies 4 and 5 result in

Table 1 A comparison of three fictitious models

	Difference in QALE vs no screening (months)				
Strategy	**Model** 1	**Model** 2	**Model** 3	**Average**	**Pessimistic**
1	-1.0	3.0	4.0	2.0	-1.0
2 (no screening)	0	0	0	0	0
3	0.9	1.1	1.0	1.0	0.6
4	1.0	1.5	2.0	1.5	1.0
5	1.2	1.2	1.2	1.2	1.2

higher QALE than (dominate) strategies 2 and 3. However, models disagree about the ranking of the non-dominated strategies 1, 4, and 5.

Choosing between the non-dominated strategies represents a trade-off. Given equally plausible mathematical models, the average of all models' assessments is the best estimate of a strategy's health impact; however, most decision makers would forgo strategy 1 despite its having the highest average assessment because it is assessed as worse than not screening by at least one model. Depending on the degree to which a decision maker is conservative concerning the most pessimistic model assessment, they might instead prefer strategies 4 or 5. In this work, we use mathematical optimization via an iterated local search heuristic to identify a set of non-dominated strategies from a large set of competing strategies, allowing decision makers to optimally trade off the average and pessimistic assessments based on their preferences.

For concreteness, we outline our methodology through an application to prostate cancer screening.Among men, prostate cancer is a major cause of death globally and a leading cause of death in developed countries [19,9]. Because early cancer is more likely to be successfully controlled, early identification and treatment of the disease may reduce prostate cancer mortality and morbidity, decrease treatment costs, and increase the length and the quality of life at a population level. At the same time, screening can lead to overdiagnosis and overtreatment of indolent disease that would never manifest clinically. Thus, improving screening strategies to optimize the tradeoff between the advantages and disadvantages of screening holds can have a huge impact on a global scale [10].

In particular, PSA-based screening for prostate cancer is controversial because of the risk of overdiagnosis and overtreatment of relatively prevalent indolent cancers that are unlikely to ever manifest clinically [10]. Five randomized controlled trials (RCTs) have compared a recommendation for PSA-based screening against a no-screening strategy [2,47,45,37,31]. Because of protocol departures and missing data, strong and untestable assumptions are required to estimate the causal effect of screening in those who actually received screening, which complicates the interpretation of the evidence base with respect to whether screening affects prostate cancer mortality [30]. For these reasons, modeling is an important tool for evaluating screening strategies [17]. Several investigators have proposed models that could be used to evaluate competing screening strategies for prostate cancer, typically modeling the natural history of the disease, the effectiveness of a specified screening strategy at detect-

ing the disease, and post-detection outcomes for patients [24,29,32,35,44,49]. Comparative modeling has previously been used to estimate the impact of PSA screening on U.S. prostate cancer mortality [18], prostate cancer overdiagnosis rates [12], and the causal effect of screening on those screened in the PLCO trial [25]. To our knowledge, comparative modeling has not been used to optimize competing screening strategies for prostate cancer. We use mathematical optimization via an iterated local search heuristic to identify all non-dominated strategies from among millions of practical strategies for PSA-based prostate cancer screening.

2 Methods

2.1 Identification of eligible models

We searched PubMed and the Tufts Cost-Effectiveness Analysis (CEA) Registry [28] to identify English-language reports of mathematical models that can evaluate PSA-based screening for prostate cancer in the general population of screening-age men. We considered eligible any model that allows the estimation of quality-adjusted life expectancy (QALE) for strategies with different screening schedules (varying start and stop ages and intra-screening intervals) and age-specific PSA positivity thresholds. We examined publications from January 1, 2010 to October 3, 2015 to identify models that are current (have been developed recently or have a longer development history, but are actively maintained). A single reviewer screened citations and full texts for eligibility. The exact search strategy is in Appendix A.

2.2 Model implementation and adaptation

For eligible models, we either re-implemented the model *de novo* if the publication provided sufficient detail to do so or otherwise obtained a software implementation of the model from the original investigators.

For each model, we extracted information on their evidence sources, the health states or events they account for (e.g., how they modeled cancer development and progression and how they model downstream effects of screening including management of screen-detected cases), how they modeled the evolution of PSA levels (e.g., accounting for within-person correlations or not), and their computational approach (e.g. discrete time Markov processes or microsimulation models). We also recorded data sources, whether model parameters were calibrated, and, as applicable, the calibration targets.

We adapted eligible models to estimate QALE in a uniform way. We assigned quality-of-life decrements for screening attendance, biopsy, cancer diagnosis, radiation therapy, radical prostatectomy, active surveillance, palliative therapy, and terminal illness, and we used the literature-based estimates of preference weights that were employed in a quality-of-life analysis of the European Randomized Study of Screening for Prostate Cancer (ERSPC) trial [29].

2.3 Identification of optimal screening strategies across all models

There are infinitely many screening strategies defined by start and stop ages, intra-screen intervals, and age-specific positivity thresholds. To select those that are practical to implement, we consider all annual and biennial screening strategies with age-specific PSA cutoffs of $0.5, 1.0, 1.5, \ldots, 6.0\ ng/mL$. Because PSA levels rise slowly with age [43], we assumed that practical screening strategies can have fixed PSA cutoffs for 5-year age ranges and that cutoffs should not decrease as patients age. In total, there are more than 10.4 million such screening strategies for men aged 40–100, which can be evaluated with each eligible model.

For each model, we computed the optimal strategy, that is, the strategy with the largest QALE improvement over no screening.

In comparative modeling we aimed to identify screening strategies that maximally improve QALE compared to no screening across the K models that are considered in the analysis. Given a screening strategy s and models' evaluations of that strategy's improvement over no screening, $m_1(s), \ldots, m_K(s)$, we define the *average assessment* of the models to be $[m_1(s)+\ldots+m_K(s)]/K$ and the *most pessimistic assessment* to be the minimum $\min[m_1(s), \ldots, m_K(s)]$. The average model assessment (model averaging) is a typical choice within the subjective expected utility framework [13]. The most pessimistic assessment is also a typical choice in a maxmin expected utility framework [21]. In the main analyses, we assigned equal weight to each model's assessment, because we chose to not favor one model over another *a priori*. Because models can differ in the range of the assessments, the equally-weighted average assessment can be influenced more by an "outlier" model that systematically yields larger assessments. In sensitivity analyses we scaled each model's assessment by the model's maximum assessment. Alternative objectives can be used, including objectives that assign different weights to each model's assessment and use different aggregation functions over models' assessments.

Because of the large number of screening strategies that must be evaluated, it is impractical to identify optimal strategies within and across models by exhaustively enumerating all possible strategies. Rather, we employed mathematical optimization using a constrained local search algorithm. Details on and validation results of the optimization procedure are provided in Appendices C and D.

We computed the *efficient frontier* of non-dominated strategies that trade off between the average and most pessimistic model assessments in a manner analogous to the example in Table 1. This is in agreement with α-maxmin expected utility frameworks [23], which have been used to analyse decision making in the context of ambiguity. Among the strategies on the efficient frontier, we define as more "conservative" those that have higher QALE assessments under the most pessimistic model. Finally, we compare the screening strategies on the efficient frontier to five expert-generated strategies from Ross et al. [44] and 17 expert-generated strategies from Gulati et al. [24].

2.4 Sensitivity analyses

We performed two sets of sensitivity analyses. First, we used one-way sensitivity analyses to examine the stability of QALE assessments for strategies on the efficient frontier, the optimal strategies according to each model, and the 22 expert-generated strategies. These analyses pertained either to parameters governing the natural course of the disease, which were specific to each model, or to quality-of-life decrements associated with various events or health states, which were common across models and obtained from [29]. We varied each parameter of each model over the range defined in the sensitivity analysis in the respective papers (Appendix B) and recorded QALE assessments.

In the second set of sensitivity analyses we repeated the main analyses using quality-of-life decrement values that least favor screening, from the sensitivity analysis ranges in [29]. Specifically, we chose the sensitivity range values that least discount time spent under palliative therapy and with terminal illness and the sensitivity range values that most discount biopsy, cancer diagnosis, radiation therapy, radical prostatectomy, and active surveillance.

3 Results

3.1 Description of eligible models

Appendix Figure 5 shows the results of the literature search. Briefly, of the 547 and 75 citations returned by PubMed and CEA Registry searches, respectively, 36 were examined in full text. Three models fulfilled the eligibility criteria and were included in this analysis; we refer to them as Z [53], U [49] and G [24]. The most common reasons for exclusion were that a publication did not pertain to a mathematical model of PSA-based screening or that a described model could not evaluate screening under alternative PSA thresholds. Notably, we did not include two CISNET models of prostate cancer natural history [48,12] because they do not model PSA values through time and therefore cannot be used to evaluate screening strategies with different PSA positivity thresholds.

Table 2 summarizes the characteristics of the three models. Model G is a CISNET microsimulation model, and models U and Z both use discrete-time Markov models of disease state. Models U and Z were developed by the same research team but have different disease states and different approaches to modeling PSA levels through time.

3.2 Optimal strategies with each model

The three models differ in their estimate of the maximum attainable improvement in QALE across the 10.4 million strategies. The maximum attainable improvements with models G, U and Z were 0.5, 4.8, and 6.7 months of QALE, respectively (Table 3). Models U and Z favored more aggressive strategies,

Table 2 Characteristics of eligible models

	Model G [24]	**Model** U [49]	**Model** Z [53]
Stated purpose			
Screening strategy evaluation	Yes	Yes	Yes
Epidemiological analysis	Yes	No	No
Population trends	Yes	No	No
Modeled interventions			
Screening	Yes	Yes	Yes
Treatment	Active surveillance, radical prostatectomy, radiation ± hormones	Radical prostatectomy	Radical prostatectomy
Structure & assumptions			
Health states or attributes	No cancer; cancer (by grade and stage); metastases; cancer death; other death	No cancer; non-metastatic cancer; metastases; cancer death; other death	No cancer; cancer; cancer death; other death
PSA modeling	Linear changepoint model for log(PSA)	PSA varies by health state and last PSA measurement	PSA varies by health state
Adherence to screening	Imperfect	Perfect	Perfect
Data sources			
Biopsy compliance/accuracy	PLCO, systematic review	Autopsy study	Autopsy study
Natural and clinical history	Fitted via calibration, SEER, life tables	Autopsy study, chart review, decision analysis, SEER, life tables	Autopsy study, retrospective cohort study, life tables
PSA growth	PLCO, PCPT	OC PSA data	OC PSA data
Screening dissemination	NHIS	—	—
Treatment dissemination	SEER	—	—
Treatment effectiveness	SPCG4, observational studies	MCRPR	MCRPR
Analytic approach			
Modeling approach	Microsimulation, time to event	Microsimulation with underlying Markov model	Discrete time Markov
Calibration	To clinical trial and registry results	To incidence observed in autopsy studies	To incidence observed in autopsy studies

ERSPC: European Randomized Study of Screening for Prostate Cancer; MCRPR: Mayo Clinic Radical Prostatectomy Repository; NHIS: National Health Interview Survey; OC: Olmsted County, Minnesota, USA; PCPT: Prostate Cancer Prevention Trial; PLCO: Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial; PSA: prostate specific antigen; SEER: Surveillance, Epidemiology and End Results; SPCG4: Scandinavian Prostate Cancer Group trial 4.

with earlier start age, annual screening, and lower PSA thresholds, while model G favored less aggressive screening with PSA thresholds that rise as people age. The models disagreed not only about the magnitude of the maximum expected improvement in QALE but also about the relative ranking of the strategies. For example, according to model G, the best-performing strategy identified with model U was worse than no screening by approximately 6 days of QALE.

Table 3 Best-performing strategies with each model

Strategy		QALE, Improvement over no-screening (months)				
Age range @ PSA threshold		**Model G [24]**	**Model U [49]**	**Model Z [53]**	**Average**	**Pessimistic**
Biennial	45-54 @ 1.5	**0.5**	1.7	5.4	2.6	0.5
	55-59 @ 2.5					
	60-64 @ 3.5					
	65-69 @ 4.0					
	70-74 @ 6.0					
Annual	40-69 @ 0.5	−0.2	**4.8**	5.8	3.5	−0.2
	70-74 @ 1.5					
Annual	40-84 @ 2.5	0.2	1.6	**6.7**	2.8	0.2
	85-89 @ 4.0					

PSA thresholds are in ng/mL.

3.3 Efficient frontier of strategies across all models

Figure 1 displays aggregates of models' assessments of select strategies. Out of 10.4 million possible strategies, 64 stand out when we considered results across all three models, in that they form an efficient frontier (empty circles in the figure): No other strategies attained higher QALE improvement over no screening both on average and according to the most pessimistic model. Model G had the most pessimistic assessment for all strategies on the efficient frontier.

For comparison, Figure 1 depicts the optimal strategies identified with models G, U, and Z (shown in Table 3) as well as the 22 expert-generated strategies. None of the expert-generated strategies were on the efficient frontier: For each expert-generated strategy, there was a strategy on the efficient frontier with at least as good of a pessimistic assessment and at least 0.5 more incremental months of QALE according to the average assessment. Further, for each expert-generated strategy, there was a strategy on the efficient frontier with at least as good of an average assessment and at least 0.05 more incremental months of QALE according to the most pessimistic assessment.

The left panel in Figure 2 displays the exact screening schedules and age-specific thresholds for select strategies that span the efficient frontier (EF_1, EF_{22}, EF_{44}, and EF_{64}). The most conservative of these, EF_1, is the optimal strategy with model G (first row in Table 3). The right panel of the figure outlines all 64 strategies on the efficient frontier. Fifty-seven strategies on the frontier are biennial strategies. The most conservative strategies (light blue in the right panel of the figure) tend also to be less aggressive, in that they tend to start screening at later ages, stop at earlier ages, and use higher PSA thresholds. The least conservative (light red) tend to be more aggressive, as they start screening at earlier ages, stop at later ages, and use lower PSA screening thresholds.

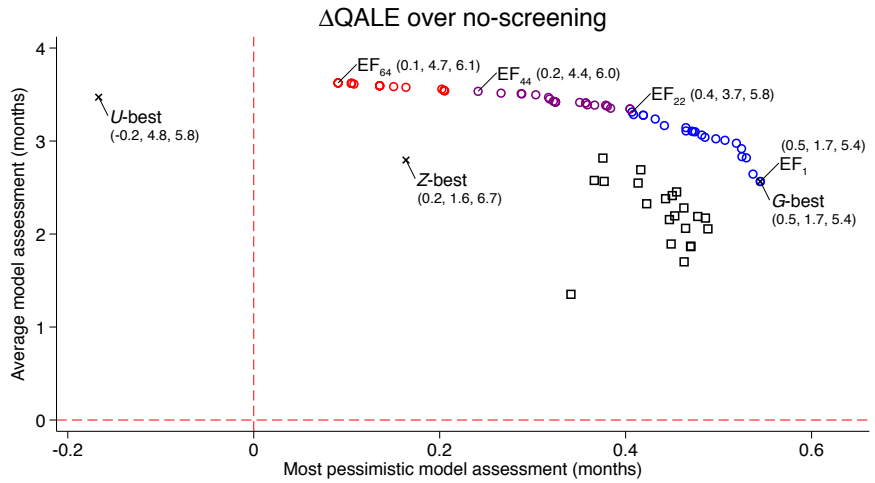

Fig. 1 Average and most pessimistic assessments of identified and expert-generated screening strategies. The 64 strategies on the efficient frontier are shown as empty circles. Strategies EF_1 to EF_{22} (most conservative tertile) are in blue, EF_{23} to EF_{44} (next less conservative tertile) in purple, and EF_{45} to EF_{64} (least conservative tertile) in red. The optimal strategies according to models G (G-best), U (U-best), or Z (Z-best) are shown with 'x' markers. The 22 expert-generated strategies are shown as empty squares. Assessments of QALE over no screening with each model are shown in parentheses for some strategies. For example, for strategy EF_1, which is also the optimal strategy with model G, the assessments of models G, U, and Z were 0.5, 1.7, and 5.4 months, respectively.

3.4 Sensitivity Analysis

In one-way sensitivity analysis we assessed each strategy plotted in Figure 1 under the 80 sensitivity scenarios specified in Appendix B. Several strategies were found to improve over no screening in all scenarios, including all 22 expert-generated strategies and 53 of the 64 strategies on the efficient frontier, including the 41 most conservative strategies, EF_1 (the strategy optimized according to model G) through EF_{41}. Eleven strategies on the efficient frontier (EF_{42}–EF_{45}, EF_{48}–EF_{49}, and EF_{56}–EF_{60}) and the strategies optimized according to models U and Z performed worse than no screening in at least one of the 80 scenarios.

When the 64 strategies on the efficient frontier were evaluated using model G under all 30 sensitivity scenarios specified in Appendix B, they had the worst average performance when the metastasis rate was set to the low value in its sensitivity range (0.18 months of incremental QALE) and when the post-recovery QALE decrement was set to its high value (0.21 months of incremental QALE), while the strategies had their best average performance when the post-recovery QALE decrement was set to its low value (0.74 months of incremental QALE) and when the palliative therapy QALE decrement was set to its high value (0.47 months of incremental QALE). When the same 64 strate-

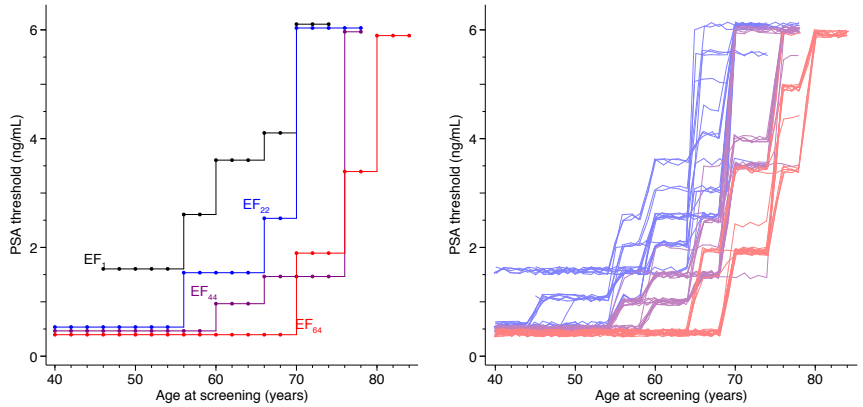

Fig. 2 Screening strategies on the efficient frontier in Figure 1 trading off average and most pessimistic QALE assessment. The left panel shows four strategies along the efficient frontier (from most to least conservative: EF_1, EF_{22}, EF_{44}, and EF_{64}). Dots indicate screenings. EF_1 optimizes the pessimistic assessment; it is also among the least aggressive strategies on the efficient frontier, in that it starts screening at a later age, ends at an earlier age, and uses higher age-specific positivity thresholds. The right panel outlines all 64 strategies on the efficient frontier, with random jitter added for visibility. EF_1 to EF_{22} (most conservative tertile of the efficient frontier) are in light blue, EF_{23} to EF_{44} (next less conservative tertile) are in light purple, and EF_{45} to EF_{64} (least conservative tertile) are in light red.

gies were evaluated using model U under all 26 sensitivity scenarios specified in Appendix B, they had the worst average performance when the incidence rate was set to its low value (0.39 months of incremental QALE) and when the metastasis probability was set to its low value (3.12 months of incremental QALE), while they had the best average performance when the death rate with metastatic cancer was set to its high value (7.18 months of incremental QALE) and when the incidence rate was set to its high value (6.90 months of incremental QALE). Finally, when the same 64 strategies were evaluated using model Z under all 24 sensitivity scenarios specified in Appendix B, they had the worst average performance when the incidence rate was set to its low value (1.12 months of incremental QALE) and when the prostate cancer death rate was set to its low value (4.30 months of incremental QALE), while they had the best average performance when the incidence rate was set to its high value (9.12 months of incremental QALE) and when the prostate cancer death rate was set to its high value (7.31 months of incremental QALE).

Figure 3 shows another set of sensitivity analyses using the most pessimistic quality-of-life decrements in the sensitivity ranges from [29] and re-optimizing to find the strategies on the efficient frontier and the most effective strategies according to each model. Of the 112 strategies on the efficient frontier, 73 improved QALE compared to no screening according to all three models. The remaining 39 performed worse than no screening according to model G. In all, 18 of the 22 expert-generated strategies and the optimal strategies obtained

with models U and Z performed worse than no screening according to model G. According to the most pessimistic assessment across the three models (strategy EF'_1 in Figure 3), patients undergo biennial screening between ages 50–69 years, with PSA positivity thresholds of 3.5 ng/mL between 50–54, 5.0 ng/mL between 55–59, and 6.0 ng/mL between 60–69.

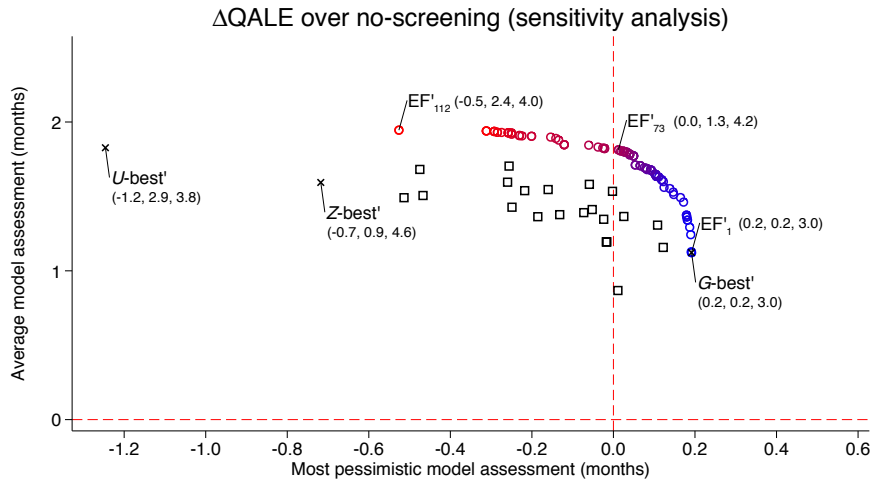

Fig. 3 Average and most pessimistic assessments of identified and expert-generated screening strategies using the most pessimistic quality-of-life decrements from [29]. The color of strategies on the efficient frontier changes from light blue to light purple to light red as one moves from more to less conservative strategies.

Figure 4 outlines all strategies on the efficient frontier in Figure 3. Strategy outlines are color-coded according to their place on the efficient frontier. As in the main analyses, the more conservative ones (light blue) tend to start screening at later ages, stop at earlier ages, and use higher PSA thresholds. The less conservative ones (light red) tend to start screening at earlier ages, stop at later ages, and use lower PSA thresholds.

Appendix E provides details of a further sensitivity analysis, normalizing each model's assessment of the QALE change compared to not screening to have a maximum value of 1. This normalization ensures that models with systematically more optimistic assessments of screening strategies are not weighted more heavily than others in the model averaging objective. This normalization results in a qualitatively different efficient frontier that is smaller and has more homogeneous screening strategies.

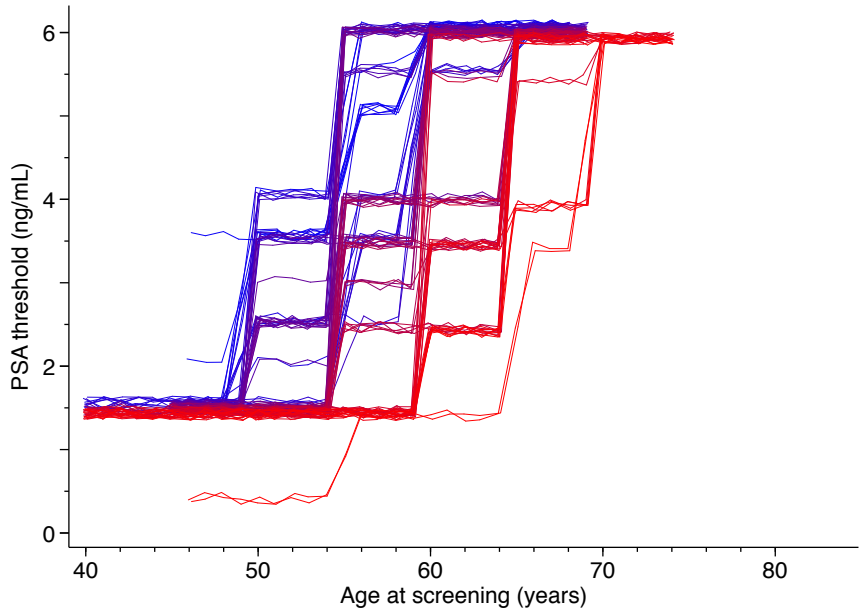

Fig. 4 Outlines of 112 strategies on the efficient frontier in Figure 3, with random jitter added for visibility. The color changes from light blue to light purple to light red as one moves from more to less conservative strategies on the efficient frontier (from EF'_1 to EF'_{112}).

4 Discussion

We describe an approach that is practical for addressing public health and clinical questions by means of comparative mathematical modeling, much like the comparative modeling used by the USPSTF to inform their recommendations about cancer screening [34,39,51]. One of the major challenges in comparative modeling pertains to dealing with and communicating the implications of conflicting model assessments [27,36]. We believe we advance typically employed comparative modeling methodologies in two ways. First, instead of exploring a relatively small set of predefined strategies (on the order of hundreds), we use mathematical optimization via an iterated local search heuristic to identify optimal strategies amongst a much larger set (in the millions) of implementable strategies. In the prostate cancer screening application we improved substantially over 22 previously published expert-generated strategies [44,24]. Second, we identify strategies that perform well under all considered models. We provide decision makers with tunable control by computing an efficient frontier of optimal strategies that trade off between performing well on average across models and performing well on the most pessimistic model. An advantage of the more pessimistic strategies is that they are more robust to parameter uncertainty: In the prostate cancer screening example, the 41 most

conservative strategies out of 64 on the efficient frontier remained beneficial over no screening across all 80 one-way sensitivity scenarios considered.

More generally, decision making is complicated by uncertainty — *aleatory uncertainty* that stems from the imprecision with which we learn from the finite empirical data in our evidence base and *epistemic uncertainty* that stems from our limited understanding of how the world works and that hinders our ability to structure and interpret said imprecise learnings [41]. Much theoretical and applied work has focused on decision making under aleatory uncertainty [41, 7], but, at least in health applications, less work has focused on negotiating epistemic uncertainty. We add to the methodological literature by providing a way to explore the impact of epistemic uncertainty, at least to the extent it is represented by a finite set of well-defined interpretations of the available evidence. In that sense, our work complements approaches such as model averaging and structural sensitivity analysis [13, 6].

The optimal strategies implied by models U and Z (Table 3 and Figure 1) are aggressive according to current clinical thinking [10], because they start early (age 40) and employ low PSA positivity cutoffs. Several key differences in the structures of the models could lead to the more aggressive screening strategies favored by models U and Z: neither of these models include clinical detection of cancer, both assume perfect screening and treatment adherence, and neither includes an active surveillance treatment option. Additionally, model Z does not model a patient's screening history of PSA test results when simulating a new PSA test result, likely overestimating how informative frequent tests will be. Beyond differences in the models, certain classes of screening strategies, such as those that take into account PSA velocity, those that vary screening strategies based on an initial baseline measurement, and those screening with other biomarkers were not considered in this work. The fact that the findings of this study cannot inform public health decision making is a limitation of the example but not of the proposed framework. Further, the large difference in the aggressiveness of the optimal strategies according to the three models nicely illustrates how our tool presents a range of decisions when models disagree in their assessments of strategies.

We adopted an approach similar to the α-maxmin expected utility framework [23] and examined objectives that are convex combinations of the average and the most pessimistic of (equally-weighted) model assessments. The average model assessment is the theory-based choice within a subjective expected utility framework, and the most pessimistic assessment is a typical choice in a maxmin expected utility framework [21]. However, the objective that a decision maker might use depends on the decision maker's preferences and on the problem at hand, and is determined by practical and aesthetic criteria and not by data. Entirely different objectives can be sensible and defensible, as long as they are consistent with the decision maker's thinking and with the context of the problem.

The approach proposed in this work could be extended in several ways. The more models one considers, the more extreme the observed range in model assessments can be. One way to reduce the observed heterogeneity in models'

assessments is to use ratios of QALE attained by different strategies versus QALE attained with no screening, instead of absolute differences [27]. More generally, however, one can aggregate model assessments in various ways to decrease the impact of extreme model assessments. In sensitivity analyses, we scaled each model's assessments by their maximum; this effectively assigns smaller weight to models U and Z and larger weight to model G in the model averaging objective, and Appendix E shows that this can qualitatively impact the efficient frontier obtained. More generally, one could use different objectives. Instead of using the most pessimistic assessment one could use a different rank-based statistic such as a weighted average of several of the most pessimistic assessments. Analogously, one could replace the average assessment with weighted averages or averages of ranks of models' assessments.

In our example we used QALE as the only decision-relevant quantity, though we could have measured and used other quantities instead, including life expectancy or costs. We conjecture that the approach outlined here can be extended to multi-criteria decision analyses with M decision-relevant quantities [20]. One way would be to compute efficient frontier surfaces in $2M$ dimensions, the total number of average and pessimistic aggregations of models' assessments for the M decision-relevant quantities.

In our analysis we examined only the average length of quality-adjusted life and did not explicitly consider the variation around it (i.e., the propagation of the aleatory uncertainty around model inputs [7]). Several ways exist to incorporate variation in models' assessments. In the computation of the efficient frontier in Figure 1 one could also include strategies that are "near" the efficient frontier in a sense that accounts for propagated uncertainty in models' assessments.

We believe that the largest practical obstacle in the routine application of the proposed approach is to identify well developed and validated models that meaningfully capture the salient aspects of the decisional problem at hand and that would be considered by the decision makers [27]. We used the PSA-based screening example for exposition and not to inform public health decision making. While we were systematic in selecting models for inclusion in the example, this example does not rise to the standard of multi-year comparative modeling exercises. For example, the identified prostate cancer models differed substantially in their inputs, calibration to external data, preference weights, and purpose, and are not as conducive to a comparative modeling exercise as a they would be had they been developed in tandem, for the same purpose. We deem that our approach is best suited in the context of already established comparative modeling consortia, in which the input sources across the models are standardized, the models' output is calibrated against the same external data, and there are iterative development steps for code verification or to identify implausible assumptions. Examples include modeling consortia on colorectal, breast, and lung cancer [51,39,34]; on tuberculosis [11]; and on human immunodeficiency virus [14].

References

1. Albertsen, P., Hanley, J., Fine, J.: 20-year outcomes following conservative management of clinically localized prostate cancer. JAMA **293**(17), 2095–2101 (2005)
2. Andriole, G.L., Crawford, E.D., Grubb, R.L., Buys, S.S., Chia, D., Church, T.R., Fouad, M.N., Gelmann, E.P., Kvale, P.A., Reding, D.J., Weissfeld, J.L., Yokochi, L.A., O'Brien, B., Clapp, J.D., Rathmell, J.M., Riley, T.L., Hayes, R.B., Kramer, B.S., Izmirlian, G., Miller, A.B., Pinsky, P.F., Prorok, P.C., Gohagan, J.K., Berg, C.D.: Mortality results from a randomized prostate-cancer screening trial. N Engl J Med **360**(13), 1310–1319 (2009)
3. Arias, E.: United States life tables, 2006. Natl Vital Stat Rep **58**(21), 1–40 (2010)
4. Aus, G., Robinson, D., Rosell, J., Sandblom, G., Varenhorst, E.: Survival in prostate carcinoma—outcomes from a prospective, population-based cohort of 8887 men with up to 15 years of follow-up. Cancer **103**(5), 943–951 (2005)
5. Bertsimas, D., Tsitsiklis, J.: Introduction to Linear Optimization. Athena Scientific (1997)
6. Bojke, L., Claxton, K., Sculpher, M., Palmer, S.: Characterizing structural uncertainty in decision analytic models: A review and application of methods. Value Health **12**(5), 739–749 (2009)
7. Briggs, A.H., Weinstein, M.C., Fenwick, E.A., Karnon, J., Sculpher, M.J., Paltiel, A.D.: Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6. Med Decis Making **32**(5), 722–732 (2012)
8. Bubendorf, L., Schöpfer, A., Wagner, U., Sauter, G., Moch, H., Willi, N., Gasser, T., Mihatsch, M.: Metastatic patterns of prostate cancer: An autopsy study of 1,589 patients. Hum Pathol **31**(5), 578–583 (2000)
9. U.S. Cancer Statistics Working Group, United States Cancer Statistics: 1999 – 2013 Incidence and Mortality Web-based Report (2016) URL http://www.cdc.gov/uscs Accessed on July 31, 2016
10. Cuzick, J., Thorat, M.A., Andriole, G., Brawley, O.W., Brown, P.H., Culig, Z., Eeles, R.A., Ford, L.G., Hamdy, F.C., Holmberg, L., Ilic, D., Key, T.J., La Vecchia, C., Lilja, H., Marberger, M., Meyskens, F.L., Minasian, L.M., Parker, C., Parnes, H.L., Perner, S., Rittenhouse, H., Schalken, J., Schmid, H.P., Schmitz-Draeger, B.J., Schroder, F.H., Stenzl, A., Tombal, B., Wilt, T.J., Wolk, A.: Prevention and early detection of prostate cancer. Lancet Oncol **15**(11), E484–E492 (2014)
11. Dowdy DW, Houben R, Cohen T, Pai M, Cobelens F, Vassall A, Menzies NA, Gomez GB, Langley I, Squire SB, White R; TB MAC meeting participants: Impact and cost-effectiveness of current and future tuberculosis diagnostics: the contribution of modelling. Int J Tuberc Lung Dis **18**(9), 1012– 1018 (2014)
12. Draisma, G., Etzioni, R., Tsodikov, A., Mariotto, A., Wever, E., Gulati, R., Feuer, E., de Koning, H.: Lead time and overdiagnosis in prostate-specific antigen screening: Importance of methods and context. Journal Natl Cancer Inst **101**(6), 374–383 (2009)
13. Draper, D.: Assessment and propagation of model uncertainty. J R Stat Soc Series B Stat Methodol pp. 45–97 (1995)
14. Eaton JW, Menzies NA, Stover J, Cambiano V, Chindelevitch L, Cori A, Hontelez JA, Humair S, Kerr CC, Klein DJ, Mishra S, Mitchell KM, Nichols BE, Vickerman P, Bakker R, Brnighausen T, Bershteyn A, Bloom DE, Boily MC, Chang ST, Cohen T, Dodd PJ, Fraser C, Gopalappa C, Lundgren J, Martin NK, Mikkelsen E, Mountain E, Pham QD, Pickles M, Phillips A, Platt L, Pretorius C, Prudden HJ, Salomon JA, van de Vijver DA, de Vlas SJ, Wagner BG, White RG, Wilson DP, Zhang L, Blandford J, Meyer-Rath G, Remme M, Revill P, Sangrujee N, Terris-Prestholt F, Doherty M, Shaffer N, Easterbrook PJ, Hirnschall G, Hallett TB: Health benefits, costs, and cost-effectiveness of earlier eligibility for adult antiretroviral therapy and expanded treatment coverage: a combined analysis of 12 mathematical models. Lancet Glob Health **2**(1), 23 – 34 (2013)
15. Eddy, D.M.: Screening for cancer: theory, analysis, and design. Prentice Hall (1980)
16. Eddy, D.M., Hollingworth, W., Caro, J.J., Tsevat, J., McDonald, K.M., Wong, J.B., Pract, I.S.M.G.R.: Model transparency and validation: A report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Value Health **15**(6), 843–850 (2012)

17. Etzioni, R., Gulati, R.: Response: Reading between the lines of cancer screening trials: Using modeling to understand the evidence. Med Care **51**(4), 304–306 (2013)
18. Etzioni, R., Tsodikov, A., Mariotto, A., Szabo, A., Falcon, S., Wegelin, J., DiTommaso, D., Karnofski, K., Gulati, R., Penson, D.F., Feuer, E.: Quantifying the role of PSA screening in the US prostate cancer mortality decline. Cancer Causes Control **19**(2), 175–181 (2008)
19. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F. GLOBOCAN 2012 v1.1, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. Lyon, France: International Agency for Research on Cancer (2014) URL http://globocan.iarc.fr Accessed on July 31, 2016
20. Figueira, J.R., Greco, S., Ehrgott, M. (eds.): Multiple criteria decision analysis. State of the art surveys. Springer (2014)
21. Gilboa I., Schmeidler D.: Maxmin expected utility with a non-unique prior. J Math Econom **18**, 141–153 (1989)
22. Ghani, K.R., Grigor, K., Tulloch, D.N., Bollina, P.R., McNeill, S.A.: Trends in reporting gleason score 1991 to 2001: changes in the pathologist's practice. Eur Urol **47**(2), 196–201 (2005)
23. Ghirardato P., Maccheroni F., Marinacci M.: Differentiating ambiguity and ambiguity attitude. J Econ Theory **31;118**(2), 133–173 (2004)
24. Gulati, R., Gore, J.L., Etzioni, R.: Comparative effectiveness of alternative PSA-based prostate cancer screening strategies. Ann Intern Med **158**(3), 145–153 (2013)
25. Gulati, R., Tsodikov, A., Wever, E.M., Mariotto, A.B., Heijnsdijk, E.A.M., Katcher, J., de Koning, H.J., Etzioni, R.: The impact of PLCO control arm contamination on perceived PSA screening efficacy. Cancer Causes Control **23**(6), 827–835 (2012)
26. Haas, G., Delongchamps, N., Jones, R., Chandan, V., Serio, A., Vickers, A., Jumbelic, M., Threatte, G., Korets, R., Lilja, H., de la Roza, G.: Needle biopsies on autopsy prostates: Sensitivity of cancer detection based on true prevalence. J Natl Cancer Inst **99**(19), 1484–1489 (2007)
27. Habbema, J.D.F., Schechter, C.B., Cronin, K.A., Clarke, L.D., Feuer, E.J.: Modeling cancer natural history, epidemiology, and control: reflections on the CISNET breast group experience. J Natl Cancer Inst Monogr **2006**(36), 122–126 (2006)
28. Center for the Evaluation of Value & Risk in Health: The Cost-Effectiveness Analysis Registry. Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, US. www.cearegistry.org (2015). Accessed on October 03, 2015
29. Heijnsdijk, E.A., Wever, E.M., Auvinen, A., Hugosson, J., Ciatto, S., Nelen, V., Kwiatkowski, M., Villers, A., Páez, A., Moss, S.M., Zappa, M., Tammela, T.L., Mäkinen, T., Carlsson, S., Korfage, I.J., Essink-Bot, M.L., Otto, S.J., Draisma, G., Bangma, C.H., Roobol, M.J., Schröder, F.H., de Koning, H.J.: Quality-of-life effects of prostate-specific antigen screening. N Engl J Med **367**(7), 595–605 (2012)
30. Ilic, D., Neuberger, M.M., Djulbegovic, M., Dahm, P.: Screening for prostate cancer. Cochrane Database Syst Rev **1**(1), CD004,720 (2013)
31. Kjellman, A., Akre, O., Norming, U., Törnblom, M., Gustafsson, O.: 15-year followup of a population based prostate cancer screening study. J Urol **181**(4), 1615–1621 (2009)
32. Kobayashi, T., Goto, R., Ito, K., Mitsumori, K.: Prostate cancer screening strategies with re-screening interval determined by individual baseline prostate-specific antigen values are cost-effective. Eur J Surg Oncol **33**(6), 783–789 (2007)
33. Kong, C.Y., Kroep, S., Curtius, K., Hazelton, W.D., Jeon, J., Meza, R., Heberle, C.R., Miller, M.C., Choi, S.E., Lansdorp-Vogelaar, I., van Ballegooijen, M., Feuer, E.J., Inadomi, J.M., Hur, C., Luebeck, E.G.: Exploring the recent trend in esophageal adenocarcinoma incidence and mortality using comparative simulation modeling. Cancer Epidemiol Biomarkers Prev **23**(6), 997–1006 (2014)
34. de Koning, H.J., Meza, R., Plevritis, S.K., ten Haaf, K., Munshi, V.N., Jeon, J., Erdogan, S.A., Kong, C.Y., Han, S.S., van Rosmalen, J., Choi, S.E., Pinsky, P.F., de Gonzalez, A.B., Berg, C.D., Black, W.C., Tammemägi, M.C., Hazelton, W.D., Feuer, E.J., McMahon, P.M.: Benefits and harms of computed tomography lung cancer screening strategies: A comparative modeling study for the U.S. Preventive Services Task Force. Ann Intern Med **160**(5), 311–320 (2014)
35. Krahn, M., Mahoney, J., Eckman, M., Trachtenberg, J., Pauker, S., Detsky, A.: Screening for prostate cancer: A decision analytic view. JAMA **272**(10), 773–780 (1994)

36. Kuntz, K.M., Lansdorp-Vogelaar, I., Rutter, C.M., Knudsen, A.B., van Ballegooijen, M., Savarino, J.E., Feuer, E.J., Zauber, A.G.: A systematic comparison of microsimulation models of colorectal cancer: the role of assumptions about adenoma progression. Med Decis Making **31**(4), 530–539 (2011)
37. Labrie, F., Candas, B., Cusan, L., Gomez, J.L., Bélanger, A., Brousseau, G., Chevrette, E., Lévesque, J.: Screening decreases prostate cancer mortality: 11-year follow-up of the 1988 Quebec prospective randomized controlled trial. Prostate **59**(3), 311–318 (2004)
38. Lee, S.J., Zelen, M.: Statistical models for screening: planning public health programs. In: C. Beam (ed.) Biostatistical Applications in Cancer Research, pp. 19–36. Springer US (2002)
39. Mandelblatt, J.S., Cronin, K.A., Bailey, S., Berry, D.A., de Koning, H.J., Draisma, G., Huang, H., Lee, S.J., Munsell, M., Plevritis, S.K., Ravdin, P., Schechter, C.B., Sigal, B., Stoto, M.A., Stout, N.K., van Ravesteyn, N.T., Venier, J., Zelen, M., Feuer, E.J.: Effects of mammography screening under different screening schedules: Model estimates of potential benefits and harms. Ann Intern Med **151**(10), 738–747 (2009)
40. Messing, E.M., Manola, J., Yao, J., Kiernan, M., Crawford, D., Wilding, G., di'SantAgnese, P.A., Trump, D.: Immediate versus deferred androgen deprivation treatment in patients with node-positive prostate cancer after radical prostatectomy and pelvic lymphadenectomy. Lancet Oncol **7**(6), 472–479 (2006)
41. National Academies of Science: Assessing the reliability of complex models: Mathematical and statistical foundations of verification, validation, and uncertainty quantification (2012)
42. National Cancer Institute: Surveillance epidemiology and end results (2008). URL http://seer.cancer.gov
43. Oesterling, J.E., Jacobsen, S.J., Chute, C.G., Guess, H.A., Girman, C.J., Panser, L.A., Lieber, M.M.: Serum prostate-specific antigen in a community-based population of healthy men. Establishment of age-specific reference ranges. JAMA **270**(7), 860–864 (1993)
44. Ross, K.S., Carter, H.B., Pearson, J.D., Guess, H.A.: Comparative efficiency of prostate-specific antigen screening strategies for prostate cancer detection. JAMA **284**(11), 1399–1405 (2000)
45. Sandblom, G., Varenhorst, E., Rosell, J., Löfman, O., Carlsson, P.: Randomised prostate cancer screening trial: 20 year follow-up. BMJ **342**, d1539 (2011)
46. Scardino, P.T., Beck, J.R., Miles, B.J.: Conservative management of prostate cancer. N Engl J Med **330**(25), 1831 (1994)
47. Schröder, F.H., Hugosson, J., Roobol, M.J., Tammela, T.L., Ciatto, S., Nelen, V., Kwiatkowski, M., Lujan, M., Lilja, H., Zappa, M., Denis, L.J., Recker, F., Berenguer, A., Mttnen, L., Bangma, C.H., Aus, G., Villers, A., Rebillard, X., van der Kwast, T., Blijenberg, B.G., Moss, S.M., de Koning, H.J., Auvinen, A.: Screening and prostate-cancer mortality in a randomized European study. N Engl J Med **360**(13), 1320–1328 (2009)
48. Tsodikov, A., Szabo, A., Wegelin, J.: A population model of prostate cancer incidence. Stat Med **25**(16), 2846–2866 (2006)
49. Underwood, D.J., Zhang, J., Denton, B.T., Shah, N.D., Inman, B.A.: Simulation optimization of PSA-threshold based prostate cancer screening policies. Health Care Manag Sci **15**(4), 293–309 (2012)
50. Weinstein, M.C., O'Brien, B., Hornberger, J., Jackson, J., Johannesson, M., McCabe, C., Luce, B.R., ISPOR Task Force on Good Research Practices–Modeling Studies: Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices–modeling studies. Value Health **6**(1), 9–17 (2003)
51. Zauber, A.G., Lansdorp-Vogelaar, I., Knudsen, A.B., Wilschut, J., van Ballegooijen, M., Kuntz, K.M.: Evaluating test strategies for colorectal cancer screening: A decision analysis for the U.S. Preventive Services Task Force. Annals of Internal Medicine **149**(9), 659–669 (2008)
52. Zelen, M., Feinleib, M.: On the theory of screening for chronic diseases. Biometrika **56**(3), 601–614 (1969)

53. Zhang, J., Denton, B., Balasubramanian, H., Shah, N., Inman, B.: Optimization of PSA screening policies: A comparison of the patient and social perspectives. Med Decis Making **32**(2), 337–349 (2012)

A Literature Review

We searched PubMed (January 1, 2010, through October 3, 2015) using the following query:

```
    ("Early Detection of Cancer"[Mesh] OR "early diagnosis"[Mesh] OR
     "Prostatic Neoplasms/Diagnosis"[Mesh] OR "Mass Screening"[Mesh] OR screening)
AND ("decision analysis" OR "decision analyses" OR
     "Decision Support Techniques"[Mesh] OR "Decision Trees"[Mesh] OR
     "decision trees" OR "Cost-Benefit Analysis"[Mesh] OR
     "cost-benefit analysis" OR "Markov Chains"[Mesh] OR
     "Computer Simulation"[Mesh] OR "computer simulation" OR
     simulate OR simulation[all fields] OR simulating OR
     "Monte Carlo Method"[Mesh] OR "monte carlo method" OR markov)
AND ("prostate cancer" OR "Prostatic Neoplasms"[Majr])
```

We also searched the Tufts Cost-Effectiveness Analysis Registry [28] (from inception to October 3, 2015) for the term "prostate". Two citations were retrieved in full text, but were also identified in the PubMed searches. Figure 5 shows the results of searches and reasons for exclusion.

One modification was required to use Model Z [53] to assess an arbitrary PSA-based screening strategy. Given a patient's cancer status, Model Z assigns a probability that the patient will have a PSA value in the ranges $[0, 1)$, $[1, 2.5)$, $[2.5, 4)$, $[4, 7)$, $[7, 10)$, and $[10, \infty)$. We assume all PSA values in a range are equally likely to occur, and we limit to PSA values between 10 ng/mL and 20 ng/mL for the highest range.

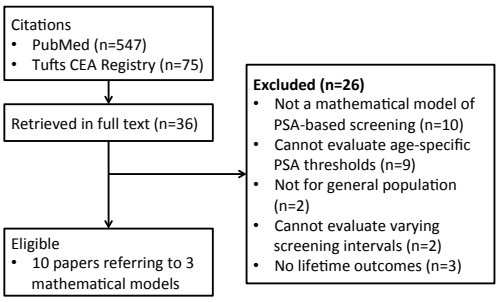

Fig. 5 Literature identification.

B Sensitivity Analysis

We varied parameters based on sensitivity ranges used in the papers describing each model. The variables names are from the respective papers.

B.1 Model G [24], parameters governing the course of the disease

We varied each of the following five parameters to the maximum and minimum value in the 100 sets of sensitivity parameters used in Gulati et al. [24].

- *grade.onset.rate*: A rate controlling how quickly patients experience prostate cancer onset.
- *grade.metastasis.rate*: A rate controlling how quickly patients with undetected prostate cancer experience metastasis.
- *grade.clinical.rate.baseline*: A rate controlling how quickly a patient's cancer is clinically detected.
- *grade.clinical.rate.distant*: A rate controlling how quickly a patient's cancer is clinically detected after metastasis.
- *low.grade.slope*: A parameter controlling the likelihood that a patient who developed cancer has a low-grade cancer.

B.2 Model U [49], parameters governing the course of the disease

We varied parameters using sensitivity ranges from [49].

- d_t: The rate of other-cause (non-prostate cancer) mortality at age t was varied ±20% from the base-case parameter values from [3, 42].
- w_t: The prostate cancer incidence rate for a man at age t was varied using sensitivity ranges from [8]. For patients aged 40–49, the sensitivity range was defined as [0.00020, 0.00501]; for patients aged 50–59, the sensitivity range was defined as [0.00151, 0.00491]; for patients aged 60–69, the sensitivity range was defined as [0.00243, 0.00852]; for patients aged 70–79, the sensitivity range was defined as [0.00522, 0.01510]; and for patients aged 80 or more, the sensitivity range was defined as [0.00712, 0.01100].
- b_t: The annual probability of metastasis among patients with detected cancer treated with radical prostatectomy was varied ±20% from the base-case parameter value of 0.006 derived from the Mayo Clinic Radical Prostatectomy Repository.
- e_t: The annual probability of metastasis among patients with undetected cancer was varied ±20% from the base-case parameter value of 0.069 from [22, 46].
- z_t: The annual probability of dying from prostate cancer among men aged t with metastatic disease was varied using the sensitivity range [0.07, 0.37] from [40, 4] around the base-case values of 0.074 for patients aged 40–64 and 0.070 for patients aged 75 and older [42].
- f: The probability of a biopsy detecting cancer in a patient with prostate cancer was varied ±20% from its base-case value of 0.8 from [26].

B.3 Model Z [53], parameters governing the course of the disease

The sensitivity analyses in [53] did not vary any parameters governing the course of the disease, and pertained only to costs and literature-derived quality-of-life decrements. Because of the similarities with model U, we used the sensitivity ranges from model U for model Z's d_t, w_t, and f parameters, additionally varying the following parameters:

- b_t: The annual probability of a man of age t with detected prostate cancer treated with radical prostatectomy dying of the disease was varied ±20% from its base-case value of 0.0067 for men aged 40–64 and 0.0092 for men aged 65 and older [42].
- e_t: The annual probability of a man of age t with undetected prostate cancer dying of the disease was varied ±20% from its base-case value of 0.033 from [1].

B.4 All models, quality-of-life decrements

Each model in this work uses the literature review-based quality-of-life decrements (utility weights) from a re-analysis of the ERSPC study from [29]. The sensitivity analysis ranges used in that work are as follows:

- *Screening attendance*: The utility estimate for the week following screening was varied in range [0.99, 1.00] from base estimate 0.99.
- *Biopsy*: The utility estimate for the three weeks following biopsy was varied in range [0.87, 0.94] from base estimate 0.90.
- *Cancer diagnosis*: The utility estimate for the month following cancer diagnosis was varied in range [0.75, 0.85] from base estimate 0.80.
- *Radiation therapy*: The utility estimate for the first two months after radiation therapy was varied in range [0.71, 0.91] from base estimate 0.73, and the utility estimate for the next 10 months after radiation therapy was varied in range [0.61, 0.88] from base estimate 0.78.
- *Radical prostatectomy*: The utility estimate for the first two months after radical prostatectomy was varied in range [0.56, 0.90] from base estimate 0.67, and the utility estimate for the next 10 months after radical prostatectomy was varied in range [0.70, 0.91] from base estimate 0.77.
- *Active surveillance*: The utility estimate for the first seven years of active surveillance was varied in range [0.85, 1.00] from base estimate 0.97.
- *Postrecovery period*: The utility estimate for years 1–10 following radical prostatectomy or radiation therapy was varied in range [0.93, 1.00] from base estimate 0.95.
- *Palliative therapy*: The utility estimate during 30 months of palliative therapy was varied in range [0.24, 0.86] from base estimate 0.60.
- *Terminal illness*: The utility estimate during six months of terminal illness was varied in range [0.24, 0.40] from base estimate 0.40.

C Building an Efficient Frontier of Screening Strategies

Given a screening strategy s, let $A(s)$ be the average assessment of the strategy across all mathematical models and let $P(s)$ be the pessimistic assessment of the strategy across all mathematical models. To construct an efficient frontier of strategies trading off the average and most pessimistic assessment, we use mathematical optimization via an iterated local search heuristic to maximize the objective function $\lambda A(s) + (1 - \lambda)P(s)$ for $\lambda \in \{0, 0.1, 0.2, \ldots, 1.0\}$ over annual screening strategies and biennial screening strategies, optimizing a total of 22 times. From the set of all screening strategies encountered during the optimization process (not just the final values identified through optimization), we construct an efficient frontier trading off the average and pessimistic assessments. Solutions encountered while optimizing the objective with parameter value λ using iterated local search may not be optimal for the objective with that λ but may still lie on the efficient frontier trading off the average and pessimistic assessments, so the final efficient frontier may contain more than 22 efficient strategies.

The key step in constructing the efficient frontier is solving $\max_{s \in S} \lambda A(s) + (1-\lambda)P(s)$, where S is the set of all feasible screening strategies. We consider strategies with age-specific PSA cutoffs limited to 0.5, 1.0, 1.5, ..., 6.0 ng/mL, fixed cutoffs for 5-year age ranges, and cutoffs that are non-decreasing in a patient's age. We consider screening from ages 40 through 99, so there are 10.4 million possible screening strategies; as a result, it would be time consuming to use enumeration to identify the strategy with the highest average incremental QALE compared to not screening. Instead, we use constrained iterated local search to identify a locally optimal strategy that cannot be improved by changing a single age-specific PSA threshold.

The central step in the iterated local search is the local search, which takes as input a screening strategy s and a single age range r and searches a small number of similar strategies to s. For each possible PSA threshold (0.5, 1.0, ..., 6.0 ng/mL), the local search procedure constructs a new strategy by modifying s to use that threshold in age range r,

additionally making the smallest possible changes to the remaining PSA thresholds in s to retain non-decreasing PSA thresholds in age. Each of these 12 screening strategies is evaluated, and if any improves over s then the one with the best objective value is selected to replace s. In the case where r is either the first or last age range in s in which patients screen, the procedure also considers a no-screening option for age range r.

As an example, consider a screening strategy for which annual screening is performed for ages 45–69, with cutoff 2.0 ng/mL from ages 45–49, 3.0 ng/mL from ages 50–54, and 5.0 ng/mL from ages 55–59, 60–64, and 64–69. We can write this screening strategy compactly as $(2, 3, 5, 5, 5)$, with each value in the vector representing the cutoff for a 5-year period. If we apply local search to the cutoff for ages 55–59, then we will consider changing the cutoff for that age range to each value in $\{0.5, 1.0, 1.5, \ldots, 6.0\}$ ng/mL, adjusting other cutoffs the smallest amount possible to ensure all cutoffs are non-decreasing in age. For instance, if the cutoff for ages 55–59 were set to 2.5 ng/mL, then the cutoff for ages 50–54 would also need to be decreased to 2.5 ng/mL in order to maintain non-decreasing cutoffs, yielding final screening strategy $(2, 2.5, 2.5, 5, 5)$. The set of all possible screening strategies considered by a local search on age range 55–59 is:

```
(0.5, 0.5, 0.5, 5, 5)
(1, 1, 1, 5, 5)
(1.5, 1.5, 1.5, 5, 5)
(2, 2, 2, 5, 5)
(2, 2.5, 2.5, 5, 5)
(2, 3, 3, 5, 5)
(2, 3, 3.5, 5, 5)
(2, 3, 4, 5, 5)
(2, 3, 4.5, 5, 5)
(2, 3, 5, 5, 5)
(2, 3, 5.5, 5.5, 5.5)
(2, 3, 6, 6, 6)
```

Among these strategies, the one resulting in the largest objective value $\lambda A(s) + (1 - \lambda)P(s)$ is the one selected by the local search.

The iterated local search begins with a strategy of never screening for prostate cancer. The procedure repeatedly loops through a random permutation of the age ranges, performing local search on an age range if it's within 5 years of an age range for which the current strategy screens with PSA. The procedure terminates when the current screening strategy cannot be improved by applying local search to any valid age range.

D Details of Optimizing Screening Strategies

The iterated local search procedure was implemented in python. The C source code for model G and the C++ source code of model U were provided by the authors of those works; model U was re-implemented in python to improve the efficiency of the procedure. Model Z was implemented in python based on the published description of that model. All procedures were tested on a Dell Precision T7600 with 128 GB RAM and two Intel Xeon E5-2687W Processors, each with 8 cores and a clock speed of 3.1 GHz.

The runtime of the iterated local search procedure for each objective function is provided in Table 4.

To validate the performance of the local search optimization approach, we computed the exact optimal solution for models Z and U by evaluating all 5.2 million feasible biennial strategies and all 5.2 million feasible annual strategies with each model, a process that required 60.1 CPU hours for model Z and 369.1 CPU hours for model U. The local search heuristic had identified the global optimal solution for models Z and U. Given the heavy computational burden of evaluating strategies with model G, we did not compute exact optimal solutions for model G or for any of the objectives used to compute the efficient frontier.

Table 4 Computation time required for iterated local search procedure

	Runtime (minutes)	
Objective	**Annual Strategies**	**Biennial Strategies**
Single-Model: Model G	183	225
Single-Model: Model U	1	1
Single-Model: Model Z	1	1
Efficient Frontier: $\lambda = 0.0$	185	227
Efficient Frontier: $\lambda = 0.1$	338	248
Efficient Frontier: $\lambda = 0.2$	213	247
Efficient Frontier: $\lambda = 0.3$	251	288
Efficient Frontier: $\lambda = 0.4$	212	171
Efficient Frontier: $\lambda = 0.5$	203	271
Efficient Frontier: $\lambda = 0.6$	161	237
Efficient Frontier: $\lambda = 0.7$	161	237
Efficient Frontier: $\lambda = 0.8$	161	226
Efficient Frontier: $\lambda = 0.9$	199	263
Efficient Frontier: $\lambda = 1.0$	266	248

E Sensitivity Analysis: Model Averaging of Normalized Assessments

As a sensitivity analysis, we reproduced the efficient frontier using a normalized version of the objective function. For each model, we normalized the QALE change compared to not screening to have a maximum value of 1, ensuring that models with systematically more optimistic assessments of screening strategies are not weighted more heavily than others in the model averaging objective.

We computed an efficient frontier as before, trading off the average and most pessimistic assessment of the normalized objective function. The efficient frontier, single-model solutions, and expert strategies are plotted in Figure 6.

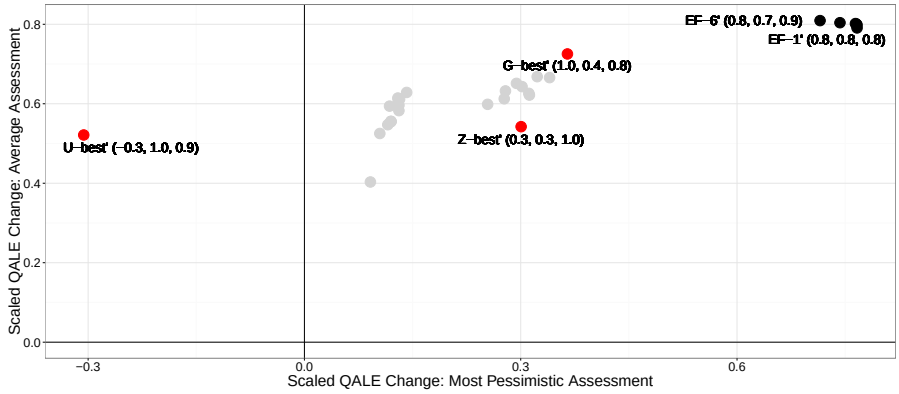

Fig. 6 Average and most pessimistic assessments of identified and expert-generated screening strategies. The 6 strategies on the efficient frontier are shown as black circles. The optimal strategies according to models G (G-best), U (U-best), or Z (Z-best) are shown as red circles. The 22 expert-generated strategies are shown as gray circles. Normalized assessments of QALE over no screening with each model are shown in parentheses for some strategies. For example strategy EF-1$'$ was assessed as 0.8 proportion of the maximum attainable QALE improvement by models G, U, and Z.

The efficient frontier with the normalized objective function is qualitatively different from the efficient frontier with the non-normalized objective. No screening strategy optimized with a single model falls on the efficient frontier, and all strategies in the efficient frontier dominate the single-model and expert-generated strategies in both the most pessimistic and the average model assessments. The efficient frontier is smaller, comprising only six screening strategies, and the strategies in the frontier are more homogeneous. The strategy optimizing the most pessimistic assessment prescribes biennial screening with threshold 0.5 ng/mL from ages 40–54, 1.5 ng/mL from ages 55–64, 4.0 ng/mL from ages 65–69, and 5.0 ng/mL from ages 70–74. The strategy optimizing the average assessment is similar, prescribing biennial screening with threshold 0.5 ng/mL from ages 40–49, 1.0 ng/mL from ages 50–59, 1.5 ng/mL from ages 60–64, 2.0 ng/mL from ages 65–69, and 6.0 ng/mL from ages 70–79. For all six strategies on the efficient frontier model U was the most pessimistic in the normalized assessment.