MAIN ARTICLE

"Saving lives or harming the healthy?" Overuse and fluctuations in routine medical screening

Özge Karanfil[a,b]* and John Sterman[c]

Abstract

Tests to screen for certain diseases—for example, thyroid cancer screening, screening mammography, and screening of high blood pressure for hypertension—are increasingly common in medical practice. However, guidelines for routine screening are contentious for many disorders and often fluctuate over time. Some tests are over- or underused compared to available evidence that justifies their use, with clinical practice persistently deviating from evidence-based guidelines. Here we develop an integrated, broad boundary feedback theory and formal model to explain the dynamics of routine population screening including fluctuations in policy-decision thresholds and the expansion of selection criteria which may lead to inappropriate use. We present a behaviorally realistic, boundedly rational model of detection and selection for medical screening that explains the potential of endogenous oscillations in practice guidelines as decision-makers—including epidemiologists, clinicians, and patients, or policymakers from guideline issuing organizations, perceive harms and benefits from potential outcomes and make trade-offs between sensitivity and specificity by altering the existing guidelines and actual practice. The model endogenously generates fluctuations in screening indications, test thresholds, test efficiency, and the target screening population, leading to long periods during which practice guidelines are suboptimal even if the underlying evidence base is constant. We use cancer screening as a motivating example, but the model is generic with a wide range of potential applications for important managerial problems in medical contexts, such as screening for hypertension, hypercholesterolemia, autism spectrum disorder, Alzheimer's disease, and related dementia. It also applies to other managerial problems in nonmedical contexts, such as airport screening, background checks, tax audits, automotive emission tests, contentious jurisdiction, or to consumers of other kinds of information who need to make a decision—on behalf of an individual, or for the whole population.

Syst. Dyn. Rev. (2020)

Additional Supporting Information may be found online in the supporting information tab for this article.

Introduction

Practice guidelines are developed for various reasons, including the emergence of new, potentially practice-changing scientific evidence or a perceived need for guidance in times of uncertainty. Tests to screen for certain disorders, for example thyroid cancer screening or screening of high blood

[a] Koç University, College of Administrative Sciences and Economics, Department of Operations Management, Istanbul, Turkey
[b] Koç University, School of Medicine, Istanbul, Turkey
[c] MIT System Dynamics Group at the MIT Sloan School of Management, Cambridge, Massachusetts, USA
* Correspondence to: Özge Karanfil. E-mail: okaranfil@ku.edu.tr; karanfil@hsph.harvard.edu
Accepted by Luis Luna-Reyes, Received 16 September 2019; Revised 5 May 2020; Accepted 30 June 2020

System Dynamics Review
System Dynamics Review
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/sdr.1661

pressure for hypertension, are increasingly common in medical practice. There is universal agreement that their guidelines should be based on the best available scientific evidence. The classical approach to setting evidence-based guidelines is based on the fundamental trade-off between Type I and Type II errors, in which policymakers seek the optimal balance between sensitivity (and thus the risk of false positives) and specificity (and thus the risk of false negatives), given the costs and benefits of different outcomes.[i]

However, guidelines for routine screening are contentious for many disorders and often fluctuate over time. Some tests are over- or underused compared to available evidence that justifies their use, while clinical practice deviates from evidence-based guidelines. Over the last few decades, the selection and detection criteria for screening and disease definitions for several important disorders have changed significantly, including biomarker thresholds dividing positive from negative test results and the recommended ages for routine screening. Major health organizations have recommended changes in several common disease definitions, often resulting in the expansion of the criteria for screening, diagnosis, and treatment, leading to increases in reported disease incidence and prevalence, which justify additional treatment or immediate action (Croswell *et al.*, 2010; Esserman *et al.*, 2014; Hoffman and Cooper, 2012).

A good example of a recent controversy is changes in the traditional definition of hypertension. In 2017, the American College of Cardiology (ACC) and the American Heart Association (AHA) revised the guidelines for management of high blood pressure, adding a new category: Stage 1 hypertension, previously termed as "prehypertension." Under this definition (which decreased the systolic blood pressure reading from 140 to 130 mmHg), the number of U.S. adults with the condition increased from 72 to 103 million, or from 32 percent to 46 percent, affecting nearly half of the adult population (Bakris and Sorrentino, 2018). Around the same time, the National Institute on Aging (NIA) commissioned a study from the National Academy of Sciences (NAS): "Preventing Cognitive Decline and Dementia: A Way Forward" to suggest three main interventions to delay or slow age-related cognitive decline, one of them being intensive treatment of blood pressure (NAP, 2017). A recently published JAMA editorial offers additional support for intensive treatment of blood pressure in reducing the risk of developing mild cognitive impairment, by shooting for a target number of 120 or lower (Yaffe, 2019), a further departure from the traditional guidelines that targeted 140 or lower.

[i]Sensitivity" is a test's ability to correctly identify the truly positive cases, defined as the fraction of true cases yielding a positive test result. "Specificity" is a test's ability to correctly identify the truly negative cases, defined as the fraction of healthy individuals whose test result is negative.

DOI: 10.1002/sdr

These trends have important repercussions in aging societies where efforts to prevent cognitive decline is increasing (Hellmuth *et al.*, 2019) and of particular concern for cancer screening: based on 2013–15 data, 38.4 percent of Americans will be diagnosed with cancer at some point during their lifetime (NCI Cancer Statistics, 2018).

Despite the importance of "getting it right," clinical practice guidelines (CPG's) for many diseases differ among different stakeholders (epidemiologists, clinicians, patients, and patient advocacy groups), including the selection criteria for routine screening such as the recommended starting age and the threshold in test results (e.g. fasting glucose) indicating referral for biopsy or treatment, potentially increasing use of medical care. Guidelines for diagnosis and indications for treatment for hypercholesterolemia, hypertension, thyroid cancer, prostate specific antigen (PSA) testing, mammography, routine pelvic exam, and neurodegenerative diseases to name just a few, vary substantially across the United States, sparking confusion and controversy for the public (Belluck, 2013; Pollack, 2013; Rabin, 2009, 2014;). See Table 1 for a summary of some recent changes in screening indications for some major conditions. This puts potential excessive testing at scale, calling for transparency in the regulatory capture and algorithms used to define these thresholds (Bakris and Sorrentino, 2018; Mandl and Manrai, 2019; Schwartz and Woloshin, 2019; Welch, 2017).

To illustrate a detailed example, Figure 1 shows changes in the guidelines for routine PSA testing in the United States promulgated by the U.S. Preventive Services Task Force (USPSTF), the American Cancer Society (ACS), the American Urological Association (AUA), and the National Comprehensive Cancer Network (NCCN), with respect to the recommended starting age. Only the ACS guidelines remain constant over the last few decades, though they switched to shared decision-making (SDM) after 2010; the other guidelines fluctuate. Note also that the USPSTF, an independent scientific volunteer panel of 16 members who are charged with evaluating the scientific evidence and independent of specialty groups such as the AUA and others, concluded that the evidence did not justify routine PSA screening in 2012 for men of all ages (Moyer, 2012; USPSTF, 2018). Some believe that their recommendations went too far, and they criticized their approach (e.g. Schröder, 2011; Wilt *et al.*, 2014). Growing concerns about their recommendations led to the USPSTF Transparency and Accountability act (Blackburn M. H.R.1151 - 114th Congress (2015–2016). USPSTF issued a statement in 2017 recommending to inform men ages 55–69 years about potential benefits and harms indicating SDM. Recent recommendations suggest SDM and no screening for men over 70 (USPSTF, 2018). A draft version of this recommendation statement was posted for public comment on the USPSTF website from April 11 to May 8, 2017.

Evidence-based guidelines are often not followed by clinicians and patients, with significant overscreening for some tests and underscreening for others: practice often does not follow the evidence. For example, prostate screening

DOI: 10.1002/sdr

Table 1. Variation in clinical practice guidelines (CPGs)

Change in CPG	Direction of Change in Breadth Selection Criteria	News Coverage
Blood Cholesterol (2018)	Broadening; bad cholesterol levels Narrowing; doctors should not put most people on cholesterol-lowering medications like statins based on cholesterol levels alone Broadening…	New Cholesterol Guidelines Abandon LDL Targets (Riordan, 2013) Don't Give More Patients Statins (Abramson and Redberg, 2013) Do Latest US Guidelines Bypass, or Spare, Millions From Statins? (Wendling. 2017) New AHA/ACC Cholesterol Treatment Guideline Expands Role of LDL Targets (Stiles, 2018) Cholesterol targets are back! (Bhatt, 2018)
Hypertension (2017)	Broadening the breadth selection criteria Narrowing the breadth selection criteria Broadening, prehypertension became Stage 1 Htn	Hypertension Guide May Affect 7.4 Million (Kolata, 2013) Hypertension Guidelines Can Be Eased, Panel Says (Kolata, 2013) Don't Let New Blood Pressure Guidelines Raise Yours (Welch, 2017) Why New Blood Pressure Guidelines Could Lead to Harm (Carroll, 2017)
Screening Mammography (2016)	Broadening; initiate mammograms at 40 Narrowing, especially for women in their 40s and 70s	New Guidelines on Breast Cancer Draw Opposition (Rabin, 2009) Panel Urges Mammograms at 50, Not 40 (Kolata, 2009) Start mammograms at age 40 not 50 (Change.org; 2020)
Prostate Screening (2018)	Broadening; screen men over 50, Narrowing; do not screen men of any age Broadening; screen men between 55 and 69 selectively	Deciphering the Results of a Prostate Test (Brody, 2007) Prostate Screening Guidelines are Loosened (Pollack, 2013) Prostate Cancer Screening Still Not Recommended for All (Parker-Pope, 2014) New Study Offers Support for Prostate Testing (Rabin, 2017)
Routine Pelvic Exam (2019)	Broadening: pelvic exam suggested for women Narrowing the breadth selection criteria	Guideline Calls Routine Pelvic Exams Unnecessary (Rabin, 2014) Many young women get unnecessary pelvic exams (Rapaport, 2020)

DOI: 10.1002/sdr

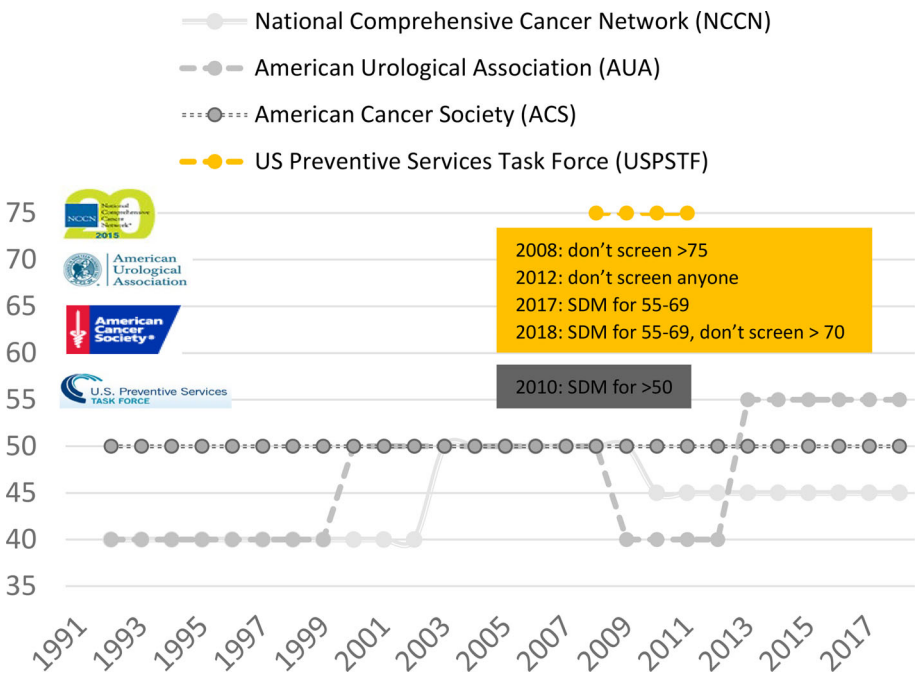

Fig 1. Recommended PSA test starting age for asymptomatic men (SDM: shared decision-making). [Color figure can be viewed at wileyonlinelibrary.com]

with less benefits is more common among men in the United States than all types of colorectal screening (fecal occult blood testing, flexible sigmoidoscopy, or colonoscopy) with well-established mortality benefit (Sirovich *et al.*, 2003). Here we develop a dynamic hypothesis to formalize the policy-formation process for practice guidelines and explain the problem of endogenous cycles and expansion of selection criteria in population screening. Our theory explains inappropriate use and oscillations in screening indications and the proposed formal model endogenously generates fluctuations in indications even when the underlying evidence base were to stay constant.

In medicine, it is known that evaluations may affect both adoption and patient selection (i.e. extent of use) (Fineberg, 1985). Homer's (1987) approach to model diffusion of medical technologies uses this principle, which has parallels to our work of modeling the formation and dissemination/adoption of CPGs. His model is particularly useful for analyzing evolving technologies, where both technology's acceptance and its extent of use are subject to change over time as a result of new performance-related information, where performance itself is affected by changes in technology or its application.

We argue that variations in screening (policy) thresholds may also arise not from changes in the benefits and harms environment of screening such as technological advances, but from the interaction of delays between

DOI: 10.1002/sdr

generating and assimilating scientific evidence with boundedly rational decision-making, including biases in judgment and decision-making arising from widely used heuristics. This is in line with Hammonds cognitive continuum theory which states "quasirationality" as an important middle ground between intuition and analysis. Accordingly, cognitive performance is dictated by the match between task properties and mode of cognition and that there is an oscillation between these two (Hammond, 2007). We develop a stylized model that is realistic enough to replicate the fluctuations in practice guidelines for, and overuse of, important medical screening tests such as mammography and PSA testing, yet generic enough to be adapted to other medical contexts, such as Alzheimer's screening, prenatal screening, early screening for autism or thyroid screening, as well as nonmedical settings such as airport screening, applicant background checks, or tax audits. We use a mix of quantitative and qualitative methods including semistructured expert interviews, a medical literature search, and empirical data collection on how screening criteria have evolved over time to formulate a system dynamics (SD) model for population screening (Forrester, 1961; Sterman, 2000).

We first review the theory relevant to judgment and decision-making and setting decision thresholds in the presence of Type I and Type II errors, then present the model structure, parametric assumptions, results, and sensitivity analysis. We conclude with implications for theory and practice and discussion of model limitations with potential extensions.

Background

We draw on a large literature in the social sciences, judgment and decision-making, psychology, marketing research, political sciences, and finally public health and medicine in which decision-makers must set a threshold for classifying a condition as positive or negative. Swets played a key role in adapting signal detection theory and specifically the Receiver Operating Characteristic (ROC) curve to the psychology of perception (Green and Swets, 1966; Swets, 1964). He was the first to describe shifting and cycling decision thresholds. In medicine, Pauker and Kassirer (1980) introduced the concept of the "therapeutic threshold"—a probability of disease that constitutes a point of indifference between treating and not treating. Later on in political science, Schlesinger (1986) proposed the concept of "regular oscillations" to describe cyclical variations in dominant political parties in his book, *The Cycles of American History*.

In his famous book on human judgment and decision-making, Hammond (1996) attempted to understand the policy-formation process. He proposed that any policy problem involving irreducible uncertainty has the potential for dual error, and policy thresholds may oscillate over time due to

DOI: 10.1002/sdr

opposing pressures coming from constituencies representing those treated unfairly. Accordingly, any imperfect test that employs a threshold would lead to some error and yield an irreducible uncertainty/inevitable error/unavoidable injustice to some constituency, namely, the false positives (incorrect rejection of a true null hypothesis) and the false negatives (incorrectly retaining a false null hypothesis). High, and especially salient and consequential, realizations of each type of error would lead to pressure on policymakers to move the threshold to reduce the error, but at the cost of increasing the other error type, eventually creating a counter pressure and causing cycling of that decision threshold over time. He argued that there are oscillations in public and professional attitudes with implicit policy thresholds and that those cycles would last about 30 years across decision domains (Hammond, 1996), concluding "If such oscillations can be shown to exist, and if they can be shown to have a definite period … then we have at hand not only a means for predicting our future political climate far in advance, but an important phenomenon that strongly invites, indeed, demands, analysis and interpretation." In his last book entitled *Beyond Rationality*: *The Search for Wisdom in a Troubled Time*, published at age 92, he characterized movement along the cognitive continuum as oscillatory or alternating between intuition and analysis, and the key to wisdom lies in being able to match modes of cognition to properties of the task (Hammond, 2007).

A related line of research suggests that physicians' decision thresholds may vary over time. Stewart and Mumpower (2004) and Swets *et al.*, (2000) and Stewart et al. (2012) document wide variation among radiologists' decisions regarding the interpretation of mammograms and the appropriate trade-off between false positives and false negatives. This conflicts with the view that clinicians' judgmental accuracy is fixed and suggests that both the thresholds suggested by formal guidelines and physicians' actual decision thresholds might fluctuate. We know that this is true for PSA screening and for its formal versus actual biopsy thresholds (Gulati *et al.*, 2010).

Stewart and Mumpower (2004) describe different domains of decision-making about mammography screening, focusing on the decisions made by radiologists in their practice, and the variation in radiologists' decisions. These domains include the decisions by women and their doctors to obtain screening, decisions by radiologists to recommend biopsy, and decisions of policymakers regarding criteria for routine screening.

In the system dynamics domain, Weaver and Richardson (2006) developed a model based on the policy threshold cycling theory of Hammond and other scholars (Hammond, 1996; Schlesinger, 1986; Swets, 1992). They first present a simplified theory of Hammond's initial insight and then develop three alternative models: one with delays in policy-maker responsiveness; one with shifts in stakeholders' constituencies in response to recent errors; and one with integral control representing the historical dissatisfaction of

DOI: 10.1002/sdr

competing constituencies. More recently, Sheldrick *et al.*, (2016) developed an SD model to explore how clinicians referral decisions may be influenced by changes in context, using developmental and behavioral screening as a case study, and Lyon *et al.*, (2016) modeled the impact of school-based universal depression screening on service capacity needs.

Some of the other SD cancer-screening modeling studies include Kivuti-Bitok *et al.* (2014) who developed a model for cervical cancer vaccination and screening interventions in Kenya; Palma *et al.* (2016) who replicated the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) for serum PSA screening for prostate cancer; and Fett (2001) who replicated the Swedish two-county trial of mammographic screening for breast cancer, which are all population-level models. For a more detailed discussion on medical screening models in SD, Darabi and Hosseinichimeh (2020) provide an extensive picture of the field.

Our model differs from these models as it is not at population level in the sense that the real U.S. male population and their screening-related metrics are replicated here in detail, but instead that populations's collective response to changing guidelines, demographics, and the policy thresholds are modeled in a stylized way together with a realistic screening test and parameters, calibrated for the PSA test for prostate cancer.

Screening Model Overview

We begin with the minimal feedback structure required to model threshold determination and resulting impact on test sensitivity and specificity (the evidence-based "core model"). We then gradually expand the model boundary to include additional feedbacks including the interpretation and implementation of formal guidelines by clinicians and patients, testing the model structure and its behavior throughout. Karanfil (2016) and Karanfil *et al.* (2017) provides an extensive discussion for the rationale and history of the policy-formation process for the population screening problem, mainly in the U.S. context, by keeping a comparative perspective between the United States and Europe and between PSA and other medical-testing problems.

Figure 2 shows the boundary of the full model, including the evaluation of evidence, benefits, and harms resulting from actual clinical practice and patient outcomes and delays in assessing and responding to these outcomes. The model boundary emerged with respect to the problem of concern, and at the core, it includes the endogenous "system variables" which include selection and detection criteria, harms and benefits of screening, disease prevalence in the target screening population, and the test diagnostics themselves. These interact to create broad boundary feedbacks that give rise to the (problematic) behavior and condition the adoption of and adherence to CPG's and

DOI: 10.1002/sdr

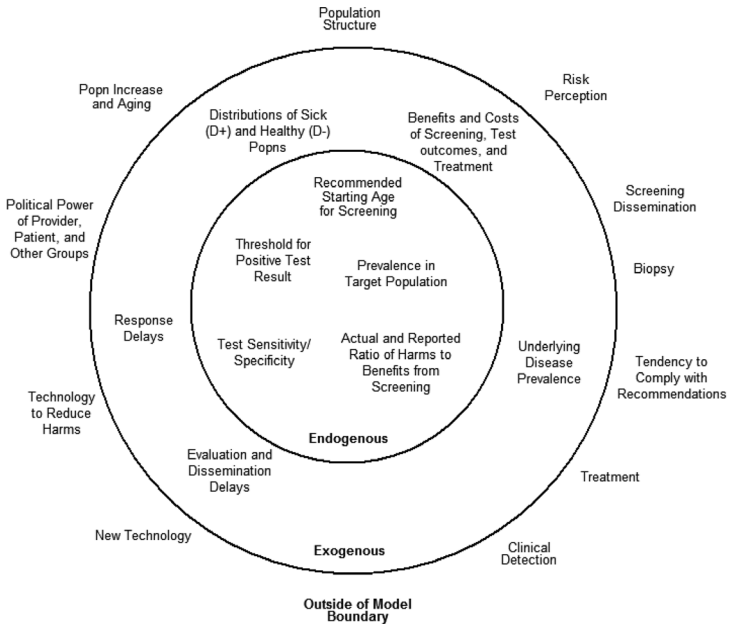

Fig 2. Model boundary diagram.

hence are modeled explicitly. Exogenous variables are the parameters and other time-series information that are used as an external input for the simulation, and other nonrelevant concepts are left intentionally outside of the model boundary. Since screening practice in the United States started in the 1980s, the time horizon of the model is roughly selected as 1980–2080 to show the time scale of the screening problem, which is in line with Hammonds suggested cycles in professional attitudes.

Classical Approach to Setting Evidence-Based Guidelines

At the core of all evidence-based guidelines, there is a decision-theoretic framework which is the first and fundamental step in all types of screening, including medical screening. Ideally, this first step relies only on evidence about the benefits and harms to patients from screening and potential follow-up. For routine medical screening, evidence includes the available options (e.g. screening or not screening), reports on the likelihoods of, and the benefits and harms associated with, the possible outcomes including true positives (e.g. cases diagnosed, lives extended), false positives (e.g. harms from unnecessary biopsy, anxiety, and consequent treatment), true negatives, and false negatives (e.g. harms from undiagnosed and untreated disease). Benefit and harm calculations should consider various potential impacts on patients including morbidity, mortality, and other formal

DOI: 10.1002/sdr

measures of disease burden such as lost quality or disability-adjusted life years (QALYs or DALYs), including the impacts arising from any follow-up treatments, along with the emotional costs of anxiety created by testing and its outcomes, or alternatively, positive responses to coping information (Kahn and Luce, 2003).

The classical approach to setting guidelines for screening is to seek an evidence-based balance between the sensitivity and specificity of a diagnostic test (Figure 3). However, the distribution of PSA levels for the D^+ and D^- populations overlap inevitably; various degrees of overlap are typical in many diagnostic settings. Hence, for any threshold, there will be nonzero rates of Type I or Type II error (or both), with higher errors as the overlap between the D^+ and D^-populations increases. PSA was first identified in 1970 as a biomarker of the prostate gland (Ablin *et al.*, 1970).

The Receiver Operating Characteristic (ROC) curve is the most commonly used tool to evaluate the diagnostic performance of a screening test (Metz, 1978) and can also be used to compare the diagnostic performance of two or more different tests (Griner *et al.*, 1981). In a ROC curve, the True Positive rate, TPR = Sensitivity, is plotted against the False Positive rate, FPR = 1 − Specificity. For any test, the higher the sensitivity (fraction of true cases detected by the test), the higher the rate of false positives.

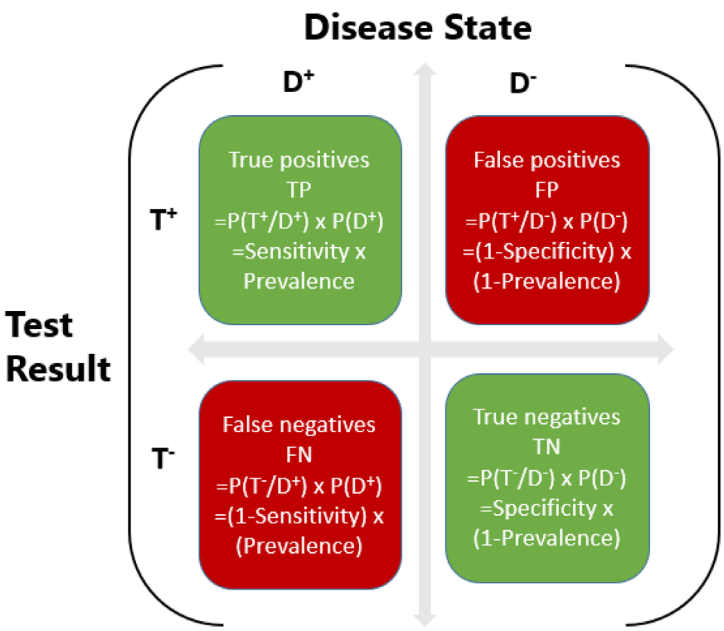

Fig 3. Decision matrix: potential results of a screening test (T^+test positive, D^+disease present). [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/sdr

Consider two groups in a population, one with a specific condition or disease (denoted D^+) and the other without the disease (D^-). Diagnostic tests are typically imperfect. For example, PSA testing for prostate cancer measures the level of PSA in the blood, with levels above a certain threshold taken to indicate a positive result (T^+) for possible cancer (leading to biopsy and possibly treatment) and levels below the threshold taken to indicate a negative result (T^-). Deciding on the threshold for an individual is different than the one for the population and depends on individual risk factors such as age and family history. Potential long-term benefits and risks of screening differ greatly in younger versus older people. In addition, the optimal threshold is shaped by individuals attitudes about available treatments and consequences, which, in turn, can be affected by education levels, cultural norms, and other socioeconomic factors (Cutler and Lleras-Muney, 2010).

Model of Practice Guidelines

Figure 4 provides an overview of the full model, including the policy structure for development of evidence-based screening (Core Model) and the

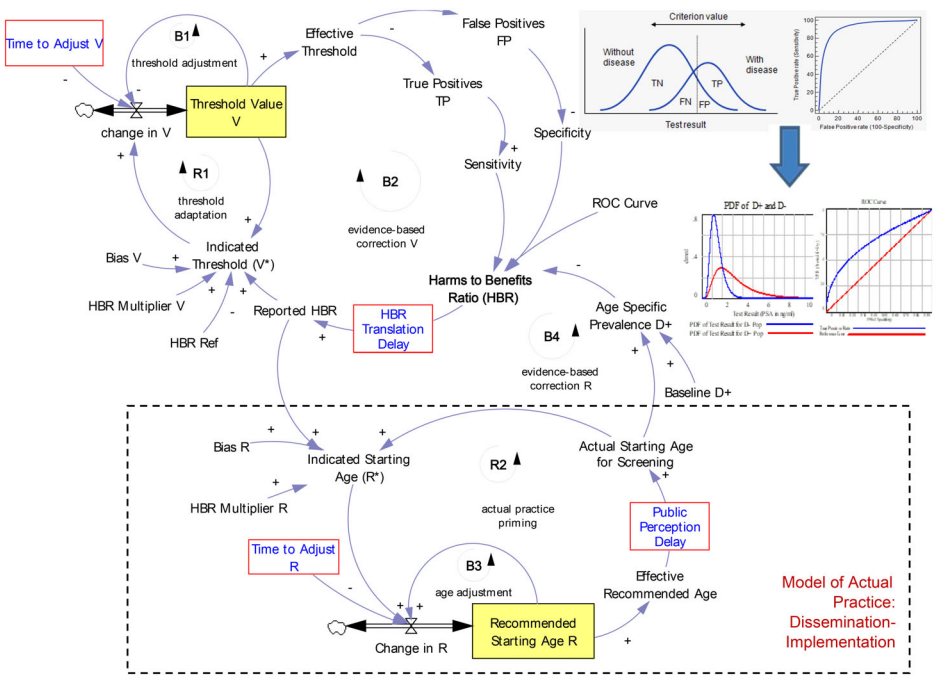

Fig 4. Overview of full model showing the key feedback structures. [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/sdr

policy structure for guidelines in use (Model of Actual Practice). The two critical state variables are the *Recommended Starting Age* (*R)* and the test *Threshold Value (V)*. Test values above *V* indicate a positive result (the possibility of disease); values below *V* indicate a negative result (no disease). The optimal values are not known but must be discovered over time. The threshold and recommended age each adjust in response to the reported ratio of benefits to harms, using a hill-climbing search procedure (Sterman, 2000).

Hill climbing is a plausible heuristic to adjust *V* when its optimal value is unknown: the guideline is gradually changed in the direction that is perceived to improve performance, which is in line with the heuristic suggested in Hammond (2007). Balancing feedback B1 adjusts the actual threshold *V* toward the indicated threshold value *V* *, the threshold implied by the reported Harm-to-Benefit Ratio (*HBR)* and, possibly, other external pressures arising from, e.g. medical and treatment providers and patient advocacy groups. The *V* * is anchored on the current *V*, so that the threshold falls (increasing sensitivity) as long as the *Reported HBR*, *or RHBR* is favorable and rises (decreasing sensitivity) as long as the *RBHR* is unfavorable, creating the reinforcing threshold-adaptation feedback R1. As long as the net effect of the pressures on the threshold goal causes *V** to exceed *V*, the threshold will grow; otherwise it will decay. The formulation for the changes in *R* is analogous.

Evidence documenting harms and benefits should lead to convergence to the appropriate threshold V and starting age *R*, proxies for breath-selection criteria. Balancing feedback B2 provides evidence conditioning decisions about increasing or decreasing the threshold based on the *RHBR*. As the threshold rises, specificity rises (fewer false positives), but sensitivity falls (more false negatives). The *HBR* will adjust, leading to further changes in the threshold. The threshold will equilibrate at the level that yields the optimal *HBR*, which may vary based on context.

However, gathering, publishing, and responding to scientific evidence takes time. There are two major time delays: the delay in collecting, evaluating, and reporting evidence on the harms and benefits, denoted the *HBR Translation Delay* (λ_t),and the delay in public's response to changing recommendations, denoted the *Public Perception Delay*, (λ_p). In addition, it takes time for the *V* and *R* to adjust given a gap between the indicated and actual values, denoted the *Time to Adjust T* (τ_V) and *Time to Adjust R* (τ_R), respectively.

Formally, the threshold value, *V*, adjusts toward the indicated value, *V**, over the threshold-adjustment time, τ_V. The indicated threshold *V** is anchored on the current value and then adjusted by a function of the *RHBR* and a bias, α_V. The bias captures possible external pressures for higher or lower thresholds that may arise from patient advocacy groups, payers, the public, or other interest groups:

DOI: 10.1002/sdr

$$\frac{dV}{dt} = \frac{(V^* - V)}{\tau_V} \tag{1}$$

$$V^* = f_V(RHBR)\alpha_V V. \tag{2}$$

Here we treat the bias α_V as a constant. The adjustment to V^* arising from evidence about the harm-benefit ratio is an increasing function of RHBR, the reported HBR, here formulated as a constant elasticity with value βv, and where the HBR_{ref} is the reference, or the optimal level of HBR, chosen as 1. In real life, the optimal level of HBR can take different values depending on the costs we put on potential harms and benefits for that disease.

$$f_V(RHBR) = 1 + \beta_V\left(RHBR - HBR_{ref}\right); \beta_V > 0. \tag{3}$$

The *RHBR* lags the actual value based on the current actual test threshold due to the delays in carrying out, evaluating, and publishing data on harms and benefits. Thus:

$$RHBR = \mathcal{L}(HBR, \lambda_t), \tag{4}$$

where $\mathcal{L}$ is the lag operator, with mean lag λ_t.Because the evidence collection and reporting process has multiple stages, we use a third-order Erlang lag. The HBR aggregates the evidence on the harms and benefits of routine screening. Possible benefits include treating true positives to prevent cancer death or increase the patient's quality of life. Possible harms of screening include failing to treat false negatives and erroneously treating false positives. Harms can also include anxiety, distress, and other psychological responses associated with the test and with false positive and false negative results, unnecessary follow-up testing, and overdiagnosis (finding cases that would not have resulted in clinically significant disease in the patient's lifetime), as described in an interview with a policy maker, a former member of a guideline committee:

> [Benefits are] saving lives, improving quality of life.…Potential harms? Well you have the harms of screening itself, so the screening test may be expensive or painful or difficult, and then you have the follow up of the false positives, which may be expensive, difficult, painful, and anxiety producing. So that's all about the test. Now once you have a positive test, a positive screen, you still may have harms because of the treatment.... [I]f they are false positives they can only get harm; they can't get benefit. If they are true positives, they can get benefit, but they can also get harm. (Policy Maker, Academic)

The harm-to-benefit ratio is defined as:

DOI: 10.1002/sdr

$$HBR = H/B, \tag{5}$$

where H and B are harms and benefits. Harms and benefits, in turn, are the sum of the harms and benefits associated with each of the four possible combinations of test outcomes:

$$H = \sum\nolimits_{T,D} h_{T,D}(p_T D)p_D \tag{6}$$

$$B = \sum\nolimits_{T,D} b_{T,D}(p_T D)p_D \tag{7}$$

$$T \in \{T^+, T^-\}; D \in \{D^+, D^-\},$$

where h and b are the harms and benefits for an individual associated with each of the four possible test outcomes; $p_T \mid D$ is the probability of each test outcome (positive, T^+, or negative, T^-), conditioned the true disease state, positive, D^+, or negative, D^-; and p_D is the probability of each disease state. We choose values for h and b such that true positive and true negative results yield net benefit, while false positive and false negative results yield net harm.

The probability of each test outcome conditional on the true disease state depends on the distribution of test values for each disease state, as explained above. For PSA screening, the distribution of PSA test values for the D^+ and D^- populations are reasonably approximated by a log-normal distribution (Karanfil, 2016):

$$\text{True Positive Rate}, TPR = \left(p_{T^+} \mid D^+\right) = 1 - \text{NCDF}\left[(\ln(V) - \mu_{D^+})/(2\sigma_{D^+})^{0.5}\right], \tag{8}$$

$$\text{False Positive Rate}, FPR = (p_{T^+} \mid D^-) = 1 - \text{NCDF}\left[\frac{\ln(V) - \mu_{D^-}}{2\sigma_{D^-}}^{0.5}\right], \tag{9}$$

where NCDF is the cumulative density function for the normal distribution, and $\mu_D$and $\sigma_D$are the location and scale parameters for the lognormal distribution (conditioned on real disease state D,D^+, *or* D^-).

The other two test outcomes, the true negative and false negative rates, are

$$\text{True Negative Rate}, TNR = (p_{T^-} \mid D^-) = 1 - FPR, \tag{10}$$

$$\text{False Negative Rate}, FNR = (p_{T^-} \mid D^+) = 1 - TPR. \tag{11}$$

The recommended starting age for screening (R) determines the fraction of the population considered to be appropriate candidates for routing screening. Just as the appropriate threshold for a positive test is not known but found over time through a search procedure in response to the harm-to-benefit ratio, the R is unknown and modeled using the same hill-climbing structure.

DOI: 10.1002/sdr

$$\frac{dR}{dt} = \frac{(R^* - R)}{\tau_R}, \tag{12}$$

$$R^* = f_R(RHBR)\alpha_R R_{act}, \tag{13}$$

$$R_{act} = \mathcal{L}(R, \lambda_p), \tag{14}$$

$$f_R(RHBR) = 1 + \beta_R(RHBR - HBR); \beta_R > 0. \tag{15}$$

Perception and implementation delays between the recommended and the actual starting age and the recommendations themselves vary between institutions within the United States. In this model, we assume only one set of guidelines, which is perfectly followed by the public with a time delay, hence the public's weight is 1.

There are two factors that motivate the change in breadth of selection and detection criteria, or the breadth of indications. First, as radiologists and practitioners adapt to new technologies that enable earlier detection of cancer, policymakers will tend to expand the criteria to include those patients for whom the inclusion appears to make an effective screening possible. Second, if benefit-and-harm evaluations reveal that the screening test's perceived benefits are lower than desired, this will cause policymakers, and consequently medical practitioners, to gradually become more selective in their screening target population; that is, they will narrow the selection criteria in order to improve future evaluations. The R2 loop "actual practice priming" represents this inclusion drive for the screened population, while the B4 loop represents the change of direction for the selection criteria based on evaluation of screening harm and benefit in a longer term.

Note that the core policy structure for the development of evidence-based screening has two implicit but major time delays embedded in these policy decisions: (a) the modification delay for the effective recommended starting age (τ_R), and (b) the evaluation or translation delay for the benefits and harms of screening (λ_t).Because this evaluation process takes time to complete, evaluations may fail to reflect the impact of the latest changes in screening guidelines. This "moving target" situation in policy thresholds does not indicate a problem in the model but is a natural result of bounded rationality inherent in the guideline development and the policy-formation process. The results can get more problematic and pronounced as τ_R becomes shorter, and λ_t becomes longer (see the section entitled "Results").

The *Actual Starting Age* (R_{act}) is formulated using a simple adaptive expectation structure, which is a realistic way to model the way people update their beliefs and perceptions. Patients are found to be mostly affected by their individual healthcare providers while making the decision to have a screening test and hence generating an update regarding recommendations

DOI: 10.1002/sdr

involves several stages of information processing. These include the response time of individual hospitals, doctors, and radiologists to adopt the guidelines and diffuse it into the system and the average time required by the public to perceive, process, and comply with the recommendations. Hence the *Public Perception Delay* (λ_p) is modeled as a third-order smoothed average of the *R*, where it reflects the total reaction time for the public to receive, process, and respond to changing guidelines.

The *R* in turn determines the average or the mean age of the target screening population and hence the average prevalence of disease in this population. The mean age of the target population, ($MeanAge_{target}$), is well approximated as a linear function of the R_{act}, with slope (δ) and intercept (∂), the mean age of the U.S. male population, using the U.S. Census Data for 2018, and limited by the maximum age *MaxAge* = ε = 100 *years*.

$$MeanAge_{target} = \mathrm{MIN}\left(MaxAge, \delta * R_{act} + \partial\right). \tag{16}$$

The *Target Screening Prevalence*, or the *Age Specific Prevalence* (D^{+}_{age}), represents the fraction or proportion of screen-detectable cancer for this target population. We assumed that the underlying real disease burden stays stable during the simulation time horizon and only increases by age. Age of asymptomatic onset (ϕ) and slope (Ω) of the increase in disease prevalence are estimated from autopsy studies and previously published models (Bell *et al.*, 2015; Haas *et al.*, 2008; Jahn *et al.*, 2015; Sanchez Chapado et al., 2003).

$$D^{+}_{age} = MIN\left(1, MAX\left(0, \left(MeanAge_{target} - \phi\right) * \Omega\right)\right). \tag{17}$$

Probability of disease (p_D) in this target population is then determined as:

$$p_{D+} = D^{+}_{age}, \tag{18}$$

$$p_{D-} = 1 - D^{+}_{age}. \tag{19}$$

For any given population prevalence for a specific target population, our decision-theoretic model calculates the HBR and the *RHBR*. Increasing levels of the *RHBR* increase the *Effect of HBR on Indicated Age for Screening*. As *RHBR* increases above its reference level HBR_{ref}, the effect of HBR becomes higher than one and hence shifts the *Indicated Starting Age* R^{*} above the *Actual Starting Age*. If the *RHBR* reaches HBR_{ref}, the R^{*} becomes equal to the *Actual Starting Age*. *R* is formulated as the output of an information delay structure, where the delay parameter is represented by the *Time to Adjust R* (τ_R,). This parameter gives the delay time constant for the adjustment time of *R*.

DOI: 10.1002/sdr

Table 2 lists model inputs and symbols used throughout the article with their base case values:

Model results

In the base case scenario, we assume that there is no variation in screening advice within the United States, that is, all practitioners and patients are complying with the recommendations derived from the evidence base only. This is the most "ideal-world" setting one can imagine regarding any population screening policy, as exemplified in this quote:

> I think it should be peer quantification of the harms and the benefits, and from authoritative panels. I'm not sure if the U.S. taskforce is that for the U.S. I think probably yes. There's politics involved, but I think nevertheless, I think we here in Europe should really believe in authoritative panels that are independent as possible and weigh the evidence that gets presented by the experts. I think that should be the situation. That should be the ideal situation. (Academic, Policymaker)

The base case simulation serves for assessment of the effects of delays and nonlinearities inherent in screening evaluations and advice. Simulations with base parameters confirm an overshoot in screening indications and expansion of criteria—similar to what we have observed in the 1990s–2000s in the United States—assuming even if there is no variation in the underlying prevalence of the disease, in screening technology, or in harms and benefits environment (Figure 5). While the overshoot persists over a wide range of parameter values, the degree and extent of the overshoot changes with changing values of the model parameters. More specifically, the overshoot of indications gets more amplified when the *Public Perception Delay* λ_p gets shorter, and the *HBR Translation Delay* λ_t takes longer.

We observe a similar overshoot in screening diagnostics, including false positive rates. Harms exceed benefits as the target screening population is expanded (Figure 5), and we see oscillations around the HBR_{ref}, while other variables (R, V, D^{+}_{age}) oscillate around a lower equilibrium. The HBR is below HBR_{ref} at the start of the simulation, meaning screening has an added value compared to doing nothing in the beginning, as it was in 1980s. Hence the threshold value and then the recommended and the actual ages fall sharply within the next decade, which causes an overshoot in indications, and a quick expansion in screening criteria, as suggested by low values of R, V, D^{+}_{age}, and/or higher proportion of false positives.

Note that the most commonly used formal criterion for biopsy referral threshold in the United States is 4 ng/ml, which is somewhat arbitrary

DOI: 10.1002/sdr

Table 2. Model inputs

	Symbol [unit]	Base Case	Description
Threshold value V	V [dmnl]	[4]	Threshold, or cutoff value, for the test outcome, defaulted at most commonly used threshold.
Recommended starting age R	R [ages]	[50]	Recommended starting age R for routine screening
Time to adjust V, R	τ_V, τ_R [year]	[1.5, 1.5]	Adjustment time constant for the rate of change of V and R
Location parameter of lognormal pdfϵR	μ^+, μ^- [dmnl]	[1, 0.3]	Location parameter of the associated normal pdf of test outcome for D^+ and D^-
Scale parameter of lognormal pdf > 0	σ^+, σ^- [dmnl]	[0.6, 0.4]	Scale parameter of the associated normal pdf of test outcome for D^+ and D^-
Baseline onset	ϕ [dmnl]	[25]	Asymptomatic age of onset for disease
Slope D^+	Ω [dmnl]	[0.012]	Rate of change in disease prevalence per age year, based on U.S. cancer trends)
Unit benefit $_{j,k}$ (j = T^+,T^-; k = D^+,D^-)	$UB_{j,k}$ [dmnl]	$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$	Nonnegative unit benefits for possible test outcome and disease state pairs
Unit harm $_{j,k}$ (j = T^+,T^-; k = D^+,D^-)	$UH_{j,k}$ [dmnl]	$\begin{bmatrix} 0 & 1.75 \\ 2 & 0 \end{bmatrix}$	Nonnegative unit harms for possible test outcome and disease state pairs
Bias V, R	α_V, α_R [dmnl]	1	Multiplier for effect of external pressures on threshold value V and R, such as advocacy groups
HBR translation delay	λ_t [year]	[12]	Time constant for translation of HBR, reporting delay
HBR multiplier	β_V, β_R [dmnl]	[0.5; 0.5]	Multiplier for effect of HBR on state variables V and R.
Public perception delay	λ_p [year]	[2]	Average time required for public to respond to changing guidelines
Max age	ε [ages]	[100]	Maximum age of a person
Mean age intercept	∂ [ages]	[34]	Mean age of the overall male population, calculated using U.S. census data for 2018
Mean age slope	δ [dmnl]	[0.6]	Rate of increase of the mean age of the male population, as a function of R.
Reference HBR	HBR_{ref} [dmnl]	[1]	Reference, or the optimal value of the HBR

(Hoffman, 2011). While the recommended "formal" biopsy threshold is shown to vary only modestly over time, the informal "practice" threshold has reportedly varied widely over the years (Thompson *et al.*, 2004), as has

DOI: 10.1002/sdr

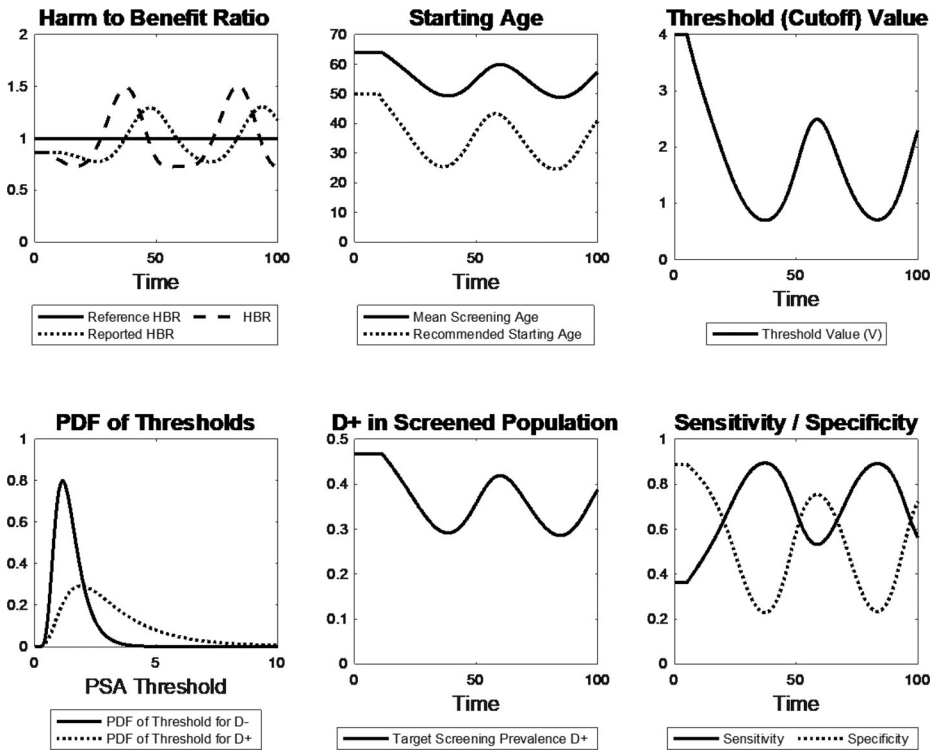

Fig 5. Simulation with base case parameter values.

the recommended starting age, similar to what we observe in model simulations The real historical pattern for the average actual biopsy threshold is unknown, but it is assumed to be 2.5 ng/ml between 1990 and 2000 (Gulati *et al.*, 2010), considerably lower than the formal value (see Figure 5). For screening to be effective, the D^+ part of the target population has to be not very low, since decreases in D^+ (as a result of expansion of breadth indications, expansion of D^-people in screened population) may cause the actual harms to exceed its benefits at population level.

Note that the huge drop of fraction of D^+ in the target population (fraction of diseased people in the screening population) coincides with an increase in test sensitivity (true positives), yet a big drop in test specificity, indicating higher rates of false positives (Remember FP = 1 − specificity!).

Also note the long phase lag between the actual and the reported HBR, which reflects the time needed to complete the evaluation process. When the benefit and harm evaluations finally revealed in the 2000s that the benefits of screening were lower than desired, policymakers gradually became more selective in defining their target population; that is, they updated their previous screening advice and narrowed their selection criteria in order to improve future evaluations. Indeed, the formal guidelines released by the

DOI: 10.1002/sdr

USPSTF in 2008 suggested that evidence was not sufficient to recommend PSA screening for men below 75, while the actual starting age for screening undershot the recommendation.

The ROC curve and "Flipping the Coin": Assuming test has no diagnostic validity

The ROC curve shows the trade-off between specificity and sensitivity but does not suggest the appropriate trade-off, that is, the optimal choice for the threshold. The choice of threshold depends on the subjective costs and benefits we assign to each of the four possible outcomes in different situations. These are conditioned both by the nature of follow-ups, treatments, and consequences of the disease, but also by the risk perceptions and preferences of individuals when faced with making an informed decision.

The 45° line represents the case where there is no ability to distinguish between the populations with and without the condition: the test has no value. An "ideal" test with perfect discrimination (where there is no overlap in the two distributions with respect to the test criterion) would have a ROC curve that passes through the upper-left corner (100 percent sensitivity and specificity). The closer the ROC curve is to the upper-left corner, the higher the overall accuracy of the test will be (Zweig and Campbell, 1993). The area under the ROC curve (AUC) is another metric that indicates how well a screening test can distinguish between the two diagnostic groups. Higher values of AUC indicate a higher discriminatory power. The empirical fitted AUC estimates with respect to discrete occurrence of prostate cancer range from 0.6 to 0.8: Our simulated ROC curve gives an AUC of 0.73 with base case parameters (Figure 6a).

A screening test may only be feasible if its diagnostic accuracy is at least slightly better than just flipping a coin. We simulated this extreme condition to test the model behavior, by perfectly overlapping the PSA distributions of the D^+ and D^- populations (not shown here). The baseline ROC curve (Figure 6a) changes, and the true positive rate falls exactly on the 45° indifference line (Figure 6b), and hence the AUC takes its minimum possible value of 0.5, indicating that no screening test can distinguish these two populations from each other, based on that particular test outcome alone. The V and R both reach unrealistically high values indicating the infeasibility of screening under this condition or indicate to a point where screening has no added value.

Sensitivity Testing

Several types of sensitivity tests are conducted by exploring the parameter space, here we present a sample. Monte Carlo simulation, also known as

DOI: 10.1002/sdr

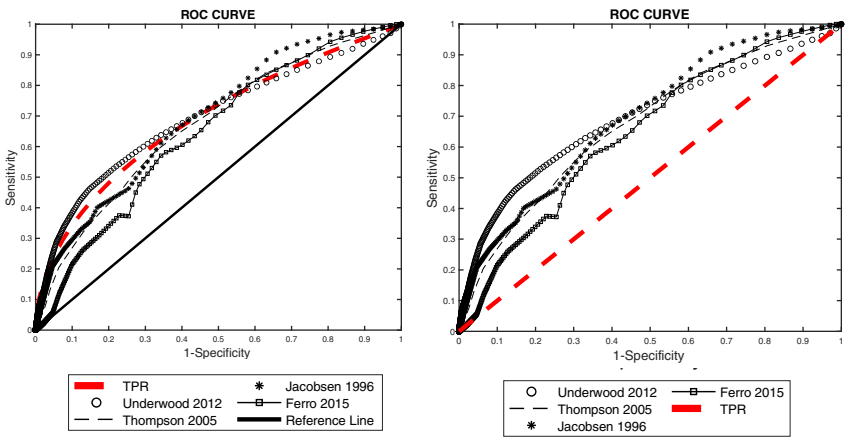

Fig 6. Simulation results for a hypothetical coin-flip test (D^+ and D^-overlap) —(a) ROC curve for base model, (b) ROC curve for flip-test. [Color figure can be viewed at wileyonlinelibrary.com]

multivariate sensitivity simulation (MVSS), is used to automate the sensitivity analysis. In each of the cases below, either one parameter or a subset of parameters is changed in certain ranges including the plausible range to see the differences in the dynamic behavior of the model.

Effect of HBR translation delay (λ_t) on screening recommendations

Figure 7 shows the change in simulation results when the λ_t varies. The HBR reaches its reference value HBR_{ref} for a wide range of λ_t,yet screening becomes inappropriate after a certain point as it gets longer, where harms of screening always exceed benefits.

Effect of HBR multiplier β (β_V and β_R) on screening recommendations

Figure 8 depicts the effect of varying the *HBR multipliers* (β_V *and* β_R) on screening recommendations. These parameters indicate the strength of HBR evaluations' impact on changing the breadth indications of screening. As β's get higher, the overshoot in breadth indications gets amplified. When the multipliers exceed a certain value, HBR evaluations override the "priming" effect of the actual practice, and screening becomes inappropriate.

Sensitivity to translation and public perception delays

The *HBR Translation Delay* (λ_t) is varied between 5 and 20 years (baseline value = 12 years) and *Public Perception Delay* (λ_p) is varied between 1 and 5 years (baseline value = 2 years). Simulation results in Figure 9 reveals that screening is feasible, but oscillations persist in most situations, except in 5 percent of the simulations where screening is inappropriate. Additional sensitivity tests were conducted by adding the two other time constants, the *Time to*

DOI: 10.1002/sdr

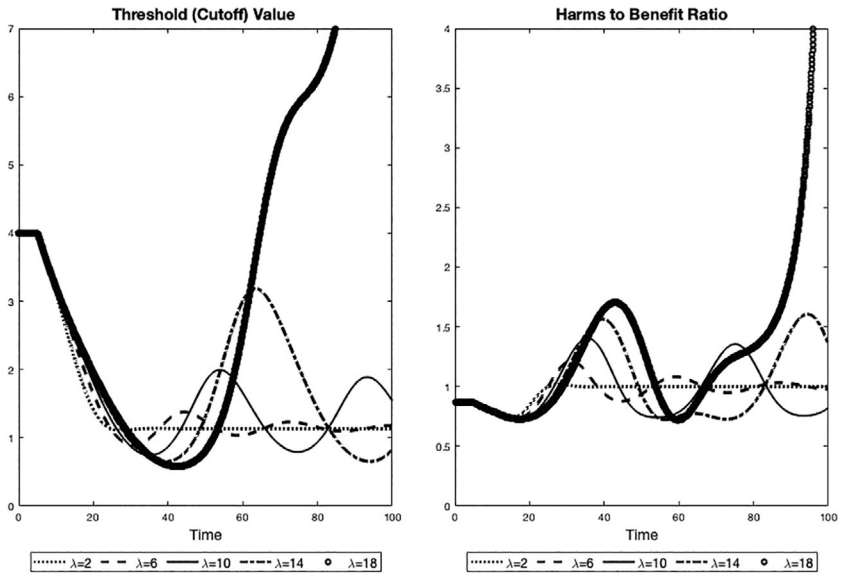

Fig 7. Effect of HBR translation delay (λ_t) on screening recommendations (varied between 2 and 18 years).

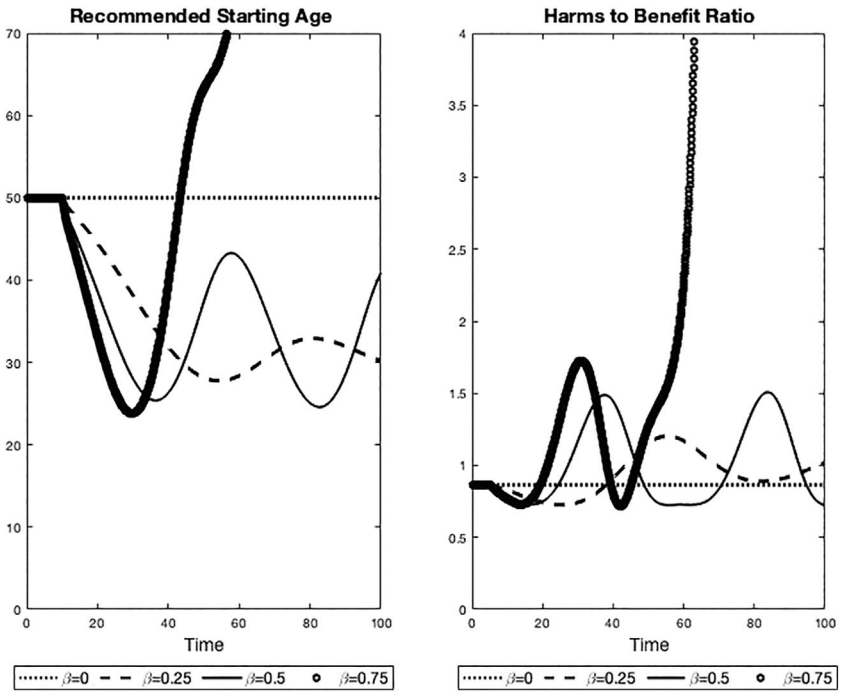

Fig 8. Effect of HBR multiplier (β_V, β_R) on selected model outputs (β_V, β_R changed at the same time between 0 and 0.75).

DOI: 10.1002/sdr

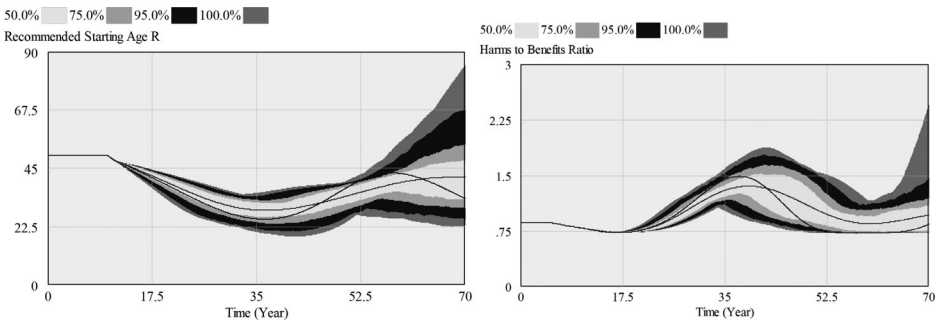

Fig 9. Effect of translation (λ_t) and public perception (λ_p) delays (λ_t; λ_p varied between 5–20 and 1–5 years).

Adjust V (τ_V) and *Time to Adjust R* (τ_R), by varying them between reasonable ranges of 0.5 and 4 years (baseline value = 1.5 years), with similar results.

Sensitivity to underlying disease prevalence $\mathbf{D^+}$ *in target screening population*

We conducted another set of simulations to test the model's behavior as the underlying real disease prevalence naturally changes for the overall population. To simulate the effect of underlying disease prevalence, we varied the Baseline Onset, or the age of asymptomatic onset (ϕ) for disease, from 20 to 35, while keeping the rate of increase in disease prevalence constant.

Figure 10 shows how the D^+ in the screening population affects test efficiency and recommendations for screening. Simulation results indicate the existence of a plausible range where screening is feasible. For screening to be effective the D^+ in the target population has to be not very low or very high since harms of screening may exceed benefits at the population level.

Sensitivity to changing the distribution of test values on test efficiency

We conducted a set of simulations to see how the efficiency of a screening test changes as the underlying distributions of the test values for disease states change relative to each other. For PSA screening, the distribution of test values for the D^+ and D^- populations are overlapping to some extent, while the mean and the standard deviation of the distribution are naturally higher for the D^+ population with respect to the test criterion. In other words, the D^+ population has a higher mean PSA test outcome, and the distribution of their test outcomes is more spread out.

We varied the location (μ_{D^+}) and the scale (σ_{D^+}) parameters of the D^+ population to simulate various combinations of test outcome distributions, which affects a test's intrinsic diagnostic efficacy to differentiate the diseased cases from the healthy ones. The μ_{D^+} is varied between 0.6 and 1.4 ng/ml (baseline value = 1), and σ_{D^+} is varied between 0.2 and 1 (baseline value = 0.6).

DOI: 10.1002/sdr

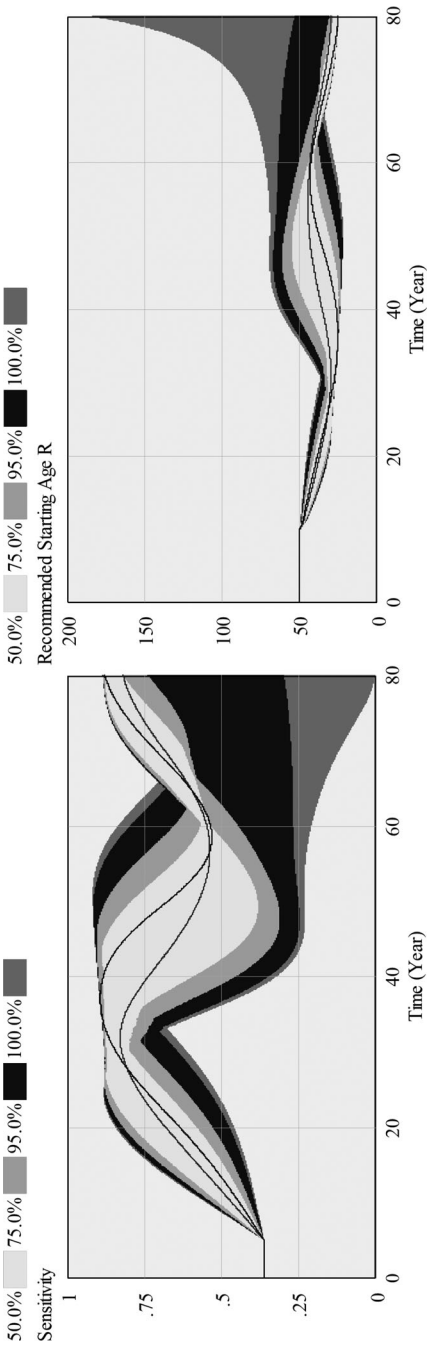

Fig 10. Effect of changing the age for asymptomatic onset (ϕ) on model outputs.

DOI: 10.1002/sdr

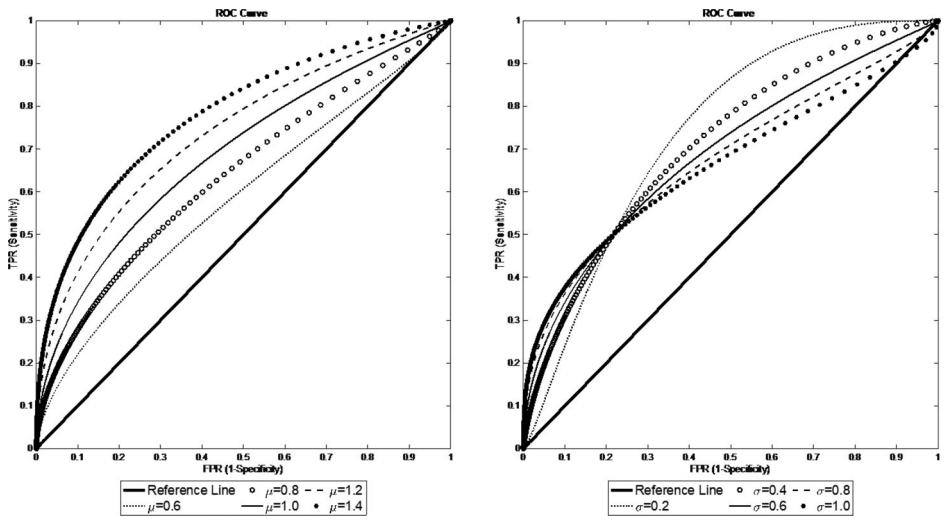

Fig 11. Effect of changing the distribution of the test values on test efficiency.

The first set of simulations show the effect of the increasing or decreasing overlap between the two distributions on the test efficiency (Figure 11). The higher the mean of the test outcome for the D^+ population, the less the overlap becomes between the two distributions, which means the groups can be separated more easily based on the test outcome alone, leading to an increased diagnostic efficiency for all sensitivity-specificity pairs. As the μ_{D^+} goes up, the AUC of the corresponding ROC curve gets closer to 1. As the means start to get closer, the distributions become more inseparable, and the AUC gets closer to 0.5, or the 45° line.

The second set of simulations show the trade-off between test sensitivity and specificity for various degrees of overlap between the two distributions, at various thresholds. As the σ_{D^+} further increases from its baseline value, the test diagnostics get better at higher threshold values, but worse for the lower thresholds. At lower values of σ_{D^+}, the test works better at lower values of the threshold, where there are less false positives for a given level of true positives. However, for lower values of σ_{D^+}, if we operate on the left side of the ROC curve at lower sensitivity levels, this can also make the screening test inappropriate, leading harms to exceed the benefits. For the PSA test, the μ_{D^+} is high, so operating on the right-hand side of the ROC curve at lower thresholds results in higher false positives.

Discussion

This work represents an endogenous theory for the formation and implementation of evidence-based guidelines and a novel attempt to explicitly model

DOI: 10.1002/sdr

the decision behavior around medical screening, including both the core issues for classical evidence-based guideline formation and the environment in which the screening decision is embedded. We think that even a very theoretical-stylized version of a dynamic model like that developed in this study (with two simple components, evidence generation and dissemination) can enhance rational decision-making and be used as a theoretical framework for studying how individuals use information from the task environment to make judgments. It may also improve the debate on policy formation by providing an analytical framework that can be used as a decision tool to aid policymakers and practitioners in multiple domains. Lessons from this debate can be more generally applied to other contentious management and high-stakes policy-decision domains in which there is, at very substantial cost, a huge benefit for a few and a small amount of harm for a larger number of people, like airport and other security or background screening against potential threats, controversial supreme court cases, tax return audits, or federal gun laws (Erev *et al.*, 1995; Kahn and Luce, 2003; Swets, 2001).

One advantage of a stylized yet dynamically complex model is the simplicity of the core dynamic at work. It becomes possible to feel it more closely, as opposed to getting lost due to the numerical complexity associated with most resource allocation and optimization research. In our study, most of the complexity comes from the structure of the system, that is, from the complexity of the intrinsic structure (delays and feedback structure). In real epidemiological studies or real managerial applications, it is easy to lose this bigger outlook amid the numeric complexity of the underlying model. We do not argue that numeric complexity is unimportant in making real-life decisions. Rather, feedback-rich and structurally complex models can provide a simpler and larger dynamic perspective and a better means to aid intuition regarding real problems of concern. The resulting complex decision aid tool can be primarily used by healthcare professionals and policymakers.

While there is a long history of research in evidence-based instruction in the medical domain, evidence-based management training is a recent addition to management education (Goodman *et al.*, 2014). Goodman *et al.* (2014) review relevant literatures from information sciences and medical education to explore how they can inform the design and practice of evidence-based management education.

Simulation models like ours can complement these efforts, by providing constructive insights and a dynamic intuition to supplement the typical empirical evidence considered to update and refine practice guidelines (such as national cancer screening programs). They can be used as formal tools to improve the "guidelines for guideline formation". Perhaps more importantly, they can help policy makers to moderate and manage the public reaction against frequently changing recommendations. Penson (2015) argued that the "swinging pendulum" in population screening needs to be stopped

DOI: 10.1002/sdr

"somewhere in the middle" and points us to the possibility that the pendulum may already be swinging back the other way in prostate screening, after decades of overly aggressive screening and treatment. Managing the public's reaction is becoming an increasingly important problem in the Covid-19 era, and in an era where the value of vaccinations is questioned more than any other time in history.

As the amplification of breadth indications gets larger in either direction, or similarly when the evaluation and reporting of benefits and harms takes longer; practitioners and the public may get more reactive to changing recommendations. Simulation results confirm that effective evaluation of the benefits and harms of screening is crucial, as well as correctly informing the public about the risks and benefits of screening, and not stampeding them to either direction, as exemplified in these quotes:

> It's an interesting piece that you're doing, because its not just a matter of looking at the evidence and saying its appropriate or not. We've taught the public that screening is so important and so vital, and you have to detect cancer at the earliest possible stage, and now were backpedaling, and people don't like that! They don't like when we tell them one thing, and then ten years later, tell them, "you know what? Maybe we've oversold screening in this country. Maybe we've oversold it a little bit. Maybe you don't need it." So, I think that's angered a lot of consumers. (Media and Science Reporter)
>
> So, there's been sort of a national pendulum in the general public viewpoint. In the beginning, people were all gung-ho for it. Everybody thought it would be great, so it was widely used and widely promoted, but then after a period of time, you begin to see that there are flaws and problems, and those begin to create a backlash, or criticism, and I think you need to report both of those. (Media and Science Reporter)

Our decision-theoretic model generates a dynamic pattern of the screening criteria that roughly matches the historical data for medical screening in the United States. The screening criteria for many diseases have clearly expanded in the past 30 years and then narrowed down, while showing little sign of rebalancing as the evidence base is ignored and overshadowed by patients, practitioners, and advocacy groups, going beyond the oscillations described in our theoretical-stylized model.

Simulation results reveal that expansion of selection criteria and perception/reaction/ and evaluation delays play an important role in screening evaluations and in the overshoot behavior of screening indications. The nonlinear feedback processes, bounded rationality, and delays inherent in evidence-based screening aggravates the suboptimality of screening guidelines and the policy-formation process. Although this study illustrates the "overshoot of indications" behavior for routine population screening, other managerial applications exist with similar potential behavior in repeated contexts.

DOI: 10.1002/sdr

Conclusion

Despite the universal nature of the scientific evidence base, major health organizations in the United States adopt different and conflicting guidelines, with significant variation in actual practice. There are additional gaps in the scientific evidence which cannot be directly addressed by empirical evidence and clinical trials on which evidence-based screening recommendations are based. Policymakers and clinicians face a tough choice and trade-offs in managing their recommendations and especially in dealing with the public reaction and resistance to frequently changing recommendations.

Policymakers and scientists increasingly employ various modeling studies to fill in these gaps, to guide the guideline-making process, and to explore the trade-offs in quality, capacity, and their cost effectiveness (Güneş and Örmeci, 2018; Güneş *et al.*, 2004; Goldie *et al.*, 2006; Pandya *et al.*, 2015) which is particularly important at a time when their trustworthiness or quality have been questioned (Ransohoff *et al.*, 2011; Ransohoff *et al.*, 2013; Ransohoff and Sox, 2013). The USPSTF started the effort to standardize the guideline-development process in the early 2000s (Harris *et al.*, 2001) and has been since using model-based insights in developing its breast (Mandelblatt *et al.*, 2009), colorectal (Zauber *et al.*, 2008), and prostate screening recommendations (Draisma *et al.*, 2009). Most recently, the World Health Organization (WHO) released a special edition on enhancing WHOs standard guideline-development methods, entitled "Complex Health Interventions in Complex Systems: Concepts and Methods for Evidence-Informed Health Decisions." The issue suggests considering issues of complexity when developing evidence-based guidelines may make WHO guidelines more relevant and have greater impact in countries (Booth *et al.*, 2019; Noyes *et al.*, 2019; Petticrew *et al.*, 2019).

While there is a proliferation of modeling studies to inform the CPG's, not many are addressing the actual guideline-making process itself. Existing studies largely ignore the broad boundary feedbacks and inherent delays in decision-making that condition the adoption of and adherence to CPG's. In most guideline frameworks, evidence reviews lead to evidence interpretation and then policy, with no feedback incorporated/considered at any of these steps. Linear thinking is prevalent in policy formation and interpretation, and broad boundary feedbacks are largely ignored in existing frameworks and mindsets.

Another widely ignored aspect in existing studies is their ubiquitous assumption for constancy in most of their variables; most system variables are treated as exogenous parameters. Earlier models assume a biopsy referral threshold which stays constant during the simulation time horizon, not an accurate reflection of the clinical practice as we discussed in this study. Hence the same thing is true for the efficiency of the test and the test

DOI: 10.1002/sdr

diagnostics, including sensitivity and specificity, and the disease prevalence in the target screening population, which are intrinsically connected and endogenously evolving over time.

Since we derive the test diagnostics directly from the underlying probability distributions of diseased and healthy people, we can show how test diagnostics and the probability of disease change over time with changing indications of screening and changing target population. Future extensions of the model can as well be used to estimate the real underlying prevalence of a disease, and addition of an "indolent" disease category may facilitate to make inferences about the real occult disease prevalence in the population. In this study, we show how the most important "system variables" interact to create broad boundary feedbacks and that they are far more important than merely focusing on parameter values. These feedbacks and delays need to be included in future studies and the guideline development and policy-formation process.

The theoretical evidence-based policy structure we present corresponds to an idealized world: we assume that there is only one set of guidelines, and they are followed by the public, while the only consideration is the evidence-based harms-benefits calculations. In this perfect world, we may indeed gradually approach the reference target level for screening; yet note that even in this idealized situation, we see an overshoot in breadth indications in most cases.

Another counterintuitive insight from the model is that trying to detect disease at earlier stages results in the expansion of the eligibility criteria for disease, decrease in the real disease prevalence in the target population as a result of this expansion, which eventually leads to a worsening in test diagnostics, in terms of reduced specificity and more false positives. In other words, the more we screen for a condition to eradicate a chronic disease with varying indications, such as autism spectrum disorder, attention deficit disorder, Alzheimer's disease, or hypertension, the more people need to be included in the target population to reach that goal. This firefighting effort in population screening may lead to a vicious cycle, labeling more people with disease, making the screening effort both ineffective and impossible to afford, which is nonintuitive and contrary to heuristic thinking (Hammond, 2007; Kahneman, 2011; Marshall, 2014; Ransohoff *et al.*, 2002; Sterman, 1989, 2006).

Acknowledgements

The project has been funded by the 2232 International Fellowship for Outstanding Researchers Program of TUBITAK (Project No: 118C327) supporting Dr. Özge Karanfil. However, all scientific contributions made in this project are owned and approved solely by the authors.

DOI: 10.1002/sdr

List of Abbreviations

ACS	American Cancer Society
ACC	American College of Cardiology
AHA	American Heart Association
AUA	American Urological Association
AUC	Area Under Curve
CPG	Clinical Practice Guidelines
COUHES	Committee on the Use of Humans as Experimental Subjects
DALY	Disability Adjusted Life Years
FNR	False Negative Rate
FPR	False Positive Rate
HBR	Harm-to-Benefit Ratio
MVSS	Multivariate Sensitivity Simulation
NCCN	National Comprehensive Cancer Network
NAP	National Academies Press
NAS	National Academy of Sciences
NIA	National Institute of Aging
PLCO	Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial
PSA	Prostate Specific Antigen
QALY	Quality Adjusted Life Years
ROC	Receiver Operating Characteristic
SDM	Shared Decision-Making
TNR	True Negative Rate
TPR	True Positive Rate
USPSTF	U.S. Preventive Services Task Force
WHO	World Health Organization

References

Ablin RJ, Soanes WA, Bronson P, Witebsky E. 1970. Precipitating antigens of the normal human prostate. *Reproduction* **22**(3): 573–574.

Abramson JD, Redberg RF. (2013). Opinion | Don't give more patients statins. *The New York Times* [Internet]. Retrieved from: https://www.nytimes.com/2013/11/14/opinion/dont-give-more-patients-statins.html

Bakris G, Sorrentino M. 2018. Redefining hypertension — assessing the new blood-pressure guidelines. *New England Journal of Medicine* **378**(6): 497–499.

Bell KJL, Del Mar C, Wright G, Dickinson J, Glasziou P. 2015. Prevalence of incidental prostate cancer: a systematic review of autopsy studies. *International Journal of Cancer* **137**(7): 1749–1757.

Belluck P. (2013, December 25). Common knee surgery does very little for some, study suggests. *The New York Times* [Internet]. Retrieved from https://www.nytimes.com/2013/12/26/health/common-knee-surgery-does-very-little-for-some-study-suggests.html.

DOI: 10.1002/sdr

Bhatt, D. (2018) The new cholesterol guidelines: what you need to know [Internet]. *Harvard Health Blog*. Retrieved from https://www.health.harvard.edu/blog/the-new-cholesterol-guidelines-what-you-need-to-know-2018112615422

Blackburn M. (2015). H.R.1151 - 114th Congress (2015-2016): USPSTF Transparency and Accountability Act of 2015 [Internet]. Retrieved from https://www.congress.gov/bill/114th-congress/house-bill/1151

Booth A, Moore G, Flemming K, Garside R, Rollins N, Tunçalp Ö *et al.* 2019. Taking account of context in systematic reviews and guidelines considering a complexity perspective. *BMJ Global Health* **4**(Suppl 1): e000840.

Brody JE. 2007. Deciphering the Results of a Prostate Test. The New York Times [Internet]. Available from: https://www.nytimes.com/2007/05/08/health/08brod.html

Carroll AE. 2017. Why New Blood Pressure Guidelines Could Lead to Harm. The New York Times [Internet]. [cited 2020 Sep 1]; Available from: https://www.nytimes.com/2017/12/18/upshot/why-new-blood-pressure-guidelines-could-lead-to-harm.html

Kolata G. 2009. Panel Urges Mammograms at 50, Not 40. The New York Times [Internet]. [cited 2020 Sep 1]; Available from: https://www.nytimes.com/2009/11/17/health/17cancer.html

Croswell JM, Ransohoff DF, Kramer BS. 2010. Principles of cancer screening: lessons from history and study design issues. *Seminars in Oncology* **37**(3): 202–215.

Cutler DM, Lleras-Muney A. 2010. Understanding differences in health behaviors by education. *Journal of Health Economics* **29**(1): 1–28.

Darabi N, Hosseinichimeh N. 2020. System dynamics modeling in health and medicine: a systematic literature review. *System Dynamics Review*, **36**: 29–73. https://onlinelibrary.wiley.com/doi/full/10.1002/sdr.1646.

Draisma G, Etzioni R, Tsodikov A, Mariotto A, Wever E, Gulati R *et al.* 2009. Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. *Journal of the National Cancer Institute* **101**(6): 374–383.

Erev I, Gopher D, Itkin R, Greenshpan Y. 1995. Toward a generalization of signal detection theory to N-person games: the example of two-person safety problem. *Journal of Mathematical Psychology* **39**(4): 360–375.

Esserman LJ, Thompson IM, Reid B, Nelson P, Ransohoff DF, Welch HG *et al.* 2014. Addressing overdiagnosis and overtreatment in cancer: a prescription for change. *The Lancet Oncology* **15**(6): e234–e242.

Fett MJ. 2001. Computer modelling of the Swedish two county trial of mammographic screening and trade-offs between participation and screening interval. *Journal of Medical Screening* **8**(1): 39–45.

Fineberg HV. 1985. Effects of clinical evaluation on the diffusion of medical technology. In *Assessing Medical Technologies*, Institute of Medicine (US) Committee for Evaluating Medical Technologies in Clinical Use (ed). Washington, DC: National Academies Press (US), 8–9. https://www.ncbi.nlm.nih.gov/books/NBK217470/

Force UPST, Grossman DC, Curry SJ, Owens DK, Bibbins-Domingo K, Caughey AB *et al.* 2018. Screening for Prostate Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA*. **319**(18): 1901–1913. https://doi.org/10.1001/jama.2018.3710.

Forrester JW. 1961. *Industrial Dynamics*. Cambridge, MA: Massachusetts Institute of Technology Press.

DOI: 10.1002/sdr

Goldie SJ, Kim JJ, Myers E. 2006. Chapter 19: Cost-effectiveness of cervical cancer screening. *Vaccine* **24**(Suppl 3): S3/164–S3/170.

Goodman JS, Gary MS, Wood RE. 2014. Bibliographic search training for evidence-based management education: a review of relevant literatures. *Academy of Management Learning and Education* **13**(3): 322–353.

Green DM, Swets JA. 1966. *Signal Detection Theory and Psychophysics*. John Wiley: Oxford. xi, 455.

Griner PF, Mayewski RJ, Mushlin AI, Greenland P. 1981. Selection and interpretation of diagnostic tests and procedures. Principles and applications. *Annals of Internal Medicine* **94**(4 Pt 2): 557–592.

Gulati R, Inoue L, Katcher J, Hazelton W, Etzioni R. 2010. Calibrating disease progression models using population data: a critical precursor to policy development in cancer control. *Biostatistics* **11**(4): 707–719.

Güneş ED, Örmeci EL. 2018. OR applications in disease screening. In *Operations Research Applications in Health Care Management*, Kahraman C, Topcu YI (eds). Switzerland: Springer International Publishing; 297–325. https://doi.org/10.1007/978-3-319-65455-3_1.

Güneş ED, Chick SE, Akşin Z. 2004. Breast cancer screening services: trade-offs in quality, capacity, outreach, and centralization. *Health Care Management Science* **7**(4): 291–303.

Haas GP, Delongchamps N, Brawley OW, Wang CY, de la Roza G. 2008. The worldwide epidemiology of prostate cancer: perspectives from autopsy studies. *The Canadian Journal of Urology* **15**(1): 3866–3871.

Hammond KR. 1996. *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*. New York: Oxford University Press. https://g.co/kgs/J7n34w.

Hammond KR. 2007. *Beyond Rationality: The Search for Wisdom in a Troubled Time*. Oxford University Press: New York, NY. xxiv, 338. https://g.co/kgs/Hig1D5.

Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM *et al.* 2001. Current methods of the US Preventive Services Task Force: a review of the process. *American Journal of Preventive Medicine* **20**(3 Suppl): 21–35.

Hellmuth J, Rabinovici GD, Miller BL. 2019. The rise of pseudomedicine for dementia and brain health. *Journal of the American Medical Association* **321**(6): 543–544.

Hoffman RM. 2011. Screening for prostate cancer. *New England Journal of Medicine* **365**(21): 2013–2019.

Hoffman JR, Cooper RJ. 2012. Overdiagnosis of disease: a modern epidemic. *Archives of Internal Medicine* **172**(15): 1123–1124.

Homer JB. 1987. A diffusion model with application to evolving medical technologies. *Technological Forecasting and Social Change* **31**(3): 197–218.

Jahn JL, Giovannucci EL, Stampfer MJ. 2015. The high prevalence of undiagnosed prostate cancer at autopsy: implications for epidemiology and treatment of prostate cancer in the Prostate-specific Antigen-era. *International Journal of Cancer* **137**(12): 2795–2802.

Kahn BE, Luce MF. 2003. Understanding high-stakes consumer decisions: mammography adherence following false-alarm test results. *Marketing Science* **22**(3): 393–410.

Kahneman D. 2011. *Thinking, Fast and Slow*. Penguin: London.

DOI: 10.1002/sdr

Karanfil Ö. (2016) Why clinical practice guidelines shift over time: A dynamic model with application to prostate cancer screening. PhD Thesis. Massachusetts Institute of Technology. Retrieved from http://dspace.mit.edu/handle/1721.1/107531

Karanfil, Ö., Rahmandad H, Homer J, Sterman J. 2017. A Dynamic Model for Understanding Long-Term Trends in Prostate Cancer Screening. *Proceedings of the 35th System Dynamics Society*. Cambridge.

Kivuti-Bitok LW, McDonnell G, Abdul R, Pokhariyal GP. 2014. System dynamics model of cervical cancer vaccination and screening interventions in Kenya. *Cost Effectiveness and Resource Allocation.* **12**(1): 26.

Kolata G. 2013a. Hypertension Guide May Affect 7.4 Million. The New York Times [Internet]. [cited 2020 Sep 1]; Available from: https://www.nytimes.com/2013/12/20/health/hypertension-guide-may-affect-7-4-million.html.

Kolata G. 2013b. Hypertension Guidelines Can Be Eased, Panel Says. The New York Times [Internet]. [cited 2020 Sep 1]; Available from: https://www.nytimes.com/2013/12/19/health/blood-pressure-guidelines-can-be-loosened-panel-says.html

Lyon AR, Maras MA, Pate CM, Igusa T, Vander Stoep A. 2016. Modeling the impact of school-based universal depression screening on additional service capacity needs: a system dynamics approach. *Administration and Policy in Mental Health* **43**(2): 168–188.

Mandelblatt JS, Cronin KA, Bailey S, Berry DA, de Koning HJ, Draisma G *et al.* 2009. Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. *Annals of Internal Medicine* **151**(10): 738–747.

Mandl KD, Manrai AK. 2019. Potential excessive testing at scale: biomarkers, genomics, and machine learning. *Journal of the American Medical Association* **321**(8): 739–740.

Marshall E. 2014. Dare to do less. *Science* **343**(6178): 1454–1456.

Metz CE. 1978. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8**(4): 283–298.

Moyer VA. 2012. Screening for prostate cancer: US Preventive Services Task Force recommendation statement. *Annals of Internal Medicine* **157**(2): 12O–123O.

National Academies of Sciences E, Division H and M, Policy B on HS, Impairment C on PD and C, Downey A, Stroud C, *et al.* 2017. *Preventing Cognitive Decline and Dementia* [Internet]. National Academies Press (US); [cited 2020 Sep 1]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK436397/.

NCI Cancer Statistics. 2018 Retrieved from https://www.cancer.gov/about-cancer/understanding/statistics

Noyes J, Booth A, Moore G, Flemming K, Tunçalp Ö, Shakibazadeh E. 2019. Synthesising quantitative and qualitative evidence to inform guidelines on complex interventions: clarifying the purposes, designs and outlining some methods. *BMJ Global Health* **4**(Suppl 1): e000893.

ODPHP. 2017. Prostate Cancer Screening Recommendations: An Update from USPSTF [Internet]. News & Events | Health.gov. [cited 2020 Sep 1]. Available from: https://health.gov/news/blog/2017/07/prostate-cancer-screening-recommendations-an-update-from-uspstf/

Palma A, Lounsbury DW, Schlecht NF, Agalliu I. 2016. A system dynamics model of serum prostate-specific antigen screening for prostate cancer. *American Journal of Epidemiology* **183**(3): 227–236.

DOI: 10.1002/sdr

Pandya A, Sy S, Cho S, Weinstein MC, Gaziano TA. 2015. Cost-effectiveness of 10-year risk thresholds for initiation of statin therapy for primary prevention of cardiovascular disease. *Journal of the American Medical Association* **314**(2): 142–150.

Parker-Pope T. 2014. Prostate Cancer Screening Still Not Recommended for All [Internet]. Well. Available from: https://well.blogs.nytimes.com/2014/08/06/prostate-cancer-screening-still-not-recommended-for-all/

Pauker SG, Kassirer JP. 1980. The Threshold Approach to Clinical Decision Making. *New England Journal of Medicine.* **302**(20): 1109–1117.

Penson DF. 2015. The pendulum of prostate cancer screening. *Journal of the American Medical Association* **314**(19): 2031–2033.

Petticrew M, Knai C, Thomas J, Rehfuess EA, Noyes J, Gerhardus A *et al.* 2019. Implications of a complexity perspective for systematic reviews and guideline development in health decision-making. *BMJ Global Health* **4**(Suppl 1): e000899.

Pollack A. Looser Guidelines Issued on Prostate Screening. The New York Times [Internet]. 2013 May 3; Available from: https://www.nytimes.com/2013/05/04/business/prostate-screening-guidelines-are-loosened.html

Rabin RC. (2009, November 16). New guidelines on breast cancer draw opposition. *The New York Times* [Internet]. Retrieved from https://www.nytimes.com/2009/11/17/health/17scre.html

Rabin RC. (2014, June 30). Guideline calls routine pelvic exams unnecessary. *The New York Times* [Internet]. Retrieved from https://www.nytimes.com/2014/07/01/health/doctors-group-advises-against-regular-pelvic-exams.html

Rabin RC. 2017. New Study Offers Support for Prostate Testing. The New York Times [Internet]. Available from: https://www.nytimes.com/2017/09/04/well/live/new-study-offers-support-for-prostate-testing.html

Rapaport, 2020. Many young women get unnecessary pelvic exams. Reuters [Internet]. Jan 6; Available from: https://www.reuters.com/article/us-health-gynecology-girls-idUSKBN1Z51WA.

Ransohoff DF, Sox HC. 2013. Guidelines for guidelines: measuring trustworthiness. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **31**(20): 2530–2531. https://doi.org/10.1200/JCO.2013.50.0462.

Ransohoff DF, McNaughton Collins M, Fowler FJ. 2002. Why is prostate cancer screening so common when the evidence is so uncertain? A system without negative feedback. *The American Journal of Medicine* **113**(8): 663–667.

Ransohoff DF, Pignone M, Russell LB. 2011. Using models to make policy: an inflection point? *Medical Decision-making* **31**(4): 527–529. https://doi.org/10.1177/0272989X11412079.

Ransohoff DF, Pignone M, Sox HC. 2013. How to decide whether a clinical practice guideline is trustworthy. *JAMA* **309**(2): 139–140. https://doi.org/10.1001/jama.2012.156703.

Riordan, 2013. New Cholesterol Guidelines Abandon LDL Targets [Internet]. [cited 2020 Sep 1]. Available from: https://www.medscape.com/viewarticle/814152

Sánchez-Chapado M, Olmedilla G, Cabeza M, Donat E, Ruiz A. 2003. Prevalence of prostate cancer and prostatic intraepithelial neoplasia in Caucasian Mediterranean males: An autopsy study. *The Prostate.* **54**(3): 238–247.

Schlesinger AM. 1986. *The Cycles of American History*. Boston/New York: Houghton Mifflin Harcourt Publishing Company. Mariner Books.

DOI: 10.1002/sdr

Schröder FH. 2011. Stratifying risk — The U.S. Preventive Services Task Force and Prostate-Cancer Screening. *New England Journal of Medicine* **365**(21): 1953–1955.
Schwartz LM, Woloshin S. 2019. Medical marketing in the United States, 1997-2016. *Journal of the American Medical Association* **321**(1): 80–96.
Sheldrick RC, Breuer DJ, Hassan R, Chan K, Polk DE, Benneyan J. 2016. A system dynamics model of clinical decision thresholds for the detection of developmental-behavioral disorders. *Implementation Science* **11**: 156–170. https://doi.org/10.1186/s13012-016-0517-0.
Sirovich BE, Schwartz LM, Woloshin S. 2003. Screening men for prostate and colorectal cancer in the United States: does practice reflect the evidence? *JAMA* **289**(11): 1414–1420.
Sterman JD. 1989. Modeling managerial behavior: misperceptions of feedback in a dynamic decision-making experiment. *Management Science* **35**(3): 321–339.
Sterman JD. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Irwin/McGraw-Hill: Boston, MA. https://g.co/kgs/N1TE85.
Sterman JD. 2006. Learning from evidence in a complex world. *American Journal of Public Health* **96**(3): 505–514.
Stewart TR, Mumpower JL. 2004. Detection and selection decisions in the practice of screening mammography. *Journal of Policy Analysis and Management* **23**(4): 908–920.
Stewart TR, Mumpower JL, James Holzworth R. 2012. Learning to make selection and detection decisions: the roles of base rate and feedback. *Journal of Behavioral Decision-making* **25**(5): 522–533.
Stiles, 2018. New AHA/ACC Cholesterol Guideline Expands Role of LDL Targets [Internet]. Medscape. [cited 2020 Sep 1]. Available from: http://www.medscape.com/viewarticle/904736
Swets JA. 1964. *Signal Detection and Recognition by Human Observers: Contemporary Readings*. New York: Wiley.
Swets JA. 1992. The science of choosing the right decision threshold in high-stakes diagnostics. *The American Psychologist* **47**(4): 522–532.
Swets JA. 2001. Signal detection theory. In *International Encyclopedia of the Social & Behavioral Sciences*, Baltes NJSB (ed). Pergamon: Oxford; 14078–14082. Retrieved from http://www.sciencedirect.com/science/article/pii/B0080430767006781.
Swets JA, Dawes RM, Monahan J. 2000. Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest* **1**(1): 1–26.
Thompson IM, Pauler DK, Goodman PJ, Tangen CM, Lucia MS, Parnes HL *et al*. 2004. Prevalence of prostate cancer among men with a prostate-specific antigen level ≤4.0 ng per milliliter. *New England Journal of Medicine* **350**(22): 2239–2246.
US Preventive Services Task Force. 2018. Final Recommendation Statement: Prostate Cancer: Screening - US Preventive Services Task Force. Available from: https://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/prostate-cancer-screening.
Weaver EA, Richardson G. 2006. Threshold setting and the cycling of a decision threshold. *System Dynamics Review* **22**(1): 1–26.

DOI: 10.1002/sdr

Wendling, 2017. Do Latest US Guidelines Bypass, or Spare, Millions From Statins? [Internet]. Medscape. [cited 2020 Sep 1]. Available from: http://www.medscape.com/viewarticle/878787

Welch HG. 2017. Cancer screening, overdiagnosis, and regulatory capture. *JAMA Internal Medicine* **177**(7): 915–916.

Wilt TJ, Scardino PT, Carlsson SV, Basch E. 2014. Prostate-specific antigen screening in prostate cancer: perspectives on the evidence. *Journal of the National Cancer Institute* **106**(3): 1–6. Retrieved from https://doi.org/10.1093/jnci/dju010.

Yaffe K. 2019. Prevention of Cognitive Impairment With Intensive Systolic Blood Pressure Control. *JAMA.* **321**(6): 548–549.

Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, Wilschut J, van Ballegooijen M, Kuntz KM. 2008. 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.

Zweig MH, Campbell G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**(4): 561–577.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's website.

Appendix S1. Supporting Information.

DOI: 10.1002/sdr