What is the Stakeholder Effect on Clinical Guideline Formation Process: An Experimental Study

Özge Karanfil [a]*, Şanser Güz [b], Orkun İrsoy [b], Mahdi Hashemian [a]

[a] College of Administrative Sciences and Engineering, Koç University, Istanbul, Turkey
[b] Department of Industrial Engineering, Boğaziçi University, Istanbul, Turkey
* Corresponding author: okaranfil@ku.edu.tr

Background

The use of practice guidelines for screening spans medical, security, and managerial contexts. Despite their importance especially in high-risk and high-stake conditions, guidelines are far from optimal. More intriguingly, they do not seem to have a consistently improving trajectory, and not followed by clinicians and the public, which made their trustworthiness and quality questionable (Ransohoff and Sox, 2013; Ransohoff, Pignone, and Sox, 2013).

Despite the consensus on setting evidence-based guidelines on the best available scientific evidence, scholars have reported significant fluctuations over time for many routine screenings (Penson, 2015; Bakris and Sorrentino, 2018). For instance, clinical practice guidelines (CPGs) for prostate cancer screenings have been varying significantly over time in the United States. Moreover, it is established that patients and most physicians misinterpret available scientific evidence related to routine screening and need supporting tools to be able to interpret them (Gigerenzer et al., 2007; Wegwarth and Gigerenzer, 2012).

Scientific community only recently recognized the complexity inherent to the guideline formation process itself and the existence of broad boundary feedbacks in which the screening decision is embedded. Researchers are increasingly invited and encouraged to explore the potential implications and help policymakers to deal with this complexity (Booth et al. 2019, Petticrew et al., 2019).

Using cancer screening as a motivating example, Karanfil and Sterman (2020) propose an endogenous theory of oscillations in policy decision thresholds of practice guidelines. Their theoretical evidence-based model structure corresponds to a stylized, or "ideal" world where the underlying scientific evidence base is constant. In the model, there is only one set of guidelines (or single evidence-based guideline-issuing organization) perfectly followed by the public. Yet even this theoretical "small" model endogenously generates fluctuations and overshoot in breadth indications of screening, persisting in most cases. In this model, policy decision thresholds cycle over time mainly due to delays in and between policy formation and implementation by boundedly rational decision-makers with cognitive- and information limitations about the underlying system (Schlesinger, 1987).

Model Replication and Extensive Sensitivity Analyses

We build on Karanfil and Sterman (2020) and use prostate cancer screening as our main example. We analyze and experimentally test the effect other stakeholders might have on guideline fluctuations, such as specialists and patient advocacy groups. We aim to provide empirical evidence on Karanfil's (2016) hypothesis that the involvement of multiple stakeholders exacerbates guideline threshold recommendations fluctuations. Advocacy groups are one of the major stakeholders in many contexts and mainly in the U.S, especially in the health arena. Therefore, we particularly focus on the cancer advocacy groups as the second of major category of stakeholder in this system. Advocacy groups often include patients, patient advocacy groups, and specialists. Focusing on advocacy groups can also be illuminating since they differ from evidence-based groups in terms of their risk perceptions on screening harms and benefits.

We replicate Karanfil and Sterman's (2020) original model, their extended case study, and formal model on PSA screening for the U.S (Karanfil, 2016; Karanfil, Homer, Rahmandad and Sterman, 2017)). The extended case study includes two main additional features (1) modeling the natural progression of the disease, respective changes in population size and composition, treatment, dissemination of screening practice, among physicians and changes in technology (2) incorporating multiple stakeholders, roughly corresponding to evidence-based and specialty-advocacy groups. We further the above research by focusing on the latter, evaluating the stakeholder effect in oscillatory dynamics of two guideline state variables; Recommended Starting Age for Screening (Starting-Age hereafter) and Recommended PSA threshold for biopsy referral (Threshold hereafter).

It is widely acknowledged that in general the benefits of cancer screening are overstated, and harms downplayed (Gigerenzer, 2016). In our study, we develop multiple testable hypotheses and conduct additional, extensive sensitivity analyses to test the effect of varying these benefit and harm parameters, in the presence of multiple stakeholders. Subsequently, we propose an experimental platform to test the proposed effects of including multiple stakeholders. Finally, we analyze the effect of including specialty-advocacy groups, juxtaposing the sensitivity and experimental results.

Quantification of the Stakeholder Effect

We choose the amplitude ratios (compared to single-stakeholder scenario) as our main outcome of interest to quantitatively study the extent of the stakeholder effect on fluctuations in screening recommendations**.** Focusing on amplification or changes in amplitudes has been customary in the literature for similar purposes. For example, in his canonical paper, Sterman (1989) uses amplification ratios to gauge the supply chain impact in the Beer Game. Here, we use amplitude ratio instead of amplification ratio because instead of a clear change in input, we have a change in the structure of the model. In line with Karanfil (2016), we find that differences in stakeholders' harm-benefit evaluations, as well as public and clinicians' opinions about stakeholders, are highly influential on model dynamics. Therefore, we select the following sets of parameters for sensitivity analyses; (1) parameters corresponding to the extent to which advocacy groups are different in interpreting and weighing potential harms and benefits of screening (2) parameters corresponding to how much clinicians and the public/patients adhere to evidence-based recommendations.

Results

Preliminary results suggest that, in most cases, we observe an increase in amplitude ratios as the gap between advocacy and evidence-based groups' perceived harm-and benefit ratios increases. We find a more significant impact when this gap is rooted in advocacy groups' view on False Negatives, a moderate impact in the case of True Positives, and a faint effect in the case of False Positives. Moreover, we find that to reduce amplification, interestingly, it is more important that clinicians and the public similarly listen to both stakeholders rather than asking clinicians to focus just on evidence-based recommendations. When a stakeholder's influence on Starting-Age or Threshold is more dominant than its effect on the other, these two state variables will be controlled by perceptions of two heterogeneous stakeholders. This mixed feedback signal leads to the emergence of a transient vicious cycle, leading one state variable to have greater oscillations while the other converges to all-in screening recommendation (i.e., an extremely low Threshold value). In short, model dynamics are heavily conditioned by stakeholders' perceptions of harms and benefits. Therefore, the inclusion of advocacy groups, especially if their risk perceptions deviate significantly from evidence-based groups, exacerbates the oscillations resulting in greater fluctuations and overshoots in breadth indications of screening.

Proposed Experimental Design

In our proposed experimental design, we first test the effect and behavior of the single stakeholder model. Subsequently, we assess the impact of including a second stakeholder (i.e., the advocacy groups). We develop an online experiment in which participants make decisions over time based on model generated data, essentially interacting with the observed behavior through the underlying model.

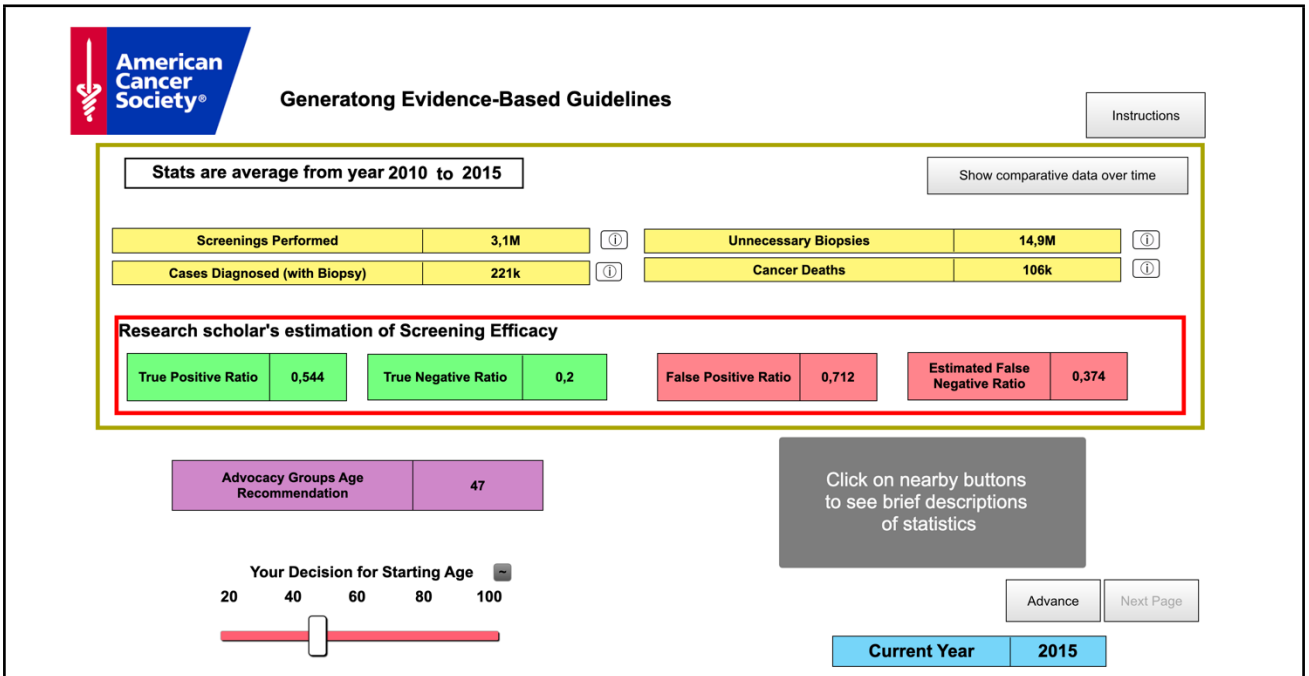

Figure A.1 Sample Policymaker Interface: The name of a real stakeholder is used for clarification purposes. The actual interface will only include a hypothetical organization.

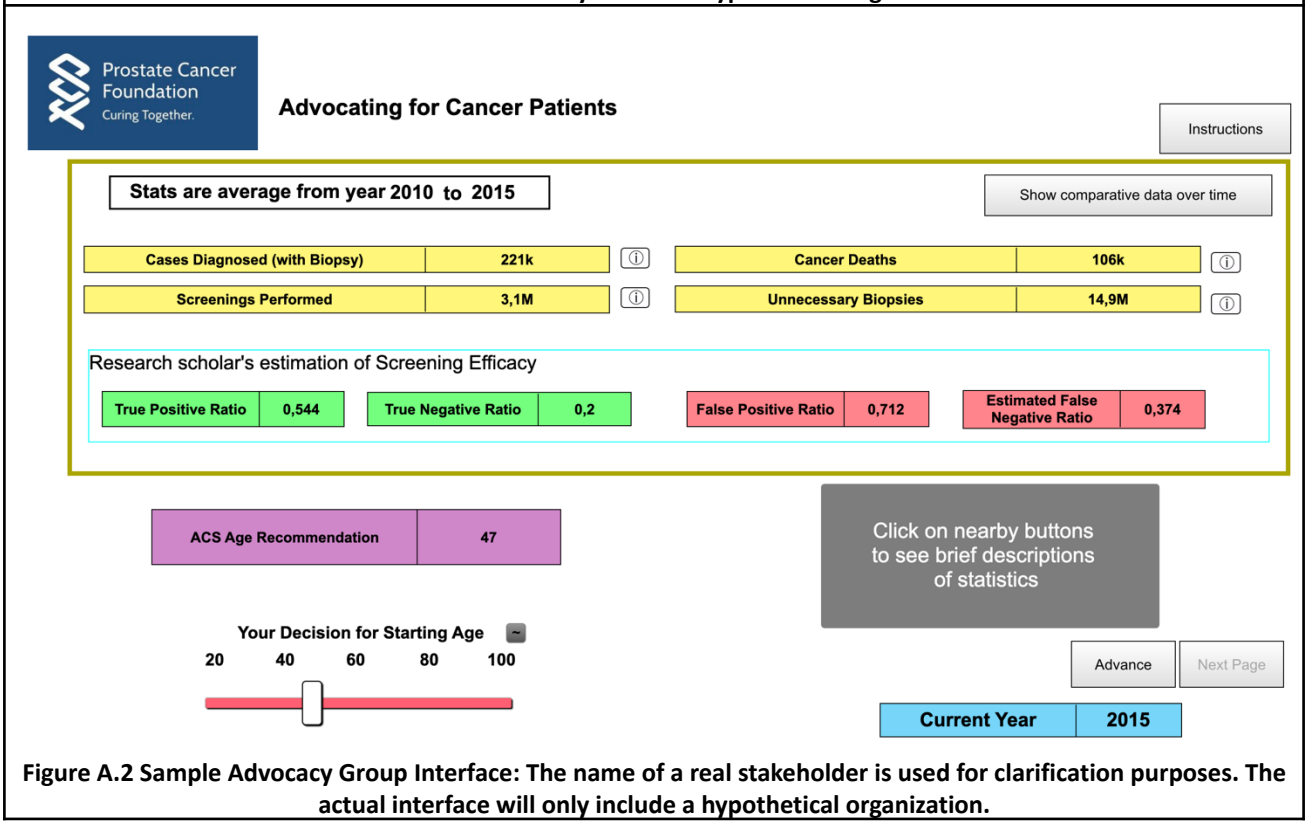

Figure A.2 Sample Advocacy Group Interface: The name of a real stakeholder is used for clarification purposes. The actual interface will only include a hypothetical organization.

Recruitment of Participants: Participants will be recruited form Amazon MTurk and first will be asked to fill an online survey. The survey is designed to collect demographic data as well as the participants a priori perceptions of cancer screening, policymakers, and advocacy groups. We plan to randomly frame participants into two groups, assuming the role of either a policymaker (i.e., representative of evidenced-based guideline issuing organizations) or an advocacy group leader. We plan to utilize a comprehensive information session that participants need to complete before playing the simulator to prime them into their roles [**See Figure A for preliminary interfaces**]. The information session familiarizes participants in each group with their respective interfaces, their decisions, and the statistics they can observe. Using a quiz, only participants with sufficient understanding will be allowed to participate in playing the simulator and complete the experiment. To reduce the mental load and avoid confusion, all participants will be asked to make recommendations only on the starting age. In contrast, in all conditions, recommended thresholds will be determined endogenously by the model.

Single- and Two-Stakeholder Models: All participants in the single stakeholder model will assume the role of an evidence-based policymaker. This enables us to directly test the endogenous theory of overuse and fluctuations in medical screening (Karanfil and Sterman, 2020). In the second experiment, we aim to focus on testing the effect of including the advocacy groups on the dynamic behavior of the main system variables, and particularly on amplifying oscillations. Using the extended model, two main groups of stakeholders can issue recommendations simultaneously.

As for the two-stakeholder model, we can use both single- and multiplayer designs. In the single-player setting, participants will assume the role of an advocacy group leader. In this condition, evidence-based guidelines will be generated by the simulation model. In the multi-player setting, pairs of participants will be randomly selected, each of whom will then be randomly assigned to take the role of a policymaker or an advocacy group leader. In this multiplayer condition, each pair will play the game synchronously.

Open-access Simulator: Upon the completion of our study, we plan to publish an open-access online simulator. The simulator can be used for two objectives: 1) for assisting policymakers in understanding the effects of their potential courses of action, 2) for pedagogical purposes by deriving practical insights on the dynamic complexity of systems dominated by endogenous feedback. (e.g., Stewart et al., 2012; Weaver and Richardson, 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.

References:

Bakris G, Sorrentino M. (2018). Redefining Hypertension — Assessing the New Blood-Pressure Guidelines. New England Journal of Medicine 378(6): 497–499.

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

Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping Doctors and Patients Make Sense of Health Statistics. Psychol Sci Public Interest. 2007 Nov;8(2):53–96.

Gigerenzer G. Full disclosure about cancer screening. BMJ. 2016 Jan 6;352:h6967.

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

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

Karanfil, Ö. (2016). *Why clinical practice guidelines shift over time: a dynamic model with application to prostate cancer screening* (Doctoral thesis, MIT, Cambridge, USA). Retrieved from https://dspace.mit.edu/handle/1721.1/10753107531

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.

Karanfil, Ö. and Sterman, J. (2020), "Saving lives or harming the healthy?" Overuse and fluctuations in routine medical screening. Syst. Dyn. Rev., 36: 294-329. https://doi.org/10.1002/sdr.1661

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.

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.

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

Penson, D. F. (2015). The pendulum of prostate cancer screening. JAMA - Journal of the American Medical Association, 314(19), 2031–2033. https://doi.org/10.1001/jama.2015.13775

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.

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, 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.

Schlesinger, Arthur Meier. (1986). The Cycles of American History. Houghton Mifflin Harcourt Publishing Company

Schwartz LM, Woloshin S. 2019. Medical marketing in the United States, 1997-2016. Journal of the American Medical Association 321(1): 80–96.

Sterman, J. (1989). Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision-Making Experiment. Management Science, 35(3), 321-339. Retrieved April 20, 2021, from http://www.jstor.org/stable/2631975

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.

Swets JA. 1992. The science of choosing the right decision threshold in high-stakes diagnostics. The American Psychologist 47(4): 522–532.

Swets JA, Dawes RM, Monahan J. 2000. Psychological science can improve diagnostic decisions. Psychological Science in the Public Interest 1(1): 1–26.

Weaver EA, Richardson G. 2006. Threshold setting and the cycling of a decision threshold. System Dynamics Review 22(1): 1–26.

Wegwarth O, Schwartz LM, Woloshin S, Gaissmaier W, Gigerenzer G. Do physicians understand cancer screening statistics? A national survey of primary care physicians in the United States. Ann Intern Med. 2012 Mar 6;156(5):340–9.