Published in final edited form as:
Med Care. 2013 April ; 51(4): 304–306. doi:10.1097/MLR.0b013e31828a7e1a.

Response: Reading between the lines of cancer screening trials: Using modeling to understand the evidence

Ruth Etzioni, PhD and
Fred Hutchinson Cancer Research Center

Roman Gulati, MS
Fred Hutchinson Cancer Research Center

Abstract

In our article about limitations of basing screening policy on screening trials, we offered several examples of ways in which modeling, using data from large screening trials and population trends, provided insights that differed somewhat from those based only on empirical trial results. In this editorial, we take a step back and consider the general question of whether randomized screening trials provide the strongest evidence for clinical guidelines concerning population screening programs. We argue that randomized trials provide a process that is designed to protect against certain biases but that this process does not guarantee that inferences based on empirical results from screening trials will be unbiased. Appropriate quantitative methods are key to obtaining unbiased inferences from screening trials. We highlight several studies in the statistical literature demonstrating that conventional survival analyses of screening trials can be misleading and list a number of key questions concerning screening harms and benefits that cannot be answered without modeling. While we acknowledge the centrality of screening trials in the policy process, we maintain that modeling constitutes a powerful tool for screening trial interpretation and screening policy development.

The article by Melnikow and colleagues (1) brings into sharp focus the essence of the policy development process and the tug-of-war between randomized controlled trials (RCTs) and other sources of evidence, in this case, the use of models. Their opinions reflect the widespread sentiments of confidence in the ability of RCTs to eliminate bias and of distrust in modeling due to its complexity and frequent lack of transparency. These comments compel us to examine closely the issues of bias and complexity in screening studies and the roles of study design and analysis in achieving unbiased interpretations of the evidence.

There is no question that the RCT paradigm represents a gold standard for evidence. But why? Because it provides a process that enables the interventions of interest to be allocated to subjects in a random, non-selective fashion. Thus, the RCT, by design, avoids one of the greatest threats to valid inference, namely selection bias. Further characteristics of the RCT process (e.g., blinding subjects and/or investigators and intention-to-treat methods) are designed to reinforce the freedom of resulting inferences from selection and related biases. But the RCT paradigm does not actually specify how those inferences are to be made and it does not provide a blueprint for the “correct” analytic model. Thus, the RCT paradigm only sets the stage for unbiased inferences; it does not guarantee them.

Correspondence to: Ruth Etzioni, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M2-B230, Seattle WA 98109-1024; Tel: +1.206.667.6561; Fax: +1.206.667.7264; retzioni@fhcrc.org.

The case of cancer screening provides a perfect example for how conventional analysis of a well-conducted RCT can yield a biased inference. The Health Insurance Plan breast cancer screening trial was a seminal RCT of mammography screening initiated in 1963 (2). Beyond the extensive ramifications of this study for clinical practice, the trial stimulated a rich statistical methodological investigation regarding appropriate methods for analyzing cancer screening trials (e.g., (3–5)). A key outcome of this work was the finding that the standard Cox proportional hazards model typically used to model disease-specific survival outcomes among clinical trial participants is not valid in the screening trial setting because the hazards (or risks) of death in the two groups are not proportional. Thus, the hazard ratio (or the often cited mortality rate ratio) is a biased estimate of the relative reduction in the risk of disease-specific death associated with screening. As Hanley (6) explains, there is invariably a delay from the start of the trial until the attainment of screening-induced mortality reductions. Analyses that merge the deaths in this early “no-reduction window” with later deaths attenuate estimates of screening benefit. He illustrates his point by examining how the mortality rate ratio in the ERSPC has changed with time since the beginning of the trial. Results indicate that after a delay of approximately 7 years the prostate cancer mortality reductions are considerably greater than the 20% reduction reported by ERSPC investigators, reaching 67% (80% confidence interval 30–89%) beginning after 12 years of follow-up.

This simple example demonstrates the complexity of quantifying the benefits of a cancer screening test. The statistical literature has clearly shown that, even in the case of a well-designed screening trial, the standard analyses that are established in the treatment trials setting must be modified to achieve valid inferences about the relative mortality reduction induced by screening. And inferences about absolute mortality reductions are even more suspect because of their clear dependence on the time horizon used to estimate them. Indeed, even if the relative mortality reduction is constant over time (i.e., the proportional hazard assumption is met), the absolute mortality reduction (lives saved by screening) will continue to grow (7). Thus, screening trials conducted over a limited time horizon cannot provide unbiased information about absolute benefit in terms of the lives that will be saved by screening over a lifetime. Similarly, attempting to use observed incidence data from trials to estimate harms such as overdiagnosis invariably produces an inflated result (7). This is because the excess incidence in the screened group relative to the control group, typically used as a proxy for overdiagnosis, consists of a mixture of overdiagnosed cases and true early detections, but we cannot distinguish these on the basis of observed data. Indeed, we cannot think of a setting where empirical data can be used to provide an unbiased estimate of overdiagnosis.

It is not our purpose to question the importance of screening trials and their necessary place in the policy development process. Our point is that making valid inferences from screening trials and correctly interpreting the evidence generally requires going beyond the observed trial results. It is possible, and indeed even likely, that using models to do this will produce estimates of harms and benefits that differ substantially from the results observed in screening trials. For example, using both a simple back-of the envelope approach (8) and a considerably more complex model (9), we projected that the ERSPC relative mortality reduction should translate into 6 lives saved per 1000 men screened in a population screening program beginning at age 50 or 55 rather than the 1 life saved based on the observed results (10). Similarly, we have estimated, using two different models of overdiagnosis, that roughly 25% of screen-detected cases are overdiagnosed in the US (11, 12) instead of the more than 50% based on excess incidence in the ERSPC trial after 9 years of follow-up (13). But it is precisely because we have used models to go beyond the trials that our results are different. When we restrict the models to the trial protocol and follow-up, our projections closely match observed relative and absolute mortality reductions after 11

years of follow-up (9). This step of validating a model against published findings is necessary to test its reliability.

The issue of model reliability is critical since no model can perfectly represent the biological complexity of disease natural history and its interaction with a screening intervention. And there is no question that there are inadequate and biased models in the literature. Some modeling studies make indefensible assumptions, do not provide sufficient information about the assumptions made, or do not adequately validate their findings against published data. But there is also a growing cadre of models that represent thoughtful abstractions of the biology, conduct mathematically coherent calibration to observed data, and provide careful and detailed documentation of clinically plausible assumptions.

The assumptions made by models are considered to be their Achilles' heel by Melnikow and colleagues (1). But even simple analyses that would not be considered "modeling" make assumptions. For example, using a mortality rate ratio to summarize the empirical reduction in the risk of prostate cancer death under screening effectively assumes that the relative risk is constant in the screening and control groups and approximates the hazard ratio in a Cox proportional hazards analysis of disease-specific survival (14). Further, statements that may appear intuitive often make implicit assumptions that are simply not acknowledged. For example, Melnikow and colleagues state that "competing causes of mortality in older men make it progressively less likely that longer follow-up will demonstrate a large absolute reduction in disease-specific mortality." While it is true that competing deaths increase as men age, so do the number of fatal prostate cancers and the potential lives that could be saved by early detection (9). This statement therefore implicitly assumes that the growth in the number of fatal cancers that could be saved by screening is outweighed by the competing risk of other-cause death among men in their seventies. As another example, the statement that the "effect of crossover [in the PLCO] is to reduce the impact of the intervention, but it cannot eliminate a benefit that is truly present" implicitly makes two assumptions. The first assumption is that if screening is beneficial compared with no screening, then more screening will save more lives than less screening. In the case of the PLCO trial, "more screening" corresponded to screening every year and "less screening" corresponded to screening approximately every other year (15, 16). However, several studies have indicated that any additional benefit of screening for prostate cancer annually versus every other year is likely to be marginal (9, 17). The second assumption is that the trial itself was precise enough to detect a difference between more versus less screening. In fact, there were far fewer deaths than expected in the PLCO trial. Using modeling we were able to show that the mortality results were noisy enough that the chance of a null or reversed result (excess mortality in the intervention group) was possible even if screening is beneficial (18). There is no question that making assumptions can be dangerous, but we believe that explicit, documented assumptions are far preferable to implicit, undocumented ones.

In recent years, modeling has become more widely accepted as a part of the policy development process. The USPSTF has been perhaps the most influential of US policy panels in the movement towards greater acknowledgment of the utility of modeling in this setting. Indeed, the USPSTF has used models in developing its most recent recommendations for both breast and colorectal cancer screening (19, 20). These models are not dissimilar to the ones we have developed and advocated in the case of the prostate screening trials. The critical question is how models should contribute to the evidence that will ultimately drive the policy decision.

We agree that randomized controlled trials provide the best opportunity to obtain the strongest evidence for clinical guidelines. The problem with randomized screening trials is that this evidence is rarely packaged in a way that permits proper interpretation. With

screening trials there is almost always a further step required to actually unlock the evidence that they are poised to provide. Modeling by itself cannot create evidence but modeling has the potential to unlock the rich repository of evidence in screening trial data. Thus, using models, we are able to conclude that a reverse result was indeed possible in the PLCO even in the presence of a clinically significant screening benefit. Using models, we can interrogate screening trial incidence patterns to learn about the lead time (time by which screening advances diagnosis) and estimate overdiagnosis, an inherently unobservable quantity. And, using models, we can project absolute lives saved implied by trial mortality results over a time horizon that matches the policy perspective. If we insist on taking randomized screening trials at face value then we run the risks of, at best, inadequately using a prime resource for information about screening outcomes and, at worst, making incorrect inferences about screening harm and benefit. Well-developed and documented models can complement, rather than contravene, empirical screening trial results and provide a more complete assessment of the net benefits of cancer screening.

Bibliography

1. Melnikow J, LeFevre ML, Wilt TJ, et al. Randomized Trials Provide the Strongest Evidence for Clinical Guidelines: The US Preventive Services Task Force and Prostate Cancer Screening. Med Care. 2013
2. Shapiro S. Evidence on screening for breast cancer from a randomized trial. Cancer. 1977; 39:2772–2782. [PubMed: 326378]
3. Aron JL, Prorok PC. An analysis of the mortality effect in a breast cancer screening study. Int J Epidemiol. 1986; 15:36–43. [PubMed: 3957541]
4. Zucker DM, Lakatos E. Weighted Log Rank Type Statistics for Comparing Survival Curves when There is a Time Lag in the Effectiveness of Treatment. Biometrika. 1990; 77:853–864.
5. Self SG, Etzioni R. A likelihood ratio test for cancer screening trials. Biometrics. 1995; 51:44–50. [PubMed: 7766795]
6. Hanley JA. Measuring mortality reductions in cancer screening trials. Epidemiol Rev. 2011; 33:36–45. [PubMed: 21624962]
7. Gulati R, Mariotto AB, Chen S, et al. Long-term projections of the harm-benefit trade-off in prostate cancer screening are more favorable than previous short-term estimates. J Clin Epidemiol. 2011; 64:1412–1417. [PubMed: 22032753]
8. Etzioni R, Gulati R, Cooperberg MR, et al. Limitations of Basing Screening Policies on Screening Trials: The US Preventive Services Task Force and Prostate Cancer Screening. Med Care. 2013
9. Gulati R, Gore JL, Etzioni R. Comparative Effectiveness of Alternative PSA-based Screening Strategies. Ann Intern Med. 2013
10. Moyer VA. on behalf of the USPSTF. Screening for Prostate Cancer: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med. 2012; 157:120–134. [PubMed: 22801674]
11. Gulati R, Wever EM, Tsodikov A, et al. What if I don't treat my PSA-detected prostate cancer? Answers from three natural history models. Cancer Epidemiol Biomarkers Prev. 2011; 20:740–750. [PubMed: 21546365]
12. Telesca D, Etzioni R, Gulati R. Estimating lead time and overdiagnosis associated with PSA screening from prostate cancer incidence trends. Biometrics. 2008; 64:10–19. [PubMed: 17501937]
13. Schroder FH, Hugosson J, Roobol MJ, et al. Screening and prostate-cancer mortality in a randomized European study. N Engl J Med. 2009; 360:1320–1328. [PubMed: 19297566]
14. Whitehead J. Fitting Cox's regression model to survival data using GLIM. J R Stat Soc Ser C Appl Stat. 1980; 29:268–275.
15. Pinsky PF, Black A, Kramer BS, et al. Assessing contamination and compliance in the prostate component of the Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer screening trial. Clin Trials. 2010; 7:303–311. [PubMed: 20571134]

16. Berg CD. The Prostate, Lung, Colorectal and Ovarian cancer screening trial: The prostate cancer screening results in context. Acta Oncol. 2011; 50 (Suppl 1):12–17. [PubMed: 21604935]

17. Ross KS, Carter HB, Pearson JD, et al. Comparative efficiency of prostate-specific antigen screening strategies for prostate cancer detection. JAMA. 2000; 284:1399–1405. [PubMed: 10989402]

18. Gulati R, Tsodikov A, Wever EM, et al. The impact of PLCO control arm contamination on perceived PSA screening efficacy. Cancer Causes Control. 2012; 23:827–835. [PubMed: 22488488]

19. Mandelblatt JS, Cronin KA, Bailey S, et al. Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. Ann Intern Med. 2009; 151:738–747. [PubMed: 19920274]

20. Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, et al. Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force. Ann Intern Med. 2008; 149:659–669. [PubMed: 18838717]