Recognizing the Limitations of Cancer Overdiagnosis Studies: A First Step Towards Overcoming Them

Ruth Etzioni, Roman Gulati

Affiliations of authors: Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA (RE, RG).

Correspondence to: Ruth Etzioni, PhD, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M2-B210, P. O. Box 19024, Seattle, WA 98109-1024 (e-mail: retzioni@fredhutch.org).

Abstract

Numerous studies have attempted to quantify the number of breast cancers that would never have been diagnosed in the absence of screening. Unfortunately, results are highly variable across studies and there is considerable disagreement about both the frequency of overdiagnosis and the validity of different methodologic approaches. In this Commentary, we review limitations of the two major approaches used in existing studies. Studies that use excess incidence as a proxy for overdiagnosis require a valid estimate of incidence in the absence of screening and sufficient follow-up to ensure the excess excludes relevant (ie, nonoverdiagnosed) cancers detected early. The requirement of sufficient follow-up applies to both population studies and clinical trials, but only certain clinical trial designs have the potential to yield unbiased results. Studies that model disease natural history to infer overdiagnosis must, in addition, examine whether their models produce valid estimates in the presence of nonprogressive cases. In this setting, limited follow-up could lead to a lack of identifiability of the parameters needed to accurately infer overdiagnosis. In a polarized research community, the excess incidence and modeling approaches are generally viewed as competitors, but we argue that they are complementary, with models being more complex but having greater potential to inform about disease natural history and the outcomes of candidate screening policies. Rather than arguing why one approach should be preferred to another, investigators should focus on developing studies that generate reliable estimates of overdiagnosis. Recognizing that both approaches have limitations, which existing studies rarely overcome, is a first step towards reconciling methodologic perspectives and achieving consensus about the real magnitude of the overdiagnosis problem.

Mammography screening for breast cancer implies both benefits and harms. The primary benefit—cancer deaths avoided—has been demonstrated in randomized controlled trials for women age 50 to 74 years (1,2). In contrast, the primary harm—overdiagnosis—is a subject of heated discussion (3–7).

Much of the discussion about overdiagnosis revolves around a simple but critical question: how many overdiagnosed tumors are there? In other words, how many screen-detected tumors would never surface in the absence of screening? This question has proven surprisingly difficult to answer. Estimates span a wide range—from less than 5% (8) to over 50% (9) of all breast cancers—with very different implications for screening decisions. Disagreement about the fraction of cases overdiagnosed has created a great deal of confusion for patients and providers. And disagreement about the methodologic approach used has polarized the screening research community. In this Commentary, we make the point that each methodologic approach has challenges and limitations; our objective is to carefully clarify these in an effort to advance ideas for the reconciliation of opposing perspectives.

There are two main approaches for quantifying overdiagnosis (10). The excess incidence (EI) approach is more direct and is based on a comparison of disease incidence with and without screening. The lead time (LT) approach is less direct; it first estimates the lead time—the time by which diagnosis is advanced by screening—and then infers the implied frequency of overdiagnosis by comparing the lead time with the time to competing death. Each approach has potential strengths and limitations, but these are not always well understood. We previously identified methodologic approaches as one of several identifiable...
reasons for the variation in published overdiagnosis estimates (10).

In this Commentary, we provide a fresh look at the limitations of the two most commonly used approaches for estimating overdiagnosis. Rather than preferring one over the other, we find reasons to be cautious about the results of either approach unless certain conditions are satisfied, and we observe that these conditions are rarely satisfied in existing studies. We issue a call to acknowledge these limitations and conclude with recommendations to improve the reliability of cancer overdiagnosis estimates.

We begin by noting that there are potentially two kinds of overdiagnosed tumors. One kind is a progressive tumor that would have reached the point of clinical presentation were it not for the limited life expectancy of the patient, perhaps because of advanced age or comorbidity. The other is a nonprogressive tumor that would have remained indolent or spontaneously regressed. Either way, the patient would have died of another cause without a cancer diagnosis were it not for screening. Different approaches for quantifying overdiagnosis account for nonprogressive tumors in different ways.

The EI Approach

The EI approach compares disease incidence with and without screening to provide a direct estimate of overdiagnosis that includes progressive and nonprogressive cancers. The EI approach has been applied to population incidence patterns (11–14) and screening trials (15,16).

In population studies, once screening has been adopted it is not possible to observe the counterfactual incidence without screening. Studies have imputed this counterfactual by extrapolating from prescreening trends (11,12,14), by using contemporaneous patterns among age groups not offered screening (11,14), or by utilizing incidence data from geographic regions that adopted screening later than the region of interest (13). In clinical trials, the control group should, in principle, provide an estimate of otherwise comparable incidence without screening. The timing of the EI calculation is also critical. Sufficient time must elapse before the calculation will produce an unbiased estimate. The timing issue pertains to estimates conducted in both population and clinical trial settings and arises because excess incidence under screening is known to include both overdiagnosed and nonoverdiagnosed cases. How long to wait is still not fully understood in general, but a recent study in the population setting suggests that once screening has been fully adopted one needs to wait at least as long as the longest lead time (17), a quantity that is directly related to disease natural history and to test sensitivity. Further research is needed to understand the minimum waiting time needed in the clinical trial setting under the various trial designs that are used for evaluating screening.

Long-term EI studies may seem to overcome the timing problem, but these must be scrutinized even more carefully for their assumptions about counterfactual incidence. For example, a recent study estimated overdiagnosis because of mammography screening in US women age 40 years and older over a period of 30 years (12). The counterfactual incidence in this study was based on incidence trends in women under age 40 years over the same period. Several responses to this study have questioned the validity of this imputed counterfactual (4,5,18). Because the study estimate that 31% of breast cancers were overdiagnosed in 2008 is based on the final year of an extrapolation over three decades, even small differences in the annual change in counterfactual incidence could substantially impact the results.

Screening trials in which there is long-term follow-up of both the intervention and control groups may seem to be optimal for overdiagnosis estimation, but even these may not permit accurate assessment of overdiagnosis. Indeed, in continuous-screen trials, where screening takes place for the full duration of follow-up, overdiagnosis estimates derived by comparing the cumulative cancers on screened and control arms are generally inflated. This is because the cancers on the screened arm will always include nonoverdiagnosed cases that would have been detected after the end of follow-up, but the corresponding cancers on the control arm are excluded from the calculation. This can potentially be remedied in stop-screen trials, where screening occurs during the first few years and follow-up continues beyond this point, but again adequate follow-up is needed and, further, careful monitoring of screening behavior on both arms is important to be confident that screening has indeed stopped on both arms. For example, in estimating overdiagnosis based on 25-year follow-up of the Canadian Breast Cancer Screening Trial (16) the screening behavior in the intervention and control arms was not monitored following the initial five-year screening period, so it is unclear whether differences in screening patterns across the trial arms persisted beyond the end of this period (19). Further, screening trials with long-term follow-up after the screening period present a conundrum with respect to the calculation of the frequency of overdiagnosis. An estimate expressed relative to all cases detected during the course of follow-up will be quite different than one expressed relative to those detected or screen-detected during the screening interval. The general problem of the sensitivity of overdiagnosis estimates to the denominator used has been previously discussed (8,20,21).

The LT Approach

The LT approach is based on a direct relationship between overdiagnosis and lead time. Mathematically, overdiagnosis arises because of the competition between the lead time and the time from screen detection to other-cause death. Indeed, the frequency of overdiagnosis is equivalent to the likelihood that the time from screen detection to other-cause death is shorter than the time to nonscreen diagnosis (i.e., the lead time). The LT approach exploits this relationship by first estimating the lead time and then using it to infer the overdiagnosis frequency.

Methods for estimating lead time are established in the statistical literature (22–24) and have been used in the context of both trials and population studies (8,25–27). Like the EI approach, the LT approach generally requires data on or an assumption about clinical incidence in the absence of screening. Unlike the EI approach, the LT approach also requires data on screening utilization.

Many LT studies applied to breast cancer focus on invasive tumors and assume that they are progressive (23,28–30). While there are exceptions (25,27), for mathematical tractability these statistical models often assume that the lead time follows an exponential distribution. Estimates of the mean lead time based on these studies (25) range from two to four years for invasive cases detected by screening; a recent methodologic study cited 40 months as a consensus value (17).

If a study estimated overdiagnosis using both the EI and the LT approaches, one could conceivably be used as a check on the other. To explore this idea, we recently considered what the mean lead time would have to be to match the published EI estimate that 31% of breast cancers in the United States were overdiagnosed in 2008 (31). We conservatively calculated that this
estimate would imply a mean lead time of nine years among invasive cases, considerably longer than 40 months. We concluded that the 31% overdiagnosis estimate seemed excessive. However, our calculations—like those that had yielded the value of 40 months for the mean lead time—were all performed under a progressive disease assumption for invasive cases (32).

The progressive disease assumption (6,7,32) is restrictive because it does not allow for separate specification and identification of the nonprogressive tumors that theoretically have an infinite lead time. Allowing for these tumors would produce a very different form for the lead-time distribution, one with a spike of values at infinity (Figure 1). A lead-time distribution that does not allow for this spike may not reflect the frequency of all (ie, progressive and nonprogressive) overdiagnosed cancers. The extent to which LT-based overdiagnosis estimates that assume progressive disease may misrepresent the true frequency of overdiagnosis has not been formally studied. We therefore designed a simple simulation exercise to explore this question.

Suppose a study assumes an exponential lead time distribution and estimates a mean of two to four years, suggesting relatively little overdiagnosis. Is it possible that the true lead time distribution is actually a mixture of finite lead times, representing progressive cancers, and infinite lead times, representing nonprogressive cancers? And if so, could the fraction of nonprogressive cancers be considerable?

To answer this question, we simulated a sample of lead times generated by such a mixture. We varied the percent of nonprogressive cancers (p) and the mean for the finite exponential lead times (m); for each combination, we used maximum likelihood to estimate the mean of the resulting sample of lead times (l) assuming an exponential distribution. We repeated this exercise for different choices of follow-up to induce different amounts of censoring in the data. Finally, we identified combinations of p and m that yielded a value for l of 40 months under each censoring time. Additional details are given in the Appendix below.

Figure 2 shows that, as expected, we recover a mean of 40 months when all cases are progressive as differences between the estimate and 40 months are small (indicated by the dark region in the figure, all lead times are finite). However, the figure also shows that there is a range of combinations of p and m consistent with the 40-month estimate (dark region in the figure for varying p). This range includes combinations with non-trivial (10%-15%) fractions of nonprogressive cancers and relatively short mean lead times for progressive cancers, particularly when follow-up is limited. Whether these combinations are biologically plausible is unclear, but if they are then this exercise suggests that ignoring nonprogressive cancers could lead to inaccuracies in overdiagnosis estimates even in an idealized setting in which lead times could be observed.

In practice, the LT approach faces even greater challenges. First, lead times are not directly observable, so a standard approach (eg, [23,28–30]) is to derive them based on estimating the sojourn time, which is the latent preclinical duration (33). In this case, the issue illustrated by this simplified example, namely the potential to misidentify the fraction of nonprogressive cases, will still apply. In practice, however, sojourn times are often interval-censored, making the identification of the fraction of nonprogressive cases potentially even more difficult. Further research is needed to understand the circumstances that will permit estimation of these models when explicitly allowing for nonprogressive cancers.

The LT approach represents an instance of a more general approach in which a model of disease natural history is estimated from screening trends and disease incidence and overdiagnosis is inferred based on the fitted model. In principle, this approach can produce estimates of overdiagnosis partitioned into the fraction due to progressive vs nonprogressive cases so long as the model used is adequately complex to represent the mixture of cases and so long as the data permit identification of the model parameters. The estimated model can then provide useful information that goes beyond the overall fraction of cases diagnosed, including how the frequency of overdiagnosis varies across subgroups with differing clinical or demographic covariates (34–36). It can also be used to predict the frequency of overdiagnosis under alternative screening trial protocols and/or alternative screening service programs, which can directly inform patient and policy decisions. The EI approach does not naturally integrate into this extended framework.

**Overcoming Limitations**

Based on our review of the limitations of the EI and LT approaches for estimating overdiagnosis, we offer several observations. First, strong opinions about the frequency of overdiagnosis based on
existing studies are not warranted. In this sense, we are inclined to agree with Bieshuevel et al. (38) that all available estimates of overdiagnosis due to mammography screening are likely to be biased. Second, studies that have reliable assessments of the incidence in the absence of screening and sufficient follow-up must be preferred. In the case of screening trials, the trial design can influence the validity of estimates of overdiagnosis; eg, cumulative estimates from continuous-screen trials are generally biased. If cumulative estimates are being used, stop-screen trials are to be preferred, but adequate follow-up is still important and the screening behavior of patients in both comparison groups must be verified for the entire duration of follow-up. Third, what is meant by sufficient follow-up remains to be fully defined, but Duffy and Parmar suggest that it should be at least as long as the time until screening stabilizes plus the maximum lead time (17). Fourth, both the EI and LT approaches, when applied in population settings, can be sensitive to assumptions about counterfactual incidence. In general, practitioners of both approaches need to be more self-critical about their assumptions and to closely scrutinize robustness of their estimates against alternative biological explanations that could have generated the same empirical data. And again, additional research is needed to better understand the magnitude of bias in various trial and population settings.

While the LT approach and related modeling methods have the potential to provide greater insights into the disease process than the EI approach, very few LT studies (for example, (27)) have attempted to capture a mixture in the natural history that explicitly involves nonprogressive cancers. Recent work from the Cancer Intervention and Surveillance Modeling Network (CISNET) includes some mixture-type models (37). Mathematically, this is a challenging problem, as the same observed incidence patterns may result from different combinations of the fraction of nonprogressive cancers and lead times among progressive cancers. This is known as a problem of identifiability. The characteristics of studies that are more or less susceptible to nonidentifiability problems have not been formally established. However, studies with adequate follow-up and that include sufficient numbers of interval cases are likely to be most tractable. Explicitly partitioning cases detected under screening into progressive and nonprogressive cancers will enhance estimation but will in general require additional knowledge or data. For example, information about frequency and characteristics of indolent or regressive cancers could eventually be based on molecular signatures that have been linked with disease outcomes (38–41). Explicitly incorporating information by cancer subtype, such as in situ cancers, in which the nonprogressive fraction may be quite high, is particularly important. To date, there are very few studies that estimate the likelihood of overdiagnosis separately for in situ and invasive cases detected by screening. Both exceptions that we are aware of rely on external information (27,42).

In conclusion, overdiagnosis is a critically important potential harm of cancer screening and understanding its frequency both in absolute terms and relative to cancer deaths prevented is a vital part of informed decision-making around early detection. However, the available approaches for estimating overdiagnosis are methodologically challenging. It is relatively straightforward to understand how they work but difficult to verify that they are working correctly. We believe that lack of clarity about the limitations of existing studies is leading to potentially incorrect assessments of the overdiagnosis risks associated with breast cancer screening. This is of particular concern because it could eventually lead to poor public policies and personal screening decisions. However, there is hope so long as researchers restrict attention to data and models that are adequate for overdiagnosis estimation and so long as the consumers of overdiagnosis studies understand that existing research studies don’t always produce the right answer.

**Funding**

This work was supported by the National Cancer Institute at the National Institutes of Health (award number U01CA157224) as part of the Cancer Intervention and Surveillance Modeling Network (CISNET). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

**Note**

The authors have no conflicts of interest to declare.

**Appendix**

We designed the following simulation exercise to examine how well estimation of the mean lead time works when lead times are a mixture of finite (for progressive cancers) and infinite (for nonprogressive cancers) lead times.

First, we generated a population of \( n = 5000 \) individual case records. A percent \( p \) of cases was assumed to have nonprogressive cancer and assigned infinite lead times. The remaining cases

![Figure 2. Simulated differences between 1) estimated mixtures of finite (for progressive cancers) and infinite (for nonprogressive cancers) lead times and 2) 40 months. Smaller absolute differences are darker. The dark region shows values for the fraction of nonprogressive cancers and the mean lead time for the progressive cancers that generate an estimated mean lead time of 40 months under a progressive (exponential) lead time assumption. Longer follow-up constrains the possible percent of lead times that are infinite and consequently the fraction of cancers that can be nonprogressive.](image-url)
were randomly assigned lead times from an exponential distribution with mean $m$ years. All infinite lead times and finite lead times exceeding five, 10, or 15 years were censored.

Second, given the simulated case records, we estimated the mean lead time assuming all lead times were generated from an exponential distribution. The score equation in this situation is:

$$\frac{d}{dm} \ln\left[p(t_c | c) \right] = \frac{d}{dm} \ln\left[\prod_{i=1}^{n} m^{-c_i} e^{(c_i / m)} \right] = m \sum_{i=1}^{n} c_i - \sum_{i=1}^{n} t_i = 0,$$

where $(t_c, c)$ is the lead time ($t_c > 0$) and censoring status ($c = 1$ for nonscreen diagnosis and $c = 0$ otherwise) for the $i$th record. Thus the maximum likelihood estimator is the ratio of total events to total person-years at risk:

$$\hat{m} = \frac{1}{\sum_{i=1}^{n} t_i / \sum_{i=1}^{n} c_i}.$$

Finally, we compared $\hat{m}$ and a published estimate of mean lead time for invasive breast cancers—40 months—for each combination of the mean of finite lead times $m$ and the percent of nonprogressive cancers $p$ over a regular grid of candidate values. We note that a closed-form analytic formula to maximize the likelihood of the mixture can be derived for large samples (details available upon request):

$$\hat{m} = \frac{p}{1-p} T(1-z) + m,$$

where $z = e^{-7/T}$ and $T$ is the censoring time for the study. Using this solution, the combinations of $m$ and $p$ that yield a value for $\hat{m}$ of 40 months lie exactly on the locus of darkest points determined by the simulation exercise in Figure 2.

References