BMJ 2015;350:g7773 doi: 10.1136/bmj.g7773 (Published 7 January 2015)

RESEARCH

Quantifying and monitoring overdiagnosis in cancer screening: a systematic review of methods

OPEN ACCESS

Jamie L Carter *resident physician*[1], Russell J Coletti *resident physician*[2], Russell P Harris *professor of medicine*[3]

[1]Department of Medicine, University of California, San Francisco, San Francisco, CA 94110, USA; [2]Division of General Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [3]Sheps Center for Health Services Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Abstract

Objective To determine the optimal method for quantifying and monitoring overdiagnosis in cancer screening over time.

Design Systematic review of primary research studies of any design that quantified overdiagnosis from screening for nine types of cancer. We used explicit criteria to critically appraise individual studies and assess strength of the body of evidence for each study design (double blinded review), and assessed the potential for each study design to accurately quantify and monitor overdiagnosis over time.

Data sources PubMed and Embase up to 28 February 2014; hand searching of systematic reviews.

Eligibility criteria for selecting studies English language studies of any design that quantified overdiagnosis for any of nine common cancers (prostate, breast, lung, colorectal, melanoma, bladder, renal, thyroid, and uterine); excluded case series, case reports, and reviews that only reported results of other studies.

Results 52 studies met the inclusion criteria. We grouped studies into four methodological categories: (1) follow-up of a well designed randomized controlled trial (n=3), which has low risk of bias but may not be generalizable and is not suitable for monitoring; (2) pathological or imaging studies (n=8), drawing conclusions about overdiagnosis by examining biological characteristics of cancers, a simple design limited by the uncertain assumption that the measured characteristics are highly correlated with disease progression; (3) modeling studies (n=21), which can be done in a shorter time frame but require complex mathematical equations simulating the natural course of screen detected cancer, the fundamental unknown question; and (4) ecological and cohort studies (n=20), which are suitable for monitoring over time but are limited by a lack of agreed standards, by variable data quality, by inadequate follow-up time, and by the potential for population level confounders. Some ecological and cohort studies, however, have addressed these potential weaknesses in reasonable ways.

Conclusions Well conducted ecological and cohort studies in multiple settings are the most appropriate approach for quantifying and monitoring overdiagnosis in cancer screening programs. To support this work, we need internationally agreed standards for ecological and cohort studies and a multinational team of unbiased researchers to perform ongoing analysis.

Introduction

Overdiagnosis, the detection and diagnosis of a condition that would not go on to cause symptoms or death in the patient's lifetime, is an inevitable harm of screening. Overdiagnosis in cancer screening can result from non-progression of the tumor or from competing mortality due to other patient conditions (that is, other conditions that would lead to the patient's death before the cancer would have caused symptoms). The consequences of overdiagnosis include unnecessary labeling of people with a lifelong diagnosis as well as unneeded treatments and surveillance that cause physical and psychosocial harm.[1] A patient who is overdiagnosed cannot benefit from the diagnosis or treatment but can only be harmed.[2]

Patients, healthcare providers, and policy makers need information about the frequency of overdiagnosis as they weigh the benefits and harms of screening. Several studies have found that patients want to factor information about overdiagnosis into their decisions about screening for breast or prostate cancer.[3-5] On a policy level, accurate measurement of the frequency of overdiagnosis is essential for monitoring the effects over time of both new screening technology (which could result in either increased or decreased overdiagnosis), new treatment, and interventions to reduce overdiagnosis.

Correspondence to: R P Harris russell_harris@med.unc.edu

Extra material supplied by the author (see http://www.bmj.com/content/350/bmj.g7773?tab=related#datasupp)

Supplemental tables: 1, modified PICOTS criteria for study eligibility; 2, search strategy; 3, standard criteria for evaluating risk of bias; 4, criteria for evaluating strength of evidence

Because it is impossible to distinguish at the time of diagnosis between an overdiagnosed cancer and one that will become clinically meaningful, measurement of overdiagnosis is not straightforward. Researchers have used various methods to indirectly quantify overdiagnosis resulting from cancer screening, but the magnitude of these estimates varies widely. We conducted a systematic review to identify and evaluate the methods that have been used for measuring overdiagnosis of cancer. We also analyzed the advantages and disadvantages of each method for providing valid and reliable estimates of the magnitude of overdiagnosis, and for monitoring overdiagnosis over time.

Methods

Key questions

We have the following key questions:

1: What research methods have been used to measure overdiagnosis resulting from cancer screening tests?

2: What are the advantages and disadvantages of each method for:

Providing a valid and reliable estimate of the frequency of overdiagnosis?

Monitoring overdiagnosis over time?

Eligibility criteria

We included studies that have quantified the frequency of overdiagnosis resulting from cancer screening in an asymptomatic adult population. We limited the scope of the review to studies of overdiagnosis in the nine types of solid tumors with the highest incidence in the United States in 2012—prostate, breast, lung, colorectal, melanoma, bladder, renal, thyroid, and uterine cancers.[6] Studies in English from any setting and time frame were included. All study designs were included except non-systematic reviews,, case reports, and case series. Systematic reviews were excluded if they simply summarized studies that had each quantified overdiagnosis (for example, by combining data from several estimates of overdiagnosis). We included systematic reviews that used data from identified studies to independently compute a new estimate of overdiagnosis.

We accepted any of three definitions of overdiagnosis, each with excess incidence attributable to screening in the numerator: (1) cancers diagnosed by screening; (2) all cancers diagnosed by any method during the screening period; and (3) all cancers diagnosed by any method over the patient's lifetime (or long term follow-up).

Study identification and selection

We conducted a systematic search of PubMed and Embase on 28 February 2014 with no limits placed on dates or study design (see appendices for search strategy). To further find relevant studies, we also hand searched reference lists of systematic and narrative reviews identified during the initial search. Abstracts and full texts were reviewed independently by two reviewers for inclusion. Any disagreements about inclusion or exclusion of these studies were resolved by consensus, and a third senior reviewer was consulted to resolve any remaining disagreements.

Data extraction

One reviewer extracted relevant data into a standardized form. These data were verified by a second reviewer, and discrepancies were resolved by consensus.

Risk of bias assessment

We created standard criteria to evaluate risk of bias for each of the four main types of studies found in this review: modeling studies, pathological and imaging studies, ecological and cohort studies, and follow-up of a randomized controlled trial. Two reviewers independently rated the risk of bias for each study, and we resolved discrepancies by consensus. We adapted the criteria for ecological and cohort studies from quality criteria used in a recent systematic review of observational studies of breast cancer screening.[7] Risk of bias criteria for randomized controlled trial follow-up and pathological and imaging studies were adapted from standard criteria used by the U S Preventive Services Task Force (USPSTF).[8] We developed a new set of criteria for evaluating modeling studies for the purpose of this review, outlined in table 1⇓.

Based on these criteria, we rated a study as having high, moderate, or low risk of bias. Studies with high risk of bias had a fatal flaw that made their results very uncertain; studies with low risk of bias met all criteria, making their results more certain. Studies that did not meet all criteria but had no fatal flaw (thus making their results somewhat uncertain) were rated as having moderate risk of bias. We give general deficiencies of the studies in each study type category in the appropriate section.

Strength of evidence assessment

We developed criteria to evaluate overall strength of evidence for the body of literature for each study type based on criteria used by the USPSTF[8] and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group.[9] Each individual study was evaluated for risk of bias, directness (see below), external validity, and precision. Ecological and cohort studies and randomized controlled trials were also rated on the appropriateness of their analysis and time frame. For these studies, analysis is a central consideration because a study can be well designed and performed with minimal bias but still provide an unreliable estimate of overdiagnosis because of a faulty analysis. Two reviewers independently determined ratings for each of these criteria, and we resolved discrepancies by consensus. We adapted criteria for evaluating external validity of individual studies from the USPSTF procedure manual.[8] Although we initially assessed the external validity of studies based on their relevance to a general US adult population, we then reassessed external validity based on relevance to a Western European population, finding no change in our conclusions.

The GRADE working group defines directness as the extent to which the evidence being assessed reflects a single direct link between the interventions of interest and the ultimate outcome.[9] In this review, we evaluated the extent to which the evidence links the screening test directly to the health outcome of excess cases of cancer attributable to screening without making assumptions. A study with good directness requires minimal assumptions to draw conclusions about the magnitude of overdiagnosis and avoids extrapolating over gaps in the evidence.

We combined the ratings for risk of bias, directness, analysis, time frame, external validity, and precision with an evaluation of the consistency of the results to determine the strength of evidence for the overall body of evidence for each study design and cancer type. Table 2⇓ outlines our definitions of these terms; a complete list of criteria used to evaluate risk of bias and strength of evidence by study design can be found in the online supplemental tables 3 and 4.

Based on the criteria above, we rated the strength of evidence for each study type as being high (that is, met all criteria), moderate (did not meet all criteria but had no fatal flaw), or low (had at least one fatal flaw that made estimates highly uncertain). We give general deficiencies of the literature for each study type studies in the appropriate section, including examples of what we regarded as fatal flaws.

Data synthesis and analysis

We performed qualitative data synthesis, organizing the results by study design and cancer type. We did not attempt to perform quantitative synthesis because of the heterogeneity of the study designs, populations, and results. Using our critical appraisal of individual studies and the body of evidence for each study design, we identified strengths and weaknesses of each study design used to measure overdiagnosis. We did not assess publication bias.

Results

We reviewed 968 abstracts and 120 full texts, including 52 individual studies. When we identified multiple reports from the same authors investigating the same population or model, we included only the most recent study. The figure⇓ shows the flow diagram of the study selection process.[10] The included studies fell into four methodological groups, which we categorized as modeling studies (n=21), pathological and imaging studies (n=8), ecological and cohort studies (n=20), and follow-up of a randomized controlled trial (n=3).

Modeling studies

Characteristics of included studies: modeling studies

We included 21 modeling studies in this review: 10 of prostate cancer,[11-20] seven of breast cancer,[21-27] three of lung cancer,[28-30] and one of colon cancer overdiagnosis.[31] In general, these studies model the way cancer would hypothetically occur without screening, and then the way cancer occurs with screening, comparing the two to determine the frequency of overdiagnosis. These studies modeled a variety of screening situations and schedules. Some studies modeled only the non-progressive disease and not the competing mortality component of overdiagnosis[23 26 27]; such studies almost certainly underestimate its magnitude. Table 3⇓ summarizes the evidence from the included modeling studies.

Risk of bias: modeling studies

Several concerns raised the risk of bias in modeling studies. First, no modeling study discussed the potential biases in the data sources used in their models. Only two modeling studies[13 28] provided a table of assumptions and data sources. No studies were supported by systematically reviewed evidence; most studies picked data inputs from a variety of sources without justification, raising the risk of bias to achieve a desired output.

Second, several studies found that mean sojourn time was a key uncertain variable for which sensitivity analyses should be performed, yet only six studies specifically varied mean sojourn time or its equivalent in univariate or probabilistic sensitivity analyses.[11 13 24 25 27 28] All other studies either performed minimal sensitivity analyses that did not directly address key uncertain variables or did not perform sensitivity analyses at all, both of which we considered fatal flaws with high risk of bias.

Third, no study validated their model using a dataset and population different from the one to which the model was calibrated. Several studies used a dataset to calibrate the model and then "validated" the model by fitting it to the same original dataset. Performing true external validation would lend more credibility to the assumptions made in the model and would make it more likely that the calibrated parameters are applicable to other populations. Furthermore, all modeling studies adjusted for mean sojourn time or lead time using model-derived estimates of these values which are obtained with overdiagnosed cancers included in the calculation, resulting in incorrectly prolonged estimates of lead time which bias the overdiagnosis results toward zero.[32]

Overall, 15 of 21 modeling studies had a high risk of bias because they had the fatal flaw of not performing key sensitivity analyses. The five studies that performed univariate sensitivity analyses for mean sojourn time were rated as having moderate risk of bias.[11 24 25 27 28]

Strength of evidence: modeling studies

We rated overall strength of evidence as low for breast, prostate, lung, and colon cancer modeling studies. We rated directness as poor for all modeling studies, as they used insufficiently supported assumptions to draw conclusions about overdiagnosis, especially progression of cancer in the absence of screening. The frequency and rate of this progression is fundamental to overdiagnosis; the estimates from such models are by nature indirect. We rated the overall risk of bias for modeling studies as high, external validity as good, and consistency as poor. Precision could not be determined.

Pathological and imaging studies

Characteristics of included studies: pathological and imaging studies

We found eight studies that drew conclusions about overdiagnosis based on a pathological or imaging characteristic.[33-40] These studies examined only overdiagnosis resulting from non-progressive disease and not competing mortality; thus, they underestimated total overdiagnosis. Table 4⇓ summarizes the evidence from the included pathological and imaging studies.

Risk of bias: pathological and imaging studies

Several problems increased risk of bias for these studies: inability to obtain complete follow-up information on the included patients,[35 36 38] non-management of potential confounders,[33 35 39 40] use of inconsistent methods for determining tumor characteristics,[38 40] and invalid or unreliable ascertainment of cause of death.[35] Overall, three lung cancer studies had a high risk of bias,[35 38] and three had a moderate risk of bias. Both prostate cancers studies had a high risk of bias.

Strength of evidence: pathological and imaging studies

We rated the strength of evidence as low for all prostate and lung cancer pathological and imaging studies. With one exception, directness was poor for all pathological and imaging studies because the validity of the study estimates was contingent on the unexamined assumption that the pathological or imaging characteristics were directly and strongly correlated with cancer related morbidity and mortality. We rated the overall risk of bias of pathological and imaging studies as moderate to high, external validity as fair to good, and consistency as poor. Precision could not be determined.

Ecological and cohort studies

Characteristics of included studies: ecological and cohort studies

We found 20 ecological and cohort studies that met our criteria, 18 of breast cancer[41-58] and two of prostate cancer.[59 60] The breast cancer studies were typically European with screening programs involving biennial mammography for women aged 50–69 years. Table 5⇓ summarizes the evidence from the included ecological and cohort studies.

Risk of bias: ecological and cohort studies

In general, ecological and cohort studies have a high risk of selection bias and confounding due to the comparison of non-randomized populations or cohorts. The included studies used several variations of unscreened reference populations with varying potential for bias. Most studies modeled the prescreening incidence trend through the study period to determine reference incidence, though this assumes that incidence would have continued at the same rate without non-linear changes. Several studies[45 47 53] used contemporary geographic areas without screening programs as the reference population; this approach could introduce confounders that are distributed differently between the two geographic areas. The use of a historical control group is complicated by potential confounders that may have changed in a substantial way between time periods. Two studies used a combination of three control groups, including a contemporary unscreened group and historical groups in the regions with and without screening.[49 51] These studies are better able to control for differences in incidence growth between regions but could still be biased by differential influence of confounders between regions. Some studies took additional steps to reduce the probability of selection bias and confounding, including adjusting for risk factors on a population level[41 48 50] and considering "extreme" scenarios.[41] Because of such additional steps, we were able to rate 18 of 20 ecological and cohort studies as having moderate risk of selection bias and confounding. We rated two studies as having high risk of bias because they compared screening attenders with non-attenders, groups with known differences in general health and health behaviors.[44 55] We rated 17 of 20 breast cancer studies as moderate risk of measurement bias because they did not discuss the validity and reliability of their data sources.

Overall, we rated 17 of 20 ecological and cohort studies as having a moderate risk of bias. Three breast cancer studies,[44 53 55] however, had a high risk of bias overall due to a high risk of confounding. Both prostate cancer studies had a moderate risk of bias overall.

Analysis: ecological and cohort studies

Several analysis issues related to measuring and calculating overdiagnosis are unique to ecological and cohort studies. Screening advances the time of diagnosis of preclinical cancers by the lead time, such that incidence is predictably increased in a screened population during the screening period. After the screening period, in the absence of overdiagnosis, cancers that would have presented clinically have already been detected by screening, so cumulative incidence tends to increase more gradually, and incidence rate declines. Because lead time varies among cancers and individuals, and with different screening strategies, there is no single lead time that correctly captures this time period for a population.

Often, overdiagnosis is calculated by determining the number of excess cases of cancer in a screened population (compared with a non-screened population) during the screening period, subtracting the deficit of cases in post-screening women compared with an unscreened reference, thus estimating the absolute difference in long term cumulative incidence attributable to screening. Studies that obtain follow-up data for a short period after screening ends may not sufficiently capture the post-screening deficit of cases and thus can overestimate overdiagnosis. Other studies instead perform a statistical adjustment for lead time as an alternative to achieving longer term follow-up. The validity of these statistical adjustments, however, is not clear. For example, adjusting for a "mean" lead time is likely biased because overdiagnosis depends not just on "average" lead time but also on the distribution of individual lead times, which is much more difficult to estimate and may be less generalizable from population to population. Also, most estimates of lead time are derived from models which include overdiagnosed cancers in their calculation of lead time, leading to underestimates of true overdiagnosis.[32]

We rated the adequacy of the time frame of included ecological and cohort studies. Because the lead time magnitude and distributions are largely unknown, we used these ratings as a general guide to highlight where biased estimation might be occurring. When studies performed a statistical adjustment for lead time[43 45 50 52] we did not rate their time frame. Two studies with no follow-up post-screening received poor ratings.[53 56] We rated as good a cohort study that achieved at least 10 years follow-up post-screening on all women[44] and a study[41] performed over a 30 year period during which screening had reached a steady state.[44] We rated the remaining ecological and cohort studies as fair, as they achieved variable amounts of follow-up time (4–14 years) post-screening. Studies that performed an unjustified statistical adjustment for lead time introduced greater uncertainty into the analysis and greater concern about bias; we thus rated their analysis as poor.[43 45 50 52]

Six ecological and cohort studies calculated overdiagnosis as the risk ratio of cumulative incidence of cancer in the screening group compared with the reference group over the screening period and a period of follow-up post-screening.[51 54 55 59 60] These studies defined overdiagnosis as the proportion of all cancers (including ones diagnosed after the screening period) that would never have caused clinical problems. The inclusion of cases diagnosed after the screening period in the denominator dilutes the estimate of overdiagnosis and makes the frequency of overdiagnosis highly dependent on the length of follow-up time. We rated the analysis of these studies as poor, because they all provided underestimates of overdiagnosis according to our definition. We rated as good studies that calculated overdiagnosis as the absolute excess of cases in the screened population divided by cases diagnosed in the screened population during the screening period.[41 44 46 47 57 58]

Directness: ecological and cohort studies

We rated directness as good for 15 of the 20 ecological and cohort studies because they directly quantified excess cumulative incidence in a screened population. The exceptions were the studies that performed an unjustified statistical adjustment for lead time,[43 45 50 52] which we rated as having poor directness because these studies require additional assumptions about cancer progression, and one study which excluded data from the prevalence screening round.[42]

Strength of evidence: ecological and cohort studies

We rated the strength of evidence as low for the overall body of ecological and cohort studies. However, five breast cancer ecological studies stood out among the others for having a moderate risk of bias, an unbiased analysis, and fair to good time frames.[41 46 47 57 58] The estimates of overdiagnosis from these studies gave greater confidence of accuracy. We rated the overall directness of ecological and cohort studies as good (n=15) or poor (n=5), external validity as good, precision as fair, consistency as poor, analysis as good (n=6) or poor (n=14), and time frame as fair.

Follow-up of a randomized controlled trial

Characteristics of included studies: follow-up of a randomized controlled trial

We included three long term follow-up studies of randomized controlled trials: one of the Malmo randomized controlled trial in Sweden,[61] which randomized women aged 44 to 69 years to several mammography rounds or no screening and followed them for 15 years; the second of the National Lung Screening Trial,[62] which randomized high risk US patients aged 55–74 years and followed them for up to seven years; and the third of the Canadian National Breast Screening trial,[63] which randomized Canadian women aged 40–59 and followed them for an average of 22 years. Table 6⇓ summarizes the evidence from the included follow-up studies of a randomized controlled trial.

Risk of bias: follow-up of a randomized controlled trial

All studies had a low risk of selection bias and confounding. We rated the risk of measurement bias as moderate in the Malmo and National Lung Screening Trial studies because the authors did not describe the validity and reliability of their data sources, particularly over the long term follow-up periods. In the National Lung Screening Trial follow-up, measurement bias was also moderate because lung cancer incidence assessment was not masked. Overall risk of bias was low for all three studies.

Strength of evidence: follow-up of a randomized controlled trial

We rated the time frame of the Malmo and Canadian studies as good because they achieved complete 15 year follow-up of all women in the study and 22 year follow-up on average, respectively. We rated the National Lung Screening Trial time frame as fair, only achieving seven years follow-up. The initial analysis of the Malmo study received a poor rating for diluting the overdiagnosis estimate, but the re-analysis performed by Welch and colleagues was unbiased.[64] The Welch re-analysis used the denominator of cases diagnosed during the screening period and a numerator of excess cases diagnosed in the screening group, resulting in 18% overdiagnosis rather than 10%. Overall strength of evidence was moderate for both cancer types, as only one or two studies represented each type. We rated the overall directness for follow-up of randomized controlled trials as good, external validity as good, precision as fair, consistency was not applicable, analysis as good, and time frame as fair to good. The overall rating was moderate.

Discussion

Principal findings of the review

This review identified four major research methods that have been used to measure overdiagnosis from cancer screening: modeling studies, pathological and imaging studies, ecological and cohort studies, and follow-up of a randomized controlled trial. Using the frameworks for evaluating risk of bias and strength of evidence, we identified strengths and weaknesses of each of these methods for providing valid and reliable estimates of the frequency of overdiagnosis and the suitability for monitoring overdiagnosis over time (table 7⇓).

For the purposes of estimation of overdiagnosis at a point in time, follow-up of a randomized controlled trial is ideal for minimizing biases and directly addressing the question of interest. However, because these studies require significant time and resources and often have limited external validity, they are less useful for monitoring overdiagnosis over time.

Modeling studies require less time than randomized controlled trials and, with the help of sometimes unexamined assumptions, are able to project through areas of uncertainty. These projections, however, do not change the fact of uncertainty. Sensitivity analyses demonstrate that varying key uncertain inputs such as the distribution of sojourn time substantially changes overdiagnosis estimates.[11] Most of the included studies made no efforts to mitigate these uncertainties with unbiased selection of data sources, sensitivity analyses, or external validation, and most had a high risk of bias. Because the effectiveness of treatment and the sensitivity of screening tests change over time (which may change the natural history of both treated and untreated cancer), models would need constant modification to provide valid monitoring of overdiagnosis over time. Finally, models would need to continually adjust for changes in competing mortality risks, which also change over time.

Pathological and imaging studies tend to over-simplify overdiagnosis, with an arbitrary cutoff of a defining characteristic such as volume doubling time. Furthermore, both modeling and pathological imaging studies are indirect because they require assumptions about cancer progression.

Some ecological and cohort studies are limited by confounding and problematic analyses, including uncertain statistical adjustments. However, when well designed and interpreted in combination with studies from other geographic areas and time periods, these studies can provide credible estimates of overdiagnosis. They are also suitable for monitoring of "real world" overdiagnosis over time.

Similar to models that require an estimate of lead time, some ecological and cohort studies have performed a statistical adjustment for lead time in their analyses, which introduces uncertainty. Lead time is not only uncertain, but actually prospectively unknowable, as it varies among individuals and by screening practice. It is possible to calculate an average lead time from randomized trials, though these estimates are biased because they include overdiagnosed cancers in the calculation if a model is used. The heterogeneity of cancer and of individuals among and between populations, as well as variation and changes in the sensitivity of screening tests and treatment effectiveness, makes estimating the lead time distribution a highly uncertain endeavor.

Ongoing ecological and cohort studies within established national or regional screening programs, however, with appropriate collection of information about cancer incidence, potential confounders, screening adherence, and treatments

used, have the ability to compare cancer incidence in areas with one type of screening program to incidence in areas with a different type of screening program. When carefully analyzed in an unbiased manner, such international ecological and cohort studies have the potential to help us better understand the effects of different screening programs on overdiagnosis, as well as trends in overdiagnosis as screening programs and treatments change over time. The potential credibility and usefulness of ecological and cohort studies is greater than modeling studies for these purposes.

Strengths and weaknesses of the review

The major strengths of this study are that it is a systematic review, that it offers specific criteria for evaluating studies measuring overdiagnosis, and that it looks broadly at studies of overdiagnosis of different types of cancer. There are several limitations of our review. We had to modify criteria for strength of evidence to fit the different research designs; readers should examine these criteria when interpreting our findings. We combined certain studies when multiple studies were available from the same authors or were using the same model and population, and it is possible that we missed some of the variability in the data available. We also limited the scope of our review to include only the nine types of solid tumors with the highest incidence in US adults, and no overdiagnosis estimates were available for melanoma and bladder, renal, thyroid, and uterine cancers.

Strengths and weaknesses in relation to other studies

To our knowledge, there are no other systematic reviews that have comprehensively identified all studies that measure overdiagnosis. Several systematic and non-systematic reviews have explored a subset of the overdiagnosis literature. Biesheuvel and colleagues systematically reviewed studies of breast cancer overdiagnosis with a focus on potential sources of bias in the estimates.[65] We disagree that statistical adjustment and exclusion of prevalence screening data, which they recommend, are adequate to manage problems of lead time. Furthermore, they advocate the "cumulative incidence method," in which overdiagnosis is calculated as a risk ratio of cumulative incidences several years after screening has ended, which has been a major source of confusion for other researchers who have referenced this review. This analysis method is problematic because it dilutes the overdiagnosis estimate and makes it dependent on the length of follow-up time.[65] Puliti and colleagues reviewed European observational studies of breast cancer overdiagnosis, making note of which studies adequately adjusted for breast cancer risk and lead time.[66] We question their assessment, as they favorably rated studies that statistically adjusted for lead time as well as studies that included post-screening follow-up years in the analysis.

Etzioni and colleagues non-systematically reviewed studies of breast and prostate cancer overdiagnosis.[67] They label ecological and cohort studies that do not statistically adjust for lead time as the "excess incidence approach" of overdiagnosis estimation and argue that these studies may yield a biased estimate if the early years of screening dissemination are included. They advocate excluding the first few years of screening data to make an overdiagnosis estimate less biased. We agree that if a study includes only the first few years of screening dissemination without any post-screening follow-up that this can lead to overestimation, but most existing studies appropriately measure incidence during a screening and post-screening follow-up period and are thus able to measure overdiagnosis without this bias.[67]

Etzioni and colleagues also discuss modeling studies for measuring overdiagnosis which they refer to as the "lead time approach." They claim that the main limitation of modeling studies is their lack of transparency, and that prior publication of the model in peer reviewed statistics literature is a positive indicator of the model's validity.[67] Rather than lack of transparency, we found that the inherent lack of directness of modeling studies and the potential for key uncertain inputs to greatly alter overdiagnosis estimates are the primary limitations of modeling studies. Prior model publication in the statistics literature is not a sufficient indicator of a model's validity, and authors of modeling studies should be encouraged to take steps to increase the validity of their study by using systematically reviewed data inputs and performing sensitivity analyses and external validation.

Finally, Etzioni and colleagues point out a dichotomy in the selected studies they chose to present, where modeling studies tended to have much lower estimates of overdiagnosis than ecological studies, particularly among breast cancer studies.[67] However, we found several breast cancer modeling studies[22 25 27] with much higher overdiagnosis estimates than the ones they presented, as well as ecological studies[51 54 55] with lower estimates than those presented. Their suggestion that all ecological and cohort studies overestimate overdiagnosis is unfounded.

Meaning of the review: implications for future practice and research

We suggest that the public health policy community begin a coordinated effort to develop an international ecological and cohort data monitoring system for cancer screening programs, including monitoring overdiagnosis. We found that well conducted ecological and cohort studies performed in a variety of settings can give accurate estimates and enable us to compare overdiagnosis among different screening programs and to monitor overdiagnosis over time. Some of this research is ongoing, especially in European countries with breast cancer screening programs, but it is not being performed in a uniform way. We suggest the formation of a group of unbiased international experts to set standards for ecological and population cohort studies, for countries to adopt these standards in their registries, and then for unbiased methodological experts to conduct ongoing studies to monitor screening and overdiagnosis over time.

These standards should include an adequate time frame that achieves sufficient follow-up post-screening, such that all participants in the post-screening age groups have previously been offered screening. Researchers should determine standard population level confounders, unique to each cancer type, that should be monitored and adjusted for. In addition to considering cancer risk factors as potential confounders, information systems should monitor screening strategies, screening adherence, treatments used, and patient outcomes (such as complications, morbidity, and mortality). Finally, standards for analysis should include calculation of overdiagnosis as an absolute excess of cases attributable to screening divided by a denominator of cases diagnosed during the screening period.

Setting up these registries and information systems may be challenging for some countries, but others have already made great strides in this direction. There will certainly need to be an initial investment of resources, but, once established, the potential benefits from these information systems are large.

These systems could examine the effects of variations in screening programs on the magnitude of harms and costs of overdiagnosis, as well as determining when a screening program is no longer effective because of improved treatment. Beyond overdiagnosis, these studies may also provide real world information about the benefits and harms of newer screening technologies, helping to make policy decisions about which programs to implement more widely. Such information systems could also provide platforms on which randomized controlled trials of new screening programs could be efficiently tested.

Conclusions

Researchers have measured overdiagnosis using four main methods. Follow-up of a randomized trial is ideal for internal validity but requires extended time, may lack external validity, and is not useful for monitoring. Modeling studies and pathological and imaging studies are simpler to perform but introduce uncertainty by lack of directness and requiring assumptions about cancer progression. Ecological and cohort studies can be limited by confounding and require careful analysis, but when performed well they can provide a more valid and reliable estimate of overdiagnosis. They are also well designed to monitor and compare screening programs over time. A group of unbiased researchers should set standards for these studies and monitor overdiagnosis and other outcomes of cancer screening programs in multiple countries. Monitoring screening programs is important not only in attempts to reduce overdiagnosis, but for maximizing the benefits of cancer screening while minimizing the harms and costs.

Contributors: All the authors were involved in conceptualizing the work, performing abstract and full text review, performing quality rating of included studies, synthesizing the results and drawing conclusions, and drafting and reviewing the manuscript. JLC performed the literature search, was involved in all aspects of the systematic review process, and was the lead author of the manuscript. RJC was heavily involved in abstract and full text review, data abstraction, quality rating, and data synthesis and analysis. RPH was the lead reviewer and editor of the manuscript in addition to being involved in the review process. JLC and RPH are the guarantors of the paper.

Funding: This project was supported by the Agency for Healthcare Research and Quality (AHRQ) Research Centers for Excellence in Clinical Preventive Services (grant No P01 HS021133). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality. The funders had no role in study design; in the collection, analysis, or interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Competing interests: All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Transparency: The lead author affirms that this manuscript is an honest, accurate and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from this study as planned have been explained.

Data sharing: Technical appendices are available from the corresponding author at russell.harris@med.unc.edu.

1 Black WC. Overdiagnosis: an underrecognized cause of confusion and harm in cancer screening. *J Natl Cancer Inst* 2000;92:1280-2.
2 Welch HG, Black WC. Overdiagnosis in cancer. *J Natl Cancer Inst* 2010;102:605-13.
3 Hersch J, Jansen J, Barratt A, Irwig L, Houssami N, Howard K, et al. Women's views on overdiagnosis in breast cancer screening: a qualitative study. *BMJ* 2013;346:f158.
4 Schwartz LM, Woloshin S, Sox HC, Fischhoff B, Welch HG. US women's attitudes to false-positive mammography results and detection of ductal carcinoma in situ: cross-sectional survey. *West J Med* 2000;173:307.
5 De Bekker-Grob EW, Rose JM, Donkers B, Essink-Bot ML, Bangma CH, Steyerberg EW. Men's preferences for prostate cancer screening: a discrete choice experiment. *Br J Cancer* 2013;108:533-41.
6 Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. *CA Cancer J Clin* 2013;63:11-30.
7 Harris R, Yeatts J, Kinsinger L. Breast cancer screening for women ages 50 to 69 years: a systematic review of observational evidence. *Prev Med* 2011;53:108-14.
8 US Preventive Services Task Force Procedure Manual. AHRQ publication no. 08-05118-EF. 2008. www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm.
9 Owens DK, Lohr KN, Atkins D, Treadwell JR, Reston JT, Bass EB, et al. Grading the strength of a body of evidence when comparing medical interventions-agency for healthcare research and quality and the effective health care program. *J Clin Epidemiol* 2010;63:513-23.
10 Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann Intern Med* 2009;151:W65-94.
11 Davidov O, Zelen M. Overdiagnosis in early detection programs. *Biostatistics* 2004;5:603-13.
12 Draisma G, Etzioni R, Tsodikov A, Mariotto A, Wever E, Gulat R, et al. Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. *J Natl Cancer Inst* 2009;101:374-83.
13 Gulati R, Gore JL, Etzioni R. Comparative effectiveness of alternative prostate-specific antigen-based prostate cancer screening strategies: model estimates of potential benefits and harms. *Ann Intern Med* 2013;158:145-53.
14 Gulati R, Inoue LY, Gore JL, Katcher J, Etzioni R. Individualized estimates of overdiagnosis in screen-detected prostate cancer. *J Natl Cancer Inst* 2014;106:djt367.
15 Heijnsdijk EAM, Der Kinderen A, Wever EM, Draisma G, Roobol MJ, De Koning HJ. Overdetection, overtreatment and costs in prostate-specific antigen screening for prostate cancer. *Br J Cancer* 2009;101:1833-8.
16 McGregor M, Hanley JA, Boivin JF, McLean RG. Screening for prostate cancer: estimating the magnitude of overdetection. *Can Med Assoc J* 1998;159:1368-72.
17 Pashayan N, Duffy SW, Pharoah P, Greenberg D, Donovan J, Martin RM, et al. Mean sojourn time, overdiagnosis, and reduction in advanced stage prostate cancer due to screening with PSA: implications of sojourn time on screening. *Br J Cancer* 2009;100:1198-204.
18 Telesca D, Etzioni R, Gulati R. Estimating lead time and overdiagnosis associated with PSA screening from prostate cancer incidence trends. *Biometrics* 2008;64:10-9.
19 Tsodikov A, Szabo A, Wegelin J. A population model of prostate cancer incidence. *Stat Med* 2006;25:2846-66.
20 Wu GHM, Auvinen A, Maattanen L, Tammela TLJ, Stenman UH, Hakama M, et al. Number of screens for overdetection as an indicator of absolute risk of overdiagnosis in prostate cancer screening. *Int J Cancer* 2012;131:1367-75.
21 De Gelder R, Fracheboud J, Heijnsdijk EAM, den Heeten G, Verbeek ALM, Broeders MJM, et al. Digital mammography screening: weighing reduced mortality against increased overdiagnosis. *Prev Med* 2011;53:134-40.
22 De Gelder R, Heijnsdijk EAM, Van Ravesteyn NT, Fracheboud J, Draisma G, De Koning HJ. Interpreting overdiagnosis estimates in population-based mammography screening. *Epidemiol Rev* 2011;33:111-21.
23 Duffy SW, Agbaje O, Tabar L, Vitak B, Bjurstam N, Björneld L, et al. Estimates of overdiagnosis from two trials of mammographic screening for breast cancer. *Breast Cancer Res* 2005;7:258-65.
24 Gunsoy NB, Garcia-Closas M, Moss SM. Modelling the overdiagnosis of breast cancer due to mammography screening in women aged 40 to 49 in the United Kingdom. *Breast Cancer Res* 2012;14:R152.
25 Martinez-Alonso M, Vilaprinyo E, Marcos-Gragera R, Rue M. Breast cancer incidence and overdiagnosis in Catalonia (Spain). *Breast Cancer Res* 2010;12:R58.
26 Olsen AH, Jensen A, Njor SH, Villadsen E, Schwartz W, Vejborg I, et al. Breast cancer incidence after the start of mammography screening in Denmark. *Br J Cancer* 2003;88:362-5.
27 Seigneurin A, Francois O, Labarere J, Oudeville P, Monlong J, Colonna M. Overdiagnosis from non-progressive cancer detected by screening mammography: stochastic simulation study with calibration to population based registry data. *BMJ* 2011;343:d7017.
28 Duffy SW, Field JK, Allgood PC, Seigneurin A. Translation of research results to simple estimates of the likely effect of a lung cancer screening programme in the United Kingdom. *Br J Cancer* 2014;110:1834-40.
29 Hazelton WD, Goodman G, Rom WN, Tockman M, Thornquist M, Moolgavkar S, et al. Longitudinal multistage model for lung cancer incidence, mortality, and CT detected indolent and aggressive cancers. *Math Biosci* 2012;240:20-34.
30 Pinsky PF. An early- and late-stage convolution model for disease natural history. *Biometrics* 2004;60:191-8.
31 Luo D, Cambon AC, Wu D. Evaluating the long-term effect of FOBT in colorectal cancer screening. *Cancer Epidemiol* 2012;36:e54-60.
32 Zahl PH, Jorgensen KJ, Gotzsche PC. Overestimated lead times in cancer screening has led to substantial underestimation of overdiagnosis. *Br J Cancer* 2013;109:2014-9.
33 Dominioni L, Rotolo N, Mantovani W, Poli A, Plsani S, Conti V, et al. A population-based cohort study of chest x-ray screening in smokers: Lung cancer detection findings and follow-up. *BMC Cancer* 2012;12:18.
34 Lindell RM, Hartman TE, Swensen SJ, Jett JR, Midthun DE, Tazelaar HD, et al. Five-year lung cancer screening experience: CT appearance, growth rate, location, and histologic features of 61 lung cancers. *Radiology* 2007;242:555-62.
35 Sobue T, Suzuki T, Matsuda M, Kuroishi T, Ikeda S, Naruke T. Survival for clinical stage I lung cancer not surgically treated: comparison between screen-detected and symptom-detected cases. *Cancer* 1992;69:685-92.
36 Sone S, Nakayama T, Honda T, Tsushima K, Li F, Haniuda M, et al. Long-term follow-up study of a population-based 1996-1998 mass screening programme for lung cancer using mobile low-dose spiral computed tomography. *Lung Cancer* 2007;58:329-41.

What is already known on this topic

Studies of cancer overdiagnosis, using various methods, have found an extremely wide range of results

It is unclear how to evaluate the methods of these studies in order to interpret the conflicting results and how to better perform such studies in the future

What this study adds

This systematic review highlights the high potential for bias and the reliance on unproven assumptions in modeling studies and studies that quantify overdiagnosis using pathological or imaging characteristics

We recommend that well done ecological or cohort studies performed by unbiased researchers be used to quantify and monitor overdiagnosis in various settings worldwide

37 Veronesi G, Maisonneuve P, Bellomi M, Rampinelli C, Durli I, Bertolotti R, et al. Estimating overdiagnosis in low-dose computed tomography screening for lung cancer: a cohort study. *Ann Intern Med* 2012;157:776-84.
38 Yankelevitz DF, Kostis WJ, Henschke CI, Heelan RT, Libby DM, Pasmantier MW, et al. Overdiagnosis in chest radiographic screening for lung carcinoma: frequency. *Cancer* 2003;97:1271-5.
39 Graif T, Loeb S, Roehl KA, Gashti SN, Griffin C, Yu X, et al. Under diagnosis and over diagnosis of prostate cancer. *J Urol* 2007;178:88-92.
40 Pelzer AE, Colleselli D, Bektic J, Schaefer G, Ongarello S, Schwentner C, et al. Clinical and pathological features of screen vs non-screen-detected prostate cancers: is there a difference? *BJU Int* 2008;102:24-7.
41 Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. *New Engl J Med* 2012;367:1998-2005.
42 Coldman A, Phillips N. Incidence of breast cancer and estimates of overdiagnosis after the initiation of a population-based mammography screening program. *CMAJ* 2013;185:E492-8.
43 Duffy SW, Tabar L, Olsen AH, Vitak B, Allgood PC, Chen THH, et al. Absolute numbers of lives saved and overdiagnosis in breast cancer screening, from a randomized trial and from the Breast Screening Programme in England. *J Med Screen* 2010;17:25-30.
44 Falk RS, Hofvind S, Skaane P, Haldorsen T. Overdiagnosis among women attending a population-based mammography screening program. *Int J Cancer* 2013;133:705-12.
45 Hellquist BN, Duffy SW, Nystrom L, Jonsson H. Overdiagnosis in the population-based service screening programme with mammography for women aged 40 to 49 years in Sweden. *J Med Screen* 2012;19:14-9
46 Jørgensen KJ, Gøtzsche PC. Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. *BMJ* 2009;339:b2587.
47 Jorgensen KJ, Zahl PH, Gotzsche PC. Overdiagnosis in organised mammography screening in Denmark. a comparative study. *BMC Women's Health* 2009;9:36.
48 Junod B, Zahl PH, Kaplan RM, Olsen J, Greenland S. An investigation of the apparent breast cancer epidemic in France: screening and incidence trends in birth cohorts. *BMC Cancer* 2011;11:401.
49 Kalager M, Adami H, Bretthauer M, Tamimi RM. Overdiagnosis of invasive breast cancer due to mammography screening. *Ann Intern Med* 2012;157:221-2.
50 Morrell S, Barratt A, Irwig L, Howard K, Biesheuvel C, Armstrong B. Estimates of overdiagnosis of invasive breast cancer associated with screening mammography. *Cancer Causes Control* 2010;21:275-82.
51 Njor SH, Olsen AH, Blichert-Toft M, Schwartz W, Vejborg I, Lynge E. Overdiagnosis in screening mammography in Denmark: population based cohort study. *BMJ* 2013;346:f1064.
52 Paci E, Miccinesi G, Puliti D, Baldazzi P, De Lisi V, Falcini F, et al. Estimate of overdiagnosis of breast cancer due to mammography after adjustment for lead time. A service screening study in Italy. *Breast Cancer Res* 2006;8:R68.
53 Peeters PHM, Verbeek ALM, Straatman H, Holland R, Hendriks JHCL, Mravunac M, et al. Evaluation of overdiagnosis of breast cancer in screening with mammography: results of the Nijmegen programmme. *Int J Epidemiol* 1989;18:295-9.
54 Puliti D, Zappa M, Miccinesi G, Falini P, Crocetti E, Paci E. An estimate of overdiagnosis 15 years after the start of mammographic screening in Florence. *Eur J Cancer* 2009;45:3166-71.
55 Puliti D, Miccinesi G, Zappa M, Manneschi G, Crocetti E, Paci E. Balancing harms and benefits of service mammography screening programs: a cohort study. *Breast Cancer Res* 2012;14:R9
56 Svendsen AL, Olsen AH, Von Euler-Chelpin M, Lynge E. Breast cancer incidence after the introduction of mammography screening: what should be expected? *Cancer* 2006;106:1883-90.
57 Zahl PH, Strand BH, Maeshlen J. Incidence of breast cancer in Norway and Sweden during introduction of nationwide screening: prospective cohort study. *BMJ* 2004;328:921-4.
58 Zahl PH, Maehlen J. Overdiagnosis of breast cancer after 14 years of mammography screening. *Tidsskr Nor Legeforen* 2012;132:414-7.
59 Ciatto S, Gervasi G, Bonardi R, Frullini P, Zendron P, Lombardi C, et al. Determining overdiagnosis by screening with DRE/TRUS or PSA (Florence pilot studies, 1991-1994). *Eur J Cancer* 2005;41:411-5.
60 Zappa M. Overdiagnosis of prostate carcinoma by screening: an estimate based on the results of the Florence screening pilot study. *Ann Oncol* 1998;9:1297-300.
61 Zackrisson S, Andersson I, Janzon L, Manjer J, Garne JP. Rate of over-diagnosis of breast cancer 15 years after end of Malmo mammographic screening trial: follow-up study. *BMJ* 2006;332:689-91.
62 Patz EF Jr, Pinsky P, Gatsonis C, Sicks JD, Kramer BS, Tammemägi MC, et al. Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Intern Med* 2014;174:269-74.
63 Miller AB, Wall C, Baines CJ, Sun P, To T, Narod SA. Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ* 2014;348:g366.
64 Welch HG, Schwartz LM, Woloshin S. Ramifications of screening for breast cancer: 1 in 4 cancers detected by mammography are pseudocancers. *BMJ* 2006;332:727.
65 Biesheuvel C, Barratt A, Howard K, Houssami N, Irwig L. Effects of study methods and biases on estimates of invasive breast cancer overdetection with mammography screening: a systematic review. *Lancet Oncol* 2007;8:1129-38.
66 Puliti D, Duffy SW, Miccinesi G, de Koning H, Lynge E, Zappa M, et al. Overdiagnosis in mammographic screening for breast cancer in Europe: a literature review. *J Med Screen* 2012;19(suppl 1):42-56.
67 Etzioni R, Gulati R, Mallinger L, Mandelblatt J. Influence of study features and methods on overdiagnosis estimates in breast and prostate cancer screening. *Ann Intern Med* 2013;158:831-8.

Accepted: 21 October 2014

Cite this as: *BMJ* 2015;350:g7773

Tables

Table 1| Criteria for evaluating risk of bias by study type among studies that quantified overdiagnosis resulting from cancer screening

Study type	Risk of bias criteria
Modeling studies	Extent to which assumptions are transparent and clearly stated Extent to which assumptions are backed with evidence Probability for bias in data used in model Sensitivity analyses performed for uncertain variables External validation of model
Pathological and imaging studies	Probability of selection bias and confounding Probability of measurement bias
Ecological and cohort studies	
Follow-up of a randomized controlled trial	

Table 2| Definitions of criteria for evaluating strength of evidence among studies quantifying overdiagnosis from cancer screening

Criterion	Definition
Risk of bias (high/moderate/low)	See table 1
Directness (good/fair/poor)	Extent to which the evidence links screening directly to differences in long term cumulative incidence between populations without making assumptions
Analysis (good/fair/poor)	Extent to which the analysis appropriately quantifies overdiagnosis, without inclusion of age groups or time frames that lack the potential to be overdiagnosed, and without statistically adjusting for lead time
Time frame (good/fair/poor)	Extent to which the time frame is sufficient to account for the effects of lead time
External validity (good/fair/poor)	Extent to which study population is similar to US general population or Western European populations in factors that are associated with cancer incidence, screening situation (such as expertise of screening radiographers, quality of screening facilities, threshold for defining a result as abnormal), medical care, and risks for competing mortality
Precision (good/fair/poor/cannot determine)	Confidence interval of difference in cumulative incidence attributable to screening between populations over an appropriate time frame should be provided. Width of confidence interval should be narrow, not wider than about 20%.
Consistency (good/fair/poor)	Extent to which the overdiagnosis measurements from the included studies have a similar magnitude

Table 3| **Summary of evidence from 21 modelling studies quantifying overdiagnosis from cancer screening.**

Study; model(s)	Modelled population: country, ages; screening schedule	Data sources: (*a*) Incidence; (*b*) Mortality; (*c*) Other	(*a*) External validation?; (*b*) Includes competing mortality?; (*c*) Includes DCIS?	Reports outcome as % of screen detected cancers?	Magnitude of overdiagnosis (95% CI)	Sensitivity analyses varying mean sojourn time or lead time?	Overall risk of bias
Prostate cancer (n=10)							
Davidov 2004[11]	US; 50–60, 70, or 80 year olds; at 5 year intervals	(*a*) SEER 1993–97; (*b*) SSA life tables	(*a*) No; (*b*) Yes; (*c*) N/A	Unclear	8.48–53.6%	Univariate. MST 5–15 years Overdiagnosis varied greatly with MST	Moderate
Draisma 2009[12]; MISCAN, FHCRC, UMichigan	US, 54–80 year olds; typical US screening patterns	(*a*) SEER 1985–2000; (*b*) Standard life tables	(*a*) No; (*b*) Yes; (*c*) N/A	Yes	MISCAN 42%; FHCRC 28%; UMich 23%	Not performed	High
Gulati 2013[13]; FHCRC	US, 40 year olds; 32 screening schedules simulated	(*a*) SEER 1975–2000; (*b*) US life tables	(*a*) No; (*b*) Yes; (*c*) N/A	No, reports lifetime risk of overdiagnosis	1.8–6%	Other sensitivity analyses performed.	Moderate
Gulati 2014[14]	US, 50–84 year olds; multiple	(*a*) SEER 1975-2005; (*b*) US life tables	(*a*) No; (*b*) Yes; (*c*) N/A	No	2.9–88.1% depending on age, Gleason score, and PSA level (% likelihood that a tumor is overdiagnosed)	Not performed	High
Heijnsdijk 2009[15]; MISCAN	Europe; 55–70 year olds every 1 or 2 years, or 55–75 year olds every 4 years	(*a*, *b*) ERSPC Rotterdam; (*c*) Cure rates by stage from Amsterdam Cancer Center	(*a*) No; (*b*) Yes; (*c*) N/A	Yes (estimated from figures)	Annual, 60%; biennial, 60%; every 4 years (to age 75), 67%	Not performed	High
McGregor 1998[16]	Quebec, 50–85 year olds; annual PSA test for ages 50–70	(*a*) Multiple; (*b*) Quebec Ministry of Health	(*a*) No; (*b*) Yes; (*c*) N/A	Yes	84%	Other sensitivity analyses performed.	High
Pashayan 2009[17]	UK; single PSA	(*a*) Eastern Cancer Registry, ProtecT study, UK Office of National Statistics; (*b*) UK Office of National Statistics	(*a*) No; (*b*) Yes; (*c*) N/A	Yes	50–54 years, 10% (7 to 11%); 55–59, 15% (12 to 15); 60–64, 23% (20 to 24); 65–69, 31% (26 to 32)	Not performed	High
Telesca 2008[18]	US; typical US screening patterns	(*a*) SEER 1973–87; (*b*) CDC Vital Statistics 1992	(*a*) No; (*b*) Yes; (*c*) N/A	Yes	White men 22.7%; black men 34.4%	Not performed	High
Tsodikov 2006[19]	US; typical US screening patterns	(*a*) SEER; (*b*) Human Mortality Database	(*a*) No; (*b*) Yes; (*c*) N/A	Yes	30%	Not performed	High
Wu 2012[20]	Finland; 55, 59, 63, 67 year olds; 3 PSA tests every 4 years until age 71	(*a*) Finnish Prostate Cancer Screening Trial, Finnish Cancer Registry; (*b*) Statistics Finland	(*a*) No; (*b*) Yes; (*c*) N/A	No	3.4% (2.4 to 5.7%) risk of overdetection during study period	Not performed	High
Breast cancer (n=7)							
De Gelder 2011 (Epi Rev)[21]; MISCAN	Netherlands; 0–100 year olds; biennial mammogram ages 49–74	(*a*) Dutch Comprehensive Cancer Centers, National Evaluation Team for Breast Cancer Screening 1990–2006	(*a*) No; (*b*) Yes; (*c*) Yes	Yes	Implementation 22.1–67.4%; extension 15.4–30.5%; steady state 8.9–15.2%	Not performed	High
De Gelder 2011 (Prev Med)[22]; MISCAN	Netherlands, 0–100 year olds; biennial film or digital mammogram	(*a*) Dutch Cancer Registry, National Evaluation Team for Breast Cancer Screening 1990-2006	(*a*) No; (*b*) Yes; (*c*) Yes	Yes	Screen film, 7.2%; digital, 8.2%	Other sensitivity analyses performed.	High
Duffy 2005[23]	Sweden; 40–74 or 39–59 year olds; mammogram every 18, 24, or 33 months	All data: Swedish 2-County RCT (1977–84) and Gothenburg RCT (1982–87) (separate analyses)	(*a*) No; (*b*) No; (*c*) Yes	Yes	Swedish: 1st screen, 3.1% (0.1 to 10.9%); 2nd, 0.3% (0.1 to 1); 3rd, 0.3% (0.1 to 1) Gothenburg: 1st screen, 4.2% (0.0 to 28.8), 2nd, 0.3% (0.0 to 2.0), 3rd, 0.3% (0.0 to 2.0)	Not performed	High

Table 3 (continued)

Study; model(s)	Modelled population: country, ages; screening schedule	Data sources: (*a*) Incidence; (*b*) Mortality; (*c*) Other	(*a*) External validation?; (*b*) Includes competing mortality?; (*c*) Includes DCIS?	Reports outcome as % of screen detected cancers?	Magnitude of overdiagnosis (95% CI)	Sensitivity analyses varying mean sojourn time or lead time?	Overall risk of bias
Gunsoy 2012[24]	UK, 40–49 year olds; annual mammogram	(*a*) England/Wales Office of National Statistics, Age RCT Control Arm; (*b*) Office of National Statistics; (*c*) parameter estimation model: age RCT	(*a*) No; (*b*) Yes; (*c*) Yes	Yes	0.70%	Univariate; varied MST and sensitivity; 0.5 to 2.9%	Moderate
Martinez-Alonso 2010[25]	Spain; 25-84 year olds; biennial mammogram ages 50–69	(*a*) Girona Cancer Registry and IARC Registry	(*a*) No; (*b*) Yes; (*c*) No	No, reported as % excess of expected incidence	1935 birth cohort, 0.4% (–8.8 to 12.2%); 1940, 23.3% (9.1 to 43.4); 1945, 30.6% (12.7 to 57.6); 1950, 46.6% (22.7 to 85.2)	Univariate; varied MST from 1 to 5. 18.3 to 51.1%	Moderate
Olsen 2006[26]	Denmark; 50-69 year olds; biennial mammogram	(*a*) Danish Cancer Registry, Breast Cancer Cooperative Group, Central Population Registry	(*a*) No; (*b*) No; (*c*) Yes	Yes	1st screen, 7.8% (0.3 to 27.5%); 2nd screen, 0.5% (0.01 to 2.2)	Other sensitivity analyses performed	High
Seigneurin 2012[27]	France, 50-69 year olds; not specified	(*a*) French population-based study by Seigneurin 2009	(*a*) No; (*b*) No; (*c*) Yes	Yes	DCIS, 31.9% (2.9 to 62.3%); invasive cancer, 3.3% (0.7 to 6.5)	Univariate; varied MST; DCIS, 17.3 to 51.7%; invasive cancer, 0 to 8.9%	Moderate
Lung cancer (n=3)							
Duffy 2014[28]	UK; 55–74 or 50–75 year olds; annual and biennial	(*a*) NLST, UKLS; (*b*) NLST, SEER	(*a*) No; (*b*) Unclear; (*c*) N/A	Yes	11%	Univariate; varying MST; 0 to 18%	Moderate
Hazelton 2012[29]	US; Heavy smokers, <5 years asbestos exposure; low dose CT	(*a*, *b*) CARET (calibration); (*c*) calibrated model applied to NYU Biomarker and Moffitt Cancer Center Trials	(*a*) No; (*b*) Yes; (*c*) N/A	Yes	Men 14.1% (11.6 to 19.7%); women 35.2% (28.9 to 39.3)	Not performed	High
Pinsky 2004[30]	US; men aged 50–75 years, heavy smokers; annual CXR and sputum cytology	All data: Mayo Lung Screening Trial (prevalence screen and screening arm only)	(*a*) No; (*b*) Yes; (*c*) N/A	Yes	13–17%	Not performed	High
Colon cancer (n=1)							
Luo 2012[31]	US; 40, 50, or 60 year olds; 5 annual or 3 biennial FOBT	(*a*) Minnesota Colon Cancer Control study (1976–82); (*b*) SSA life tables	(*a*) No; (*b*) Yes; (*c*) N/A	Yes (reported for age 50)	Women 6.65% (2.56 to 20.49%); men 6.15% (1.92 to 44.69)	Not performed	High

SEER=Surveillance, Epidemiology, and End Results database; SSA=Social Security Administration; MST=mean sojourn time; MISCAN= Microsimulation Screening Analysis model; FHCRC=Fred Hutchinson Cancer Research Center; IARC=International Registry for Research on Cancer; DCIS=ductal carcinoma in situ; NLST=National Lung Screening Trial; UKLS= UK Lung Screening pilot trial; CARET=Carotene and Retinol Efficacy Trial; CT=computed tomography; CXR=chest x ray; FOBT=Fecal Occult Blood Test.

Table 4| Summary of evidence from 8 pathological and imaging studies quantifying overdiagnosis from cancer screening

Study; study period	Country; No of cancers; screening test	Overdiagnosis definition	Results	Magnitude of overdiagnosis (%)	Overall risk of bias
Lung cancer (n=6)					
Dominioni 2012[33]; 1997–2011	Italy; 21; CXR	VDT >300 days	1/21 cancers had VDT >300 days	"Minimal"	High
Lindell 2007[34]; 1999–2004	US; 61; CT	VDT >400 days	13/48 cancers had VDT >400 days	27	Moderate
Sobue 1992[35]; 1976–89	Japan; 42; CXR	Dying from a cause other than lung cancer in patients diagnosed with clinical stage 1 disease	20% of screen detected patients died from cause other than lung cancer	"Minimal"	High
Sone 2007[36]; 1996–98	Japan; 45; CT	Expected age of death (calculated from VDT) greater than average Japanese life expectancy	6/45 cases had expected death age greater than Japanese life expectancy	13.3	Moderate
Veronesi 2012[37]; 2004–10	Italy; 120; LDCT	VDT >400 days	31/120 cases had VDT >400 days	25.8 (95% CI 18.3 to 34.6)	Moderate
Yankelevitz 2003[38]; not provided	US; 87; CXR or sputum cytology	VDT >400 days	4/87 cases had VDT >400 days	5	High
Prostate cancer (n=2)					
Graif 2007[39]; 1989–2005	US; 2126; PSA	Tumor volume <0.5 cm^3, Gleason score <7, organ-confined disease in RRP specimen with clear surgical margins	4.5% met criteria for overdiagnosis compared with 27% meeting criteria for underdiagnosis	4.5	High
Pelzer 2008[40]; 1999–2006	Austria; 997 (806 screened, 161 not screened); PSA	Gleason score <7, pathological stage of pT2a, and negative surgical margins	16.8% of screened group and 7.9% of unscreened met overdiagnosis criteria	16.8%	High

CXR=chest x-ray; VDT=volume doubling time; CT=computed tomography; LDCT=low dose computed tomography; RRP=radical retropubic prostatectomy.

Table 5| Summary of evidence from 20 ecological and cohort studies quantifying overdiagnosis from cancer screening

Study; study design	Study population: country; ages; time period	Reference population	Adjustment for confounders	Management of lead time	Calculation of overdiagnosis	Magnitude of overdiagnosis (95% CI)	Risk of bias; time frame; analysis
Breast cancer (n=18)							
Bleyer 2012[41]; ecological	US; ≥40 year olds; 1976–2008	Pre-screening trend (1976–78)	HRT, baseline increasing incidence	Steady-state screening	(Excess cases)/ (observed cases) during screening	31%	Moderate; good; good
Coldman 2013[42]; ecological	British Columbia; 40–89 year olds; 2005–09	(*a*) Screening non-attenders; (*b*) pre-screening trend (1970–79)	Age, baseline increasing incidence	Including women aged up to 89 in incidence, with up to 10 years FU post-screening	(Excess cases)/ (observed cases during screening and post-screening)	(*a*) 17.3% (*b*) 6.7% (–21 to 30%)	Moderate; fair; poor
Duffy 2010[43]; cohort and ecological	Sweden; 50–60 year olds; 1977–98 UK; 47–73; 1989–2003	Sweden from Swedish 2-county RCT as control UK pre-screening trend (1974–89)	Swedish: unclear UK: baseline changes in incidence	Swedish: excluded prevalence screen UK: unclear	Based on complex calculation	Swedish: 12%* UK: 2.3 per 1000 screened for 20 years	Moderate; NA; poor
Falk 2013[44]; cohort	Norway; 50–69 year olds; 1995–2009	Screening program non-attenders	Age, county, calendar year	10 year FU post-screening	(Excess cases)/ (expected cases) during screening	19.4% (11.8 to 27.0%)	High; good; good
Hellquist 2012[45]; ecological	Sweden; 40-49 year olds; 1986–2005	Contemporary counties without screening	Differences in baseline incidence trends	Statistical adjustment	(excess cases)/ (expected cases) during screening	1% (–6 to 8%) (16% without lead time adjustment)	Moderate; NA; poor
Jorgensen 2009 (*BMJ*)[46]; ecological	UK; 50–64 year olds; 1993–99 CA; 50–69; 1995–2005 NSW; 50–69; 1996–2002 Sweden; 50–69; 1998–2006 Norway; 50–69; 2000–06	Pre-screening trend (UK 1971–84, CA 1970–78, NSW 1972–87, Sweden 1971–85, Norway 1980–94)	Baseline increasing incidence	Up to 7 years FU post-screening	(Excess cases)/ (expected cases) during screening	UK: 57% (53 to 61%) CA: 44% (25 to 65) NSW: 53% (44 to 63) Sweden: 46% (40 to 52) Norway: 52% (36 to 70) Meta-analysis: 52% (46 to 58)	Moderate; fair; good
Jorgensen 2009 (*BMC*)[47]; ecological	Denmark; 50–69 year olds; 1991–2003	Contemporary counties without screening	Age and differences in baseline incidence trends	Up to 10–12 years FU post-screening	(Excess cases)/ (expected cases) during screening	33%	Moderate; fair; good
Junod 2011[48]; ecological	France; 50–79 year olds; 1995–2005	Age-matched historical cohorts from 1980–90	HRT, alcohol intake, obesity	Unclear	(Excess cases)/; (expected cases) during screening	Ages 50–64: 76% (67 to 85%)† Ages 65–79: 23% (15 to 31)†	Moderate; fair; poor
Kalager 2012[49]; ecological	Norway; 50–79 year olds; 1996–2005	Contemporary counties without screening; historical cohorts in screening region without screening	Differences in baseline incidence trends	Including women up to age 79 in incidence with up to 10 years FU post-screening	(Excess cases)/ (observed cases) during screening period, including women up to age 79	Entire country: 25% (19 to 31%)† County with 10 years FU: 18% (11 to 24)†	Moderate; fair; poor
Morrell 2010[50]; ecological	NSW; 50–69 year olds; 1991–2001	Pre-screening trend (1972–90)	HRT, obesity, nulliparity	Statistical adjustment	(excess cases)/ (expected cases) during screening	30%†	Moderate; NA; poor
Njor 2013[51]; cohort	Denmark Copenhagen; 56–79 year olds; 1991–2009 Funen; 59–78; 1993–2009	Contemporary counties without screening; historical cohorts in screening region without screening	Differences in baseline incidence trends	Up to 8 years FU post-screening	(Excess cases)/ (expected cases) during screening and 8 years post-screening	Copenhagen: 6% (–10 to 25%) Funen: 1% (–7 to 10) Pooled: 2.3% (–3 to 8)	Moderate; fair; poor

Table 5 (continued)

Study; study design	Study population: country; ages; time period	Reference population	Adjustment for confounders	Management of lead time	Calculation of overdiagnosis	Magnitude of overdiagnosis (95% CI)	Risk of bias; time frame; analysis
Paci 2006[52]; cohort	Italy; 50–74 year olds; 1986–2006 (10 year period)	Pre-screening trend	Age	Statistical adjustment	(Excess cases)/ (expected cases) during screening	4.6% (2 to 7%) after adjustment for lead time 36.2% (34 to 39) before adjustment for lead time	Moderate; NA; poor
Peeters 1989[53]; ecological	Netherlands; ≥35 year olds; 1975–86	Contemporary county without screening	None	Did not	(Excess cases)/ (expected cases) during screening	11%	High; poor; poor
Puliti 2009[54]; cohort	Italy; 60–69 year olds; 1990–2005	Pre-screening trend (forced to 1.2% growth)	Age	5–10 years FU post-screening	(Excess cases)/ (expected cases) during screening and 5 years post-screening	1% (–5 to 7%)	Moderate; fair; poor
Puliti 2012[55]; cohort	Italy; 60–69 year olds; 1991–2007	Screening non-attenders	Age, marital status, area-level socioeconomic status	5–14 years FU post-screening	(Excess cases)/ (expected cases) during screening and 5–14 years post-screening	10% (–2 to 23%)	High; fair; poor
Svendsen 2006[56]; ecological	Denmark; 50–69 year olds; 1991–2001	Pre-screening trend (1979–90)	Age	Did not	Not calculated	"None"	Moderate; poor; poor
Zahl 2004[57]; ecological	Norway (N); 50–74 year olds; 1995–2000 Sweden (S); 50–70 year olds; 1986–2000	N: pre-screening period (1991) S: pre-screening trend (1971–85)	Age	Up to 4 (N) and 14 (S) years FU post-screening	(Excess cases)/ (expected cases) during screening	N: 56% (42 to 73%) increased incidence with no post-screening drop S: 45% (41 to 49%) increased incidence with 12% drop	Moderate; poor (N) fair (S); good
Zahl 2012[58]; ecological	Norway; 50–79 year olds; 1995–2009	Pre-screening trend (1991–95)	Age, county, population growth, baseline incidence trend	Up to 14 years FU post-screening	(Excess cases)/ (expected cases) during screening	Confirmed 50% incidence growth from Zahl 2004, with non-significant drop of 7% in women aged 70–74	Moderate; fair; good
Prostate cancer (n=2)							
Ciatto 2005[59]; cohort	Italy; 60–74 year olds; 1991–2000	Contemporary counties without screening	Age	7–9 years FU post-screening	(Excess cases)/ (expected cases) during screening and 9 years post-screening	66% (40 to 100%)	Moderate; fair; poor
Zappa 1998[60]; cohort and modeling	Italy; 60 or 65 year olds; not provided	Contemporary counties without screening	None	4 years FU post-screening	(Excess cases)/ (expected cases) during screening and 4 years post-screening	Age 60: 25% (19 to 32%) Age 65: 65% (58 to 73%)	Moderate; fair; poor

HRT=hormone replacement therapy; FU=follow up; CA=Canada; NSW=New South Wales, Australia.

*Unclear if Duffy 2010 estimates of overdiagnosis include ductal carcinoma in situ.

†Does not include ductal carcinoma in situ.

Table 6| Summary of evidence from 3 randomized controlled trial follow-up studies quantifying overdiagnosis from cancer screening

Study	Study population: country; age; time period	Post-study length of follow-up	Calculation of overdiagnosis	Magnitude of overdiagnosis (95% CI)	Risk of bias; time frame; analysis
Lung cancer (n=1)					
Patz 2013[62]	US high risk; 55–74 year olds; 2002–09	Up to 7 years	(Excess cases)/(screen detected cases)	18.5% (5.4 to 30.6%)	Low; fair; good
Breast cancer (n=2)					
Miller 2014[63]	Canada; 40–59 year olds; 1980–2005	22 years (average)	(Excess cases)/(screen detected cases)	22%	Low; good; good
Zackrisson 2006[61]	Sweden; 55–69 year olds; 1976–86	15 years	(Excess cases)/(control cases) during trial and 15 years follow-up	10% (1 to 18%)*	Low; good; poor†

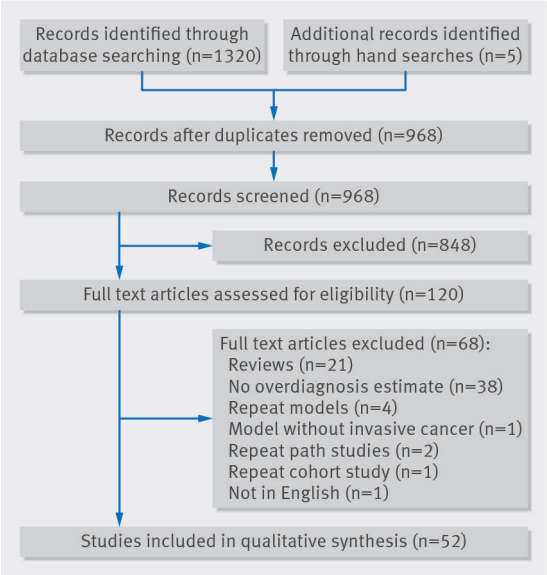
*Welch et al re-analysis[64] found overdiagnosis of 15% as percentage of cases diagnosed during screening period; overdiagnosis of 24% as percentage of screen detected cases.

†Welch et al re-analysis rated as good.

Table 7| Strengths and weaknesses of the main research methods used to quantify overdiagnosis from cancer screening

Research method	Strengths	Weaknesses
Follow-up of randomized controlled trials	Best able to minimize biases Directly answers question of interest	Substantial time and resource requirements Limited external validity Not useful for monitoring over time
Modeling	Can project through areas of uncertainty Not limited by time constraints Can evaluate multiple screening situations Can be used for monitoring over time	Validity of results depends on assumptions (poor directness) Needs constant updating of model constraints to reflect changing nature of cancer diagnosis and treatment Small changes in assumptions and model can lead to large changes in estimates Difficult to critically appraise (a "black box") Need long follow-up to determine overdiagnosis, yet uncertainty increases with time in models May give false sense of precision, insufficient attention to uncertainty
Pathological and imaging studies	Can be used for monitoring over time One of the simplest approaches	Validity of results depends on assumptions (poor directness) Not able to account for competing mortality Need to be sure all diagnosed cases are ascertained, and that data are collected in same way
Ecological and cohort studies	Directly answers question of interest Provides "real world" view of overdiagnosis Able to compare results from different settings Can be used for monitoring over time	Potential for confounding factors related to diagnosis, treatment, and health status between populations Moderate time requirements Needs investment in population registries, full and accurate ascertainment of all cases, and full and accurate collection of potential confounders

Figure



Flow diagram of study selection process