Quantifying and Monitoring Overdiagnosis in Cancer Screening: A Systematic Review of Methods

By

Jamie Carter

A Master's Paper submitted to the faculty of
the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements for
the degree of Master of Public Health in
the Public Health Leadership Program

Chapel Hill

2013

_____

Advisor

_____

Date

_____

Second Reader

_____

Date

Abstract

Introduction: To reduce overdiagnosis, we need accurate methods to quantify and monitor this phenomenon over time.

Aims: To systematically review the methods that have been used for measuring overdiagnosis from cancer screening; to evaluate the strengths and weaknesses of each method.

Methods: We searched PUBMED, EMBASE, and the Cochrane Library for primary research studies of any design that quantified overdiagnosis from cancer screening. We abstracted relevant data and appraised study design and methods using established criteria.

Results: 49 studies met inclusion criteria. We grouped studies into four methodologic categories and found strengths and weaknesses with all designs. (1) Follow-up of a well-designed RCT (n=1) is theoretically an ideal method but requires substantial time, may not be generalizable, and is not suitable for monitoring. (2) Pathologic/imaging studies (n=8) that draw conclusions about overdiagnosis by examining the range of biological or behavioral characteristics among cancers are simpler in design but assume that these characteristics are highly correlated with progression. (3) Modeling studies (n=19) can be done in a shorter time frame but require complex mathematical equations simulating the natural history of screen-detected cancer, which is the fundamental unknown question. (4) Ecologic studies (n=21) are limited by a lack of agreed-upon standards, by variable data quality, by inadequate follow-up time, and by the potential for population-level confounders. Some ecologic studies, however, have used excellent methods; several of these studies from different geographic areas may together provide the best overall estimate of overdiagnosis and are ideal for monitoring it over time.

Conclusions: Well-conducted ecologic studies in multiple settings should be used for quantifying and monitoring overdiagnosis in cancer screening programs. To support this work, we need internationally agreed-upon standards for ecologic studies and a multi-national team of unbiased researchers to perform analysis.

Introduction

A critical part of medical decision-making regarding whether or not to be screened for cancer is determining the balance of benefits and harms of screening programs. An appropriate cancer screening program is one for which the potential benefits outweigh the potential harms. While the exact point for which benefits outweigh harms or vice-versa is based on a complex judgment and may differ among individual patients, health care providers and policy-makers, this determination requires an accurate assessment of the magnitude of such benefits and harms.[1]

A harm of cancer screening that is increasingly being recognized and reported is overdiagnosis. Overdiagnosis refers to the diagnosis of a condition that would have otherwise not resulted in any symptoms or death during the patient's lifetime. In cancer, there are different scenarios that can lead to overdiagnosis. First, overdiagnosis can occur because of characteristics of the tumor and its potential for growth and regression. A tumor can either grow so slowly that the patient never would have developed symptoms or in some cases it can actually regress. Alternatively, overdiagnosis can result from the diagnosis of a cancer that progresses at a rate such that the patient ends up dying from another cause before the cancer becomes symptomatic. Consideration of competing mortality is important in cases of this second type of overdiagnosis, as the degree of medical comorbidity may contribute to the likelihood of death from another cause.[2]

Many consider overdiagnosis to be the most serious side effect of cancer screening, as overdiagnosis results in erroneous labeling of the patient with an incorrect life-long diagnosis. The resulting treatments and surveillance as well as the label of cancer itself cause physical and psychosocial harm to patients.[3] Physicians cannot distinguish between a patient with a cancer destined to cause harm and an overdiagnosed cancer at the time of diagnosis, so essentially all cancers are treated. A patient who is overdiagnosed with cancer cannot benefit from this diagnosis or treatment but instead can only be harmed.[4]

Potential for Overdiagnosis: Early Stage Disease Reservoir

The existence of overdiagnosis requires that there be a reservoir of non-progressive or slowly progressive disease in the population that can be detected with diagnostic tests, or a subset of cancers that are currently present that will later regress, or likely both. There are biologic mechanisms that enable cancers to be non-progressive, such as a cancer that outgrows its blood supply, that is recognized and successfully contained by the host immune system, or that simply lacks typical aggressive characteristics.[4] Autopsy studies have been used to investigate whether a reservoir of undetected cancer exists in patients who died from other causes.

A review of autopsy studies of women not known to have breast cancer during their lives found the median prevalence of breast cancer in seven studies among women of all ages was 1.3%, and the median prevalence of ductal carcinoma in situ (DCIS) was 8.9%. The prevalence of breast cancer or DCIS among women of screening age ranged from 7% to 39%. The authors observed that prevalence was correlated with the degree of rigor of the observation, as studies that used more slides to examine the tissues found a higher prevalence, which may explain the large variation in estimates.[5] Studies that follow patients who were initially misdiagnosed with benign lesions who were actually confirmed later to have DCIS also provide the some direct evidence about the progression of DCIS to invasive cancer. These studies report a range of 14-53% progression of DCIS to invasive cancer over a period of ten years or more.[6] These studies provide evidence that a reservoir of undetected non-progressive breast cancer and DCIS is likely present in some women, though the exact prevalence is uncertain.

Other autopsy studies provide further evidence of a disease reservoir in other cancer types as well. A retrospective study in Australia of over 13,000 autopsy reports of people who died from natural causes found 47 cases of incidental lung cancer, 86% of which were stage one.[7] Another study that compared autopsy evaluation for pulmonary nodules with detection of the same nodules on CT scan in the two months prior to death suggests this may be an underestimate. This study investigated whether claims

of low levels of overdiagnosis could be accurately made from autopsy studies, noting that prior studies had found very low rates of malignant lung cancers in autopsy subjects. Researchers found that 32% of patients with pulmonary nodules identified on a thoracic CT within 2 months of deaths did not have these same pulmonary nodules detected on autopsy, thus concluding that autopsy likely underestimates the prevalence of clinically insignificant lung cancer.[8] The age-independent frequency of histological prostate cancer in an autopsy study of 212 patients without a history of prostate cancer was 18.8% and ranged from 0% to 56% among different age groups.[9] In another study, the overall prevalence of prostate cancer was 37.3% in Russian men and 34.6% in Japanese men who died from causes other than prostate cancer. Prevalence was greater than 40% in men over 60 and greater than 60% in men over 80, demonstrating a large early-stage disease reservoir that increase with age.[10] Finally, a study of incidental thyroid cancer in Finland found a papillary thyroid carcinoma prevalence of 35.6% among autopsy patients. Many of the carcinoma specimens were smaller than the width of each slide (2.5mm) leading to the conclusion that the investigators had likely missed cases.[11]

Potential for Overdiagnosis: Cancer Regression

The ability of tumors to regress has been documented in cases reports throughout the medical literature for a variety of cancer types. In breast cancer, a 1999 review identified 32 cases of reported spontaneous remission of breast cancer[12], and another case report of has been published since the review.[13] A 2009 review found 76 reported cases of spontaneous regression of metastatic melanoma since 1866.[14] During a randomized controlled trial of interferon-beta for metastatic renal cell carcinoma, six patients of 99 in the control group achieved remission, including three whose cancer had completely regressed.[15] Spontaneous regression of hepatocellular carcinoma has been documented 75 times in the medical literature.[16] In a Japanese screening program, patients with localized, low-risk neuroblastoma were offered the option of observation instead of treatment. Eleven patients were identified for observation and all eleven of the tumors decreased in size.[17] In another study of neuroblastoma, 93 patients with low-risk disease were observed, and 44 of these tumors regressed with complete regression

in 17 patients by 20 months following diagnosis.[18] Other studies indicate the potential for lesions typically considered as "pre-cancerous" to regress, including polyps[19] and cervical low-grade squamous intraepithelial lesions.[20]

Another study demonstrated the potential for regression of breast cancer using different methods. This study compared breast cancer incidence in four counties of Norway in a group of women invited for three rounds of screening from 1996-2001 to an age-matched control group that was monitored from 1992-1997 and then offered one-time screening at the end of the observation period. Potential confounders including educational attainment, family income, reproductive history and screening attendance were all closely matched between the two groups. Before the control group was invited for screening, the cumulative incidence of breast cancer was significantly higher in the screened group (RR 1.57, 95% CI 1.44-1.70). However, contrary to what would be expected, the cumulative incidence of breast cancer remained elevated by 22% in the screened group compared to the control group after the control group underwent prevalence screening after the observation period (RR 1.22, 95% CI 1.16-1.30). This finding suggests that the natural history of some screen-detected breast cancers is to regress, as some of the cancers detected on repeated mammography would not have been detectable at the end of the 6-year screening period. The authors of this study also noted that the Canadian randomized controlled trial of breast cancer screening in women aged 40 to 49[21] also reported a 22% excess of incidence in the screened group that was not detected in the control group despite four years of screening at the end of the trial.[22]

Increasing Incidence of Early Stage Cancers

A pattern of increasing incidence of early stage cancer, especially that corresponds temporally with screening, is indicative of possible overdiagnosis. In the United States, several types of cancer have had dramatic increases in incidence over the past decades. Melanoma incidence increased 2.5-fold in patients over 65 between 1986 and 2001. After accounting for a possible increase in the true incidence of melanoma for alternative reasons, researchers found that the increase in incidence was associated with the

increase in biopsy rate over the same time frame and that the majority of new cases were confined to early stage disease.[23] Other studies found similar growth, but some also found an increase in the incidence in late-stage, thicker melanomas and thus concluded that the increase in incidence cannot be entirely attributed to increased detection of early-stage lesions.[24] Similarly, thyroid cancer incidence increased 2.6-fold from 1973 to 2006. The greatest growth, of 441%, was seen in the smallest subset of papillary thyroid cancer tumors less than 1centimeter. However, larger tumors also increased leading some researchers to argue that the increasing incidence of thyroid cancer is not entirely due to an increase in early lesions.[25] Some types of overdiagnosis may not be due to screening, but actually to the increased use of imaging tests such as abdominal CT scans. Incidence rates of localized renal cell carcinoma more than doubled from 1988 to 2006, from 3.8 per 100,000 person-years to 8.2 per 100,000 person-years.[26]

Some types of cancer are currently decreasing in incidence despite having increased greatly over the past few decades. Prostate cancer incidence has dropped slightly in recent years after increasing dramatically from 1986 to 1992 with the introduction of PSA screening. However, the relative incidence rates of prostate cancer in 2005 relative to 1986 varied widely by age group, with relative rates of 3.64 in men ages 50 to 59 and 7.23 in men younger than 50, so in some demographics incidence is still on the rise.[27] Breast cancer incidence but sustained a sharp increase between 1980 and 1987 corresponding with increasing use of screening mammography. During this period, incidence of breast cancers smaller than one centimeter quadrupled, from 9 per 100,000 to 36 per 100,000, and incidence of DCIS more than tripled from 4 per 100,000 to 15 per 100,000.[28] Breast cancer incidence then decreased in the 1990's and early 2000's but since 2006 has been again increasing.[29] Another recent study spanning a longer time frame found that the incidence of early stage breast cancer more than doubled from 112 cases per 100,000 to 234 per 100,000 from 1976 to 2008.[30]

Patterns of cancer incidence and mortality in the United States demonstrate that overdiagnosis is occurring. A cancer with both increasing incidence and increasing mortality represents a true increase in cases of that type of cancer. However, a cancer with increasing incidence but mortality that remains unchanged over the same time period likely indicates overdiagnosis.[4] This pattern has been seen over the

past 30 years in the United States with melanoma, breast cancer, prostate cancer, kidney cancer, and thyroid cancer. An alternative explanation for this pattern is that improvements in diagnosis and treatment of the cancer are causing an improvement in mortality that exactly counteracts the increase in incidence. However, this explanation involves more assumptions than are required for the explanation involving overdiagnosis and as such is less likely.[4]

Addressing Overdiagnosis

Experts have suggested a variety of strategies for addressing the problem of overdiagnosis in cancer. Several have proposed raising the threshold for labeling a test result or image abnormal and consider monitoring lesions over time to assess growth instead of jumping straight to a bipsy.[4,31,32] Others advise replacing the term cancer with another term that suggests the more benign nature of much of the spectrum of cancer to represent early-stage lesions.[32] Eventually, we may have biomarkers that can distinguish between indolent and more aggressive cancers such that therapies can be targeted towards those cancers most likely to be fatal.

Education about overdiagnosis will also be crucial for minimizing its harms[4,31], and this includes education for medical students, residents, current physicians, and the public. Medical practitioners currently receive mixed messages regarding overdiagnosis, however, as its coverage and emphasis in the medical literature, as well as estimates of its magnitude, have varied widely. A 2007 study found unequal attention given to benefits and harms in articles on screening mammography, which was related to the professional affiliation of the author. Benefits were mentioned more often than harms (96% versus 62%), with 38% of articles mentioning only benefits. Overdiagnosis was mentioned in only 40% of articles on screening mammography and was more likely to be downplayed or rejected by authors that worked specifically in screening (40%) than by authors in screening-affiliated specialties (like breast cancer surgery or radiology) who were not working directly with screening (17%) or by authors in an unrelated specialty (7%).[33] Within the medical community, overdiagnosis is a polarizing issue, and the lack of clarity in communication about overdiagnosis or understanding of its magnitude is a barrier to effectively addressing it.

Most importantly, we need to develop an understanding of patient and societal values regarding overdiagnosis, which likely will follow an effort to educate patients so they are able to make informed decisions. Currently, knowledge of overdiagnosis among the public appears to be minimal. In a cross sectional study (Schwartz et al) of US women's attitudes regarding potential consequences of breast cancer screening, no women identified the detection or treatment of a non-progressive breast cancer as a potential harm of screening, and only 7% were aware of the existence of non-progressive breast cancer.[34] A recent qualitative study[35] of women's values regarding overdiagnosis reported that women had minimal awareness of overdiagnosis prior to participation in the study, and some expressed surprise at being informed of it. Despite low awareness, women from various socioeconomic and educational backgrounds could understand information presented on overdiagnosis and valued this information in making their screening decisions.[35] Other studies confirm that the public values information on overdiagnosis as a harm of screening. In the Schwartz et al study, 60% of women wanted to factor information on non-progressive cancers into their decisions for pursuing mammography.[34] A 2013 study of men's preferences for prostate cancer screening found that men considered risk of unnecessary treatment and biopsy as a factor in decision-making.[36]

In the preliminary Hersch et al study, the magnitude of overdiagnosis also appeared to be valuable to women making decisions about screening mammography. When the estimate of overdiagnosis was 50%, some women expressed that they would much more carefully consider the decision to be screened, with some women expressing that they would likely forego screening altogether, be less concerned about achieving a rigid screening interval, or consider delaying screening until a later age. At lower estimates, women were less concerned about overdiagnosis. With overdiagnosis of 1 to 10%, some women expressed that this was minimal and would not at all affect their intentions for screening. With overdiagnosis of 30%, women acknowledged concern with this large number and effect on people's lives, but many indicated that they would probably still continue to undergo screening.[35]

Both the existence and magnitude of overdiagnosis are important to patients, and its magnitude can be critical for decision-making on a population level as well. In evaluating screening programs, experts argue that benefits and harms should be weighed using an outcomes table which depicts all the possible outcomes of a screening test and their relative likelihoods among an eligible cohort[1]. Critical policy decisions regarding provision of screening are made based on an often delicate balance of benefits and harms, and a change in the magnitude of overdiagnosis can shift this balance one way or another. Thus, accurate measurement of overdiagnosis is important for both individual and population-level decision-making.

Unfortunately, because it is impossible to distinguish at the time of diagnosis between an overdiagnosed cancer and one that would have become clinically meaningful, the measurement of overdiagnosis is not straight-forward. Researchers have used various methods to indirectly quantify overdiagnosis resulting from cancer screening, but the magnitudes of such estimates have varied widely. This systematic review attempts to identify and evaluate the methods that have been used for measuring overdiagnosis resulting from cancer screening and analyze the advantages and disadvantages of each method. We will also determine which methods for measuring overdiagnosis are most suitable for monitoring it over time, as monitoring will be key to preventing overdiagnosis and tracking our progress with this endeavor. A better understanding of methods for measuring overdiagnosis will aid future researchers in designing studies to accurately measure this phenomenon. In turn, more reliable and accurate measurements of overdiagnosis resulting from cancer screening will enable better representation of the benefits and harms of such tests, ultimately providing the tools for patient-centered medical decision-making. Finally, as we develop interventions to try to decrease overdiagnosis, we need to be able to perform surveillance and monitor overdiagnosis over time, which requires accurate and reliable methods for measurement.

Methods

Key Questions

This review aims to answer the following key questions:

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

Key Question 2: What are the advantages and disadvantages of each method of measuring overdiagnosis?

Key Question 3: What methods would be most suitable for monitoring overdiagnosis over time?

Eligibility Criteria

We designed inclusion and exclusion criteria for the review to include all studies that have attempted to measure, quantify, or estimate the amount of overdiagnosis resulting from a cancer screening test in an asymptomatic adult population. We used modified PICOTS criteria (see Table 1) whereby the population of interest was studies that measure overdiagnosis. 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, as these cancer types likely have the highest potential for overdiagnosis due to a high rate of cases diagnosed. These cancer types are prostate, breast, lung, colorectal, melanoma, bladder, renal, thyroid, and uterine cancer.[37] Studies reporting overdiagnosis not relating to cancer were excluded, as were studies that addressed the potential for overdiagnosis but did not report a quantity. For example, autopsy studies reporting on the prevalence of low grade or early stage cancer in patients who died of other causes are important for demonstrating the principles of overdiagnosis but were ineligible for this review. Studies that examined biologic or behavioral characteristics of detected tumors and then drew conclusions about an amount of overdiagnosis were included. Outcomes of interest included the estimated magnitude from each type of measurement, as well as the way the measurement was calculated.

Studies not providing a numerical estimate for the magnitude of overdiagnosis were excluded. Studies from any setting and time frame were included. Because measuring overdiagnosis is possible through a variety of study designs, and because the comparison of methods was the key intention of this review, a wide range of study designs were eligible including randomized controlled trials, prospective or retrospective cohort studies, ecologic studies, and case control studies. When multiple publications that measured overdiagnosis with modeling used the same model and populations to determine the overdiagnosis estimate, only the most recent publication was included. Non-systematic and systematic reviews, case reports and case series were excluded. Only studies in English were included.

Search Strategy

We conducted a systematic search of PUBMED, EMBASE, and the Cochrane library on February 22, 2013. A research librarian helped with the development of the search terms and the adaptation of the terms to the different databases. The search terms used to search PubMed were as follows: "(cancer*[tw] OR neoplasms[MeSH]) AND (Screening*[tw] OR early diagnos*[tw] OR early detect*[tw]) AND (overdiagnos*[tw] OR over diagnos*[tw] OR overdetect*[tw] OR over detect*[tw])". We placed no date or language limitations on studies to avoid missing studies that had not yet been indexed. We also hand-searched reference lists of included systematic reviews and other narrative reviews identified during the initial search for additional relevant studies.

Table 1: Modified PICOTS Criteria for Study Eligibility

	Include	Exclude
Population	Studies that attempt to measure, estimate, or quantify the amount of overdiagnosis resulting from a cancer screening test in an asymptomatic population Cancer types eligible for inclusion: prostate, breast, lung, colon, melanoma, bladder, renal, thyroid, uterine Studies that look at biologic or behavioral characteristics of tumors (i.e., grade, doubling time) and draw conclusions about an amount of overdiagnosis	Studies reporting overdiagnosis not related to cancer screening or related to a type of cancer screening not listed in the inclusion criteria Studies addressing the potential for overdiagnosis but that do not draw conclusions regarding an amount of overdiagnosis (for example, studies that report on prevalence of early stage cancer detected at autopsy) Studies investigating different thresholds for tumor markers that comment on implications for overdiagnosis Studies performed in a symptomatic population
Intervention	Method for measuring, estimating, or quantifying overdiagnosis	
Outcome	Magnitude of overdiagnosis	Studies that do not report a magnitude of overdiagnosis
Time Frame	Studies performed over any time frame	
Setting	Any setting	
Study Design	Randomized controlled trials, prospective or retrospective cohort studies, ecologic studies, case control studies, modeling studies	Non-systematic and systematic reviews, case reports, case series

Study Selection

Titles and abstracts of all studies identified by the search were reviewed independently by two reviewers for inclusion based on the criteria discussed and listed in Table 1. Any article that was identified for inclusion by either reviewer or for which there was not enough information available in the abstract had its full text reviewed for the same eligibility criteria. Two reviewers independently determined whether the identified studies could be verified for inclusion by analyzing the full texts. 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 on a spreadsheet that was standardized for each type of study. These data were then verified by a second reviewer, and discrepancies were resolved by consensus. For all study types, information on study design, study population, time period, screening test, screening schedule, threshold for labeling a result abnormal, length of follow-up, magnitude of overdiagnosis, and conclusions were extracted. If studies used modeling to measure overdiagnosis, information on the name and type of model, data sources, and sensitivity analyses were extracted. For cohort and ecologic studies, information on the reference population and statistical adjustment for confounders and lead time were noted. When randomized trials were followed-up to measure overdiagnosis, the type of statistical analysis performed and the baseline characteristics of the two groups were extracted. Some studies used information on pathologic or imaging characteristics of cancers to draw conclusions about overdiagnosis, and the details of these characteristics were documented for these studies.

All studies were assessed for reporting of the preferred outcome, which was overdiagnosis defined as excess of cancer cases diagnosed during the screening period divided by total number of cases detected by screening. This method is the most appropriate way to calculate overdiagnosis, as overdiagnosis is an outcome of screening and can only occur in asymptomatic patients diagnosed by screening. Using a different denominator, such as the total number of cancer cases detected including interval cancers diagnosed by the presence of symptoms, dilutes the overdiagnosis measurement and makes its implications less clear. The timeframe over which overdiagnosis is measured can also affect its magnitude, and we made note of this timeframe in our data extraction. We extracted the preferred outcome from all studies that reported it, and we calculated it using data provided in the paper when it

was available. Otherwise, we extracted the overdiagnosis measurement that was reported but made note of the way the calculation was performed.

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, which were cohort and ecologic studies, pathologic and imaging studies, follow-up of randomized controlled trials, and modeling studies. We rated each individual study for risk of bias using ratings of high, moderate or low. Two reviewers independently rated the risk of bias for each study, and we resolved discrepancies by consensus. The four sets of complete criteria used for risk of bias assessment can be found in Appendix A.

We adapted the criteria for cohort and ecologic studies from quality criteria used in a recent systematic review of observational studies of breast cancer screening.[38] These adapted criteria include the potential for selection bias, measurement bias, and confounding with a focus on the use of comparable groups with regards to potential confounders. Risk of bias criteria for randomized controlled trial follow-up studies were adapted from standard criteria used by the USPSTF.[39] Pathologic and imaging studies typically did not have a control group, so the risk of bias assessment focused on the validity and reliability of the measurements performed. The appropriateness of the time frame was evaluated for all study types, as the measurement of overdiagnosis requires consideration of the lead time of the cancers studied, and measuring overdiagnosis without adequate time for follow-up can affect its magnitude.

We developed a new set of criteria for evaluating modeling studies for the purpose of this review. Authors of modeling studies were expected to discuss the probability for biases in the data used in the model and to choose data that had low probability for biases. Assumptions were to be clearly stated, ideally in a table of assumptions, and all assumptions made in the model needed to be backed with evidence that was ideally identified and assessed for quality in a systematic review. Sensitivity analyses should have been performed for any uncertain variables, ideally with probabilistic multivariate sensitivity

analyses rather than single-variable analyses. Finally, models should also have been validated using population data different from the data used to calibrate the model.

Strength of Evidence Assessment

To evaluate the strength of evidence, we developed a set of criteria from other criteria used by the USPSTF[39] and by the GRADE working group.[40] Each individual study was evaluated for risk of bias, directness, external validity, and precision. We rated each study as high, moderate or low for risk of bias and good, fair, poor, or "cannot determine" for the other criteria. Among ecologic and cohort studies, each study was also reviewed for the appropriateness of the analysis and its ability to provide an unbiased overdiagnosis estimate, which will be discussed further in the results section. Two reviewers individually determined ratings for each of these criteria, and we resolved discrepancies by consensus.

The GRADE working group defined directness as the extent to which the intervention relates the evidence to health outcomes.[40] In this review, we evaluated the extent to which the evidence links the screening test directly to health outcomes without making certain assumptions, including assumptions regarding the progression of a screen-detected cancer to cancer-related morbidity and mortality, and assumptions regarding the association of pathologic or behavioral characteristics of a cancer with cancer progression, morbidity and mortality. A study with good directness requires minimal assumptions to draw conclusions about the magnitude of overdiagnosis and directly measures excess cases of cancer.

We adapted criteria for evaluating external validity of individual studies from the USPSTF procedure manual.[39] Studies were assessed for their relevance to a general US adult population. We considered the extent to which the study population was similar to the general US population in factors associated with cancer incidence and in the quality of medical care and risks for competing mortality. We also assessed the similarity of the screening situation in each study to the way screening is performed in the US, including the expertise of the radiographers, the quality of the screening facilities, and the

threshold used to label a result abnormal. All of these factors affect the way cancer is diagnosed and thus can affect the degree of overdiagnosis present.

We combined the ratings for risk of bias, analysis, directness, external validity, and precision with an evaluation of the consistency of the study results to determine of the strength of evidence for the overall body of evidence. We evaluated strength of evidence for each study design and cancer type. The risk of bias, directness, external validity, and precision of each of the individual studies was used to assess the aggregate risk of bias, directness, external validity, and precision for the body of evidence. We used the GRADE working group's definition to evaluate consistency by looking at the degree to which the overdiagnosis measurement from all the included studies of that cancer type and study design had the same magnitude. The complete list of criteria used to evaluate strength of evidence can be found in Appendix B.

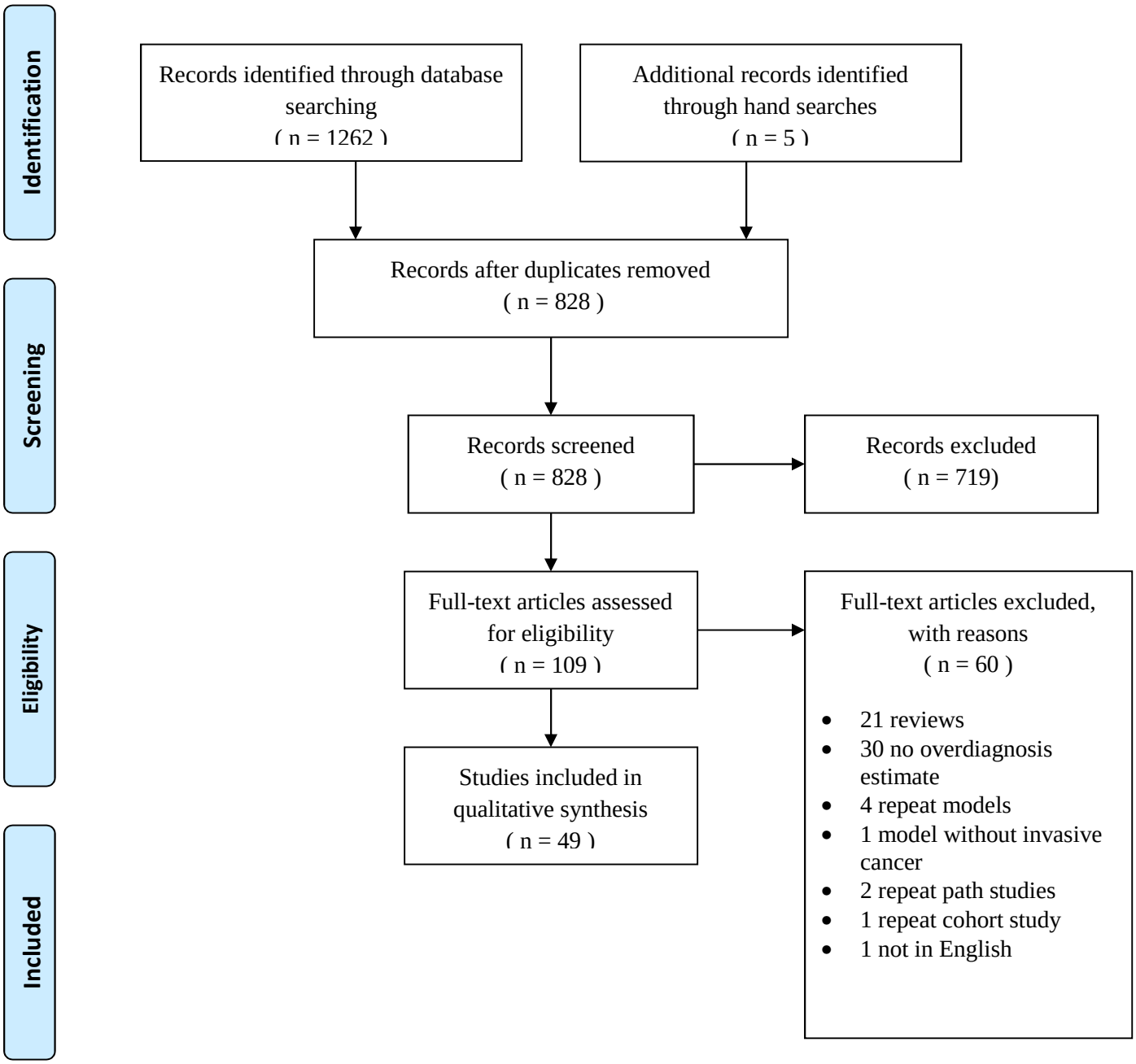
Data Synthesis and Analysis

We performed qualitative data synthesis, organizing the results by study design and cancer type. We did not attempt to perform quantitative analysis because of the heterogeneity of the study designs, populations, and results. We identified strengths and weaknesses of each study design used to measure overdiagnosis, based on the criteria used to evaluate the strength of the body of evidence. We did not attempt to assess publication bias.

Results

Our search, performed on February 22, 2013, yielded a total of 1262 studies from PubMed, EMBASE, and the Cochrane library. After duplicates were removed, 823 studies remained, and five additional studies were later identified by hand searching reference lists of relevant reviews. We reviewed 828 abstracts and identified 109 studies that were eligible for full-text review. During the full-text review process, 21 studies were excluded for wrong study design as they were non-systematic and systematic reviews, 30 were excluded for not providing an estimate of the magnitude of overdiagnosis,

Figure 1: PRISMA Flow Diagram of Study Selection Process



and one was excluded for not being in English. All studies that used modeling to measure overdiagnosis were again reviewed to see if there were any studies that used the same model and population as another study. We combined four of these modeling studies because they met these criteria, and in each case the most recent publication was included. We excluded one additional modeling study for not including invasive cancer. These modeling studies that provided an estimate of overdiagnosis magnitude but were excluded for the reasons discussed are listed in Appendix C. Three studies of prostate cancer overdiagnosis by the same group of authors[41-43] used essentially identical methods and the same population to study overdiagnosis defined based on pathologic criteria. Of these three studies, only the most recent publication[41] was included. Finally, a cohort study of breast cancer overdiagnosis in Florence[44] was excluded because two other more recent studies by overlapping authors[45,46] were available. A total of 49 studies were included for qualitative synthesis. Figure 1 shows the flow diagram of the study selection process based on PRISMA recommendations.[47]

The included studies fell into four categories. Some studies used mathematical and statistical models to measure overdiagnosis. Other studies examined pathologic or imaging characteristics of tumors and used information about these characteristics to draw conclusions about overdiagnosis. A large group of studies analyzed cancer incidence in either two populations or two cohorts to measure overdiagnosis. Finally, one study followed participants of a randomized-controlled trial for fifteen years after the trial ended. We categorized studies as modeling studies, pathologic/imaging studies, ecologic or cohort studies, and follow-up of randomized controlled trials.

Characteristics of Included Studies: Modeling Studies

We included nineteen modeling studies in this review, including nine models of prostate cancer, seven models of breast cancer, two models of lung cancer and one model of colon cancer overdiagnosis. The characteristics of the included modeling studies are listed in Table 2, and full details are provided in Appendix Table 2. These studies modeled a variety of screening situations and schedules, with one

study[48] modeling 32 different hypothetical prostate screening schedules, for example. Not all studies modeled both the non-progressive disease and competing mortality components of overdiagnosis, as three breast cancer studies[49,50,51] did not include overdiagnosis resulting from competing mortality in their models. One breast cancer model did not include DCIS.[52] It is important to note that these studies provide an incomplete look at overdiagnosis and almost certainly underestimate its magnitude to some degree.

In contrast to the other study types, most modeling studies reported the preferred outcome, which was the percentage of excess cancers divided by the total number of screen-detected cancers. However, a few studies instead reported an overdiagnosis measure that dilutes the estimate, such as lifetime risk of overdiagnosis of prostate cancer reported by Gulati and colleagues.[48] There was considerable variability within and among modeling studies in estimates of overdiagnosis, with estimates of overdiagnosis as a percentage of screen-detected prostate cancers ranging from 8.48% to 67% depending on model parameters and screening and population details. Many studies provided multiple estimates depending on these various parameters to demonstrate the variability of the results. Among studies of breast cancer, estimates of overdiagnosis as a percentage of screen-detected cancers ranged from 0.3% to 67.4%, though these estimates were provided for specific time points such as the prevalence screen or the second screen in a series of screens, which makes them difficult to interpret.

Table 2: Summary Evidence Table of Modeling Studies

Study; Cancer type; Model(s)	Modeled population: Country, ages; Screening schedule	Data Sources a. Incidence b. Mortality c. Other	a. External Validation? b. Includes Competing Mortality? c. Includes DCIS?	Reports Preferred Outcome?	Magnitude of Overdiagnosis (95% CI)	Sensitivity Analyses varying mean sojourn time or lead time?	Overall Risk of Bias
Davidov 2004[53]; Prostate	US; 50 to 60, 70 or 80 at 5-year intervals	a. SEER1993-7 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[54]; Prostate; MISCAN, FHCRC, UMichigan	US, 54-80; 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[48]; Prostate; FHCRC	US, 40; 32 screening schedules simulated	a. SEER 1975-2000 b. US life tables	a. No b. Yes c. n/a	No. Reports lifetime risk of over-diagnosis*	*1.8 to 6%**	Other sensitivity analyses performed.	Moderate
Heijnsdijk 2009[55]; Prostate; MISCAN	Europe; 55-70 every 1 or 2 years or 55-75 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[56]; Prostate	Quebec, 50-85; Annual PSA 50-70	a. see appendix table b. Quebec Ministry of Health	a. No b. Yes c. n/a	Yes	84%	Other sensitivity analyses performed.	High
Pashayan 2009[57]; Prostate	UK; Single PSA	a. Eastern Cancer Registry, ProtecT study, UK Office of National Statistics b. UK Office of National Stats	a. No b. Yes c. n/a	Yes	50-4: 10% (7-11) 55-9: 15% (12-15) 60-4: 23% (20-24) 65-9: 31% (26-32)	Not performed	High
Telesca 2008[58]; Prostate	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

Study; Cancer type; Model(s)	Modeled population: Country, ages; Screening schedule	Data Sources a. Incidence b. Mortality c. Other	a. External Validation? b. Includes Competing Mortality? c. Includes DCIS?	Reports Preferred Outcome?	Magnitude of Overdiagnosis (95% CI)	Sensitivity Analyses varying mean sojourn time or lead time?	Overall Risk of Bias
Tsodikov 2006[59]; Prostate	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[60]; Prostate	Finland, 55, 59, 63, 67; 3 PSA tests every 4 years until 71	a. Finnish Prostate Cancer Screening Trial, Finnish Cancer Registry b. Statistics Finland	a. No b. Yes c. n/a	No*	*3.4% (2.4-5.7) risk of overdetection during study period**	Not performed	High
De Gelder 2011 (Epi Rev)[61]; Breast; MISCAN	Netherlands, 0-100; Biennial mammo 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)[62]; Breast; MISCAN	Netherlands, 0-100; Biennial screen film or digital mammo	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[49]; Breast	Sweden, 40-74/ 39-59; Mammo 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% (0.1-10.9), 2nd: 0.3% (0.1-1), 3rd: 0.3% (0.1-1). Gothenburg: 1st: 4.2% (0.0-28.8), 2nd: 0.3% (0.0-2.0), 3rd: 0.3% (0.0-2.0)	Not performed	High
Gunsoy 2012[63]; Breast	UK, 40-49; Annual mammo	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

Study; Cancer type; Model(s)	Modeled population: Country, ages; Screening schedule	Data Sources a. Incidence b. Mortality c. Other	a. External Validation? b. Includes Competing Mortality? c. Includes DCIS?	Reports Preferred Outcome?	Magnitude of Overdiagnosis (95% CI)	Sensitivity Analyses varying mean sojourn time or lead time?	Overall Risk of Bias
Martinez-Alonso 2010[52]; Breast	Spain, 25-84; Biennial mammo 50-69	a. Girona Cancer Registry and IARC Registry	a. No b. Yes **c. No**	No. Reported as percent excess of expected incidence. *	*1935 birth cohort: 0.4% (-8.8 to 12.2)* *1940: 23.3% (9.1-43.4%)* *1945: 30.6% (12.7-57.6%)* *1950: 46.6% (22.7-85.2%)**	Univariate. Varied MST from 1 to 5. 18.3 to 51.1%	Moderate
Olsen 2006[50]; Breast	Denmark, 50-69; Biennial mammo 50-69	a. Danish Cancer Registry, Breast Cancer Cooperative Group, Central Population Registry	a. No **b. No** c. Yes	Yes	1st screen 7.8% (0.3-27.5) 2nd screen: 0.5% (0.01-2.2%)	Other sensitivity analyses performed	High
Seigneurin 2012[51]; Breast	France, 50-69; Not specified	a. French population-based study by Seigneurin 2009	a. No **b. No** c. Yes	Yes	DCIS: 31.9% (2.9-62.3) Invasive cancer: 3.3% (0.7-6.5)	Univariate. Varied MST. DCIS: 17.3-51.7% Invasive: 0-8.9%	Moderate
Hazelton 2012[64]; Lung	Heavy smokers, <5yrs asbestos exposure; Low dose CT	a,b. CARET (calibration) c. Calibrated model applied to NYU Biomarker Center Trial and Moffitt Cancer Center Trial	a. No b. Yes c. n/a	Yes	Men: 14.1% (11.6-19.7) Women: 35.2% (28.9-39.3)	Not performed	High
Pinsky 2004[65]; Lung	Men 50-75, heavy smokers; Annual CXR and sputum cytology 50-75	All data: Mayo Lung Screening Trial (prevalence screen and screening arm only)	a. No b. Yes c. n/a	Yes	13 to 17%	Not performed	High
Luo 2012[66]; Colon	Cohort age 40, 50 or 60; 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)	Females: 6.65% (2.56-20.49) Males: 6.15% (1.92-44.69%)	Not performed	High

Abbreviations: MST, mean sojourn time; FHCRC, Fred Hutchinson Cancer Research Center; mammo, mammography; IARC, International Registry for Research on Cancer; CARET, Carotene and Retinol Efficacy Trial; MLT, Mayo Lung Trial

Risk of Bias: Modeling Studies

For this review, we developed new criteria for evaluating risk of bias of modeling studies. An ideal modeling study would clearly state its assumptions and data sources in a table, and all assumptions would be supported by evidence identified and quality rated in a systematic review. Only one modeling study[48] provided a table of assumptions, and none were supported by systematically-reviewed evidence. Instead, most studies picked different data inputs from a variety of sources without justification for the use of such diverse sources. This raises the risk of manipulation of the model to achieve a desired output and thus the risk for bias. An alternative to performing a systematic review to inform the model with high quality evidence would be to use all data inputs from a well-done randomized controlled trial. One breast cancer modeling study[49] did use all data from the Swedish 2-County and Gothenburg randomized controlled trials and was given credit for its use of consistent unbiased data sources, but this study had a fatal flaw in that it did not perform sensitivity analyses. Three other studies[55,63,66] used some data from a randomized controlled trial but also pulled data from other sources.

We also attempted to rate modeling studies on the probability for biases in the data used in the model. We expected that authors would choose data sources in an attempt to minimize bias as well as discuss the potential biases in their choices in an effort to convince the reader of the validity of their results. However, none of the modeling studies provided any information or discussion on potential biases in their data.

A major component of the risk of bias assessment for modeling studies was the performance of sensitivity analyses. The ideal study would perform probabilistic multivariate sensitivity analyses for key uncertain parameters including mean sojourn time or lead time. Only four studies specifically varied mean sojourn time in univariate sensitivity analyses[51-53,63] and one other study varied rates of disease onset, metastasis and clinical detection,[48] which is likely equivalent. 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.

Several studies used a data set to calibrate the model and determine uncertain model parameters, such as transition probabilities between different states in the model, and then "validated" the model by fitting it to the same original data set. A study which truly externally validated its model would use one data set to calibrate the model and determine parameters and then externally validate it to another data set in a different population. Performing this 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 more than just the modeled population. None of the modeling studies included in this review performed external validation of their models and thus did not achieve this degree of credibility.

We rated the majority of modeling studies as having a high risk of bias because they had a fatal flaw of not performing sensitivity analyses for key uncertain variables, in addition to the other potential biases already discussed. The five studies that performed univariate sensitivity analyses as described above were rated as having moderate risk of bias, as none performed external validation or informed the model with systematically reviewed evidence or data from a single randomized trial.

Strength of Evidence: Modeling Studies

We assessed strength of evidence for modeling studies grouped by cancer type, with other criteria in addition to risk of bias being directness, external validity, precision and consistency. Our ratings for each study are available in Appendix Table 3. Directness for all modeling studies was rated as poor, because by nature the models used to draw conclusions about overdiagnosis require assumptions about progression of cancer from early, preclinical stages to later stages. The nature of this progression is fundamental to the question of overdiagnosis and its magnitude, so in many ways it is inappropriate that models attempting to answer such questions would require such assumptions. Ratings for external validity for prostate cancer modeling studies were generally good, as these studies tended to be based on US data and based on typical US screening patterns. In contrast, all the breast cancer models were based on European populations and screening situations, which differ from the US in having a much lower

diagnostic rate of DCIS, and thus were rated as fair for external validity. We were often unable to determine the precision of the overdiagnosis estimates because confidence intervals were not provided in many cases, but in most other cases precision was fair to poor. Finally, we rated consistency for both breast and prostate cancer modeling studies as poor. Strength of evidence was rated as low for breast, prostate, lung and colon cancer modeling studies.

Characteristics of Included Studies: Pathologic and Imaging Studies

We included eight studies that drew conclusions about overdiagnosis based on a pathologic or imaging characteristics, six of lung cancer overdiagnosis and two of prostate cancer overdiagnosis. Table 3 highlights the characteristics of these included studies, and full details are available in Appendix Table 4. The lung cancer studies were typically small studies that retrospectively looked at volume doubling time of patients diagnosed with lung cancer by screening chest x-ray or CT scan. These studies included a total of 376 cancers. The definition for overdiagnosis was typically set at a volume doubling time of 400 days, though one study used 300 days[67], and one study used information on volume doubling time to calculate the patient's expected time of death which was then compared with the typical life expectancy in Japan to determine if the cancer was overdiagnosed.[68] Another study with unique methods among those of lung cancer followed patients diagnosed with screen-detected clinical stage 1 lung cancer who did not undergo surgical treatment and defined overdiagnosis as death from a cause other than lung cancer.[69] Estimates of lung cancer overdiagnosis varied from "minimal" to 27% in these studies.

The two pathologic/imaging prostate cancer studies involved a total of 3093 patients and used similar definitions of overdiagnosis based on Gleason score, negative surgical margins and other criteria.[41,70] Patients in both studies were undergoing radical retro-pubic prostatectomy for prostate cancer detected in various screening situations. Estimates of overdiagnosis were 4.5%[70] and 16.8%[41].

Table 3: Summary Evidence Table of Pathologic and Imaging Studies

Study Cancer Type Study Period	Country # of cancers Screening test	Overdiagnosis Definition	Results	Magnitude of Overdiagnosis (95% CI)	Overall Risk of Bias
Dominioni 2012[67] Lung 1997-2011	Italy 21 CXR	VDT> 300 days	1/21 cancers had VDT > 300 days	"minimal"	Moderate
Lindell 2007[71] Lung 1999-2004	US 61 CT	VDT>400 days	13/48 cancers had VDT>400 days	27%	Moderate
Sobue 1992[69] Lung 1976-1989	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[68] Lung 1996-1998	Japan 45 CT	Expected age of death (calculated from VDT) greater than average Japanese life expectancy	6 of 45 cases had expected death age greater than Japan life expectancy	13.3%	High
Veronesi 2012[72] Lung 2004-2010	Italy 120 LDCT	VDT>400 days	31/120 cases had VDT> 400 days	25.8% (18.3-34.6)	Moderate
Yankelevitz 2003[73] Lung Not provided	US 87 CXR/sputum cytology	VDT> 400 days	4/87 cases had VDT> 400 days	5%	High
Graif 2007[70] Prostate 1989-2005	US 2126 PSA	tumor volume <0.5 cm^3, Gleason <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[41] Prostate 1999-2006	Austria 997 (806 screened, 161 unscreened) PSA	Gleason <7, pathologic stage of pT2a and negative surgical margins	16.8% of screened group and 7.9% of unscreened met overdiagnosis criteria	16.8%	High

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

Risk of Bias: Pathologic and Imaging Studies

As most pathologic and imaging studies did not have control groups, traditional selection bias and confounding were not the key internal validity issues for these studies. However, many of the lung cancer studies were unable to obtain complete follow-up information on their initial set of diagnosed cancers and omitted certain patients from the analysis, increasing the risk for bias by arbitrarily cutting down an already limited sample. The Sobue et al [69] study, which followed patients with stage 1 lung cancer to determine causes of death, compared screen-detected patients to a control group of patients with symptom-detected lung cancer who were matched by age within 5 years, sex and year of diagnosis. This study had a high risk for selection bias and confounding because it did nothing to mitigate confounding beyond matching, in addition to having low numbers and omitting several patients from the analysis. Likewise, the Pelzer 2008 et al[41] study of prostate cancer compared overdiagnosis based on RRP specimens from a screened group compared to an unscreened referred cohort without controlling for any confounders. In a similar study, Graif and colleagues created a study group of "screened" individuals based on three different sets of screening and biopsy criteria.[70]

While many studies had problems with incomplete data, selection bias, and confounding, measurement bias was the major flaw in many of these pathologic and imaging studies. In the Sobue et al study of lung cancer, verification of the cause of death was not convincingly valid and reliable as only 35 of 42 patients who died of lung cancer even had progression of their lung cancer verified in their medical records.[69] Yankelevitz et al[73] calculated volume doubling time for lung cancers with data obtained from two different studies. For one study, individual data on tumor size was available, but for the other study only the frequency distribution of tumor dimension and disease stage at the time of diagnosis was available without data on individual tumor size, so the authors assumed the smallest tumors were the stage 1 malignancies. We rated these two studies as having a high risk of measurement bias. The Graif et al study of prostate cancer also had a high risk for measurement bias as different procedures were used during the study to determine tumor volume.[70] The other lung cancer studies were rated as moderate risk

of measurement bias because they provided minimal information on how volume doubling time was calculated and whether it was done in a valid and reliable way. Likewise, the final prostate cancer study provided minimal information on the uniformity of the RRP procedures producing the study specimens and had only one pathologist reading biopsies.[41]

Overall, three lung cancer studies had a high risk of bias[68,69,73], and three had a moderate risk of bias.[67,71,72] Both prostate cancers studies had a high risk of bias.[41,70]

Strength of Evidence: Pathologic and Imaging Studies

Ratings for Risk of Bias and Strength of Evidence criteria for pathologic and imaging studies are available in Appendix Table 5. Directness was poor for all pathologic and imaging studies, with one exception, because the validity of the conclusions of the studies was contingent on the assumption that the pathologic or imaging characteristics were directly correlated with cancer-related morbidity and mortality. With the exception of Veronesi and colleagues[72], no authors attempted to explain the linkage between the pathologic or imaging characteristic and cancer progression or to justify the somewhat arbitrary cutoff they had chosen as the definition of overdiagnosis. In contrast, directness was good for the Sobue et al study[69] because this study followed untreated early stage cancer patients until death from cancer or another cause, directly examining the link between cancer diagnosis and cancer death. Unfortunately, the study's methodologic flaws limit its usefulness. External validity was fair for the majority of studies either due to European settings or to use of screening tests such as chest x-ray that are no longer relevant to current screening discussions. Authors mostly did not provide confidence intervals for overdiagnosis estimates so we were unable to determine precision. Based on a moderate to high aggregate risk of bias along with poor directness, fair external validity, questionable precision and lack of consistency, we rated the strength of evidence as low for both prostate and lung cancer pathologic and imaging studies.

Characteristics of Included Studies: Ecologic and Cohort Studies

A total of 21 ecologic and cohort studies were included in this review, the full details of which are available in Appendix Table 6. Table 4 provides the summary details of 19 of these studies, 17 of breast and two of prostate cancer overdiagnosis. Two additional prostate cancer cohort studies[74,75] are not included in the summary table because they do not share many of the relevant issues, but these studies are listed in the full appendix table. Of the breast cancer studies, one took place in the United States[30] and one in New South Wales, Australia[76], with the rest being performed in European countries. The majority of these studies were ecologic studies in European countries, but several were cohort studies that took advantage of population-based registries to track large numbers of individuals for their screening experiences and cancer outcomes. The screening programs were fairly comparable between studies, tending to involve biennial mammography most commonly for women ages 50 to 69 years. One study looked only at younger women ages 40 to 49 years[77], and a few included some screening extended to women in their seventies.[78,79]

The breast cancer studies used several variations of unscreened reference populations in comparison with the screened populations studied. Most studies modeled the continuation of the pre-screening period incidence trend throughout the screening period with linear regression as the reference. A few studies[77,80,81] took advantage of the fact that screening programs were introduced gradually throughout certain countries and were able to use contemporary counties where screening had not yet been introduced as the reference population. One study used historical age-matched cohorts as the reference.[79] Another two studies used a combination of three control groups, including a contemporary unscreened group and historical groups in the regions with and without screening, in an effort to control for differences in baseline incidence trends between geographic areas.[82,83] Finally, two studies compared screening program attenders to non-attenders as the reference group.[46,84] The magnitude of overdiagnosis reported in breast cancer ecologic and cohort studies was highly variable and ranged from 1% to 76%, depending on age group and the way overdiagnosis was calculated.

Table 4: Summary Evidence Table of Ecologic and Cohort Studies

Study Cancer Type 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) **does not include DCIS	A. Risk of Bias B. Timeframe C. Analysis
Bleyer 2012[30] Breast Ecologic	US 40+ 1976-2008	Pre-screening trend (1976-8)	HRT, baseline increasing incidence	Steady-state screening	(excess cases)/ (observed cases) during screening	31%	A. Moderate B. Good C. Good
Duffy 2010[78] Breast Cohort and Ecologic	Sweden 50-60 1977-98; UK 47-73 1989-2003	Calculated from Swedish 2-County 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	A. Moderate B. NA C. Poor
Falk 2013[84] Breast Cohort	Norway 50-69 1995-2009	Screening program non-attenders	Age, county, calendar year	10-year FU post-screening	(excess cases)/ (expected cases) during screening	19.4% (11.8-27.0)	A. High B. Good C. Good
Hellquist 2012[77], breast Ecologic	Sweden 40-49 1986-2005	Contemporary counties w/o screening	Differences in baseline incidence trends	Statistical adjustment	(excess cases)/ (expected cases) during screening	1% (-6 to 8%) [16% w/o lead time adjustment]	A. Moderate B. NA C. Poor
Jorgensen 2009 (BMJ)[85] Breast Ecologic	UK: 50-64 (1993-1999) CA:50-69 (1995-2005) NSW: 50-69 (1996-2002) Sweden: 50-69 (1998-2006) Norway:50-69 (2000-2006)	Pre-screening trend (UK 1971-84, CA 1970-78, NSW 1972-87, Sweden 1971-85, Norway 1980-94)	Baseline increasing incidence	Up to 7-year FU post-screening	(excess cases)/ (expected cases) during screening	UK: 57% (53-61%) CA: 44% (25-65%) NSW: 53% (44-63%) Sweden: 46% (40-52%) Norway: 52% (36-70%) Meta-analysis: 52% (46-58%)	A. Moderate B. Fair C. Good
Jorgensen 2009(BMC)[80] Breast Ecologic	Denmark 50-69 1991-2003	Contemporary counties w/o screening	Age and differences in baseline incidence trends	Up to 10-12 years FU post-screening	(excess cases)/ (expected cases) during screening	33%	A. Moderate B. Fair C. Good
Junod 2011[79] Breast Ecologic	France 50-79 1995-2005	Age-matched historical cohorts from 1980-90	HRT, alcohol and obesity	Unclear	excess cases)/ (expected cases) during screening	Ages 50-64: 76% (67-85%) ** Ages 65-79: 23% (15-31%) **	A. Moderate B. Fair C. Poor

Study **Cancer Type** **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)** ****does not include DCIS**	**A. Risk of Bias** **B. Timeframe** **C. Analysis**
Kalager 2012[83] Breast Ecologic	Norway 50-79 1996-2005	Contemporary counties w/o screening, and historical cohorts in screening region and w/o screening	Differences in baseline incidence trends	Including women up to 79 in incidence, w/ up to 10 years FU post-screening	(excess cases)/ (observed cases) during screening period, including women up to age 79	Entire country: 25% (19-31%) ** County w/10yrs follow-up: 18% (11-24%) **	A. Moderate B. Fair C. Poor
Morrell 2010[76], breast Ecologic	NSW, Aust. 50-69 1991-2001	Pre-screening trend (1972-90)	HRT, obesity, and nulliparity	Statistical adjustment	(excess cases)/ (expected cases) during screening	30% **	A. Moderate B. NA C. Poor
Njor 2013[82] Breast Cohort	Denmark (Copenhagen /Funen) 56-79/59-78 1991/93-2009	Contemporary counties w/o screening, and historical cohorts in screening region and w/o screening	Differences in baseline incidence trends	Up to 8 years follow-up 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%)	A. Moderate B. Fair C. Poor
Paci 2006[86] Breast Cohort	Italy 50-74 1986-2006 (10-year period)	Pre-screening trend	Age	Statistical adjustment	(excess cases)/ (expected cases) during screening	4.6% (2-7%) after adjustment for lead time 36.2% (34-39%) before adjustment for lead time	A. Moderate B. NA C. Poor
Peeters 1989[81] Breast Ecologic	Netherlands 35+ 1975-86	Contemporary county w/o screening	None	Did not	(excess cases)/ (expected cases) during screening	11%	A. High B. Poor C. Poor
Puliti 2009[45] Breast Cohort	Italy 60-69 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%)	A. Moderate B. Fair C. Poor
Puliti 2012[46] Breast Cohort	Italy 60-69 1991-2007	Screening non-attenders	Age, marital status, and area-level socio-economic status	5-14 years FU post-screening	(excess cases)/ (expected cases) during screening and 5-14 years post-screening	10% (-2 to 23%)	A. High B. Fair C. Poor

Study Cancer Type 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) **does not include DCIS	A. Risk of Bias B. Timeframe C. Analysis
Svendsen 2006[87] Breast Ecologic	Denmark 50-69 1991-2001	Pre-screening trend (1979-90)	Age	Did not	Not calculated	"None" (See appendix table) **	A. Moderate B. Poor C. Poor
Zahl 2004[88] Breast Ecologic	Norway (N) 50-74 1995-2000 Sweden (S) 50-70 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% (42-73%) increased incidence with no post-screening drop ** S: 45% (41-49%) increased incidence with 12% drop **	A. Moderate B. Poor (N)/ Fair (S) C. Good
Zahl 2012[89] Breast Ecologic	Norway 50-79 1995-2009	Pre-screening trend (1991-5)	Age, county, population growth and 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 70-74	A. Moderate B. Fair C. Good
Ciatto 2005[90] Prostate Cohort	Italy 60-74 1991-2000	Contemporary counties w/o screening	Age	7-9 year FU post-screening	(excess cases)/ (expected cases) during screening and 9-years post-screening	66% (40-100%)	A. Moderate B. Fair C. Poor
Zappa 1998[91] Prostate Cohort/ Modeling	Italy 60 or 65 Not provided	Contemporary counties w/o screening	None	4 years FU post-screening	(excess cases)/ (expected cases) during screening and 4-years post-screening	age 60: 25% (19-32%) age 65: 65% (58-73%)	A. Moderate B. Fair C. Poor

Abbreviations: HRT, hormone replacement therapy; FU, follow up; w/o, without; CA, Canada; NSW, New South Wales;
[¥]Unclear if Duffy 2010 estimates of overdiagnosis include DCIS

The prostate cancer studies included in the summary table were both set in Italy and were based on different variations of a PSA screening schedule. Both studies used contemporary unscreened counties as the reference populations. The magnitude of overdiagnosis reported in these studies ranged from 25 to 66%, depending on age and method of calculation.[90,91]

Risk of Bias: Ecologic and Cohort Studies

Ecologic and cohort studies have an elevated risk for selection bias and confounding due to the comparison of non-randomized populations or cohorts. By choosing a control group that is as similar as possible to the study group in factors associated with incidence of cancer, a study has the best chance of minimizing the biases of confounding. However, all potential reference populations that were used in the included breast cancer ecologic and cohort studies, as discussed above, are problematic in certain ways. Modeling the pre-screening incidence trend through the screening period requires the assumption that incidence will continue growing at the same rate without any non-linear changes. The use of contemporary counties without screening programs as the reference population can introduce other confounders that are distributed differently between the two geographic areas. The use of a historical control group is complicated by confounders that have changed between time periods which are likely substantial, but the study that did this adjusted for differences in breast cancer risk factors between eras, thus moderating some of the bias.[79] Studies that used three control groups[82,83] as described above are better able to account for differences in incidence growth between regions but still could be biased by differential influence of confounders between regions.

We rated studies that used these various types of reference populations as having a moderate risk of selection bias and confounding. Within this group of moderately-rated studies, some had a higher risk for selection bias and confounding than others, and some studies took certain actions that increased their credibility. Three studies adjusted for breast cancer risk factors on a population level including hormone replacement therapy use, nulliparity and obesity.[30,76,79] In addition to performing an adjustment for

hormone therapy use, Bleyer and Welch considered two "extreme" scenarios where incidence growth was much greater than the rate predicted based on the pre-screening trend and were able to demonstrate substantial overdiagnosis even given these extremes of incidence growth.[30] It is helpful when authors acknowledge and consider the uncertainties in their studies and perform analyses like these to demonstrate a range of possible results. We rated two breast cancer cohort studies that compared screening attenders and non-attenders[46,84] as having a high risk of selection bias and confounding, because non-attenders of health screenings and services are known to be much different from attenders in terms of general health and other health behaviors. Finally, an early ecologic study by Peeters and colleagues was rated as high risk for confounding because it used a contemporary county without screening as the reference population without consideration for any confounders including age.[81] Our ratings for risk of bias for individual studies are available in Appendix Table 7.

Measurement bias was less of an issue for ecologic and cohort studies. Most studies received a rating of moderate risk of measurement bias because they did not discuss the validity and reliability of their data sources, particularly for cancer incidence data. Two studies did discuss the completeness and accuracy of country registries with regards to cancer incidence data and screening information, when relevant, and we rated these as low risk of measurement bias.[83,84]

We rated both the prostate cancer studies that were included in the summary table as having a moderate risk of selection bias and confounding. Both these studies[90,91] used contemporary unscreened regions of Italy as the reference population, without adjustment for confounders other than age. Neither study discussed the validity and reliability of their data sources, thus we rated both as moderate risk of measurement bias.

We rated the majority of the breast cancer ecologic and cohort studies as moderate in terms of overall risk of bias, as most had a moderate risk of selection bias, confounding, and measurement bias. Three breast cancer studies[46,81,84] had a high risk of bias overall, due to a high risk of confounding and a

moderate risk of measurement bias. Both prostate cancer studies included in the summary table had a moderate risk of bias overall.[90]

Strength of Evidence: Ecologic and Cohort Studies

In addition to risk of bias, ecologic and cohort studies were rated on analysis, directness, external validity, precision, and consistency for an overall evaluation of strength of evidence within each cancer type. Several analysis issues related to measuring and calculating overdiagnosis are unique to ecologic and cohort studies. The first is the appropriateness of the time frame over which overdiagnosis is measured. Related to this is the consideration of lead time, or the time that screening advances the diagnosis of cancer in the screened group, such that the study population should be expected to have an elevated incidence in comparison to the reference population over a certain time frame. Studies need to appropriately consider the lead time in their analysis or they risk under- or over-estimating overdiagnosis. Finally, details of the overdiagnosis calculation itself, including the timeframe over which overdiagnosis is calculated and who is included in the calculation, can affect the overdiagnosis magnitude and thus should be performed in the most appropriate way.

Screening advances the time of diagnosis of preclinical cancers by a period of time known as the lead time. Because of this, cancer incidence is predictably increased in the screened study population during the screening period because cancers that would have presented clinically during and after the screening period are detected earlier by screening. After the screening period, in the absence of overdiagnosis, these cancers that would have presented clinically have already been detected by screening, so incidence decreases in women in the post-screening period. The duration of the drop in incidence should equal the lead time, though typically there is a distribution of lead times that is largely uncertain. Overdiagnosis in cohort and ecologic studies is typically investigated by looking for an increase in incidence during screening in a screened population and a subsequent decrease in incidence in the post-screening years, in comparison to a reference population that remains unscreened. Often,

overdiagnosis is calculated by determining the excess cases of cancer in the screening group during the screening period that are not balanced out by a deficit of cases in post-screening women. Studies that obtain follow-up data for only a few years after screening ends may not sufficiently capture the post-screening drop in incidence, often referred to as the "compensatory drop", and thus can overestimate overdiagnosis. Some studies avoid the need for waiting for this extra follow-up by performing a statistical adjustment for lead time, though we will discuss the downsides to this type of adjustment.

We rated the adequacy of the time frame of studies as one component of the analysis considerations. Time frames were rated as good, fair or poor. Because the lead time and lead time distributions of cancers are largely unknown and thus the true time needed for follow-up in these studies is unknown, these ratings were used as a general guide to highlight where overestimation of overdiagnosis might be occurring. When studies performed a statistical adjustment for lead time[76-78,86] we did not rate their time frame. Two studies did not have any follow-up of women post-screening and these were rated as having a poor time-frame for evaluating overdiagnosis, [81,87] though both of these studies reported low estimates of overdiagnosis that were thus not likely overestimates, and both also had other methodologic flaws. A cohort study by Falk and colleagues[84] achieved at least ten years of follow-up for all women in the study post-screening, which we rated as a good time frame for evaluating overdiagnosis. Another study by Bleyer and Welch[30] did not look at the overall incidence deficit in post-screening women but rather the deficit of late-stage cases, and we rated this study's time frame as good because it was performed over a 30-year period during which screening had reached a steady-state. The remaining breast and prostate cancer ecologic and cohort studies achieved variable amounts of follow-up time post-screening, from four years of complete follow-up on all study participants to up to fourteen years of follow-up on the oldest subset of study participants. Many studies fulfilled a long amount of follow-up time such that a post-screening drop could reasonably be expected to be completely seen in its oldest age groups, but younger age groups were still undergoing screening or had only completed a few years of post-screening follow-up. We rated these studies as having a fair time frame.

As mentioned previously, several studies performed a statistical adjustment for lead time as an alternative to following up study participants in the post-screening period. Conceptually, performing the adjustment for lead time in its simplest form is equivalent to adding "n" years, where "n" equals the lead time, to the age of each woman in the screening group at the time of a cancer diagnosis. Women in the screening group with this age adjustment are then compared with women in the control group for the assessment of overdiagnosis. Performing this type of adjustment for lead time introduces a high degree of uncertainty into the analysis because the mean lead time is largely unknown, and because there is likely a wide distribution of lead times such that this type of simple adjustment is a gross over-simplification. Because of this, we rated the analysis of studies that performed a statistical adjustment for lead time as poor.[76-78,86]

There are many complexities to the calculation of overdiagnosis, and we listed how the overdiagnosis estimate was calculated for each study in the Summary Evidence Table 4. Much of the discussion in past reviews on overdiagnosis has revolved around the denominator for the overdiagnosis calculation[61], and we agree that the choice of denominator can greatly influence the magnitude of overdiagnosis that is reported. Study authors differ on what choice of denominator they think is most appropriate, and sometimes estimates can only be presented in certain ways due to the types of data that are available. Because an overdiagnosed cancer can only be one that was diagnosed by screening and not an interval cased diagnosed by symptoms or a case diagnosed outside of a screening period, we believe that overdiagnosis should be reported as the excess cases divided by the total screening-detected cases. None of the included ecologic and cohort studies reported overdiagnosis in this way, however. The majority of studies reported overdiagnosis as the percentage of cases expected in the absence of screening, which was determined from the reference population. Two studies reported overdiagnosis as the percentage of cases that were observed in the screening group.[30,83] We felt that the use of any of these denominators could be appropriate and justifiable as long as the calculation was clearly explained.

Because only cancer cases that are diagnosed by screening, and thus during the screening period, can be overdiagnosed, it is only appropriate for overdiagnosis to be calculated over the screening period and not over a screening period and an extended period of follow-up. Many ecologic and cohort studies have provided an estimate of overdiagnosis as the risk ratio of the cumulative incidence of cancer in the screening group compared to the reference group over a period of follow-up after the screening period. The following example illustrates why this analysis is problematic.

Table 5: Overdiagnosis Analysis Example 1

	(A) Reference Population	(B) Screening Population (No Overdiagnosis)	(C) Screening Population (with Overdiagnosis)
Year	Cancer Incidence	Cancer Incidence	Cancer Incidence
2000-2004 (screening period)	30 per 1000	60 per 1000	60 per 1000
2005-2009 (post-screening period)	40 per 1000	10 per 1000	30 per 1000

Using the information from the table above, when comparing the reference population (A) to the screening population (B) which does not have overdiagnosis, overdiagnosis can be calculated in two ways. First, overdiagnosis can be calculated as the ratio of the cumulative incidences in the screening and reference populations over the combined screening and post-screening periods, (70/1000)/(70/1000)= 1.00 or 0% overdiagnosis. Second, overdiagnosis can be calculated as the excess cases diagnosed in the screening period that are not compensated for by a deficit of cases in the post-screening period, divided by the expected cases during the screening period, [60-(40-10))/1000]/[30/1000]= 1.00 or 0% overdiagnosis. An equivalent way to think about this is the absolute excess of cases diagnosed in the screened group over the entire screening and post-screening period, divided by the expected cases during the screening period, or (70-70)/(30/1000)= 0% overdiagnosis. Although different denominators can be used, and we would argue that screen-detected cases diagnosed during the screening period would be the most appropriate denominator, the main point is that the time frame of the denominator should represent the screening period and should not include any post-screening cases. In the case of comparing two

populations where there is no overdiagnosis, the method of calculation is irrelevant and the resulting overdiagnosis estimate is the same as demonstrated by these calculations.

However, when we compare the reference population (A) with a screened population (C) that does have overdiagnosis present, the method of calculation becomes important. Performing the calculation as the risk ratio of cumulative incidences over the combined screening and post-screening periods yields (90/1000)/(70/1000)= 1.29 or 29% overdiagnosis. Performing the analysis over just the screening period, by calculating the excess cases not balanced by post-screening deficit cases, yields [(60-(40-30))/1000]/(30/1000)= 1.67, or 67% overdiagnosis. Equivalently, the absolute excess of cases in the screened population (20), divided by the expected cases during the screening period (30), gives the same percentage of overdiagnosis, 67%. It is inappropriate to calculate overdiagnosis using the first method, by including cases diagnosed in the screening and post-screening periods, because cases diagnosed in the post-screening period have no potential to be overdiagnosed. The inclusion of these cases dilutes the estimate of overdiagnosis.

Furthermore, the calculation of overdiagnosis by inclusion of cases diagnosed in the screening and post-screening periods results in a measure that is highly dependent on the length of follow-up time, as illustrated by the following example.

Table 6: Overdiagnosis Analysis Example 2

	(A) Reference Population	(C) Screened Population (with Overdiagnosis)
Year	Cancer Incidence	Cancer Incidence
2000-2004 (screening period)	30 per 1000	60 per 1000
2005-2009 (post-screening period)	40 per 1000	30 per 1000
2010-2014 (post-screening period)	40 per 1000	40 per 1000

First, calculating overdiagnosis as the ratio of cumulative incidences in the screened and reference populations over the screening and five year post-screening periods yields (90/1000)/(70/1000)= 1.29 or 29% overdiagnosis. Similarly, calculating overdiagnosis as the ratio of cumulative incidences over the screening and ten year post-screening periods yields (130/1000)/(110/1000)= 1.18 or 18% overdiagnosis.

However, if the analysis is restricted to the screening period, the overdiagnosis estimate remains stable regardless of the amount of post-screening follow-up. Calculating overdiagnosis as the excess cases in the screening period not compensated by a deficit in the 5-year post-screening period, divided by expected cases during screening, yields [(60-(40-30))/1000]/[30/1000]= 1.67 or 67% overdiagnosis. Calculating overdiagnosis as the excess cases in the screening period not compensated by a deficit in the 10-year post-screening period, divided by expected cases during screening, yields [(60-(80-70))/1000)]/[30/1000]= 1.67 or 67% overdiagnosis. Using the second method of analysis, overdiagnosis is not a function of the follow-up time, making this method much more appropriate for evaluating the true extent of overdiagnosis and comparing results across studies.

Many of the breast and prostate cancer ecologic and cohort studies included in this review calculated overdiagnosis as a risk ratio of cumulative incidences of the screened and reference populations over the screening and follow-up periods.[45,46,82,90,91] As discussed, this is inappropriate because the overdiagnosis measure becomes a function of the length of follow-up time, and the cases included in the calculation from the post-screening period dilute the true amount of overdiagnosis. Thus, we rated the analysis of these studies as poor, because they all likely provided underestimates of overdiagnosis. Similarly, Kalager and colleagues included women up to age 79 in their calculation of overdiagnosis as a ratio of cumulative incidences, even though screening was only offered to women through age 69.[83] This presents similar a problem of diluting the true amount of overdiagnosis because of inclusion of cancer cases from women who could not have possibly been overdiagnosed, so we rated this analysis as poor. In another study by Junod and colleagues[79], it was unclear if a post-screening period was analyzed for a compensatory drop in incidence, or if lead time was managed in another way. This study possibly overestimated overdiagnosis, and we rated its analysis as poor. Finally, we rated the analysis as poor of two remaining studies that did not have any follow-up of women post-screening.[81,87] All other studies received good ratings for analysis because they limited their overdiagnosis calculations

to the screening period and managed lead time by quantifying a deficit of incidence in post-screening women[80,84,85,88,89] or a deficit of late-stage disease.[30]

We rated directness for all ecologic and cohort studies. Because none of these studies were able to identify which individual women were overdiagnosed, and because these studies did not attempt to link cancer diagnosis with cancer-related morbidity and mortality, we rated directness as fair for the majority of studies. For several studies that performed a statistical adjustment for lead time[76-78,86], we rated directness as poor, because these studies required an additional assumption about the progression of cancer from preclinical to clinical stages. External validity was fair for the vast majority of studies as they were performed in European populations. Two exceptions were the Bleyer and Welch study[30] of breast cancer overdiagnosis in the US which received a good rating, and the Hellquist et al study[77] of overdiagnosis in British women ages 40 to 49 which received a poor rating for the limited age group. Precision was fair for the majority of studies, although many others did not provide confidence intervals. Consistency was poor for both breast and prostate cancer overdiagnosis estimates. Based on aggregate risk of bias, analysis, directness, external validity, precision, and consistency, we rated the strength of evidence as low for both breast cancer and prostate cancer ecologic and cohort studies. However, a few breast cancer ecologic studies stood out among the body of evidence for providing a clearer view of the magnitude of overdiagnosis, with a moderate risk of bias, an unbiased analysis, and fair time frames (with the exception of Bleyer and Welch[30], which had a good time frame).[30,80,85,88,89]

Characteristics of Included Studies: Follow-Up of Randomized Controlled Trial

Only one study that met inclusion criteria measured overdiagnosis by following up a randomized controlled trial. This study was a 15-year follow-up of the Malmo randomized controlled trial of mammography in Sweden, the characteristics of which are listed in Table 7. The full details of this trial are listed in Appendix Table 8. In the original trial from 1976 to 1986, over 40,000 women ages 44 to 69 were randomized to either 5 to 6 rounds of mammography every 18 to 24 months or to no screening, and

all women were subsequently followed for 15 years with the aid of Swedish population and cancer registries. There was likely substantial contamination screening of the control group during the follow-up period, as 24% of control participants underwent screening during the study period. The reported magnitude of overdiagnosis was 10% (95% CI 1% to 18%), which was calculated by including all cases over the screening period and the fifteen years of follow-up.[92] A letter to the editor by Welch and colleagues noted that calculating overdiagnosis more appropriately over a denominator of cases detected during the screening period leads to an overdiagnosis magnitude of 15%. Welch goes further to calculate our preferred outcome of overdiagnosis as a percentage of screen-detected cases in the Malmo trial and found this to be 24%.[93]

Risk of Bias: Follow-Up of Randomized Controlled Trial

The follow-up of the Malmo randomized controlled trial was rated as having a low risk for selection bias and confounding. Although the details of the randomization procedures, allocation concealment, baseline distribution of characteristics among groups, and attrition were not described in either the original trial report[94] or the overdiagnosis follow-up report[92], a recent Cochrane review of mammographic screening for breast cancer found these to be adequate with more extensive contact with the authors.[95] We rated the risk of measurement bias as moderate because the authors did not describe the validity and reliability of their data sources, particularly over the fifteen-year follow-up period.[92] Overall, the study's risk of bias was low.

Table 7: Summary Evidence Table of Randomized Controlled Trial Follow-Up Studies

Study; Cancer Type	Study Population: Country, Age Time Period	Post-Study Length of Follow-Up	Calculation of Overdiagnosis	Magnitude of Overdiagnosis (95% CI)	A. Risk of Bias B. Time Frame C. Analysis
Zackrisson 2006[92]; Breast	Sweden, 55-69 1976-1986	15 years	(excess cases)/ (control cases) during trial and 15 years follow-up	10% (1 to 18%)*	A. Low B. Good C. Poor[¥]

*Welch et al re-analysis[93] 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

Strength of Evidence: Follow-Up of Randomized Controlled Trial

Similarly to ecologic and cohort studies, we considered the analysis, directness, external validity, and precision in the evaluation of strength of evidence. The ratings for all risk of bias and strength of evidence criteria are listed in Appendix Table 9. We rated the time frame of this study as good because it achieved complete 15-year follow-up of all women in the study, which is the most of any study included in the review. The analysis, however, received a poor rating, because overdiagnosis was calculated over the entire fifteen-year follow-up period instead of over the screening period, which is problematic as discussed previously in the ecologic and cohort study results section. The re-analysis performed by Welch and colleagues received a good rating. Like ecologic and cohort studies, directness was fair, as this type of study is unable to directly identify overdiagnosed cases. External validity was also fair due to the European population and time period. Precision was fair. Based on a combination of these factors, we rated overall strength of evidence for breast cancer randomized controlled trial follow-up studies as moderate.

Discussion

Key Points

This review identified four major types of studies that have been used to measure overdiagnosis: modeling studies, pathologic and imaging studies, ecologic 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.

Modeling studies are not hindered by the constraints of time and are able to project through areas of uncertainty. However, sensitivity analyses from several of these studies demonstrated that varying key uncertain variables like mean sojourn time or lead time could substantially change the overdiagnosis estimate. Furthermore, directness is poor for these studies as they require assumptions about the progression of cancer from preclinical to clinical stages. The majority of 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.

Pathologic and imaging studies can be simple to perform and interpret but also are typically an over-simplification of overdiagnosis, usually involving an arbitrary cutoff of a characteristic such as volume doubling time to serve as the definition for overdiagnosis. Directness is poor with these studies as they require the assumption that the pathologic or imaging characteristic correlates with cancer progression. Furthermore, because these studies only deal with the non-progression aspect of overdiagnosis and not the competing mortality component, they underestimate overdiagnosis and provide a lower bound to estimates of it.

Ecologic and cohort studies can take advantage of the natural screening experiments taking place in certain countries with gradual implementation of screening programs. These studies must manage confounders and often require a significant time commitment to adequately account for lead time by following women through the post-screening period. Randomized controlled trials have the advantage of equally distributing confounders between study groups but probably entail an even longer time frame between the study and follow-up period.

Comparison with Existing Literature

To our knowledge, there are no other existing systematic reviews that have comprehensively identified all studies that measure overdiagnosis. Several systematic and non-systematic reviews exist that explore a subset of the overdiagnosis literature. In 2012, the UK convened a panel of experts to examine the benefits and harms of mammography and published their conclusions in a review. This review included a meta-analysis of overdiagnosis from the Malmo trial and the two Canadian trials of mammography, where overdiagnosis was represented as a percentage of the total cases diagnosed during the screening period and was found to be 19% overall. They also performed a meta-analysis of

overdiagnosis estimates expressed as a percentage of cases diagnosed over the entire follow-up period, which we would argue is a flawed analysis that should not be considered.[96]

Biesheuvel and colleagues systematically reviewed studies of breast cancer overdiagnosis with a focus on potential sources of bias in the estimates. We agree that differences in risk profiles between study groups, differences in participation in mammography, and inadequate consideration of lead time in determining follow-up can affect overdiagnosis magnitude. However, we disagree that statistical adjustment and excluding prevalence screening data are appropriate ways to manage lead time. Furthermore, Biesheuvel and colleagues advocate the "cumulative incidence method" for calculating overdiagnosis which appears to be a major source of confusion for other researchers who have referenced this review. Using this method, overdiagnosis is calculated as the ratio of cumulative incidences in the screened and unscreened groups at least several years after screening has ended, which is problematic and dilutes the true amount of overdiagnosis as we have previously discussed.[97] In another non-systematic review, Moss discussed randomized controlled trials of mammography with a focus on overdiagnosis but did not draw clear conclusions and also did not recognize the analysis flaw of including an extended follow-up period. Moss also attempted to calculate overdiagnosis for trials where screening was offered to the control group at the end of the trial which was inappropriate.[98] Puliti and colleagues reviewed European observational studies of breast cancer overdiagnosis, making note of which studies they felt adequately and did not adequately adjust for breast cancer risk and lead time.[99] We disagree with 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 and discussed features of studies that influence the estimates of overdiagnosis including the definition, measurement, study design, context, and estimation approaches. They label ecologic 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 of overdiagnosis if the

early years of screening dissemination are included. They claim that the observed excess incidence is not an unbiased estimate of overdiagnosis and provide a misleading example which seems to 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 be an overestimate of overdiagnosis, but very few of the existing ecologic and cohort studies actually calculate overdiagnosis in this way. Instead, most existing studies appropriately measure incidence during the entire screening period and during a period of post-screening follow-up and in so doing are able to accurately measure overdiagnosis provided that their post-screening follow-up is sufficiently long. Etzioni and colleagues have misrepresented the existing ecologic and cohort study literature, and we disagree with their contention that ecologic and cohort studies that use the "excess incidence approach" are inherently biased towards overestimation of overdiagnosis.[100]

Etzioni and colleagues also discuss modeling studies for measuring overdiagnosis which they refer to as the "lead time approach". They point out that choices about the model structure and assumptions can affect the overdiagnosis estimates, which we agree with. However, 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.[100] Rather than lack of transparency, we believe that the inherent lack of directness of modeling studies and the ability of key uncertain inputs to greatly affect 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 ecologic studies, particularly among breast cancer studies.[100] Had they performed a systematic review, however, they would have found several breast cancer modeling studies[51,52,61] with much higher overdiagnosis

estimates than the ones they chose to present, as well as ecologic studies[45,46,82] with lower estimates than those presented. The suggestion by Etzioni that all ecologic and cohort studies overestimate overdiagnosis is unfounded.

Limitations of the Literature

There are many barriers that make it challenging to perform a high quality study measuring overdiagnosis. Time is a major factor, as many countries that have the resources to perform these ecologic or large-scale cohort studies have not established screening programs for long enough to draw clear conclusions about overdiagnosis. Many of the included studies had fair time frames for examining overdiagnosis but had not adequately achieved follow-up to rule out overestimating overdiagnosis. Maintaining a true reference population for ecologic and cohort studies is also a challenge. More and more people are being screened for cancer as technology spreads and awareness grows, limiting the ability of researchers to make a direct comparison with an unscreened population. The reference population had contamination screening in many of the included studies. In places like the United States, researchers have to use modeling to determine expected incidence in the absence of screening because there is no unscreened population available to examine. Many other randomized controlled trials that could have been followed for evaluation of overdiagnosis offered screening to the control group at the end of the trial. Management of confounding is always a challenge with ecologic and cohort studies. Finally, many of the available ecologic and cohort studies had analysis problems that limited the interpretability of their results. The primary limitations of the modeling and pathologic/imaging literature were the inherent uncertainties and lack of directness of the studies.

Limitations of this Review

There were several limitations of our review. Because of the large number of included studies, we were unable to focus on the details of the individuals studies that were included. Much of this information should have been captured in the full evidence tables located in the appendices but was

unable to be fully discussed. We also combined certain studies when multiple studies were available from the same authors or using the same model and population, and it is possible that we missed some of the variability in the data available from these studies. We limited the scope of our review to include only the nine types of solid tumors with the highest incidence in US adults, so we may have missed some studies on overdiagnosis that were performed on other types of cancer. However, our search was not limited to these cancer types and we did not come across any other overdiagnosis studies within another cancer type during the abstract review process, with the exception of a few studies on neuroblastoma. While we did our best to make our assessment of risk of bias and strength of evidence as rigorous and objective as possible by using standard criteria, these ratings involved some subjectivity and it is possible that different raters would have reviewed the evidence differently.

Implications for Future Practice and Research

Despite the major limitations of much of the overdiagnosis literature, there is an emerging picture of a substantial amount of breast and prostate cancer overdiagnosis. Several breast cancer ecologic studies that had a moderate risk for bias with an adequate time frame and a sound analysis suggest that overdiagnosis is above 30% of breast cancer cases expected in the absence of screening.[30,80,85,88,89] The few breast cancer modeling studies that looked at the typical screening age range and performed sensitivity analyses also found overdiagnosis ranging from 20 to 50%, and both of these studies were under-estimates because they did not model either DCIS or competing mortality.[51,52] The prostate cancer literature is more uncertain as there are many fewer ecologic and cohort studies, but the majority of available studies suggest substantial overdiagnosis of prostate cancer as well.

The medical community needs to take action against overdiagnosis of cancer. Steps that have been suggested include raising the threshold for labeling a screening test result as abnormal, developing new biomarkers that can distinguish between more indolent and aggressive cancers, changing medical language to better reflect the benign nature of many of cancer diagnoses, and developing better education

programs about overdiagnosis for physicians, residents, medical students and the public. In some cases it may be appropriate to re-evaluate decisions to screen or to offer screening altogether. We also need to begin a coordinated effort to monitor overdiagnosis in screening programs worldwide. This review attempted to answer the question of how best to measure overdiagnosis in order to first help patients, physicians, and policy-makers make decisions about the benefits and harms of screening programs. Second, being able to measure overdiagnosis will enable us to monitor it over time. We believe that ecologic studies performed by unbiased researchers in a variety of settings will provide the most accurate view of the magnitude of overdiagnosis as well as the best way to monitor it over time. Some of this research is already ongoing, especially in European countries with breast cancer screening programs, but it is not being performed in a uniform way. Standards should be developed for these studies that can then be applied to data from different countries and can be used in a consistent way to monitor overdiagnosis as we implement interventions to reduce it.

These standards should include an adequate time-frame that achieves sufficient follow-up of women post-screening, such that all women in the post-screening age groups included in the study have previously been offered screening. Researchers should determine standard population-level confounders that should be monitored and adjusted for. The standards should include some considerations of uncertainty, such as the use of multiple control groups or the performance of sensitivity analyses. There should also be standards for analysis that require overdiagnosis to be calculated as an absolute excess of cases (or other variations on this numerator) divided by a denominator of cases diagnosed during the screening period only (screen-detected cases, ideally). Many authors make the mistake of including cases diagnosed during the post-screening follow-up period or of representing overdiagnosis as the ratio of cumulative incidences of the screened and unscreened populations over the screening and follow-up periods, both of which dilute the true amount of overdiagnosis. The uniform application of these standards to data sets from different countries over different time frames and with a variety of screening schedules and populations should move us closer to understanding the magnitude of overdiagnosis present.

Conclusions

Researchers have measured overdiagnosis using four main study types. Modeling studies use statistical models to measure overdiagnosis and are able to bypass areas of uncertainty but are limited by their assumptions and indirectness. Pathologic and imaging studies measure overdiagnosis based on a tumor's pathologic or imaging characteristic. These studies lack directness and oversimplify cancer progression and overdiagnosis. Randomized controlled trials that do not offer screening to the control group at the end of the trial period can follow subjects and measure overdiagnosis by comparing incidence after long periods of time. These studies best manage confounding but require a significant time commitment. Ecologic and cohort studies measure overdiagnosis by comparing cancer incidence in a screened and unscreened population or cohort over a certain time frame. Although these studies can be limited by confounding and require adequate time frames and careful analysis, when performed well they can provide a clear view of overdiagnosis. This view can be clarified further as ecologic studies from multiple settings and time frames are compared and demonstrate similar magnitudes of overdiagnosis. We recommend that unbiased researchers use standards to perform such ecologic studies to monitor overdiagnosis in cancer screening programs worldwide.

References

1. Harris R, Sawaya GF, Moyer VA, Calonge N. Reconsidering the criteria for evaluating proposed screening programs: Reflections from 4 current and former members of the US preventive services task force. *Epidemiol Rev*. 2011;33(1):20-35.

2. Black WC. Randomized clinical trials for cancer screening: Rationale and design considerations for imaging tests. *J Clin Oncol*. 2006;24(20):3252-3260.

3. Black WC. Overdiagnosis: An underrecognized cause of confusion and harm in cancer screening. *J Natl Cancer Inst*. 2000;92(16):1280-1282.

4. Welch HG, Black WC. Overdiagnosis in cancer. *J Natl Cancer Inst*. 2010;102(9):605-613.

5. Welch HG, Black WC. Using autopsy series to estimate the disease" reservoir" for ductal carcinoma in situ of the breast: How much more breast cancer can we find? *Ann Intern Med*. 1997;127(11):1023.

6. Erbas B, Provenzano E, Armes J, Gertig D. The natural history of ductal carcinoma in situ of the breast: A review. *Breast Cancer Res Treat*. 2006;97(2):135-144.

7. Manser RL, Dodd M, Byrnes G, Irving LB, Campbell DA. Incidental lung cancers identified at coronial autopsy: Implications for overdiagnosis of lung cancer by screening. *Respir Med*. 2005;99(4):501-507.

8. Dammas S, Patz Jr. EF, Goodman PC. Identification of small lung nodules at autopsy: Implications for lung cancer screening and overdiagnosis bias. *Lung Cancer*. 2001;33(1):11-16.

9. Stamatiou K, Alevizos A, Agapitos E, Sofras F. Incidence of impalpable carcinoma of the prostate and of non-malignant and precarcinomatous lesions in greek male population: An autopsy study. *Prostate*. 2006;66(12):1319-1328.

10. Zlotta AR, Kuk C, Kovylina M, et al. Prevalence of prostate carcinoma and its precursor lesions in russian caucasian and japanese men in autopsy specimens. *J Urol*. 2012;187(4):e65-e66.

11. Harach HR, Franssila KO, Wasenius V. Occult papillary carcinoma of the thyroid. A "normal" finding in finland. A systematic autopsy study. *Cancer*. 2006;56(3):531-538.

12. Larsen SU, Rose C. Spontaneous remission of breast cancer. A literature review. *Ugeskr Laeger*. 1999;161(26):4001-4004.

13. Dussan C, Zubor P, Fernandez M, Yabar A, Szunyogh N, Visnovsky J. Spontaneous regression of a breast carcinoma: A case report. *Gynecol Obstet Invest*. 2008;65(3):206-211.

14. Kalialis LV, Drzewiecki KT, Klyver H. Spontaneous regression of metastases from melanoma: Review of the literature. *Melanoma Res*. 2009;19(5):275.

15. Elhilali M, Gleave M, Fradet Y, et al. Placebo-associated remissions in a multicentre, randomized, double-blind trial of interferon γ-1b for the treatment of metastatic renal cell carcinoma. *BJU Int*. 2000;86(6):613-618.

16. Huz JI, Melis M, Sarpel U. Spontaneous regression of hepatocellular carcinoma is most often associated with tumour hypoxia or a systemic inflammatory response. *HPB*. 2012.

17. Yamamoto K, Hanada R, Kikuchi A, et al. Spontaneous regression of localized neuroblastoma detected by mass screening. *Journal of clinical oncology*. 1998;16(4):1265-1269.

18. Hero B, Simon T, Spitz R, et al. Localized infant neuroblastomas often show spontaneous regression: Results of the prospective trials NB95-S and NB97. *Journal of clinical oncology*. 2008;26(9):1504-1510.

19. Hofstad B, Vatn M, Andersen S, et al. Growth of colorectal polyps: Redetection and evaluation of unresected polyps for a period of three years. *Gut*. 1996;39(3):449-456.

20. Moscicki A, Shiboski S, Hills NK, et al. Regression of low-grade squamous intra-epithelial lesions in young women. *The Lancet*. 2004;364(9446):1678-1683.

21. Miller AB, To T, Baines CJ, Wall C. The canadian national breast screening study-1: Breast cancer mortality after 11 to 16 years of follow-up. *Ann Intern Med*. 2002;137(5 pt 1):305-312.

22. Zahl P, Mæhlen J, Welch HG. The natural history of invasive breast cancers detected by screening mammography. *Arch Intern Med*. 2008;168(21):2311.

23. Welch HG, Woloshin S, Schwartz LM. Skin biopsy rates and incidence of melanoma: Population based ecological study. *BMJ*. 2005;331(7515):481.

24. Linos E, Swetter SM, Cockburn MG, Colditz GA, Clarke CA. Increasing burden of melanoma in the united states. *J Invest Dermatol*. 2009;129(7):1666-1674.

25. Cramer JD, Fu P, Harth KC, Margevicius S, Wilhelm SM. Analysis of the rising incidence of thyroid cancer using the surveillance, epidemiology and end results national cancer data registry. *Surgery*. 2010;148(6):1147-1153.

26. Sun M, Thuret R, Abdollah F, et al. Age-adjusted incidence, mortality, and survival rates of stage-specific renal cell carcinoma in north america: A trend analysis. *Eur Urol*. 2011;59(1):135-141.

27. Welch HG, Albertsen PC. Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986-2005. *J Natl Cancer Inst*. 2009;101(19):1325-1329.

28. Garfinkel L, Boring CC, Heath CW. Changing trends: An overview of breast cancer incidence and mortality. *Cancer*. 2009;74(S1):222-227.

29. Age-adjusted SEER breast cancer incidence rates all ages, all races, female (2000–2009). *J Natl Cancer Inst*. 2013;105(8):512.

30. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. *New Engl J Med*. 2012;367(21):1998-2005.

31. Elmore JG, Fletcher SW. Overdiagnosis in breast cancer screening: Time to tackle an underappreciated harm. *Ann Intern Med*. 2012;156(7):536-537.

32. Esserman L, Thompson I. Solving the overdiagnosis dilemma. *J Natl Cancer Inst*. 2010;102(9):582-583.

33. Jorgensen KJ, Klahn A, Gotzsche PC. Are benefits and harms in mammography screening given equal attention in scientific articles? A cross-sectional study. *BMC Med*. 2007;5.

34. 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(5):307.

35. Hersch J, Jansen J, Barratt A, et al. Women's views on overdiagnosis in breast cancer screening: A qualitative study. *BMJ (Online)*. 2013;346(7892).

36. de Bekker-Grob EW, Rose JM, Donkers B, Essink-Bot M-, Bangma CH, Steyerberg EW. Men's preferences for prostate cancer screening: A discrete choice experiment. *Br J Cancer*. 2013.

37. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. *CA: A Cancer Journal for Clinicians*. 2013;63(1):11-30.

38. Harris R, Yeatts J, Kinsinger L. Breast cancer screening for women ages 50 to 69years a systematic review of observational evidence. *Prev Med*. 2011;53(3):108-114.

39. U.S. preventive services task force procedure manual. AHRQ publication no. 08-05118-EF
. http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm. Updated July 2008. Accessed April/21, 2013.

40. Owens DK, Lohr KN, Atkins D, 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(5):513-523.

41. Pelzer AE, Colleselli D, Bektic J, et al. Clinical and pathological features of screen vs non-screen-detected prostate cancers: Is there a difference? *BJU Int*. 2008;102(1):24-27.

42. Pelzer AE, Colleselli D, Bektic J, et al. Over-diagnosis and under-diagnosis of screen- vs non-screen-detected prostate cancers with in men with prostate-specific antigen levels of 2.0-10.0 ng/mL. *BJU Int*. 2008;101(10):1223-1226.

43. Pelzer AE, Bektic J, Akkad T, et al. Under diagnosis and over diagnosis of prostate cancer in a screening population with serum PSA 2 to 10 ng/ml. *J Urol*. 2007;178(1):93-97.

44. Paci E, Warwick J, Falini P, Duffy SW. Overdiagnosis in screening: Is the increase in breast cancer incidence rates a cause for concern? *J Med Screen*. 2004;11(1):23-27.

45. 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(18):3166-3171.

46. 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(1).

47. Liberati A, Altman DG, Tetzlaff J, 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(4):W-65-W-94.

48. 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(3):145-153.

49. Duffy SW, Agbaje O, Tabar L, et al. Estimates of overdiagnosis from two trials of mammographic screening for breast cancer. *Breast Cancer Res*. 2005;7(6):258-265.

50. Olsen AH, Agbaje OF, Myles JP, Lynge E, Duffy SW. Overdiagnosis, sojourn time, and sensitivity in the copenhagen mammography screening program. *Breast J*. 2006;12(4):338-342.

51. 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 (Online)*. 2012;344(7839).

52. Martinez-Alonso M, Vilaprinyo E, Marcos-Gragera R, Rue M. Breast cancer incidence and overdiagnosis in catalonia (spain). *Breast Cancer Res*. 2010;12(4).

53. Davidov O, Zelen M. Overdiagnosis in early detection programs. *Biostatistics*. 2004;5(4):603-613. doi: 10.1093/biostatistics/kxh012.

54. Draisma G, Etzioni R, Tsodikov A, et al. Lead time and overdiagnosis in prostate-specific antigen screening: Importance of methods and context. *J Natl Cancer Inst*. 2009;101(6):374-383.

55. 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(11):1833-1838.

56. McGregor M, Hanley JA, Boivin J-, McLean RG. Screening for prostate cancer: Estimating the magnitude of overdetection. *Can Med Assoc J*. 1998;159(11):1368-1372.

57. Pashayan N, Duffy SW, Pharoah P, 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(7):1198-1204.

58. Telesca D, Etzioni R, Gulati R. Estimating lead time and overdiagnosis associated with PSA screening from prostate cancer incidence trends. *Biometrics*. 2008;64(1):10-19.

59. Tsodikov A, Szabo A, Wegelin J. A population model of prostate cancer incidence. *Stat Med*. 2006;25(16):2846-2866.

60. Wu GH-, Auvinen A, Maattanen L, et al. Number of screens for overdetection as an indicator of absolute risk of overdiagnosis in prostate cancer screening. *Int J Cancer*. 2012;131(6):1367-1375.

61. 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(1):111-121.

62. de Gelder R, Fracheboud J, Heijnsdijk EAM, et al. Digital mammography screening: Weighing reduced mortality against increased overdiagnosis. *Prev Med*. 2011;53(3):134-140.

63. 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(6).

64. Hazelton WD, Goodman G, Rom WN, et al. Longitudinal multistage model for lung cancer incidence, mortality, and CT detected indolent and aggressive cancers. *Math Biosci*. 2012;240(1):20-34.

65. Pinsky PF. An early- and late-stage convolution model for disease natural history. *Biometrics*. 2004;60(1):191-198.

66. Luo D, Cambon AC, Wu D. Evaluating the long-term effect of FOBT in colorectal cancer screening. *Cancer Epidemiol*. 2012;36(1):e54; e60.

67. Dominioni L, Rotolo N, Mantovani W, 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.

68. Sone S, Nakayama T, Honda T, 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(3):329-341.

69. 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(3):685-692.

70. Graif T, Loeb S, Roehl KA, et al. Under diagnosis and over diagnosis of prostate cancer. *J Urol*. 2007;178(1):88-92.

71. Lindell RM, Hartman TE, Swensen SJ, et al. Five-year lung cancer screening experience: CT appearance, growth rate, location, and histologic features of 61 lung Cancers1. *Radiology*. 2007;242(2):555-562.

72. Veronesi G, Maisonneuve P, Bellomi M, et al. Estimating overdiagnosis in low-dose computed tomography screening for lung cancer: A cohort study. *Ann Intern Med*. 2012;157(11):776-784.

73. Yankelevitz DF, Kostis WJ, Henschke CI, et al. Overdiagnosis in chest radiographic screening for lung carcinoma: Frequency. *Cancer*. 2003;97(5):1271-1275.

74. Hugosson J, Aus G, Becker C, et al. Would prostate cancer detected by screening with prostate-specific antigen develop into clinical cancer if left undiagnosed? A comparison of two population-based studies in sweden. *BJU Int*. 2000;85(9):1078-1084.

75. Tornblom M, Eriksson H, Franzen S, et al. Lead time associated with screening for prostate cancer. *Int J Cancer*. 2004;108(1):122-129.

76. 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(2):275-282.

77. 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(1):14-19. doi: 10.1258/jms.2012.011104; 10.1258/jms.2012.011104.

78. Duffy SW, Tabar L, Olsen AH, 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(1):25-30. doi: 10.1258/jms.2009.009094; 10.1258/jms.2009.009094.

79. Junod B, Zahl P-, 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.

80. Jorgensen KJ, Zahl P-, Gotzsche PC. Overdiagnosis in organised mammography screening in denmark. A comparative study. *BMC Women's Health*. 2009;9.

81. Peeters PHM, Verbeek ALM, Straatman H, et al. Evaluation of overdiagnosis of breast cancer in screening with mammography: Results of the nijmegen programmme. *Int J Epidemiol*. 1989;18(2):295-299.

82. Njor SH, Olsen AH, Blichert-Toft M, Schwartz W, Vejborg I, Lynge E. Overdiagnosis in screening mammography in denmark: Population based cohort study. *BMJ: British Medical Journal*. 2013;346.

83. Kalager M, Adami H-, Bretthauer M, Tamimi RM. Overdiagnosis of invasive breast cancer due to mammography screening [4]. *Ann Intern Med*. 2012;157(3):221-222.

84. Falk RS, Hofvind S, Skaane P, Haldorsen T. Overdiagnosis among women attending a population-based mammography screening program. *Int J Cancer*. 2013. doi: 10.1002/ijc.28052; 10.1002/ijc.28052.

85. Jørgensen KJ, Gøtzsche PC. Overdiagnosis in publicly organised mammography screening programmes: Systematic review of incidence trends. *BMJ: British Medical Journal*. 2009;339.

86. Paci E, Miccinesi G, Puliti D, 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(6).

87. 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(9):1883-1890.

88. Zahl P-, Strand BH, Maeshlen J. Incidence of breast cancer in norway and sweden during introduction of nationwide screening: Prospective cohort study. *Br Med J*. 2004;328(7445):921-924.

89. Zahl P, Maehlen J. Overdiagnosis of breast cancer after 14 years of mammography screening. *Tidsskr Nor Legeforen*. 2012;132:414-417.

90. Ciatto S, Gervasi G, Bonardi R, et al. Determining overdiagnosis by screening with DRE/TRUS or PSA (florence pilot studies, 1991-1994). *Eur J Cancer*. 2005;41(3):411-415.

91. Zappa M. Overdiagnosis of prostate carcinoma by screening: An estimate based on the results of the florence screening pilot study. *Ann Oncol*. 1998;9(12):1297-1300.

92. 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. *Br Med J*. 2006;332(7543):689-691.

93. Welch HG, Schwartz LM, Woloshin S. Ramifications of screening for breast cancer: 1 in 4 cancers detected by mammography are pseudocancers. *BMJ: British Medical Journal*. 2006;332(7543):727.

94. Andersson I, Aspegren K, Janzon L, et al. Mammographic screening and mortality from breast cancer: The malmö mammographic screening trial. *BMJ: British Medical Journal*. 1988;297(6654):943.

95. Gøtzsche PC, Nielsen M. Screening for breast cancer with mammography. *Cochrane Database Syst Rev*. 2009;4(1).

96. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: An independent review. *Lancet*. 2012;380(9855):1778-1786. doi: 10.1016/S0140-6736(12)61611-0; 10.1016/S0140-6736(12)61611-0.

97. 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(12):1129-1138.

98. Moss S. Overdiagnosis in randomised controlled trials of breast cancer screening. *Breast Cancer Res*. 2005;7(5):230-234.

99. Puliti D, Duffy SW, Miccinesi G, et al. Overdiagnosis in mammographic screening for breast cancer in europe: A literature review. *J Med Screen*. 2012;19 Suppl 1:42-56. doi: 10.1258/jms.2012.012082.

100. 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(11):831-838.

Appendix A: Standard Criteria for Evaluating Risk of Bias, by Study Design

Cohort and Ecologic Studies (adapted from Harris et al, 2011[38])

A. Risk of Bias *(rate overall as high/moderate/low)*
    i. Probability of selection bias and confounding *(rate as high/moderate/low)*
        i. Unbiased creation of comparable groups (at least after adjustment), especially with regard to factors associated with cancer incidence
        ii. Maintenance of comparable groups. No large in or out migration during study period; no large drop-outs or differential drop-outs. No differential changes in factors associated with cancer incidence.
        iii. Adequate identification of potential confounders and control of potential confounding by exclusion, stratification, statistical adjustment, other
    ii. Probability of measurement bias *(rate as high/moderate/low)*
        i. Measures of exposure to screening, potential confounders (especially factors related to cancer incidence), and cancer incidence are equally applied between comparison groups
        ii. Measures of exposure to screening, potential confounders, and cancer incidence are valid, including blinding where appropriate.
        iii. Measures of exposure to screening, potential confounders, and cancer incidence are reliable

Follow-up of Randomized Controlled Trial (adapted from the USPSTF Procedure Manual[39])

A. Risk of Bias *(rate overall as high/moderate/low)*
    i. Probability of selection bias *(rate as high/moderate/low)*
        i. Unbiased creation of comparable groups, including adequate randomization, allocation concealment, and equal distribution of potential confounders among both groups
        ii. Maintenance of comparable groups. No large drop-outs or differential drop-outs. Appropriate adherence and minimal contamination or cross-overs.
    ii. Probability of measurement bias *(rate as high/moderate/low)*
        i. Measures of exposure to screening, potential confounders, and cancer incidence are equal between groups
        ii. Measures of exposure to screening, potential confounders, and cancer incidence are valid, including blinding where appropriate
        iii. Measures of exposure to screening, potential confounders, and cancer incidence are reliable
    iii. Potential for confounding *(rate as high/moderate/low)*
        i. Equal distribution of potential confounders among two groups, without changes in group composition throughout follow-up.

Pathologic and Imaging Studies

A. Risk of Bias *(rate overall as high/moderate/low)*
    i. Probability of selection bias and confounding *(rate as high/moderate/low)*
        i. No large drop-outs or inadequate follow-up of selected members of study population

ii. If control group present: unbiased creation and maintenance of comparable groups
iii. If control group present: adequate identification of potential confounders and control of potential confounding by exclusion, stratification, statistical adjustment, other

ii. Probability of measurement bias *(rate as high/moderate/low)*
i. Measures of pathologic or behavioral characteristics are valid, including blinding where appropriate and avoiding differential follow-up
ii. Measures of pathologic or behavioral characteristics are reliable

Modeling Studies

A. Risk of Bias *(rate overall as high/moderate/low)*
i. Extent to which assumptions made in the model are transparent and clearly stated *(rate as good/fair/poor)*
ii. Extent to which assumptions made in the model are backed up with evidence *(rate as good/fair/poor)*
i. ideally systematically-reviewed evidence that was critical appraised with quality ratings
iii. Probability for biases in the data used in the model *(rate as good/fair/poor/cannot determine)*
i. Measurement of outcomes in data used in model are valid and reliable
ii. Adequate measurement of and control for potential confounders in data used in model
1. This information should be presented and discussed by authors so that readers can appraise the study.
iv. Extent to which sensitivity analyses are performed for any uncertain variables *(rate as good/fair/poor)*
i. ideally probabilistic multivariate sensitivity analyses
v. Validation: model has been validated using population data different from the population data used to calibrate the model

Appendix B: Criteria for Evaluating Strength of Evidence

A. Risk of Bias *(rate as high/moderate/low)* (specific criteria listed in Appendix A)

B. Analysis *(rate as good/fair/poor)* (Ecologic and Cohort, RCT follow-up studies only)
i. Extent to which the analysis appropriately quantifies overdiagnosis, without inclusion of age groups or time frames that lack the potential to be overdiagnosed, and with appropriate consideration for lead time
ii. Extent to which the time frame is appropriate sufficient to account for the effects of lead time

C. Directness *(rate as good/fair/poor)*
i. Extent to which the evidence links the screening test directly to health outcomes with minimal assumptions regarding:
i. The progression of a screen-detected cancer to a cancer that causes morbidity and mortality

ii. The association of pathologic or behavioral characteristics of a cancer with cancer progression and cancer-related morbidity and mortality

D. External Validity *(rate as good/fair/poor)*
   i. Extent to which study population is similar to US general population in factors that are associated with cancer incidence
   ii. Extent to which the screening situation (e.g., expertise of the screening radiographers, quality of screening facilities, threshold for labeling a result as abnormal) in the study is comparable to the screening situation in the US general population
   iii. Extent to which medical care and risks for competing mortality in the study are similar to medical care in the US general population

E. Precision *(rate as good/fair/poor/cannot determine)*
   i. Confidence interval on magnitude of overdiagnosis should be provided. Width of confidence interval should be narrow.

F. Consistency *(rate as good/fair/poor)*
   i. Degree to which the overdiagnosis measurement from the included studies has a similar magnitude, within the same cancer type and study design

Appendix C: Results

Appendix Table 1: Excluded Modeling Studies with Reasons for Exclusion

Study	Reason for Exclusion
Draisma G, Boer R, Otto SJ, et al. Lead times and overdetection due to prostate-specific antigen screening: Estimates from the European randomized study of screening for prostate cancer. *J Natl Cancer Inst*. 2003;95(12):868-878	Same model and population as Heijnsdijk 2009 (MISCAN model fitted to ERSPC Rotterdam data)
Etzioni R, Cha R, Cowen ME. Serial prostate specific antigen screening for prostate cancer: A computer model evaluates competing strategies. *J Urol*. 1999;162(3 I):741-748	same model and population as Etzioni 2002 and Telesca 2008
Etzioni R, Penson DF, Legler JM, et al. Overdiagnosis due to prostate-specific antigen screening: Lessons from U.S. prostate cancer incidence trends. *J Natl Cancer Inst*. 2002;94(13):981-990	Same model and population as Telesca 2008
Gulati R, Inoue L, Katcher J, Hazelton W, Etzioni R. Calibrating disease progression models using population data: A critical precursor to policy development in cancer control. *Biostatistics*. 2010;11(4):707-719	Same model and population as Gulati 2013
Yen M, Tabar L, Vitak B, Smith RA, Chen H-, Duffy SW. Quantifying the potential problem of overdiagnosis of ductal carcinoma in situ in breast cancer screening. *Eur J Cancer*. 2003;39(12):1746-1754	Does not include invasive cancers, only looks at overdiagnosis of DCIS

Appendix Table 2: Evidence Table of Modeling Studies

Study, Cancer Type	Model Name/Type	Modeled Population, Time Period	Screening Schedule/ Details	Data Sources
Davidov 2004, prostate	Not provided	US men Not provided	5-year intervals starting at 50 and ending at 60, 70 or 80	• Incidence: SEER 1993-1997 • Mortality: SSA life tables (1997 males)
Draisma 2009, prostate	• MISCAN (microsimulation) • Fred Hutchinson Cancer Research Center (microsimulation) • UMich (statistical mixed)	US men 54-80 Not provided	Not specified. Screening patterns based on typical US screening	• Incidence: SEER 1985-2000 • Mortality: standard life tables • PSA screening patterns: NHIS 2000 and SEER-linked Medicare claims • PSA growth curves: Prostate Cancer Prevention Trial (FHCRC) • Biopsy sensitivities: literature review (FHCRC) • Biopsy compliance rates (40%): PLCO trial (FHCRC)
Gulati 2013, prostate	Fred Hutchinson Cancer Research Center (microsimulation)	US men aged 40 Not provided	32 different screening strategies based on variations of: • starting at 40 or 50 • stopping at 69 or 74 • annual vs biennial • 4 thresholds for biopsy referral	• Incidence: SEER 1975-2000, men 50-84 • Mortality: US life tables • PSA screening patterns: NHIS 2000 and SEER-Medicare linked claims • PSA growth curves: control group from Prostate Cancer Prevention Trial • Prostate cancer treatment patterns: SEER 2005 • Baseline prostate cancer survival: SEER data on untreated patients (1983-1986) • Survival benefit for radical prostatectomy: SPCG-4 trial • Survival benefit for radiotherapy: CaPSURE trial • Survival benefit for early detection: ERSPC trial • Biopsy compliance rates: PLCO trial • Biopsy sensitivity: literature review
Heijnsdijk 2009, prostate	MISCAN (microsimulation)	European Standard Population 2003, 2008-2033	55-70 every 1 or 2 years or 55-75 every 4 years, 3ng/ml biopsy threshold	• Disease and treatment-specific parameters: Rotterdam ERSPC • Cure rates: estimated from 10-year relative survival by clinical stage from Comprehensive Cancer Center Amsterdam (1985-2005)

Study, Cancer Type	**A.** Includes DCIS? **B.** Includes Competing Mortality? **C.** External Validation? **D.** Reports Preferred Outcome?	Magnitude of Overdiagnosis (95% CI)	Sensitivity Analyses?	Results of Sensitivity Analyses	Conclusions
Davidov 2004, prostate	A.N/A B.Yes C.No D. Unclear	8.48-53.6%, depending on mean sojourn time	Varied mean sojourn time (5, 7.5, 10, 12.5, 15 years) and sensitivity (0.3, 0.7, 0.9)	Overdiagnosis did not vary with sensitivity but did vary greatly with mean sojourn time	The probability of overdiagnosis is remarkably high.
Draisma 2009, prostate	A.N/A B.Yes C.No D.Yes	• MISCAN: 42% • FRCHC: 28% • UMich: 23%	Not performed	N/A	The precise definition and population used to estimate overdiagnosis can be important drivers of study results.
Gulati 2013, prostate	A.N/A B.Yes C.No D.No	Lifetime probability of overdiagnosis ranges from 1.8-6%	Varied rates of disease onset, metastasis, and clinical detection in incidence model and extent of screening effect in mortality model. Varied survival effect of screening from no effect to effect consistent with stage-shift model.	Results not reported. Authors report that varying inputs produced little variation and overall conclusions about tradeoffs across strategies are robust.	Screening strategies that use higher biopsy thresholds for older men and screen men with low PSA levels less frequently achieve similar benefits in mortality with fewer false positives and cases overdiagnosed.
Heijnsdijk 2009, prostate	A.N/A B.Yes C.No D.Yes (estimate from figures)	• Screening every 4 years: 67% • Screening annually: 60% • Screening biennially: 60%	Not performed	N/A	Implementation of PSA screening will double total costs for prostate cancer, most of which are due to diagnosis and treatment, especially of overdiagnosed cases.

Study, Cancer Type	Model Name/Type	Modeled Population, Time Period	Screening Schedule/ Details	Data Sources
McGregor 1998, prostate	Not provided	Quebec men 50-85 Not provided	Annual PSA from 50-70 with 4 ng/ml biopsy threshold	• Rate of death from prostate cancer in 1980s assumed to equal rate of lethal prostate cancer entering the population • Prostate cancer mortality: Quebec Ministry of Health (1988-1992) • Proportion of lethal cancer detectable by PSA estimated from literature review as 85%
Pashayan 2009, prostate	Not provided	UK men Not provided	Single PSA test with 3ng/ml biopsy threshold	• Incidence: Eastern Cancer Registry and Information Centre • Prevalent cases: ProtecT study • Interval cancers: UK Office of National Statistics 2002-2005 • Mortality: UK Office of National Statistics male life tables • PSA sensitivity by age: determined by linear regression from values obtained from literature review
Telesca 2008, prostate	Stochastic simulation model	US men 50+ Not provided	Not specified. Screening patterns based on typical screening in US	• Incidence: SEER 1973-1987 • Mortality: CDC's Vital Statistics of the US 1992 • PSA screening patterns: NHIS 2000 and SEER-Medicare linked claims • Probability of a positive test among men screened, biopsy frequency given positive test, and PPV obtained from literature search
Tsodikov 2006, prostate	Not provided	US men Not provided	Not specified. Screening patterns based on typical screening in US	• Incidence: SEER • Mortality: Human Mortality Database • PSA screening patterns: NHIS 2000 and SEER-Medicare linked claims • PSA sensitivity = 100%
Wu 2012, prostate	Multistep epidemiological model/ five-state Markov model	Finnish men ages 55, 59, 63, 67 Not provided	Up to 3 PSA tests every 4 years until 71 with 4 ng/ml threshold for DRE, ultrasound and biopsy. If PSA 3-3.9 ng/ml, referred for DRE or free/total PSA ratio	• Incidence: Finnish prostate cancer screening trial (largest arm of ERSPC), 1996-2005 • Interval cancers: Finnish Cancer Registry • Mortality: Statistics Finland

Study, Cancer Type	**A.** Includes DCIS? **B.** Includes Competing Mortality? **C.** External Validation? **D.** Reports Preferred Outcome?	Magnitude of Overdiagnosis (95% CI)	Sensitivity Analyses?	Results of Sensitivity Analyses	Conclusions
McGregor 1998, prostate	A.N/A B.Yes C.No D.Yes	84%	Varied detection rates of lethal cancer with PSA from 75% to 95%	78 to 87%	Of every 100 men with screen-detected prostate cancer, only 16 on average could have their lives extended by surgery since the cancer would not cause death before age 85 in the remaining 84.
Pashayan 2009, prostate	A.N/A B.Yes C.No D.Yes	• Ages 50-54: 10% (7-11) • Ages 55-59: 15% (12-15%) • Ages 60-64: 23% (20-24%) • Ages 65-69: 31% (26-32%)	Not performed	N/A	The benefit of screening in reducing advanced stage disease is limited by overdiagnosis, which is greater at older ages.
Telesca 2008, prostate	A.N/A B.Yes C.No D.Yes	• White men: 22.7% • Black men: 34.4%	Authors considered a constant secular incidence trend and a decreasing trend to account for decline in use of TURP for BPH. As sensitivity analysis, also considered an increasing trend.	Results not reported	Likelihood-based approach allows authors to make formal inferences about lead time and overdiagnosis associated with PSA screening in US
Tsodikov 2006, prostate	A.N/A B.Yes C.No D.Yes	About 30%	Not performed	N/A	
Wu 2012, prostate	A.N/A B.Yes C.No D.No	Absolute risk for overdetection during study period 3.4% (2.1-5.7%)	Not performed	N/A	Authors estimated that for every 100 men screened, 3.4 cases would be overdetected during 3 screening rounds in the Finnish trial.

Study, Cancer Type	Model Name/Type	Modeled Population, Time Period	Screening Schedule/ Details	Data Sources
De Gelder 2011 (Epi Reviews), breast	MISCAN (micro-simulation)	Dutch female population 0-100 in 1989, Not provided	Biennial screening began in 1990, gradually spread to all of Netherlands by 1997 for women 49-69. Extended to women 49-74 between 1998-2001	• Incidence: Dutch Comprehensive Cancer Centers and National Evaluation Team for Breast Cancer Screening in the Netherlands, 1990-2006
De Gelder 2011 (Preventive Medicine), breast	MISCAN (micro-simulation)	Dutch female population 0-100 in 1989, 1990-2020	30-year period of biennial screening with either screen film mammography or digital mammography	• Incidence: Dutch Cancer Registry and National Evaluation Team for Breast Cancer Screening, 1990-2006 • Screening patterns: 82% participation rate based on NETB data • 100% sensitivity of digital mammography for DCIS
Duffy 2005, breast	Not provided	• Swedish 2-County Trial: women 40-74 • Gothenburg Trial: women 39-59, Not provided	• Swedish 2-County: single-view mammography biennially (40-49) and every 33 months (50-74) • Gothenburg: 2-view mammography at first screen, # of views then dependent on breast density, every 18 months	• All Data: Swedish 2-Country RCT (1977-84) and Gothenburg RCT (1982-1987) (separate analyses)
Gunsoy 2012, breast	Markov models for screening parameter and overdiagnosis estimation	UK women 40-49, Not provided	Annual screening mammography from ages 40-49	• Parameter estimation model: all data from Age RCT (1991-2010) • Overdiagnosis model: • Incidence: England and Wales Office of National Statistics 2008 (women 40-54), control arm of Age RCT (women 40-49) • Mortality: England and Wales Office of National Statistics 2008 • Mean sojourn time: estimated from literature review (in addition to parameter estimation model)

Study, Cancer Type	**A.** Includes DCIS? **B.** Includes Competing Mortality? **C.** External Validation? **D.** Reports Preferred Outcome?	Magnitude of Overdiagnosis (95% CI)	Sensitivity Analyses?	Results of Sensitivity Analyses	Conclusions
De Gelder 2011 (Epi Reviews), breast	A.Yes B.Yes C.No D.Yes	• Implementation: 22.1-67.4% • Extension: 15.4-30.4% • Steady state: 8.9-15.2%	Not performed	N/A	Overdiagnosis depends on year calculated (earlier in screening programs find higher rates) and denominator
De Gelder 2011 (Prev Med), breast	A.Yes B.Yes C.No D.Yes	• Screen film mammography: 7.2% • Digital mammography: 8.2%	Considered two alternative models: • Progressive model, where all tumors pass through screen-detectable DCIS phase and none regress • Nonprogressive model, where no tumors pass through screen-detectable DCIS stage and majority of preclinical DCIS regresses	• Progressive model: film- 4.6%, digital-5.0% • Nonprogressive model: film-19.2%, digital-25.2%	Modeling predicted that digital mammography screening would further reduce breast cancer mortality by 4.4% at a 21% increased overdiagnosis rate. Outcomes are sensitive to underlying assumptions on natural history of DCIS.
Duffy 2005, breast	A.Yes B.No C.No D.Yes	• Swedish: 1st screen 3.1% (0.1-10.9), 2nd screen 0.3% (0.1-1), 3rd screen 0.3% (0.1-1). • Gothenburg: 1st screen 4.2% (0.0-28.8), 2nd screen 0.3% (0.0-2.0), 3rd screen 0.3% (0.0-2.0)	Not performed	• N/A	Overdiagnosis in mammography screening is a minor phenomenon.
Gunsoy 2012, breast	A.Yes B.Yes C.No D.Yes	0.70%	Varied sensitivity and mean sojourn time in combinations: high MST, low MST, high sens, low sens, high MST with low sens, high MST with high sens	0.5% to 2.9% Increasing MST had a greater impact than did increasing sensitivity.	In UK women 40-49, a small proportion of breast cancers were overdiagnosed due to screening.

Study, Cancer Type	Model Name/Type	Modeled Population, Time Period	Screening Schedule/ Details	Data Sources
Martinez-Alonso 2010, breast	Poisson regression age-cohort model and probabilistic model	Catalan women ages 25-84, 1980-2004	Biennial mammography from ages 50-69	• Incidence: Girona Cancer Registry (1980-9, 1991-2004) and International Agency for Research on Cancer (IARC) registry for Tarragona (1983-97) • Incidence model includes completed fertility rate where relative risk of breast cancer 0.85 per child born based on literature review • Mean sojourn time ranges from 2-4, depending on age • Mammography sensitivity ranges from 0.35 to 0.8, depending on age
Olsen 2006, breast	Not provided	Copenhagen women 50-69, Not provided	Biennial mammography from ages 50-69	• Incidence: Danish Cancer Registry and Danish Breast Cancer Cooperative Group • Interval cancers: Danish Cancer Registry and Central Population Registry • Screening patterns: Copenhagen mammography database • Mammography sensitivity: set at 100% in primary analysis
Seigneurin 2012, breast	Approximate Bayesian computation analysis with a stochastic simulation model	Women 50-69 in Isere, France, 1991-2006	Not specified	• Incidence: French population-based study by Seigneurin 2009 • Screening patterns: surveys for 1991-2 and 2005-6, assumed that screening mammography dissemination followed linear trend from 1991-2006, opportunistic screening assumed to be a maximum of 40% • Mean sojourn time: 2-4 years based on literature review • Mammography sensitivity: mean assumed to be 90%

Study, Cancer Type	**A.** Includes DCIS? **B.** Includes Competing Mortality? **C.** External Validation? **D.** Reports Preferred Outcome?	Magnitude of Overdiagnosis (95% CI)	Sensitivity Analyses?	Results of Sensitivity Analyses	Conclusions
Martinez-Alonso 2010, breast	A.No B.Yes C.No D.No	• 1935 birth cohort: 0.4% (-8.8 to 12.2) • 1940: 23.3% (9.1-43.4%) • 1945: 30.6% (12.7-57.6%) • 1950: 46.6% (22.7-85.2%) Calculated as excess divided by expected incidence	Varied mean sojourn time to be 1 and 5 Set mammography sensitivity to 0.9	Varying MST from 1 to 5 changed overdiagnosis from 51.1 to 18.3%. Increasing sensitivity did not greatly affect estimates.	Results support the existence of overdiagnosis in Catalonia attributable to mammography.
Olsen 2006, breast	A.Yes B.No C.No D.Yes	• 1st screen 7.8% (0.3-27.5) • 2nd screen: 0.5% (0.01-2.2%)	Varied sensitivity to be 80% and 90%	• Sensitivity 90%: first screen 8.3% (0.3-29.1), second screen 0.6% (0.02-2.4) • Sensitivity 80%: first screen 8.6% (0.3-28.4), second screen 0.6% (0.02-2.3)	A modest overdiagnosis was estimated from the Copenhagen screening program, almost exclusively from first screen.
Seigneurin 2012, breast	A.Yes B.No C.No D.Yes	• DCIS: 31.9% (2.9-62.3) • Invasive cancer: 3.3% (0.7-6.5)	• Varied sojourn time distribution and parameters • Excluded the 1991-5 study period which had more prevalent screens at the beginning of the screening program • Varied distribution of non-progressive carcinoma in-situ	• Varying sojourn time: overdiagnosis of DCIS varied from 17.3% (7.8-28.5) to 51.7% (15.4-81.8), invasive cancer from 0.0% (0.0-0.1) to 8.9% (0.5-24.0) • Excluding 1991-5 study period did not greatly affect results. • Uniform prior distribution of non-progressive CIS greatly altered point estimate of overdiagnosis proportion (7.5% instead of 28.0%) but did not improve precision (95% CI 0-45.1%)	Overdiagnosis from the detection of non-progressive disease by screening mammography was limited in 1991-2006 in Isere.

Study, Cancer Type	Model Name/Type	Modeled Population, Time Period	Screening Schedule/ Details	Data Sources
Hazelton 2012, lung	Longitudinal multistage model	Current or former heavy smokers with < 5 years asbestos exposure, Not provided	Low dose CT	• Incidence: Carotene and Retinol Efficacy Trial (calibration) • Mortality: CARET (calibration) • CT screening data: Pittsburgh Lung Screening Study (calibration) • Calibrated model applied to NYU Biomarker Center trial and Moffitt Cancer Center trial • Literature search for number of stem cells in lungs and density of malignant stem cells per unit volume • Data on pulmonary tumor size at incidence and death from Geddes 1979 review
Pinsky 2004, lung	Early- and late-stage convolution model	Men 50-75 with smoking patterns similar to those of Mayo Lung Trial participants, Not provided	CXR and sputum cytology annually from ages 50-75	• All data: Mayo Lung Cancer Screening Trial (prevalence screen and screening arm)
Luo 2012, colon	Not provided	Hypothetical cohorts with ages 40, 50 or 60 at first screening, Not provided	5 annual or 3 biennial FOBT tests, starting at ages 40, 50 or 60	• Incidence: Minnesota Colon Cancer Control study (1976-1982) • Mortality: actuarial life tables from US SSA

Study, Cancer Type	**A.** Includes DCIS? **B.** Includes Competing Mortality? **C.** External Validation? **D.** Reports Preferred Outcome?	Magnitude of Overdiagnosis (95% CI)	Sensitivity Analyses?	Results of Sensitivity Analyses	Conclusions
Hazelton 2012, lung	A.N/A B.Yes C.No D.Yes	• Detection of indolent nodules is 7.0% (4.9-11.7%) for men and 33.0% (26.9-36.9%) for women. • Additional 2.2% (2.0-2.4%) for women and 7.1% (6.7-8.0%) for men are overdiagnosed	Not performed	N/A	Significant gender differences in progression of lung cancer, where female patients have much higher rates of indolent cancers.
Pinsky 2004, lung	A.N/A B.Yes C.No D.Yes	13-17%	Not performed	N/A	
Luo 2012, colon	A.N/A B.Yes C.No D.Yes	• Females, ages 40, 50, 60: 6.50%, 6.65%, 7.33% (95% CI 2.56- 20.49%) • Males, ages 40, 50, 60: 5.61%, 6.15%, 7.48% (95% CI 1.92-44.69%)	Not performed	N/A	Probability of overdiagnosis among screen-detected cases is not as high as previously thought.

Appendix Table 3: Criteria for Evaluating Risk of Bias and Strength of Evidence for Modeling Studies

Study	Cancer Type	Assumptions Transparent and Clearly Stated	Assumptions Backed with Evidence	Probability for Biases in Model Data	Sensitivity Analyses	External Validation	**Overall Risk of Bias**	**Overall External Validity**	**Precision**
		G/F/P	*G/F/P*	*H/M/L/CD*	*G/F/P*	*Y/N*	***H/M/L***	***G/F/P***	***G/F/P/CD***
Davidov 2004	prostate	F	F	CD	F	N	**M**	**G**	**CD**
Draisma 2009	prostate	F	F	CD	P	N	**H**	**G**	**CD**
Gulati 2013	prostate	G	F	CD	F	N	**M**	**G**	**CD**
Heijnsdijk 2009	prostate	F	F	CD	P	N	**H**	**F**	**CD**
McGregor 1998	prostate	F	F	CD	P	N	**H**	**F**	**CD**
Pashayan 2009	prostate	F	F	CD	P	N	**H**	**F**	**G**
Tsodikov 2006	prostate	F	F	CD	P	N	**H**	**G**	**CD**
Wu 2012	prostate	F	F	CD	P	N	**H**	**F**	**F**
De Gelder 2011 (ER)	breast	F	F	CD	P	N	**H**	**F**	**P**
De Gelder 2011 (PM)	breast	F	F	CD	P	N	**H**	**F**	**CD**
Duffy 2005	breast	F	G	CD	P	N	**H**	**F**	**F**
Gunsoy 2012	breast	F	F	CD	F	N	**M**	**F**	**CD**
Martinez-Alonso 2010	breast	F	F	CD	F	N	**M**	**F**	**P**
Olsen 2006	breast	F	F	CD	P	N	**H**	**F**	**F**
Seigneurin 2012	breast	F	F	CD	F	N	**M**	**F**	**P**
Hazelton 2012	lung	F	F	CD	P	N	**H**	**G**	**G**
Pinsky 2004	lung	F	F	CD	P	N	**H**	**F**	**CD**
Luo 2012	colon	P	F	CD	P	N	**H**	**F**	**P**

Criteria used in strength of evidence evaluation are bolded. Abbreviations: G, good; F, fair; P, poor; CD, cannot determine; H, high; M, moderate; L, low; ER, Epidemiologic Reviews; PM, Preventive Medicine.

Appendix Table 4: Evidence Table of Pathologic and Imaging Studies

Study, Cancer Type, Study Period	Study Population	Comparison Group?	Screening Situation	Pathologic/Behavioral Outcome
Dominioni 2012, Lung, 1997-2011	21 Italians with screen-detected cancer during study period, with >10 pack-year smoking history, 45-75, fit for thoracotomy, asymptomatic	No	Baseline CXR with annual repeat screen for 4 years	Volume doubling time (VDT). Overdiagnosis defined as VDT>300 days
Lindell 2007, Lung, 1999-2004	20+ pack-year smoking history Cancers studied were 61 tumors in 59 US patients, 24 men/37 women, ages 53-79 (mean 65)	No	Chest CT performed at baseline and every 12 months thereafter for 5 years	Volume doubling time. Overdiagnosis defined as VDT >400 days
Sobue 1992, Lung, 1976-1989	Patients screened in Japanese Lung Cancer Screening Research Group from 1976-81 who did not undergo surgical treatment (42 cases)	Symptom-detected controls matched by age, sex, and year of diagnosis (27 cases)	Chest X-ray	Overdiagnosis defined as dying from a cause other than lung cancer in patients diagnosed with clinical stage 1 disease (all study patients)
Sone 2007, Lung, 1996-1998	45 patients ages 40-74 (smokers and non-smokers) with lung cancer detected in CT screening program for whom repeat CT images were available, in rural Japan from 1996-1998	No	Low-dose CT scan, with immediate work-up for suspicious lesions and 3-month delayed workup for nodules <3mm	Calculated patients' expected time of death by calculating age when tumor would reach 30 mm, based on tumor size at detection and VDT, then adding 2 years. If expected age of death was higher than average life span in Japan (78.64 for males, 85.59 for females), cancer was considered overdiagnosis.
Veronesi 2012 Lung 2004-2010	120 Italian patients with incident lung cancer detected in CT screening program (COSMOS) study, all with > 20 pack-years, 70% men, ages 50+	No	annual low-dose CT scan for 5 consecutive years, without notification for nodules <5mm	Cancer with VDT <400 days defined as fast-growing, VDT between 400-599 days defined as slow-growing, and VDT>600 days defined as indolent. Considered VDT>400 days to be overdiagnosis

Study, Cancer Type	Outcome Assessment	Results	Preferred Outcome?	Magnitude of Overdiagnosis	Conclusions
Dominioni 2012, Lung	VDT based on tumor size measurements from sequential CXRs, read by one of 3 senior radiologists. If tumor not evident retrospectively, assigned dimension of 6mm.	Median VDT 80 days (range 44-318). Only one cancer had VDT>300	No	"Minimal"	Low median volume doubling time suggests that overdiagnosis was minimal.
Lindell 2007, Lung	CTs reviewed retrospectively by one radiologist who measured 2 diameters with electronic calipers to calculate VDT with modified Schwartz equation for all tumors with at least 2 CT scans available.	4 tumors became smaller at some point during the study. 13/48 tumors (27%) had VDT longer than 400 days. 11/13 (85%) with VDT>400 days in women.	No	27%	Overdiagnosis, especially in women, may be a substantial concern in lung cancer screening.
Sobue 1992, Lung	Death certificates and medical records reviewed	20% of screen-detected and 19% of symptom-detected patients died from cause other than lung cancer	No	"Minimal"	Overdiagnosis bias would be minimal in screen-detected lung cancer cases detected by CXR.
Sone 2007, Lung	Tumor growth assessed on high resolution CT, based on largest tumor diameter. VDT calculated with Schwartz formula	13.3% (6 of 45 cases) overall were considered overdiagnosed (17.9% of male patients, 5.9% of female patients, and 40% of non-solid lesions)	No	13.30%	Estimated rate of possible overdiagnosis was 13% in total.
Veronesi 2012 Lung	VDT calculated using Lesion Management Solutions-lung software. If software unable to estimate volume, electronic calipers used to measure largest axial diameter and VDT calculated with formula from Yankelevitz 2000. To estimate VDT for a new cancer, nodule assumed to be 2mm the previous year.	58.3% (49.0-67.3%) of incident cases were fast growing and 25.8% (18.3-34.6%) were slow-growing or indolent. 15.0% (9.1-22.7%) slow-growing and 10.8% (5.9-17.8%) indolent	No	25.8% (18.3-34.6%)	VDT analysis suggests at least 75% of cases were aggressive, downsizing problem of overdiagnosis. Among the 25% of slow-growing or indolent cases, many are likely to be overdiagnosed.

Study, Cancer Type, Study Period	Study Population	Comparison Group?	Screening Situation	Pathologic/Behavioral Outcome
Yankelevitz 2003, Lung, Not provided	Patients diagnosed with stage 1 lung cancer from the screening arms of the NY Lung Cancer Detection Project at Memorial-Sloan Kettering (MSK, 43 cases) and Mayo Lung Project (MLP, 44 cases).	No	MSK: sputum cytology every 4 months with annual CXR MLP: sputum cytology and CXR every 4 months	Volume doubling time. VDT> 400 days considered overdiagnosis
Graif 2007, Prostate, 1989-2005	2126 US men with clinical stage T1c prostate cancer treated with RRP.	No	screening situation differed by "era" of study. Era 1: PSA threshold 4ng/ml, with quadrant biopsy for abnormal results. Era 2: PSA threshold 2.5 ng/ml, sextant biopsy for abnormal. Era 3: men who were referred for RRP, no standard biopsy strategy.	Overdiagnosis defined as tumor volume <0.5 cm^3, Gleason <7, organ-confined disease in RRP specimen with clear surgical margins
Pelzer 2008, Prostate, 1999-2006	997 patients with prostate cancer undergoing RRP from Tyrol, Austria and nearby areas. Patients treated with hormonal, chemo or radiotherapy were excluded. 806 of these were Tyrolean screening volunteers ("screened group")	191 patients were referred for RRP from outside Tyrol ("referral group")	PSA testing with elevated PSA leading to referral for biopsy with 10 cores until 2000 and 15 from 2000-2006	Overdiagnosis defined as Gleason <7, pathologic stage of pT2a and negative surgical margins

Study, Cancer Type	Outcome Assessment	Results	Preferred Outcome?	Magnitude of Overdiagnosis	Conclusions
Yankelevitz 2003, Lung	• MSK: Tumor length and width obtained for at least 2 time points or with documented invisibility at prior time. Visibility threshold calculated to be 0.6cm. • MLP: Used frequency distribution of tumor dimension and disease stage at time of diagnosis and number of months tumor visible in retrospect, assuming stage 1 malignancies were the smallest (no dimension and stage information available for individual tumors)	• MSK: 3/43 (7%) of cases had VDT>400 days • MLP: 1/44 (2%) of cases had VDT>400 days • 5% of overall cases had VDT>400 days	No	5%	The hypothesis that early-stage lung tumors on chest radiography during lung cancer screening may frequently be overdiagnosed, indolent cases needs to be rejected.
Graif 2007, Prostate	No info on RRP procedures. Visual estimation of percent cancer in prostate gland used in majority of cases, but some cases used grid morphometric method. Tumor volume calculated from % cancer multiplied by volume of RRP specimen	4.5% of men undergoing RRP during study period met criteria for overdiagnosis compared with 27% meeting criteria for underdiagnosis. 5-year progression-free survival of men meeting overdiagnosis criteria was 100%.	No	4.50%	Underdiagnosis of prostate cancer continues to occur more frequently than overdiagnosis
Pelzer 2008, Prostate	No info on RRP procedures. Each biopsy core reviewed by one pathologist	Overdiagnosis was 16.8% in the screening group and 7.9% in the referral group	No	• Screening group: 16.8% • Referral group: 7.9%	Other reported estimates of overdiagnosis due to screening are exaggerated, and underdiagnosis should be the primary concern.

Appendix Table 5: Criteria for Evaluating Risk of Bias and Strength of Evidence for Pathologic and Imaging Studies

Study	Cancer Type	Probability for Selection Bias and Confounding	Probability for Measurement Bias	**Overall Risk of Bias**	Explanation of Link Between Pathologic/ Imaging Characteristic and Cancer Progression	**Directness**	**External Validity**	**Precision**
		H/M/L	*H/M/L*	***H/M/L***	*G/F/P*	***G/F/P***	***G/F/P***	***G/F/P/CD***
Dominioni 2012	lung	L	M	**M**	P	**P**	**F**	**P**
Lindell 2007	lung	M	M	**M**	P	**P**	**F**	**CD**
Sobue 1992	lung	H	H	**H**	NA	**G**	**P**	**P**
Sone 2007	lung	H	M	**H**	P	**P**	**F**	**CD**
Veronesi 2012	lung	L	M	**M**	F	**P**	**F**	**F**
Yankelevitz 2003	lung	H	H	**H**	P	**P**	**F**	**CD**
Graif 2007	prostate	H	H	**H**	P	**P**	**F**	**CD**
Pelzer 2008	prostate	H	M	**H**	P	**P**	**F**	**CD**

Criteria used in strength of evidence evaluation are bolded. Abbreviations: H, high; M, moderate; L, low; G, good; F, fair; P, poor; CD, cannot determine

Appendix Table 6: Evidence Table of Ecologic and Cohort Studies

Study; Cancer Type; Study Design	Study Population; Time Period	Description of Methods	Reference Population	Screening Schedule
Bleyer 2012; Breast; Ecologic	US women 40+ 1976-2008	Using stage-specific incidence from SEER, calculated excess cases of early-stage cancer and reduction in number of late-stage cancers by projecting baseline incidence pre-screening and comparing to incidence each subsequent year. Excess of early-stage diagnoses not balanced by reduction in late-stage diagnoses considered overdiagnosed.	baseline incidence of breast cancer determined from SEER 1976-1978	trend data on proportion of women undergoing mammography estimated from NHIS
Duffy 2010; Breast; Cohort and Ecologic	Swedish 2-County trial: women 50-69; 1977-88 (screening), 1989-98 (post-screening) UK women 47-73; 1974-89 (pre-screening), 1989-2003 (screening)	Swedish: estimated expected incidence based on incidence trend in control group during first 6 years of trial. Used several equations to calculate rates of overdiagnosis. UK: Determined expected incidence in 1989-2003 based on trends from pre-screening period. Took into account non-linear trends by dividing expected numbers by relative excess for <45 age group (unscreened). Overdiagnosis calculated as excess of observed cases in ages 45-64 minus deficit in 65+	Swedish: control arm of trial before screening offered UK: women from 1974-88, before introduction of screening	Swedish: women 40-49 offered mammography every 24 months, 50-74 every 33 months. UK: mammography every 3 years
Falk 2013 Breast; Cohort	Norwegian women 50-69; 1995-2009	Incidence rate ratios calculated for women attending screening compared to non-attenders, stratified by county, calendar year and age. Reference rates determined from reference population, multiplied by IRR of screening participants to find excess of cancer diagnoses during screening program and deficit afterwards. Overdiagnosis calculated as excess not compensated by deficit in post-screening.	Reference rates by age based on observed rates of invasive breast cancer 1980-4, chosen to minimize influence of HRT, standardized to life table population from Statistics Norway based on 2010 morality rates.	biennial screening with two-view mammography from ages 50-69

Study; Cancer type	Screening in Reference Population	Management of Potential Confounders	Management of Lead Time	Calculation of Overdiagnosis	A. Includes DCIS? B. Reports preferred outcome?	Magnitude of Overdiagnosis (95% CI)
Bleyer 2012; Breast	Few cases of DCIS detected during 1976-8	Adjusted for HRT use by truncating estimate of observed incidence from 1990-2005 if greater than incidence from 2006-2008 Adjusted for baseline increasing incidence based on the increase in incidence of women <40 (0.25% per year)	Long follow-up/ steady-state screening	(excess cases)/ (observed cases) during screening period	A. Yes B. No	In 2008, 31% of all breast cancers were overdiagnosed.
Duffy 2010; Breast	Not discussed	UK: Adjusted for baseline increase in incidence, including non-linear increases	Swedish: excluded prevalence screen in calculations UK: unclear. Screening program being expanded to ages 47-73 at time of study	Based on complex calculation	A. Unclear B. No	Swedish: 12% of all cancers are overdiagnosed UK: 2.3 cases overdiagnosed per 1000 screened for 20 years
Falk 2013; Breast	Formal screening program began in 1995	Calculation of IRRs stratified by county, calendar year, and age. Authors suggest that other confounders like use of HRT almost entirely accounted for by these variables.	10-year follow-up post-screening program	(excess cases)/ (expected cases) during screening	A. Yes B. No	19.4% (11.8-27.0)

Study; Cancer type	Considerations of Uncertainty/ Sensitivity Analyses	Results of Sensitivity Analyses	Conclusions
Bleyer 2012; Breast	Assumed baseline incidence increasing by 0.5% per year (twice that seen in women <40) ("extreme" assumption) and with revision of baseline incidence of late-stage cancer using highest incidence observed in data ("very extreme" assumption)	"Extreme" assumption: 26% of all cases overdiagnosed "Very extreme" assumption: 22% of all cases overdiagnosed	Mammography has substantially increased diagnosed of early-stage cancer while minimally reducing late-stage diagnoses, suggesting substantial overdiagnosis accounting for a third of all diagnosed cases.
Duffy 2010; Breast	Not performed		Worthwhile benefit of mammography in terms of lives saved that significantly exceeds any harm in the form of overdiagnosis that may occur.
Falk 2010; Breast	Used two other reference populations: a modeled population of 40-year-old women in 1993-95, and a historical cohort of women born in 1903-1907 Also explored if there were differences in breast cancer incidence among attending and non-attending women by comparing IRRs of women not yet invited to screening to those not attending.	Modeled reference: 19.6% (12.1-27.1) historical reference: 16.5% (9.1-23.9) No differences between non-attenders and women not yet invited below age 55, but above 55 non-attenders more likely to have higher incidence of DCIS and invasive cancer	Results highlight the need for individual data with longitudinal screening history and long-term follow-up as a basis for estimating overdiagnosis.

Study; Cancer Type; Study Design	Study Population; Time Period	Description of Methods	Reference Population	Screening Schedule
Hellquist 2012; Breast; Ecologic	Swedish women 40-49; 1986-2005	Calculated risk ratio for breast cancer among women in counties offered screening compared to those not offered screening, excluding cancers detected during the prevalence screens and then adjusting for "trend bias" with a calculation that includes increasing baseline incidence of breast cancer and lead time	Swedish women 40-49 in counties not invited to screening. In some counties only certain years within 1986-2005 were selected to achieve similar follow-up in study and control groups	Not provided
Jorgenson 2009 (BMJ); Breast; Ecologic (systematic review)	UK: 50-64 (1993-1999) Manitoba, CA: 50-69 (1995-2005) New South Wales: 50-69 (1996-2002) Sweden: 50-69 (1998-2006) Norway: 50-69 (2000-2006)	Systematic review to identify incidence trends where incidence data on screening-age and older women was available 7 years before and after screening fully implemented. If compensatory drop in post-screening ages absent, calculated rate ratio between observed incidence for last observation year (determined by linear regression over time period when screening fully implemented) to expected incidence calculated from pre-screening incidence trend. If compensatory drop present, calculated the size of the drop as a rate ratio, and then determined the absolute deficit of breast cancer cases per 100,000 women. Calculated the percentage of excess cases uncompensated by drop in post-screening ages. Included DCIS when not reported by assuming that it would contribute 10% of the diagnoses	Used linear regression to calculate expected incidence in study population in absence of screening from pre-screening trends. Pre-screening period defined as 1971-84 in UK, 1970-78 in Manitoba, 1972-87 in New South Wales, 1971-85 in Sweden, 1980-94 in Norway	Formal mammography screening programs by country
Jorgenson 2009 (BMC) Breast; Ecologic	Danish women 50-69 in Copenhagen and Funen; 1991-2003	Compared incidence rates in screened and unscreened areas of Denmark, using Poisson regression to adjust for age and geographical differences in incidence in the pre-screening period (1971-1991). Quantified excess of cancers detected in screened counties not balanced by decrease in cancers in women post-screening age.	Other counties in Denmark without screening programs	Biennial mammo-graphy, 63-88% attendance
Junod 2011; Breast; Ecologic	French women 50-64 and 65-79; 1995-2005	Incidence compared for middle-aged and elderly women from 1995-2005 to age-matched reference cohorts from 1980-90. Incidence change attributable to HRT use, alcohol use, and obesity was removed, and the remaining incidence difference was considered overdiagnosis	Age-matched historical birth cohorts from 1980-1990	Mammo-graphy from 50-69, extended to 74

Study; Cancer type	Screening in Reference Population	Management of Potential Confounders	Management of Lead Time	Calculation of over-diagnosis	A. Includes DCIS? B. Reports preferred outcome?	Magnitude of Overdiagnosis (95% CI)
						1% (-6.0-8.0%)
Hellquist 2012; Breast	Certain counties excluded if screening formally introduced.	Adjusted for differences in baseline incidence between study and control groups	statistical adjustment for lead time= 1.2 years	(excess cases)/ (expected cases) during screening	A. Yes B. No	Crude estimate (including prevalence screening w/o lead time adjustment): 16% (9-23%)
Jorgenson 2009 (BMJ); Breast	Looked for abrupt increase in DCIS in pre-screening years to indicate opportunistic screening and chose a different reference if present	Accounted for baseline increasing incidence with linear regression. looked at incidence in women too young to be screened to see if any incidence growth not attributable to screening present	Up to 7 year follow-up post-screening	(excess cases)/ (expected cases) during screening	A. Yes B. No	UK: 57% (53-61%) Manitoba: 44% (25-65%) NSW: 53% (44-63%) Sweden: 46% (40-52%) Norway: 52% (36-70%) Meta-analysis: 52% (46-58%)
Jorgenson 2009 (BMC); Breast	Expect that <10% of women participated in opportunistic screening	Used poisson regression to adjust for differences in age and geographical variation in pre-screening incidence	Up to 10 to 12 years follow-up post-screening	(excess cases)/ (expected cases) during screening	A. Yes B. No	33%
Junod 2011; Breast	Opportunistic screening is substantial	Subtracted expected changes in incidence due to HRT, alcohol, and obesity. Did not account for baseline increase in incidence between the historical and contemporary cohorts	Unclear	(excess cases)/ (expected cases) during screening	A. No B. No	Ages 50-64: 76% (67-85%) Ages 65-79: 23% (15-31%)

Study; Cancer type	Considerations of Uncertainty/ Sensitivity Analyses	Results of Sensitivity Analyses	Conclusions
Hellquist 2012; Breast	Varied lead time from 1.0-1.5 years	Lead time 1 year: 2% Lead time 1.5 years: -1%	Found no significant overdiagnosis for women 40-49 in Swedish service screening mammography program.
Jorgenson 2009 (BMJ); Breast	Not performed		The increase in incidence of breast cancer was closely related to introduction of screening and little was compensated for by a drop in incidence in previously-screened women. One in three breast cancers in a population offered screening is overdiagnosed.
Jorgenson 2009 (BMC); Breast	Not performed		One in four breast cancers diagnosed in the screened age group in Danish women is overdiagnosed. Estimate is lower than for comparable countries because of lower uptake, recall rates and diagnosis of CIS.
Junod 2011; Breast	Not performed		Substantial increase in breast cancer incidence in France without a corresponding increase in mortality, which largely reflects an increase in overdiagnosis

Study; Cancer Type; Study Design	Study Population; Time Period	Description of Methods	Reference Population	Screening Schedule
Kalager 2012; Breast; Ecologic	Norwegian women 50-79; 1996-2005 (screening) 1986-2005 (pre-screening)	Calculated incidence rate ratios (current screening/historical screening)/(current nonscreening/historical nonscreening) including women up to age 79, expecting to see a compensatory deficit in incidence in this age group. Also analyzed just one county that had 10 years of follow-up after screening started.	Contemporary control group in counties not offered screening (1996-2005) and a historical non-screening group (1986-1995), and a historical group offered screening (1986-1995)	2-view mammography biennially from 50-69, 77% participation rate
Morrell 2010; Breast; Ecologic	New South Wales women 50-69; 1991-2001 (screening) 1972-1990 (pre-screening)	Calculated expected incidence rates with two methods, by interpolating incidence of women 50-69 from women <40 and >80 not offered screening, and by assuming continuation of incidence trend in women 50-69 from pre-screening period. Adjusted expected incidence for HRT use, nulliparity, obesity rates. Compared expected to observed rates.	Used linear regression to model expected incidence in women 50-69 based pre-screening trends and based on women<40 and >80 not offered screening	Biennial mammography, 60% participation
Njor 2013; Breast Cohort	Women 56-79 in Copenhagen and 59-78 in Funen; 1991/1993-2009	Compared cumulative incidence in screening regions to expected incidence in absence of screening, including women up to 8 years post-screening in overall incidence estimates. Expected incidence determined from historical screening-region control group adjusted for the change in incidence from historical to contemporary non-screening control groups, and using an interaction term to account for differences in baseline incidence trends between regions (interaction between region and period).	3 control groups: a contemporary control group from non-screening regions, a historical control group from current screening regions, and a historical control from current non-screening regions	Biennial mammography ages 50-69, 71-84% first round participation
Paci 2006; Breast Cohort	Women 50-74 in Italy; 10-year period between 1986-2006, dates vary by region	Used Poisson regression to predict incidence rates in six screening regions that had at least 5 years of screening implementation based on incidence in the pre-screening period. Corrected the observed cases statistically for lead time and compared to expected	Modeled expected incidence from pre-screening trends	Not provided

Study; Cancer type	Screening in Reference Population	Management of Potential Confounders	Management of Lead Time	Calculation of Overdiagnosis	A. Includes DCIS? B. Reports preferred outcome?	Magnitude of Overdiagnosis (95% CI)
Kalager 2012; Breast	Not provided	Current screening and non-screening incidences compared to historical screening and non-screening incidences to account for temporal trends	Including women up to 79 years in incidence estimate, with up to 10 years follow-up post-screening	(excess cases)/ (observed cases) during screening, including women up to 79	A. No B. No	Entire country: 25% (19-31%) County w/10yrs follow-up: 18% (11-24%)
Morrell 2010; Breast	Not provided	Population attributable fractions to HRT use, obesity, and nulliparity were calculated and combined, assuming the lowest level of risk factor before the study period and the highest during the study period, and expected incidence was adjusted	Statistical adjustment for lead time of 2.5 and 5 years	(excess cases)/ (expected cases) during screening	A. No B. No	Interpolation model: 42% Extrapolation model: 30% (5-year lead time)
Njor 2013; Breast	In 2000, 3% of women 50-69 in non-screening regions had screening or diagnostic mammograms	Controlled for differing temporal trends in incidence between screening and non-screening regions with interaction term in regression analysis	Up to 8 years follow-up post-screening	(excess cases)/ (expected cases) during screening and 8-years post-screening	A. Yes B. No	Copenhagen: 6% (-10 to 25%) Funen: 1% (-7 to 10%) Pooled: 2.3% (-3 to 8%)
Paci 2006; Breast	Not provided	Adjusted for age in poisson regression model	statistical adjustment, using MST of 3.7 years for ages 50-59 and 4.2 years for 60-74	(excess cases)/ (expected cases) during screening	A. Yes B. No	36.2% (34-39%) before adjustment for lead time 4.6% (2-7%) after adjustment for lead time

Study; Cancer type	Considerations of Uncertainty/ Sensitivity Analyses	Results of Sensitivity Analyses	Conclusions
Kalager 2012; Breast	Second analysis to account for lead time in a different way, excluded all cases detected at first screening round and compared incidence in current screening group with women 2 and 5 years older in historical screening groups	Lead time 5 years: 15% (8-23%) Lead time 2 years: 20% (13-28%)	Mammography screening entails a substantial amount of overdiagnosis
Morrell 2010; Breast	Also used 2.5 year lead time	Interpolation model: 51% Extrapolation model: 36% (lead time 2.5yrs)	Overdiagnosis of invasive breast cancer attributable to mammography is substantial, and estimates similar to recent estimates from other screening programs
Njor 2013; Breast	Performed analyses with only women with >8 years follow-up, and corrected the analyses to estimate overdiagnosis in participants only	8+ years follow-up: Copenhagen: 3.4% (-14 to 25%) Funen: 0.7% (-8 to 12%) Participants only: Copenhagen: 8% Funen: 2%	Overdiagnosis most likely 2.3%, and study indicates that at least 8 years of follow-up were needed to compensate for the excess during screening.
Paci 2006; Breast	performed sensitivity estimate using 95% CI upper limit of lead time (4.8 years)	2.80%	Remaining excess of cancers after individual correction for lead time was <5%.

Study; Cancer Type; Study Design	Study Population; Time Period	Description of Methods	Reference Population	Screening Schedule
Peeters 1989; Breast; Ecologic	Women in Nijmegen, Netherlands; 1975-1986	Compared incidence during six screening rounds among women in Nijmegen to neighboring city with no mass screening program.	Women in neighboring city, Arnhem	Biennial mammography for women 35+, 65-85% participation
Puliti 2009; Breast; Cohort	Women 60-69 in Florence; 1990-2005	Calculated expected incidence based on pre-screening incidence trends in the screened cohort using Poisson regression and forcing the parameter for annual percentage change to 1.2%. Compared expected incidence with observed incidence	Modeled expected incidence from pre-screening trends	Biennial mammography for women 50-69, 60-70% participation
Puliti 2012; Breast; Cohort	Women 60-69 in Florence; 1991-2007	Compared incidence rates in screening attenders and non-attenders using poisson regression models adjusted for 5-year age group, marital status and deprivation class. Incidence rate ratio for women 60-69 determined overdiagnosis	Screening non-attenders in the first two screening rounds	Biennial mammography for women 50-69, 56-70% participation
Svendsen 2006; Breast; Ecologic and Cohort	Women 50-69 in Copenhagen and Fyn, Denmark; 1991-2001 (screening), 1979-1990 (pre-screening)	Compared age-standardized incidence rates in Copenhagen and Fyn to the rest of Denmark and in the pre-screening period of the screening regions. Calculated confidence intervals for the pre-screening incidence rates with quadratic and linear regression.	Women in the rest of Denmark without organized screening programs, and pre-screening trends in the screening counties	Biennial mammography for women 50-69
Zahl 2004; Breast; Ecologic	Women 50-74 in Akershus, Oslo, Rogaland, Hordaland, Norway; 1991-5 (pre-screening), 1995-2000 (screening) Women 50-79 in Sweden; 1971-85 (pre-screening), 1986-2000 (screening)	Norway: Used Poisson regression to compare incidence rates during screening in 2000 to rates in the same regions in 1991, after demonstrating no increase in incidence during the pre-screening period 1991-5 Sweden: Used Poisson regression to compare incidence in entire country during screening to pre-screening period	Women in screening regions in pre-screening periods	Biennial mammography 50-69, 75% average participation in both countries

Study; Cancer Type	Screening in Reference Population	Management of Potential Confounders	Management of Lead Time	Calculation of Overdiagnosis	A. Includes DCIS? B. Reports preferred outcome?	Magnitude of Overdiagnosis (95% CI)
Peeters 1989; Breast	Not provided	Compared breast cancer incidence in Nijmegen and Arnhem pre-screening and mortality from 1970-79 and found them to not be statistically different, thus concluded Arnhem was an appropriate control	Did not	(excess cases)/ (expected cases) during screening	A. Yes B. No	overall: 11% 1975-79: 30% 1972-82: 3% 1983-86: 1%
Puliti 2009; Breast	Not provided	Adjusted for age in poisson regression model	5-10 year follow-up post-screening for the oldest cohorts (ages 60-69)	(excess cases)/ (expected cases) during screening and 5-years post-screening	A. Yes B. No	1% (-5 to 7%)
Puliti 2012; Breast	16% of control group attended screening during study period	adjusted for age, marital status, and deprivation index (representing area-level socioeconomic status)	5-14 year follow-up post-screening for women 60-69	(excess cases)/ (expected cases) during screening and 5-14 years post-screening	A. Yes B. No	10% (-2 to 23%)
Svendsen 2006; Breast	Opportunistic screening in Denmark is minimal	Age-standardized incidence rates	Did not	Not calculated	A. No B. No	None (screening incidence w/in 95% CI of pre-screening trends)
Zahl 2004; Breast	Norway: not discussed Sweden: in some regions women offered screening at ages 70-74	Adjusted for age, demonstrated no increase in incidence during pre-screening period needing adjustment	Norway: up to 4 years of follow-up post-screening in ages 70-74 Sweden: up to 14 years of follow-up post-screening (ages 70-79)	(excess cases)/ (expected cases) during screening	A. No B. No	Norway: 56% (42-73%) increased incidence with no post-screening drop Sweden: 45% (41-49%) increased incidence with 12% drop

Study; Cancer type	Considerations of Uncertainty/ Sensitivity Analyses	Results of Sensitivity Analyses	Conclusions
Peeters 1989; Breast	Not performed		There is no evidence that screening programs using mammography constitute a significant risk for overdiagnosis of breast cancers
Puliti 2009; Breast	Performed sensitivity estimate assuming no trend of increasing baseline incidence	13% (7-19%)	Although estimate of overdiagnosis was very sensitive to pre-screening trend estimates, data show that the degree of overdiagnosis was nearly zero and most likely less than 13%
Puliti 2012; Breast	Re-analyzed with exclusion of 34 women who were non-attenders who had a diagnosis of breast cancer within 6 months of invitation	15%	Overall cost to save one life corresponds to no more than one overdiagnosed tumor
Svendsen 2006; Breast	Not performed		The early detection program of this study did not result in any major overdetection
Zahl 2004; Breast	Not performed		Without screening one third of all invasive breast cancers in women 50-69 would not have been detected in the patients' lifetime.

Study; Cancer Type; Study Design	Study Population; Time Period	Description of Methods	Reference Population	Screening Schedule
Zahl 2012; Breast; Ecologic	women 50-79 in Norway counties Akershus, Oslo, Rogaland, and Hordaland; 1991-5 (pre-screening), 1995-2009 (screening)	Used Poisson regression to estimate changes in incidence of breast cancer in screening regions compared to pre-screening period in same regions. Looked for a decrease in incidence in women 70-74 post-screening	Women in screening regions in pre-screening period	Biennial mammography 50-69
Ciatto 2005; Prostate; Cohort	Italian men 60-74; 1991-4 (screening) 1995-2000 (post-screening)	Compared observed cancers in screened cohort to number expected with standardized incidence ratios. Expected cancers calculated based on Tuscan Cancer Registry incidence rates	Entire population in Tuscany Cancer Registry, including study cohort	Two biennial screening rounds of DRE+TRUS or PSA, with random sextant biopsy for PSA>10ng/ml and directed biopsy for suspicious findings on DRE/TRUS
Hugosson 2000; Prostate; Cohort	Swedish men 64-66; 1980-1996	Determined incidence in study cohort of men 64-66 in 1995 who had PSA>3 on screening test. Compared with incidence in cohort of men age 67 in 1980 who were followed until 1995, who also had PSA>3 on sample drawn in 1980 that was not analyzed until 1996.	Swedish men born in 1913, age 67 when PSA sample was drawn but not analyzed until 1980, then followed for cancer incidence until 1995	One-time PSA with threshold 3ng/ml for referral to DRE, TRUS and sextant biopsies

Study; Cancer type	Screening in Reference Population	Management of Potential Confounders	Management of Lead Time	Calculation of Overdiagnosis	A. Includes DCIS? B. Reports preferred outcome?	Magnitude of Over-diagnosis (95% CI)
Zahl 2012; Breast	Some opportunistic screening occurred in pre-screening period	Adjusted for age, county, population growth and baseline incidence trend. Demonstrated no change in baseline incidence in women too young for screening. Discussed how HRT use as indicated by sales of HRT was stable from 1991-3 to 2007-9 while the number of potential users increased by 33%, concluding that HRT not an important cause of breast cancer in Norway	Up to 14 years of follow-up post-screening (ages 70-79)	(excess cases)/ (expected cases) during screening	A. Yes B. No	Confirmed 50% incidence growth from Zahl 2004, with non-significant drop of 7% in women 70-74
Ciatto 2005; Prostate	Reference population includes screened study cohort. Small amount of opportunistic screening also noted in Florence.	Adjusted for age	7-9 year follow-up post-screening	(excess cases)/ (expected cases) during screening and 9-years post-screening	A. n/a B. No	66% (40-100%)
Hugosson 2000; Prostate	n/a	Not performed	n/a	n/a	A. n/a B. No	None. Men in screened cohort had lower cancer incidence (22%) at one-time PSA screening than did men in control cohort (32.9%) followed over 15 years.

Study; Cancer type	Considerations of Uncertainty/ Sensitivity Analyses	Results of Sensitivity Analyses	Conclusions
Zahl 2012; Breast	Not performed		After 14 years of mammography screening, there is a 50% increase incidence of breast cancer that cannot be explained by early diagnosis or use of HRT, but that must instead be due to overdiagnosis.
Ciatto 2005; Prostate	Not performed		High rates of overdiagnosis are confirmed for a screening experience adopting a non-aggressive protocol.
Hugosson 2000; Prostate	Not performed		Underdiagnosis rather than overdiagnosis is the case at least with one-time screening.

Study; Cancer Type; Study Design	Study Population; Time Period	Description of Methods	Reference Population	Screening Schedule
Tornblom 2010; Prostate; Cohort	Men 55-70 in Stockholm, Sweden; 1988-2000	Compared cumulative incidence in men screened with PSA who then underwent DRE and TRUS (in 1988) to a reference population of men who had a PSA sample taken in the 1980 that was analyzed later.	Men age 67 in Gothenburg participating in the Study of Men Born in 1913	PSA, DRE and TRUS with biopsies for abnormal findings on DRE, TRUS or PSA>10ng/ml
Zappa 1998; Prostate; Cohort/ Modeling	Men 60/65 otherwise similar to men participating in CSPO study in Florence; Not provided	Calculated expected age-specific incidence based on rates from Tuscany Cancer Registry 1990-1, assuming 2% increase per year. Compared to incidence in hypothetical cohort of men aged 60 based on incidence of screen-detected and interval cancers in two rounds of screening in CSPO study	Men in Tuscan Cancer Registry	Five biennial PSA tests with 4 ng/ml threshold

Study; Cancer type	Screening in Reference Population	Management of Potential Confounders	Management of Lead Time	Calculation of Overdiagnosis	A. Includes DCIS? B. Reports preferred outcome?	Magnitude of Over-diagnosis (95% CI)
Tornblom 2010; Prostate	PSA screening in Sweden was infrequent	Not performed	Follow-up beyond when incidence in reference population equaled that of screened population	Not calculated	A. n/a B. No	None (incidence in reference population surpassed that of screened population)
Zappa 1998; Prostate	Opportunistic screening in Florence is negligible before and through the study period	Not performed	4 years follow-up post-screening	(excess cases)/ (expected cases) during screening and 4-years post-screening	A. n/a B. No	age 60: 25% (19-32%) age 65: 65% (58-73%)

Study; Cancer type	Considerations of Uncertainty/ Sensitivity Analyses	Results of Sensitivity Analyses	Conclusions
Tornblom 2010; Prostate	Not performed		The early detection program of this study did not result in any major overdetection
Zappa 1998; Prostate	Also considered constant annual incidence over time	Age 60: 51% (44-59%) Age 65: 93% (85-101%)	Screening for prostate cancer associated with a relevant risk of overdiagnosis.

Appendix Table 7: Criteria for Evaluating Risk of Bias and Strength of Evidence for Ecologic and Cohort Studies

Study	Cancer type	Probability of Selection Bias and Confounding	Probability of Measurement Bias	Considerations of Uncertainty	**Overall Risk of Bias**	Time Frame	**Analysis**	**Directness**	**External Validity**	**Precision**
		H/M/L	*H/M/L*	*G/F/P*	***H/M/L***	*G/F/P*	***G/F/P***	***G/F/P***	***G/F/P***	***G/F/P/CD***
Bleyer 2012	breast	M	M	G	**M**	G	**G**	**F**	**G**	**CD**
Duffy 2010	breast	M	M	P	**M**	NA	**P**	**P**	**F**	**CD**
Falk 2013	breast	H	L	F	**H**	G	**G**	**F**	**F**	**F**
Hellquist 2012	breast	M	M	P	**M**	NA	**P**	**P**	**P**	**F**
Jorgenson 2009 (BMJ)	breast	M	M	P	**M**	F	**G**	**F**	**F**	**G**
Jorgenson 2009 (BMC)	breast	M	M	P	**M**	F	**G**	**F**	**F**	**CD**
Junod 2011	breast	M	M	P	**M**	F	**P**	**F**	**F**	**F**
Kalager 2012	breast	M	L	F	**M**	F	**P**	**F**	**F**	**F**
Morrell 2010	breast	M	M	F	**M**	NA	**P**	**P**	**F**	**CD**
Njor 2013	breast	M	M	P	**M**	F	**P**	**F**	**F**	**F**
Paci 2006	breast	M	M	F	**M**	NA	**P**	**P**	**F**	**F**
Peeters 1989	breast	H	M	P	**H**	P	**P**	**F**	**P**	**CD**
Puliti 2009	breast	M	M	F	**M**	F	**P**	**F**	**F**	**F**
Puliti 2012	breast	H	M	P	**H**	F	**P**	**F**	**F**	**F**
Svendsen 2006	breast	M	M	P	**M**	P	**P**	**F**	**F**	**CD**
Zahl 2004	breast	M	M	P	**M**	F	**G**	**F**	**F**	**F**
Zahl 2012	breast	M	M	P	**M**	G	**G**	**F**	**F**	**CD**
Ciatto 2005	prostate	M	M	P	**M**	F	**P**	**F**	**F**	**P**
Hugosson 2000	prostate	H	M	P	**H**	G	**G**	**F**	**F**	**CD**
Tornblom 2010	prostate	H	H	P	**H**	G	**G**	**F**	**F**	**CD**
Zappa 1998	prostate	M	M	F	**M**	F	**P**	**F**	**P**	**F**

Criteria for evaluating Strength of Evidence are bolded. Abbreviations: H, high; M, moderate; L, low; G, good; F, fair; P, poor; CD, cannot determine.

Appendix Table 8: Evidence Table of Randomized Controlled Trial Follow-Up Studies

Study Cancer Type	Study Population Time Period	Post-Study Length of Follow-Up	Screening Schedule	Baseline Characteristics of Study Groups	Contamination Screening in Control Group
Zackrisson 2006 Breast	Swedish women 55-69 in Malmo 1976-1986	15 years	5-6 rounds of mammography every 18-24 months, 70-74% attendance	Not reported	During trial period, 24% of control group underwent mammography

Study Cancer Type	ITT analysis?	Includes DCIS?	Reports Preferred Outcome?	Calculation of Overdiagnosis	Magnitude of Overdiagnosis (95% CI)	Conclusions
Zackrisson 2006 Breast	Not reported	Yes	No	(excess cases)/ (control cases) during trial and 15 years follow-up	10% (1 to 18%)	Fifteen years after the Malmo trial the rate of overdiagnosis was 10% in women 55-69.

Appendix Table 9: Criteria for Evaluating Risk of Bias and Strength of Evidence for Follow-Up of a Randomized Controlled Trial

Study	Cancer Type	Probability of Selection Bias	Probability of Measurement Bias	Potential for Confounding	**Overall Risk of Bias**	Time Frame	**Analysis**	**Directness**	**External Validity**	**Precision**
		H/M/L	*H/M/L*	*H/M/L*	***H/M/L***	*G/F/P*	***G/F/P***	***G/F/P***	***G/F/P***	***G/F/P***
Zackrisson 2006	breast	L	M	L	**L**	G	**P**	**F**	**F**	**F**

Criteria for evaluating Strength of Evidence are bolded. Abbreviations: H, high; M, moderate; L, low; G, good; F, fair; P, poor.