Overdiagnosis and Overtreatment of Prostate Cancer and Breast Cancer Due to Screening

By

Jeremy Littleton

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
2007

Russ Harris, Advisor
Linda Kinsinger, 2nd Reader

4/23/07
Introduction

Screening

Screening is the testing of asymptomatic individuals for a given disease with no signs or symptoms of the disease in question. Questions, physical exam, laboratory tests, and imaging are commonly used screening modalities.¹ Substantial benefits may be associated with screening, ideally a reduction in mortality if screening diagnoses disease at an earlier, more treatable stage. Earlier diagnosis of disease can also lead to decreased morbidity for these patients. In addition, for patients who worry about having a particular disease, a negative screening test may help them achieve peace of mind.

Yet screening is also associated with potential harms, including false positives, side effects, and overdiagnosis. A false positive test occurs when a positive screening result occurs in a patient who does not actually have the disease in question. False positives may lead to unnecessary biopsies, further work-up, and unnecessary anxiety for the patient. A recent review of studies that assessed the long-term effects of false positive mammograms found that women receiving a false positive diagnosis of breast cancer by mammogram were significantly more likely to experience worry, intrusive thoughts, and anxiety specific to breast cancer than women who received normal diagnoses.² Screening itself can lead to side effects, for example, in prostate cancer screening with digital rectal exam, the test itself can lead to a great deal of anxiety, whether results are positive or negative. In colon cancer screening with colonoscopy, scoping results in a ruptured colon in about 1-2 in every 1000 screened patients.³⁻⁴ A third risk of screening, particularly screening in older populations, is overdiagnosis.
**Overdiagnosis**

Overdiagnosis is the diagnosis by screening of cancers that would have never become clinically meaningful in the absence of screening. This can occur in two ways: either screening detects cancer that is non-progressive or progresses so slowly that it never causes patient morbidity, or screening detects cancer at a time in the patient’s life when the patient is old enough or sick enough that he/she will die of another disease before the cancer causes significant morbidity. Similar to false positives, overdiagnosis can lead to unnecessary biopsies, unnecessary treatments, and adverse effects. All of these excess diagnostic and therapeutic measures lead to increased costs that affect both the patient and the health care system as a whole. The risk of overdiagnosis is higher in the elderly because they have more comorbidities which contribute to overall mortality.

*Figure 1: Number of Cases of Disease in the Presence and Absence of Screening***
Overdiagnosis due to screening can be measured by comparing rates of disease in screened and unscreened cohorts of patients over a period of time. Figure 1 shows a hypothetical study in which an identical number of patients from a given population were randomly selected for yearly screening or no screening. Both groups initially have 1000 cases of disease per year. This model assumes that the baseline incidence of disease is not changing in the population, and thus the unscreened group continues to have 1000 cases of disease per year. However, when the screening program is initiated in year 3, a large peak in the number of incident cases occurs in the screened group due to screening finding many of the prevalent cases of disease in the population. These cases developed over the lifetime of the patients being screened, and after this initial increase in incidence, the number of cases declines as fewer and fewer prevalent cases are picked up by screening.

This decline continues until a period of time equivalent to the lead time has passed since the initiation of screening. The lead time is the amount of time by which screening advances diagnosis. Thus if screening can detect disease 5 years before detectible symptoms of the disease develop, the disease has a lead time of 5 years. At the end of the lead time from initial screening, the number of cases in the screened group should be identical to the number of cases in the unscreened group: all of the prevalent cases should have been detected by this time, and thus screening should be detecting each case as it arises 5 years before it would have become clinically symptomatic. Yet in figure 1, the number of cases continues to be higher in the screened group than the unscreened group after the 5-year lead time has passed. This discrepancy of 500 cases is due to overdiagnosis: these cases are those that would have never come to clinical attention because the patient would have died of comorbid diseases whether they had been screened or not before the disease in question could become symptomatic. These
patients would have never been diagnosed in the absence of screening and any treatment or further diagnostic tests for their disease would thus be unnecessary.

Yet figure 1 is a simplified example of overdiagnosis; in reality, the situation can be much more complicated. First of all, the baseline incidence of most diseases does not remain stable over time. The incidence of breast cancer, for example, gradually increased from 1980 to 2000 and has since leveled off. If the baseline incidence were increasing over time in the above example, the incidence of disease in the unscreened group would increase over time. The incidence of disease in the screened group would increase by this baseline amount plus whatever increased diagnosis occurred due to detection of prevalent cases and overdiagnosis. Another assumption that the above example does not take into account is the possibility of contamination in the unscreened group. Many screening tests are so widespread that it is difficult to prevent opportunistic screening to occur in the control group. In this case, the incidence in the control group would seem higher than the actual incidence, and thus the amount of overdiagnosis left over at the end of the study would be underestimated. In addition, the lead time of a screening test is hard to estimate with accuracy. Diseases that have a range of clinical manifestations, such as those in which some cases are aggressive and others are less so, tend to have a range of lead times. More aggressive disease has a shorter lead time than less aggressive disease. Most diseases have a range of possible lead times depending on the aggressiveness of the disease, which makes it difficult to predict the point at which excess cases above baseline are due to overdiagnosis alone.

Why is Overdiagnosis a Significant Problem

Overdiagnosis is a particular concern in cancer screening for this reason. Many cancers are so slow growing that it will take many years for the patient to develop clinical disease.
However, other cancers are extremely aggressive and clinical disease develops in a short period of time. Unfortunately, screening disproportionately detects slower growing cancers due to the longer lead time during which the patient has disease. This phenomenon is known as length-time bias.\(^1\) An extreme form of length-time bias is in effect in the type of overdiagnosis in which a screen-diagnosed cancer would never have been diagnosed because the cancer is growing so slowly that it would have never caused symptoms. Patients with slow growing cancers, therefore, are at high risk of overdiagnosis and subsequent overtreatment. This risk is enhanced in the elderly, who are at high risk for many types of cancers. The risk of overdiagnosis is concerning for two types of cancer in particular—prostate cancer and breast cancer—because each has a significant proportion of slow growing cancers, screening is fairly widespread, and the elderly are those who are most commonly screened.

**Prostate Cancer Screening**

Prostate cancer is the most common cancer in US males; the American Cancer Society estimates that 218,890 new cases will occur in 2007. It is the second leading cause of cancer death in males with 27,050 deaths estimated by the ACS to occur in 2007.\(^6\) Currently, screening for prostate cancer occurs by serum prostate-specific antigen (PSA) level and digital rectal exam (DRE). Thus far, the evidence is inconclusive as to whether the benefits of screening outweigh the risks. In their most recent review of the literature, the United States Preventive Services Task Force (USPSTF) concluded that the evidence was insufficient to recommend for or against routine PSA or DRE screening.\(^7,8\) The primary potential benefits of screening include decreased morbidity and mortality due to prostate cancer. Previous studies have shown conflicting results as to whether screening decreases prostate cancer mortality.\(^9-12\) Two large randomized controlled trials on the topic, the European Randomized Study of Screening for Prostate Cancer
(ERSPC) and the U.S. National Cancer Institute “Prostate, Lung, Colorectal, and Ovary” Trial (PLCO), are currently underway, but data on mortality will not be available for a few years.

Screening is associated with potential harms as well. PSA screening for prostate cancer is less specific in men at high risk for benign prostatic hyperplasia, thus false positives are one potential risk, especially in older populations. This can lead to unnecessary biopsies, increased anxiety, and even unnecessary treatment with prostatectomy, radiation, or androgen deprivation and the adverse effects that result from these therapies. Overdiagnosis is another worrisome risk of prostate cancer screening.

Overdiagnosis is particularly concerning in prostate cancer screening for several reasons. Elderly men are at high risk of developing prostate cancer, yet this population commonly has many other comorbidities that limit life expectancy. Thus screening in this population would lead to diagnosis of many patients who would never have developed clinical disease, leading to unnecessary biopsies, costs, and possibly treatment. In addition, much uncertainty exists regarding the natural history of untreated or minimally treated prostate cancer, but it is likely that a significant proportion of prostate cancer patients do not benefit from treatment. A prospective cohort study in Sweden found that 15 years after diagnosis of localized prostate cancer, survival rates in 223 patients who deferred treatment were similar to those who received initial treatment (81% 15-yr survival in both groups).

A follow up of this study at 20 years found that in the untreated patients, prostate-cancer specific survival fell from 79% at 15 years to 54% at 20 years, indicating that early treatment of patients with a life expectancy longer than 15 years may be beneficial. However, another recent retrospective cohort study of 767 men diagnosed with early stage prostate cancer and treated with observation or androgen deprivation therapy alone found that prostate cancer specific mortality rates dropped from 33 per 1000 patients at 15 years
to 18 per 1000 patients after 15 years. In this study, higher Gleason scores were correlated with higher prostate cancer specific mortality rates, ranging from 6 per 1000 person years for low grade cancers to 121 per 1000 person-years for high grade cancers.\textsuperscript{15}

Despite the inconsistencies across these studies, they all indicate that there is a significant proportion of men diagnosed with prostate cancer who do not benefit from early treatment. Because screening preferentially detects low-grade cancers by length-time bias and treatment does not seem to help patients with early stage disease at 10 to 15 years of follow-up, it is likely that a significant amount of overdiagnosis and overtreatment occur in patients diagnosed with prostate cancer by screening. Moreover, it is nearly impossible to predict which patients will develop comorbidities and die of causes other than prostate cancer 15 years down the line. Thus the overall potential for overdiagnosis due to prostate cancer screening seems to be high.

**Breast Cancer Screening**

Statistically, breast cancer is comparable to prostate cancer in terms of cases and mortality. The ACS predicts 178,480 new cases among US women in 2007, which makes breast cancer the most common cancer diagnosis. In addition, it will also be the second most common cause of cancer-related death among women, with 40,460 predicted deaths in 2007.\textsuperscript{6} Screening modalities for breast cancer include mammography, clinical breast exam (CBE), and breast self-examination (BSE). In their most recent review, the USPSTF performed a meta-analysis of 7 high quality trials and found a relative risk (RR) of dying of breast cancer of 0.84 (95% CI: 0.77-0.91) for women screened with mammography compared to those unscreened.\textsuperscript{16} The USPSTF thus recommends screening mammography every 1-2 years for all women aged 40 and older. However, they found insufficient evidence to recommend for or against CBE alone for breast cancer screening.\textsuperscript{17} Thus, the major benefit of breast cancer screening is a 16% relative risk
reduction in breast cancer mortality. Harms of screening include anxiety, discomfort, cost, and false positives leading to unnecessary biopsies and treatment. In addition, overdiagnosis leading to overtreatment, particularly in DCIS patients, is a concern.

Some of the breast cancer cases diagnosed by screening are ductal carcinoma in situ (DCIS), which is a localized, slowly progressive form. One large study of over 650,000 mammograms found that approximately 18% of breast cancer cases diagnosed by screening mammography were DCIS.\textsuperscript{18} This significant proportion of women diagnosed with breast cancer would be at higher risk for overdiagnosis than women diagnosed with invasive cancers due to the more slowly progressive nature of DCIS. In addition, the rate of diagnosis of DCIS by mammography seems to increase with age with over 1 in every 1000 women aged 70-84 screened by mammography diagnosed with DCIS as compared to about 1 in every 1800 women aged 40-49 screened by mammography diagnosed with DCIS.\textsuperscript{18}

Some argue that treatment of DCIS leads to a reduction in the incidence of invasive cancers, but data on the natural progression of untreated DCIS are scarce. One study of 80 women with untreated DCIS found that 14% developed invasive cancer after a mean follow-up of 17.5 years.\textsuperscript{19} Smaller studies have suggested that up to 50% of DCIS progresses to invasive breast cancer, but that half of these lesions do not show up within the lifetime of the diagnosed woman because a significant proportion of DCIS cases are elderly.\textsuperscript{20,21} Overall, it is not clear what proportion of DCIS progresses to invasive breast cancer. However, in 1999 over 28% of women with DCIS had mastectomy, 64% had lumpectomy, and 33% had radiation.\textsuperscript{22} Thus, because it is slowly progressive in most cases and it is more commonly diagnosed in the elderly, a significant proportion of women with screen-detected DCIS appear to be at risk for overdiagnosis and subsequent overtreatment. In addition, it is unknown what proportion of
women who are screen diagnosed with invasive breast cancer, particularly elderly women with short life expectancies due to other comorbidities, are overdiagnosed and how much overtreatment occurs as a result.

**Focused Clinical Question**

Overall, overdiagnosis and overtreatment due to breast cancer and prostate cancer screening are undoubtedly present. The question is how much overdiagnosis and overtreatment actually occur. Estimates of the proportion of overdiagnosed cases due to screening have ranged widely, from 15.4% to 84% for prostate cancer and from 5% to 33% for breast cancer. Thus, the true magnitude of overdiagnosis of both prostate cancer and breast cancer is unknown. This review will attempt to answer the following questions:

1. To what extent does screening for prostate cancer with serum PSA and DRE lead to overdiagnosis and overtreatment?
2. To what extent does screening for breast cancer with mammography and CBE lead to overdiagnosis and overtreatment?
3. What conclusions can be drawn about the problems of overdiagnosis and overtreatment due to screening from a comparison of prostate cancer and breast cancer screening?

**Methods**

**Background**

Many different definitions have been used for overdiagnosis and overtreatment, and thus it is necessary to define the term that I used for this review. I defined overdiagnosis as the diagnosis by screening of a case of disease that would never have come to clinical attention had that patient not been screened. Overtreatment is thus the unnecessary treatment of one of these
overdiagnosed cases, leading to unnecessary adverse effects of treatment. I consider any intervention that occurs after the initial overdiagnosis of the case by screening overtreatment. For example, although a prostate biopsy is technically a diagnostic test, for the purposes of this review, a biopsy of an overdiagnosed case is considered overtreatment. I defined the population, interventions, and outcomes of this review and developed the list of questions noted above to focus my search. Tables 1 and 2 list all inclusion and exclusion criteria used in the review.
<table>
<thead>
<tr>
<th>Selection Criteria</th>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>• Population-based focus on asymptomatic adult males screened for prostate cancer</td>
<td>• Screening high-risk subsets of the population</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>• Population-based focus on asymptomatic adult females screened for breast cancer</td>
<td></td>
</tr>
<tr>
<td><strong>Intervention</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>• Screening Serum PSA and/or digital rectal exam</td>
<td></td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>• Screening mammography and/or clinical breast exam</td>
<td></td>
</tr>
<tr>
<td><strong>Outcome</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>• Defines overdiagnosis as the diagnosis by screening of a case of disease that would have never come to clinical attention in the absence of screening</td>
<td>• Strictly pathological definition of overdiagnosis</td>
</tr>
<tr>
<td></td>
<td>• Any measure of overdiagnosis or overtreatment (% rate/1000 pts screened, etc.)</td>
<td>• No explicit numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>• Overtreatment due to prostate biopsy or treatment with radical prostatectomy, EBRT, ADT, or watchful waiting</td>
<td></td>
</tr>
<tr>
<td><strong>Study design</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Systematic reviews, randomized controlled trials, prospective and retrospective cohort studies, nested case-control studies</td>
<td>• Narrative Reviews</td>
</tr>
<tr>
<td></td>
<td>• Modeling strategies based on the above types of studies or service screening studies</td>
<td>• Case reports and case series</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Autopsy studies</td>
</tr>
</tbody>
</table>

*Note: EBRT = External Beam Radiation Therapy, ADT = Androgen Deprivation Therapy.*
Search Strategy

After developing the questions and criteria noted above, I formulated a search strategy using key search terms. I searched MEDLINE, the Cochrane Database, BIOSIS, and ISI Web of Science using the searches outlined in Tables 2 and 3 and reviewed the titles and abstracts of all English-language articles or translations. I pulled the articles which had a chance to fit the search criteria from the abstract list and read the full-text of these articles to further analyze them for possible inclusion. A second, independent reviewer read all titles and abstracts that I felt did not meet search criteria and pulled any remaining articles that he felt had a chance to meet the search criteria. In addition, I reviewed the reference lists of relevant review articles identified in the database search, the references of articles that were pulled in the database search, and the references of the National Cancer Institute’s Physician Query Database (PDQ) summaries on cancer screening, breast cancer screening, and prostate cancer screening. I accessed the full-text of any articles identified in this way and read the articles in depth to assess whether they fit inclusion criteria. Figure 2 shows a flow chart of the inclusion process for prostate cancer studies, and Appendix 1 includes reasons for exclusion of pulled prostate cancer studies. Figure 3 shows a flow chart of the inclusion process for breast cancer studies, and Appendix 2 includes reasons for exclusion of pulled breast cancer studies.
<table>
<thead>
<tr>
<th>Database</th>
<th>Search Used</th>
<th>Total Articles Identified (3/19/07)</th>
<th>Articles Pulled for Further Review</th>
<th>Articles used for Reference List Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDLINE</td>
<td>(prostate cancer OR PSA OR digital rectal exam) AND (overdiagnos* OR overtreat* OR over-treat* OR overtreat* OR overtreat* OR overdetect*)</td>
<td>159 Results</td>
<td>26 (^{23-48})</td>
<td>21 (^{49-59})</td>
</tr>
<tr>
<td>Cochrane Database</td>
<td>(prostate cancer OR PSA OR digital rectal exam) AND (overdiagnos* OR over-treat* OR over-treat* OR over-treat* OR over-treat* OR overdetect*)</td>
<td>15 Results</td>
<td>1 (^{50})</td>
<td>0</td>
</tr>
<tr>
<td>ISI Web of Science</td>
<td>(&quot;prostate cancer*&quot; OR &quot;prostate neoplasm*&quot; OR PSA OR &quot;digital rectal exam&quot; OR DRE) AND (overdiagnos* OR overtreat* OR overdetect*)</td>
<td>107 Results</td>
<td>11 (^{51-81})</td>
<td>0</td>
</tr>
<tr>
<td>BIOSIS</td>
<td>(&quot;prostate cancer*&quot; OR &quot;prostate neoplasm*&quot; OR PSA OR &quot;digital rectal exam&quot; OR DRE) AND (overdiagnos* OR overtreat* OR overdetect*)</td>
<td>57 Results</td>
<td>0</td>
<td>1 (^{83})</td>
</tr>
<tr>
<td>Database</td>
<td>Search Used</td>
<td>Total Articles Identified (3/19/07)</td>
<td>Articles Pulled for Further Review</td>
<td>Articles used for Reference List Search</td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-------------------------------------</td>
<td>-------------------------------------</td>
<td>-----------------------------------------</td>
</tr>
<tr>
<td>MEDLINE</td>
<td>(breast cancer OR breast cancers OR mammogra* OR clinical breast exam) AND (overdiagnos* OR over-diagnos* OR overtreat* OR over-treat* OR overdetect* OR over-detect*)</td>
<td>229 Results</td>
<td>345, 267, 47, 85, 113</td>
<td>955-714-124</td>
</tr>
<tr>
<td>Cochrane Database</td>
<td>(breast cancer OR breast cancers OR mammogra* OR clinical breast exam) AND (overdiagnos* OR over-diagnos* OR overtreat* OR over-treat* OR overdetect* OR over-detect*)</td>
<td>32 Results</td>
<td>5122-126</td>
<td>0</td>
</tr>
<tr>
<td>ISI Web of Science</td>
<td>(&quot;breast cancer*&quot; OR &quot;breast neoplasm*&quot; OR mammogra* OR &quot;clinical breast exam&quot; OR CBE) AND (overdiagnos* OR overtreat* OR overdetect*)</td>
<td>129 Results</td>
<td>8127-134</td>
<td>0</td>
</tr>
<tr>
<td>BIOSIS</td>
<td>(&quot;breast cancer*&quot; OR &quot;breast neoplasm*&quot; OR mammogra* OR &quot;clinical breast exam&quot; OR CBE) AND (overdiagnos* OR overtreat* OR overdetect*)</td>
<td>69 Results</td>
<td>4135-138</td>
<td>1139</td>
</tr>
</tbody>
</table>
Figure 2: Flow Chart of Prostate Cancer Search and Inclusion

339 Total Prostate Cancer Abstracts

159 MEDLINE abstracts

15 Cochrane abstracts

107 ISI Web of Science abstracts

58 BIOSIS abstracts

All Unique Studies

7 Unique from MEDLINE

43 Unique from MEDLINE & Cochrane

6 Unique from MEDLINE, Cochrane, & ISI

26 Studies Pulled

1 Study Pulled

11 Studies Pulled

0 Studies Pulled

10 Studies Included*

0 Studies Included

0 Studies Included

0 Studies Included

10 Studies Included For Review†

*See Appendix 1 for reasons for exclusion of pulled studies
†No relevant studies were identified in reference list search
Figure 3: Flow Chart of Breast Cancer Search and Inclusion

459 Total Breast Cancer Abstracts

229 MEDLINE abstracts

32 Cochrane abstracts

129 ISI Web of Science abstracts

69 BIOSIS abstracts

All Unique Studies

21 Unique from MEDLINE

46 Unique from MEDLINE & Cochrane

13 Unique from MEDLINE, Cochrane, & ISI

34 Studies Pulled

5 Studies Pulled

8 Studies Pulled

4 Studies Pulled

11 Studies Included*

0 Studies Included

0 Studies Included

0 Studies Included

11 Studies Included For Review†

*See Appendix 2 for reasons for exclusion of pulled studies
† No relevant studies were identified in reference list search
Quality Assessment

The quality of studies was largely assessed using the U.S. Preventive Services Task force criteria in Table 4. However, because there were a number of modeling studies included in the analysis, it was necessary to develop criteria with which to assess the internal validity of these studies (Table 5). For modeling studies which used data from previous studies, the quality of both the modeling process and the study whose data the model used was assessed. In addition, external validity was assessed using USPSTF defined criteria by examining studies for biologic plausibility, similarity between the population in the study and primary care patients, similarity of the intervention used to the screening intervention commonly used in clinical practice, and clinical or social environmental circumstances that could give different results than those that would be expected in a primary care environment.

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Criteria</th>
</tr>
</thead>
</table>
| Systematic Reviews           | • Comprehensiveness of sources/search strategy used  
|                              | • Standard appraisal of included studies  
|                              | • Validity of conclusions  
|                              | • Recency and relevance  
| Case-Control Studies         | • Accurate ascertainment of cases  
|                              | • Nonbiased selection of cases/controls with exclusion criteria applied equally to both  
|                              | • Response rate  
|                              | • Diagnostic testing procedures applied equally to each group  
|                              | • Appropriate attention to potential confounding variables  
| Randomized Controlled Trials and Cohort Studies | • Initial assembly of comparable groups:  
|                              | For RCTs: adequate randomization, including concealment and whether potential confounders were distributed equally among groups  
|                              | For cohort studies: consideration of potential confounders with either restriction or measurement for adjustment in the analysis; consideration of inception cohorts  
|                              | • Maintenance of comparable groups (includes attrition, crossovers, adherence, contamination)  
|                              | • Important differential loss to follow-up or overall high loss to follow-up  
|                              | • Measurements: equal, valid, and reliable (includes masking of outcome assessment)  
|                              | • Clear definitions of interventions  
|                              | • All important outcomes considered  
|                              | • Analysis: adjustment for potential confounders for cohort studies, or intention-to-treat analysis for RCT's  

140
### Table 5: Criteria for Grading the Internal Validity of Overdiagnosis Modeling Studies

<table>
<thead>
<tr>
<th>Type of Study</th>
<th>Criteria</th>
</tr>
</thead>
</table>
| Modeling based on RCT, cohort, or service screen data | • Data used for modeling comes from high quality studies based on Table 3 criteria  
• All relevant variables should be included in the model  
• Estimates for model parameters come from quality studies (lead time, transition probabilities in Markov models)  
• Follow-up should be long enough to outlast the prevalence peak due to lead time  
• Comparison groups should be applicable and as similar to the intervention group as possible for service screening modeling  
• Contamination of the comparison group should be minimal  
• Authors should discuss the role of lead time and the impossibility of completely distinguishing it from overdiagnosis  
• A probabilistic sensitivity analysis should be performed to assess the effects of varying different parameters, including lead time |

### Data Extraction

Studies that were included after the initial search were reviewed in more depth to determine whether they fit all inclusion criteria. Data from each study was extracted into a standard table to allow for easy comparison with inclusion criteria (Tables 5 and 6). In addition, the adherence of each study to the above quality criteria was assessed during the data extraction process.

### Table 5: Example Data Extraction Table—Study Characteristics

<table>
<thead>
<tr>
<th>Study and Date Extracted</th>
<th>Study Design</th>
<th>Definition of Overdiagnosis and Formula</th>
<th>Sample Characteristics (size, selection, LTFU)</th>
<th>Method. Quality</th>
<th>Interventions</th>
<th>Outcome Assessmt.</th>
</tr>
</thead>
</table>

### Table 6: Example Data Extraction Table—Results

<table>
<thead>
<tr>
<th>Study and Date Extracted</th>
<th>Length of Follow-Up</th>
<th>Dropouts</th>
<th>Results</th>
</tr>
</thead>
</table>
Data Synthesis

Data synthesis was performed primarily in a descriptive manner, without the use of quantitative methods because heterogeneity of outcomes prevented the use of this modality. By qualitatively examining the heterogeneity of the results, it was possible to come to some conclusions about overdiagnosis and overtreatment. In addition, the quality of the studies and some of their resulting limitations were noted. Although some of the smaller databases and reference lists were searched for full-text articles, publication bias is still a concern as it is with any systematic review.

Results

Prostate Cancer

Study Selection

Of the 38 prostate cancer articles in Table 2 that were pulled from the literature search for further review, 10 articles fulfilled all selection criteria. See appendix 1 for a reasons for inclusion and exclusion of studies. If a study was borderline for fulfilling selection criteria, it was left in the review. Five of the included studies are models validated by prostate cancer incidence in the SEER database, three are models based on service screening programs, and two use the same Markov model validated by RCT data.

Three tables follow that give a general idea of the included studies. Table 6 lists specific details on the characteristics of these studies, including study design, stated definition and expression of overdiagnosis, sample characteristics, intervention, and length of follow up.

Because many of the modeling studies incorporated data from previous studies into their model.
or were validated by other studies, it was necessary to include some information about these underlying studies in table 6 as well. Table 7 reports the results of these studies in three ways: as reported in the study, converted to overdiagnosed cases / 1000 patients screened (if possible), and converted to overdiagnosed cases / 1000 cases diagnosed. Lastly, table 8 includes details about the quality of the included studies.

Results

Overall, rates of overdiagnosis ranged from 8 to 99 per 1000 patients screened and 73 to 708 per 1000 patients with screen-detected prostate cancer and depended on the age at diagnosis, race, Gleason score, or mean lead time estimate used in the study. It was not possible to estimate the rate of overdiagnosis per 1000 patients screened for 5 of the studies due to insufficient data reported in the studies.\textsuperscript{25,30,33,35,41} The quality of the studies was variable, but two of the studies validated by SEER data were fair quality, one of the service screening studies was fair quality, and the two studies validated by RCT data were particularly good quality.

Of the studies based on SEER data, Etzioni et al\textsuperscript{41} and Telesca et al\textsuperscript{25} had the best overall study quality. Etzioni et al fit models separately for white and black patients, and found that for black patients, the model fit better to incidence data using a mean lead time of 7 years, whereas for white patients, the model fit better when it used a mean lead time of 5 years. This corresponded to overdiagnosis rates of 288 per 1000 white screen-detected patients and 438 per 1000 black screen-detected patients. These rates varied slightly depending on what proportion of PSA’s the authors assumed were diagnostic versus screening. Telesca et al actually used the same model as Etzioni et al, but used a likelihood model for secular trend in prostate cancer incidence to model it concurrently with lead time. They found that overdiagnosis rates were slightly less when estimated in this way: 227 per 1000 white screen-detected patients and 344
per 1000 black screen-detected patients. A caveat to these studies is that they do not model a specific screening regimen; they model the rate of PSA screening between 1988 and 1998 based on Medicare claims data and assume in the model that screening occurs at this baseline rate.

Other models using SEER incidence data as validation include Etzioni et al.,43 Davidov et al.,35 and Tsodikov et al.33 Etzioni et al found a similar overdiagnosis rate of 247 per 1000 screen-detected stage A1 cancers for annual PSA screening with a cutoff of 4, but expressed overdiagnosis as a proportion of stage A1 (low stage) cancers, which leaves out the potential for overdiagnosis in high stage cancers. Rates were lower for less frequent screening and when using age-specific PSA cutoffs. Davidov et al found that rates for 50 to 70 year olds screened with PSA (cutoff not stated) every 5 years had an overdiagnosis rate of 307 per 1000 screen-detected cases, assuming a lead time of 10 years. Rates were higher for longer sojourn times and were also higher for wider age ranges screened. Tsodikov et al found overdiagnosis rates of 100 to 500 cases depending on the age at which screening occurred. The authors in this study assumed that screening was a random process; they did not model the impact of a specific screening interval. The results are not reported in a useful format in this study; the values above were extrapolated from a graph of overdiagnosis versus age.

The best quality model based on a service screening population was Ciatto et al.,34 which found the rates of overdiagnosis to be higher. Using two biennial screening rounds of PSA with a cutoff of 10, DRE and TRUS over 9 years of follow up, the authors estimated that overdiagnosis would occur in 399 patients per every 1000 screen-detected cases and in 17 individuals per every 1000 patients screened. Other service screening models included Zappa et al.45 and Hamashima et al.28 Zappa et al found that for patients screened with biennial PSA with a cutoff of 4 starting at age 60, overdiagnosis occurred in 201 patients per 1000 screen-detected
cases and 6 patients per 1000 patients screened, assuming a 2% increase in baseline prostate
cancer incidence per year. Rates were higher if screening were started at age 65 and if the model
assumed a constant baseline incidence. Follow up of the patients in the service screening study
was only 4 years in this model. Hamashima et al found that patients screened with one-time PSA
with a cutoff of 2.7 had a rate of overdiagnosis of 708 cases per 1000 screen-detected cases and 8
per 1000 patients screened. However, this study only had a follow up of 1 year, thus much of the
increased incidence detected in the study is likely prevalent cases of prostate cancer. A problem
in all these screening studies is that they assume that all excess incidence occurs due to
overdiagnosed cases when there are undoubtedly some cases of prostate cancer that were caught
evitably by screening that would have presented clinically later. The longer the follow up in the
study, the less of a concern this is.

Two studies were based on a single model developed by Draisma et al,\textsuperscript{38} which validated
the model using both population level incidence data from The Netherlands in 1991 and data
from a randomized controlled trial that randomized over 42,000 patients to screening or no
screening and had 6 years of follow up data available. Draisma et al found that for annual PSA
screening with a cutoff of 3 of men 55-75, overdiagnosis rates were higher: 560 cases per 1000
screen-detected cases and 79 cases per 1000 men screened. Rates of overdiagnosis in this study
were lower for single screening, slightly lower for a smaller age range screened (55-67 year olds),
and slightly lower if using a four year screening interval instead of annual screening. Parker et
al\textsuperscript{30} used the same model to estimate overdiagnosis rates based on stage-specific detection rates
from a prospective cohort study of untreated or minimally treated prostate cancer.\textsuperscript{143} The authors
estimated overdiagnosis rates in 60-64 year olds were 278 per 1000 screen-detected cases of
prostate cancer with a Gleason score of 7. Rates of overdiagnosis were higher for lower Gleason scores and lower for younger age groups screened.

**Breast Cancer**

**Study Selection**

Of the 51 breast cancer studies in Table 3 that were pulled from the literature search for further review, 11 articles fulfilled all selection criteria. Appendix 2 contains reasons for inclusion and exclusion of all studies. If a study was borderline for fulfilling selection criteria, it was left in the review. Eight of the included studies are based on population-based service screening programs and three are based on RCT data.

Three tables follow that give a general idea of the included studies. Table 9 lists specific details on the characteristics of these studies, including study design, stated definition and expression of overdiagnosis, sample characteristics, intervention, and length of follow up. This table also contains information about the analysis used to calculate overdiagnosis. Table 7 reports the results of these studies in three ways: as reported in the study, converted to overdiagnosed cases / 1000 patients screened (if possible), and converted to overdiagnosed cases / 1000 cases diagnosed. Lastly, table 8 includes details about the quality of the included studies.

**Results**

Overall, rates of breast cancer overdiagnosis ranged from 4 to 14 overdiagnosed cases per 1000 patients screened and 22 to 600 overdiagnosed cases per 1000 patients with breast cancer. These values varied in some studies depending on the sensitivity of mammography and whether
invasive and in situ cases were considered together or separately. It was not possible to estimate the rate of overdiagnosis per 1000 patients screened for 8 of the studies due to insufficient data reported in the studies.\textsuperscript{83, 86, 89, 91, 99, 102, 107, 112} The quality of the studies was variable, with two of the studies validated by SEER data were fair quality, one of the service screening studies was fair quality, and the two studies validated by RCT data were particularly good quality.

Two of the studies based on service screen data, de Koning et al\textsuperscript{83} and Olsen et al\textsuperscript{88} used the same basic model of chronic disease progression\textsuperscript{144} to calculate expected incidence and compare this to the observed incidence in the screened cohort. This model fits a Poisson distribution to the data to estimate the values of parameters such as lead time, sensitivity, incidence of preclinical screen detectable cancers, and incidence of overdiagnosed cancers using predefined equations. De Koning et al used 8 years of service screen data and found that 80 cases of breast cancer would be overdiagnosed per 1000 screen-detected cases of breast cancer. Olsen et al used 6 years of service screen data and found that 78 cases of breast cancer would be overdiagnosed per 1000 screen-detected cases of breast cancer at the first screen.

Paci et al\textsuperscript{89} also used a Poisson regression to model expected incidence based on service screening data, but they took a different approach to modeling the effect of lead time. The authors calculated the contribution of lead time to excess incidence found in the service screening study by assuming an exponential distribution of lead times and using a model to generate a separate lead time for each individually screened patient. The probabilities that these cases would come to clinical detection each year after screen detection were summed to generate the number of screen-detected cases that would have arisen clinically each year in the absence of screening. The authors subtracted this number from the observed number of cases to correct for lead time, and compared the corrected number of cases with the expected number of cases from
the Poisson regression model. They found that 34 overdiagnosed cases would occur per 1000 breast cancers diagnosed. A similar, earlier model by Paci et al\textsuperscript{99} based on Florence service screening data found that less than 50 overdiagnosed cases would occur per 1000 breast cancers diagnosed.

Zahl et al\textsuperscript{102} is another model based on service screening data that uses Poisson regression to calculate expected age-specific incidence rates of breast cancer in the absence of screening. The authors compare this to the observed incidence from the service screening data to generate the excess incidence due to screening, which the authors assume to be all a result of overdiagnosis. This model does not include an attempt to distinguish lead time from overdiagnosis. The follow up in the model is 5 years in the Norway service screen and 15 years for Swedish service screen. The authors found overdiagnosis rates of 351 overdiagnosed cancers per 1000 cases in Norway and 310 overdiagnosed cancers per 1000 cases in Sweden. Anttila et al\textsuperscript{107} used similar methods for a cohort of women in Helsinki, Finland with 8.5-11.5 years of follow up and found that approximately 150 overdiagnosed cancers occurred for every 1000 breast cancer cases. Peeters et al\textsuperscript{112} also used similar methods for a screened cohort in The Netherlands and found that 99 overdiagnosed cases occurred per 1000 breast cancer cases diagnosed. Hamashima et al\textsuperscript{28} was a low quality study that found an excess incidence of 600 cases per every 1000 screen detected cancers after only one year of follow up in a small Japanese screened cohort.

The studies based on RCT data also differed from each other in methods and results. Gotzsche et al\textsuperscript{86} performed a systematic review of 7 RCT's of breast cancer screening and concluded that only two of them, the Malmo trial\textsuperscript{145} and the Canada trial,\textsuperscript{146} were adequately randomized. The authors expressed overdiagnosis as the percent excess incidence in the
screened group as compared to the control group. After 7-9 years of follow up, the authors found an excess incidence of approximately 230 cases per 1000 cancers diagnosed in the two trials, which they assumed was all a result of overdiagnosis. Zackrisson et al\textsuperscript{91} performed a similar analysis with longer follow up: 10 years of trial data and 15 years of follow up past the end of the trial. They found an excess incidence of just 91 cases per 1000 cancers diagnosed, which they assumed was all due to overdiagnosis.

Duffy et al\textsuperscript{5} used RCT data in a Poisson regression to model the expected incidence, and they used the model of Day and colleagues\textsuperscript{144} (the same model used in de Koning et al and Olsen et al above) to estimate the values for lead time, sensitivity, and incidence of screen-detectable and overdiagnosed cancers. Taking these parameters into account, they found a rate of overdiagnosis at the first screen of 32 cases per 1000 screen-detected cancers in the Swedish Two County Trial and 42 cases per 1000 screen-detected cancers in the Gothenburg Trial. Thus, after taking lead time into account, estimates of overdiagnosis are less than in the studies above that did not take lead time into account in the model, even in studies with longer follow ups such as Zackrisson et al.

**Discussion**

**Overdiagnosis in Prostate Cancer Screening**

Thus, for prostate cancer, estimates of overdiagnosis from the highest quality studies noted above range from 227 to 560 overdiagnosed cases per 1000 screen-detected cases of prostate cancer. This does not take into account the very high rates of overdiagnosis found by Parker et al for patients over 65 years of age with Gleason scores of 7 or less—up to 766 overdiagnosed cases per 1000 patients screened. Some factors associated with a higher rate of
overdiagnosis include black race (due to longer lead time), lower Gleason score, and wider and older age ranges screened. More frequent screening, i.e. a shorter screening interval, was associated with a slightly increased rate of overdiagnosis in some studies. In addition, using higher mean lead time values and assuming a constant or declining secular trend in incidence were associated with higher rates of overdiagnosis in these studies. The two higher quality studies for which overdiagnosed cases per 1000 patients screened could be estimated had overdiagnosis rates of 17 and 79 cases per patient screened. The PSA cutoffs used in these studies differed widely, with a cutoff of 10 in Ciatto et al and a cutoff of 3 in Draisma et al. This difference could lead to the lower overdiagnosis rates in the Ciatto study due to the detection of larger, more advanced cancers that would have both secreted more PSA and been more likely to present clinically.

The studies had a great deal of heterogeneity in many baseline characteristics, including intervention modeled in the study. All studies modeled PSA, yet the screening interval and cutoff for the test varied widely between studies. Three studies did not choose a specific screening interval to model as the intervention; these studies estimated current PSA screening rates from a linkage between the SEER database and Medicare claims and thus modeled the current screening situation. However, when doing this, difficulties arose distinguishing between screening PSA’s and diagnostic PSA’s from the database. Some of the studies introduced parameters to help control for this. For example, Etzioni et al incorporated a parameter “p” representing the proportion of all PSA-associated cases (cases of prostate cancer diagnosed within 3 months of a PSA test) that were screen-detected. They performed a sensitivity analysis using high, moderate and low values for p, and found that the difference in overdiagnosis using different values for “p” was less than 10 cases of overdiagnosis per 1000 patients with screen-
detected cancer, but tended to favor slightly increased overdiagnosis rates for low values for $p$. Other studies, including Draisma et al, modeled multiple interventions. Four studies modeled some type of biennial PSA screening; one of these used a cutoff of 10, two of these used a cutoff of 4, and the other used a cutoff of 3. Two studies modeled screening every 5 years and two modeled screening every 4 years. In addition, one study modeled PSA screening using an age-specific cutoff defined in a previous study.\textsuperscript{147} Thus, a great deal of heterogeneity is present in intervention modeled that makes it difficult to analyze study results.

In addition, the age distribution of the population of interest varied a great deal between the studies. Some of the age ranges modeled in these studies include all men over 50, men 50-84, men 55-74, men 60-74, men over age 55, men 60-84, men 30 to 95, men over age 60, and men over age 65. Overall, indications are that for higher age screened, the rates of corresponding overdiagnosis are also higher, which makes sense given the increasing risk of dying of a comorbid condition with increasing age. Some studies, such as Parker et al\textsuperscript{10} and Draisma et al,\textsuperscript{38} modeled multiple different age groups of interest and found that older patients did in general have higher overdiagnosis rates than younger patients.

The overall quality of the modeling studies varied, as seen in table 8. The internal validity of modeling studies depends on a number of factors, including completeness of the model, accuracy of parameters, similarity of comparison groups, assumptions, and length of follow up. All modeling studies make assumptions when combining different parameters in the study, and some of these modeling studies made more assumptions than others. Length of follow up is an important factor for studies based on service screening cohorts that affects the internal validity of these studies. The longer the length of follow up, the less likely any excess incidence beyond the expected incidence is due to prevalent cases.
One of the most common assumptions made was that the distribution of lead times in a given population followed a given distribution, such as an exponential distribution. It is not possible to know with certainty what the distribution of lead times is in a given population for prostate cancer because the data is not there. More aggressive cancers will have short lead times and less aggressive cancers will have longer lead times, but we cannot whether the distribution of these lead times is normal, exponential, bimodal, or skewed. It is more likely that the distribution of lead times does not follow a defined type of distribution. Studies that do not assume that lead time follows a specific distribution do not take it into account at all, and thus the quality of their estimate is highly dependent on length of follow up. Another common assumption was that parameters used in the study could be generalized to one another in one model even though they were determined in different populations. Finally, many models assume that the all-cause mortality in patients with screen-detected prostate cancer is similar enough to that of the population as a whole to use census mortality rates when generating date of death.

Follow up and contamination were two major quality issues in the studies based on service screening populations. Hamashima et al had follow up of 1 year and Ciatto et al had follow up of 9 years. Zappa et al had follow up of 4 years in the screened population, but modeled 14 years worth of follow up. When calculating an observed to expected incidence ratio in these studies based on service screening data, it is important that the cohort used to estimate the expected incidence is both similar to the observed cohort and free of contamination by opportunistic screening. Zappa et al calculated expected incidence based on age specific mortality rates from the Tuscany Cancer Registry in 1990-1991, a period in which they say screening was negligible. However, they offer no data to back up this claim. Ciatto et al calculated expected incidence based on age specific mortality rates from the Tuscany Cancer
Registry during the period of screening (1991-1994), a period in which they say screening was negligible. They mention that the screened cohort actually made up 3 percent of the individuals in the Tuscany Cancer Registry. Thus, 3 percent of the individuals in the expected group were contaminated by screening, and so overdiagnosis may be slightly underestimated in this study. In Hamashima et al, the expected incidence is calculated from incidence rates in 11 Japanese cancer registries. The authors do not mention the potential for opportunistic screening of these patients, thus the risk of contamination in this group is high.

The generalizability of these studies was greater in studies validated by population level data such as the SEER database because they were based on a wider spectrum of individuals, but some internal validity is sacrificed without validating the model against a randomized controlled trial or controlled cohort study. Generalizability suffers when patients were excluded from validation studies because of exclusion criteria or a narrow age range for screening. Studies that modeled blacks and whites separately have better external validity to these specific populations than those that modeled all races together.

**Implications for Overtreatment of Prostate Cancer**

The results of these modeling studies indicate that approximately 23 to 56% of all screen-detected prostate cancer cases will be overdiagnosed in a population-based screening program. Rates are higher in older patients and those with low stage disease. Overdiagnosed patients will be likely to undergo further diagnostic testing, including prostate biopsy, and also treatment. Little data exists about the proportion of overdiagnosed patients who are treated, but data does exist that focuses on the treatment of prostate cancer by stage and grade. It is difficult to estimate the likelihood of treatment of overdiagnosed cases because overdiagnosed individuals may have cancers of any clinical stage and grade. However, overdiagnosed cases are more
commonly lower risk cancers because they have longer lead times during which patients may die of comorbid conditions. Studies show that 76-85% of screen-detected cases are Gleason score < 7, the patients at greatest risk for overdiagnosis. It would make sense for treatment patterns to be more conservative in patients with low-risk cancers because it is less likely that they will die of their prostate cancer. However, current data suggests that this is likely not the case.

A retrospective cohort study based on the SEER registry estimated the proportion of men with "lower-risk" prostate cancers who were overtreated, assuming that any treatment other than expectant management was overtreatment. They defined "lower risk" as men with Gleason score 2-4 cancers or men over 70 years old with Gleason score 5-7 cancers. The authors found that 10 percent of 24,825 men with low risk cancers were overtreated with surgery and 45% of the men were overtreated with radiation therapy. This does not take into account the men who may have been treated unnecessarily with androgen deprivation therapy. The CaPSURE study, a survey of over 8,000 men with biopsy-proven prostate cancer, found that of the approximately 5,000 men with localized disease, only 5.5 percent of them claimed watchful waiting as primary disease management from 1998 to 2000. This proportion had actually declined from 7.5% in 1989-1991 and 9.5% in 1992-1994. The authors defined localized disease as clinical stage T3a or less with no evidence of lymph node involvement or metastasis. The results of the CaPSURE study indicate that as the likelihood of active treatment for prostate cancer increased throughout the 1990s, which is the same time period during which the introduction of PSA screening occurred. From 1999 to 2001, the authors found that radical treatment occurred in 77.5%, 75.8%, and 47% of low-, intermediate-, and high-risk patients, respectively. Thus, low risk patients were much more likely to receive radical treatment than patients with high-risk cancers.
Bill-Axelson et al. performed a randomized controlled trial directly comparing radical prostatectomy and watchful waiting in 695 men under 75 years of age with a life expectancy of 10 years with newly diagnosed stage T1b, T1, or T2 cancers. The authors found that fewer men died in the surgery group than in the watchful waiting group (30 vs 50; p = 0.01). However, this study excludes individuals at highest risk of overdiagnosis by excluding men over 65 and those with a life expectancy of less than 10 years. In addition, the authors found that men over 65 years of age had a lower cumulative incidence of prostate cancer specific death. Parker et al. used data from a 15 year cohort study of patients with conservatively treated prostate cancer in a model to estimate the effect of radical treatment on survival. They included in their model an estimate of overdiagnosis in these patients based on the model used by Draisma et al as well as the hazard ratio from Bill-Axelson et al. They found that survival benefit was greatest in patients with high grade disease, which conflicts with the current patterns of care from the CaPSURE study. Thus patients with lower grade disease seem to be at high risk for invasive treatment with little survival benefit, placing them at high risk of overtreatment.

Significant side effects can occur as the result of treatment for prostate cancer, including impotence, urinary incontinence, and pain. These problems occur more often in older patients, thus many overdiagnosed patients are at higher risk of treatment-related adverse effects. The Prostate Cancer Outcomes Study found that older men were more likely to have high levels of incontinence after surgery (14% of 75-79 year olds as compared to 0.7-4% of younger men). In addition, men over the age of 60 were more likely to have problems with impotence after surgery than younger men (78-85% vs. 61%). Similar problems can occur after external beam radiotherapy and brachytherapy.
Active surveillance is one treatment option for patients with low risk cancers that could help limit the extent of overtreatment and adverse effects. Active surveillance is an expectant management protocol with frequent monitoring of PSA doubling time, velocity, or amplitude and symptomatology with the option of curative therapy should progression occur. Some algorithms include repeat biopsies of the prostate to monitor progression. Active surveillance is inherently a more proactive approach than watchful waiting because patients are monitored so closely with biochemical and clinical parameters. An example of an active surveillance protocol, developed by Choo et al,$^{151}$ stated the criteria for progression as a PSA doubling time $< 2$ years based on 3 separate measurements over 6 months, a final PSA $> 8$ or a Gleason score $> 7$ on repeat biopsy of the prostate at 12-18 months. In addition, patients were considered to progress if they had an increase of more than double of the maximum perpendicular diameters of the lesion, local progression requiring TURP, ureteric obstruction, or evidence of metastases. Active surveillance protocols are one way in which treatment can be tailored to the clinical situation in individual patients to avoid overtreatment.

Thus according to current estimates, up to 56% of prostate cancers seem to be overdiagnosed, with higher rates of overdiagnosis in older patients and those with low stage disease, the same patients who seem to be at greatest risk for overtreatment according to the results of the above studies. Thus the risk of overdiagnosis and overtreatment of screen-detected prostate cancer are high. In addition, the risk of side effects from treatment is highest in older patients who are at greater risk of overdiagnosis due to comorbidities.

Yet problems with study quality, including lead time distribution assumptions and other assumptions inherent to the modeling process, limit estimates of overdiagnosis. Even the highest quality studies make significant assumptions about lead time, expected incidence, and screening
vs. diagnostic PSA tests. Without high quality randomized controlled trial data with lengthy follow up, it is difficult to assess the true amount of overdiagnosis that occurs. In addition, it is extremely difficult to assess to what extent overtreatment occurs in these patients without following them on an individual basis for a long period of time, which would likely require 15 or 20 years of follow up. The ongoing European Randomized Study of Screening for Prostate Cancer (ERSPC) and the U.S. National Cancer Institute “Prostate, Lung, Colorectal, and Ovary” Trial (PLCO) will eventually provide a much better estimate of the true amount of overdiagnosis. Until then, it is likely that prostate cancer screening with PSA leads to a substantial proportion of overdiagnosed cases, with our best estimates at 23 to 56% of all screen-detected cancers, with higher rates for older patients and low stage cancers. Active surveillance is a good option for patients with low stage disease to avoid overtreatment.

**Overdiagnosis in Breast Cancer Screening**

Estimates of overdiagnosis in breast cancer screening were highly dependent on whether the study included a method to adjust for lead time and test sensitivity. The studies which did adjust for lead time found rates of overdiagnosis ranging from 31 cases per 1000 cases of breast cancer to 80 cases per 1000 cases of breast cancer. One of these studies based its model on randomized controlled trial data and the other four based their models on service screening data. Studies which did not adjust for lead time found higher overdiagnosis rates, ranging from 91 cases per 1000 cases of breast cancer to 351 cases per 1000 cases of breast cancer. Two of these studies based their models on RCT data, and the other two studies based their models on service screening data. I excluded two studies from the above analysis due to poor quality.
The highest quality studies, Zackrisson et al\textsuperscript{91} and Gotzsche et al,\textsuperscript{86} were based on randomized controlled data and neither of these studies included a measure with which to adjust for lead time. Zackrisson et al found an overdiagnosis rate of 99 per 1000 patients screened, and Gotzsche et al found a higher rate of 230 per 1000 patients screened. Zackrisson et al followed patients for 15 years after the end of the Malmo trial, whereas follow up in the studies analyzed by Gotzsche et al ranged from 7 to 9 years. The longer follow up in the Zackrisson study may help eliminate some cases of breast cancer that are contributing to excess incidence due to lead time. However, the Zackrisson trial also assumes that no 55-69 year old women in the control group received mammography after the end of the trial, which certainly leads to an underestimation of overdiagnosis rates. The shorter follow up in the Gotzsche study is concerning in that some cases of breast cancer, particularly DCIS, may still be cases that were destined to present clinically, even after 7-9 years of follow up. The Gotzsche estimate also has potential for contamination in the control group that may lead to an underestimation of overdiagnosis rates. A random sample of 500 women in the Malmo trial control group found a contamination rate of 24 percent during the trial. Contamination is difficult to avoid in these studies due to the widespread use of mammography by the time many of the trials were conducted.

Studies which incorporate methods of adjusting for lead time found lower rates of overdiagnosis than the above studies. Duffy et al,\textsuperscript{5} de Koning et al,\textsuperscript{83} and Olsen et al\textsuperscript{88} used the model of Day and colleagues to adjust for lead time in their estimates of overdiagnosis. Duffy et al was based on two randomized controlled trials and de Koning et al and Olsen et al were based on service screening data. Rates of overdiagnosis were lower in these studies, ranging from 31 to 80 cases per 1000 breast cancer cases. The primary faults with these studies are the multiple
assumptions needed in the model to adjust for lead time. Similar to many of the prostate cancer models, this model assumes an exponential distribution of lead times in the study population. It is not possible to define the distribution of lead times in the population, and assuming a specific distribution of lead times may lead to an underestimation of overdiagnosis if the distribution of lead times is more bimodal, with a high proportion of both short and long lead times. Patients with longer lead times would be more at risk of overdiagnosis.

The two studies by Paci et al\textsuperscript{89,135} also include methods for adjusting for lead time. Yet these studies also assume an exponential distribution of lead times with mean lead time durations of 3.4 years in the 2004 study and 3.7 or 4.2 years in the 2006 study, depending on the patient age. In addition, these studies are based on service screening populations for which the potential for contamination in the “prescreening” group from which the authors estimated expected incidence remains unclear. All studies based on service screening populations also assume that predicted trends in incidence can be estimated from prescreening data, when in fact external forces such as the increasing use of hormone replacement therapy, may have led to an increase in breast cancer diagnosis rates independent of screening.

Zahl et al\textsuperscript{102} used service screening data to estimate overdiagnosis rates without adjusting for lead time and found rates of 310 overdiagnosed cases per 1000 breast cancer cases in Sweden and 351 overdiagnosed cases per 1000 breast cancer cases in Norway. Follow up was 5 years for Norway and 10 years in Sweden, which may still be somewhat short in order to avoid the effect of lead time on excess incidence. In addition, the authors assume that there is no baseline increase in breast cancer incidence rates due to factors other than screening, and thus assume that all excess incidence is a direct result of overdiagnosis. In addition, increased reporting of breast cancer cases to their nationwide registry could have led to falsely elevated rates of overdiagnosis.
The authors do not discuss the potential for contamination and do not include DCIS in their study, which could both lead to an underestimation of overdiagnosis rates. Similar to the other service screening studies above, the multiple assumptions used in this study make it difficult to figure out what the results mean.

Differences in the study characteristics within the different included studies also make it difficult to generalize the study results to a specific population. Different study populations and different interventions were used in the studies. All of the studies included some type of mammography screening, but screening intervals varied. Some studies included 2 view mammography, some included one view mammography, and others did not specify what type of mammography they used in the study. Many studies modeled biennial mammography, but some studies used annual mammography, and one study used mammography every 18 to 24 months. In addition, the age intervals screened differed between the studies. Screening began at age 50 in five studies, at age 45 in two studies, at age 40 in four studies, and at age 35 in one study. Most studies stopped screening women at age 69, but a few studies continued to screen women until age 74 or 79. These studies would be expected to show higher rates of overdiagnosis because of the higher risk of comorbid disease in elderly women.

Generalizability of the studies is most likely better in the more population-based service screening studies because they patients were not excluded like they were in some of the randomized controlled trial data. However, the women in these studies are volunteers who are inherently different in unknown ways from the population as a whole. In addition, most of the service screening studies were conducted in fairly small countries with ethnically uniform populations, and thus their results may not be generalizable to more diverse populations such as the United States.
Implications for Overtreatment of Breast Cancer

The highest quality study above\(^8\) concluded that overdiagnosis occurred in about 23% of all screen-detected cases of breast cancer. Other estimates of overdiagnosis are as low as 3 percent when adjustment for lead time occurs. However, the amount of contamination in these studies is unknown, and increases in the baseline incidence of breast cancer by causes other than screening are not always accounted for. Thus, overdiagnosis is likely a significant problem for breast cancer screening with mammography, and even more of a problem for DCIS due to its longer lead time. This problem is compounded by the high proportion of women with breast cancer treated invasively, which places many overdiagnosed women at substantial risk for overtreatment.

The primary treatment options for breast cancer include lumpectomy, lumpectomy with radiation, and mastectomy. Unlike prostate cancer, watchful waiting has historically not been an option for women with breast cancer. In 1999, over 28% of women with DCIS had mastectomy, 64% had lumpectomy, and 33% had radiation.\(^2\) Thus, there is potential for a substantial amount of overtreatment as a result of DCIS. Ernster et al performed a retrospective study of over 540,000 women who underwent mammography and found that 18 percent of screen-detected cancers were cases of DCIS.\(^1\) Although none of the included studies above specifically modeled the rate of overdiagnosis of DCIS alone, one study that modeled the incidence of non-progressive and progressive DCIS using incidence data from the Swedish Two County randomized controlled trial and multiple service screening programs found that 37% of DCIS cases detected by screening were nonprogressive.\(^10\) The authors defined nonprogressive DCIS as having no propensity to progress to invasive disease during the lifetime of the host. They estimated that at prevalence screen 30 in every 100,000 women screened would be diagnosed
with non-progressive DCIS, as compared to 572 women per 1000,000 women screened that would by diagnosed with a progressive DCIS or invasive cancer. This seems to indicate that DCIS due to screening is a minor problem in screening. One caveat to this study, however, is that it only included women up to age 69 years of age and thus excludes elderly women who have the highest rates of screen-detected DCIS. 18

In addition, other studies have shown that DCIS, progressive or not, is unlikely to lead to death in women aged 70 years of age and older. One modeling study found that 65 cases of DCIS will be detected per every 10,000 women aged 70 and older screened for 10 years. 152 If all of these cases were treated with surgery or, only 1 death from invasive breast cancer would be averted. Thus the other 64 women would be overtreated and subject to unnecessary side effects of surgery such as anxiety, infection, and even death. Thus, although the detection of clinically relevant DCIS cases is a relatively minor phenomenon in breast cancer screening, these women are at low risk of death and thus high risk of overtreatment.

Regardless of the proportion of screen-detected DCIS, it is clear that the vast majority of screen-detected breast cancers are treated surgically. The invasive treatment of any type of breast cancer, DCIS or invasive, in an overdiagnosed patient would be overtreatment. Thus, if we assume that 23% of screen-detected breast cancers are overdiagnosed with higher risks of overdiagnosis in older patients, then all of these individuals will be overtreated with surgical therapy and possibly radiation and will be at risk of the side effects of treatment. One option for decreasing the risks of overdiagnosis and overtreatment due to breast cancer screening would be setting a stopping point for screening at a certain age. The risks of screening, including the potential for overdiagnosis and overtreatment, begin to outweigh the benefits as patients progress into their seventies. It remains unclear whether mammography leads to a reduction in mortality
in women over seventy years of age, and continuing screening past age 70 in women at low risk of breast cancer has been found to result in a small gain in life expectancy while having low cost-efficacy.\textsuperscript{152}

The true percentage of overdiagnosis in mammography screening remains unclear, but the highest quality estimate at this time is 23 percent of detected cases.\textsuperscript{86} This study had relatively short follow up and substantial contamination (about 24% of the control group in one study), but it does not suffer from the multiple assumptions that many of the modeling studies based on service screen populations have. Better estimates of overdiagnosis are unlikely to be made in the future because of the high prevalence of mammography screening around the world. It is unlikely that control patients in the randomized controlled trials of breast cancer screening will remain free of mammography screening. Many of the trials invited the control group for screening at the end of the trial. This contamination makes it extremely difficult to measure the expected incidence in the absence of screening, even with extremely long follow up. As follow up increases, the likelihood of control patients being screened by mammography increases. It is not ethical to withhold mammography screening from control group women because RCT’s have shown that it reduces mortality.\textsuperscript{16} Therefore, it is unlikely that more accurate measurements of the prevalence of overdiagnosis in RCT’s of mammography screening will be made.

**Comparison of Screening for Prostate Cancer with Screening for Breast Cancer**

Overall, overdiagnosis seems to be a smaller problem for breast cancer screening than prostate cancer screening, but the rates of overdiagnosis are substantial in both diseases. Good randomized controlled trial data is present in breast cancer screening and is absent in prostate cancer screening, thus most estimates of overdiagnosis in prostate cancer screening come from modeling studies based on service screens and SEER database studies. The one study that was
Based on preliminary RCT data found an overdiagnosis rate of 56% of screen-detected cases. Nearly all of the prostate cancer studies included an adjustment for lead time, assuming an exponential distribution of lead times, because there was no long-term follow up of high-quality RCT data. The breast cancer studies showed fairly dichotomous results when lead times were modeled based on an exponential distribution as compared to when there was no adjustment for lead times. Adjustment for lead times led to lower rates of overdiagnosis, from 3.1 to 8.0% of breast cancer cases, as compared to 9.9 to 35.1% of breast cancer cases when authors did not adjust for the effect of lead time. Overall results of the highest quality studies were overdiagnosis rates of 23 to 56% of prostate cancer cases and 10 to 23% of breast cancer cases.

Rates of overdiagnosis in prostate cancer screening tended to be higher in older patients and those with lower stage disease, whereas rates of overdiagnosis in breast cancer screening also tended to be higher in older patients. It is likely that rates of overdiagnosis are higher in patients with DCIS, although this is unclear because none of the studies modeled DCIS separately. However, in multiple studies the rates of overdiagnosis of invasive and in situ cancers were higher than the rates of invasive cancers alone, indicating that patients with DCIS are at higher risk of overdiagnosis. In addition,

A high percentage of screen-detected cases in both conditions are treated invasively, with higher rates of invasive treatment occurring in breast cancer. Invasive treatment for breast cancer by surgery or radiation is nearly universal because there is not a treatment algorithm similar to watchful waiting in prostate cancer. This is most likely because DCIS makes up only 18% of screen-detected breast cancer cases, whereas low risk prostate cancer may occur in as many as 47% of screen-detected patients. A higher proportion of patients with screen-detected prostate cancer thus have low risk disease that would never present clinically. Yet although the
risk of overdiagnosis seems to be higher for prostate cancer, the risk of overtreatment of breast cancer is higher because essentially all screen-detected breast cancers are treated invasively with surgery or radiation, whereas the study by Miller et al. found that only 10% of localized prostate cancers are overtreated by surgery and 45% by radiation.

Thus, of 1000 screen-detected cases of prostate cancer, 230 to 560 cases would be overdiagnosed. If 560 cases were overdiagnosed, and we assume that these are all localized prostate cancers, then 56 cases would be overtreated by surgery and 252 cases by radiation for a total of 308 overtreated cases per 1000 screen-detected cases. Of 1000 screen-detected cases of breast cancer, 230 would be overdiagnosed based on the highest quality estimate. All of these cases would be overtreated by surgery or radiation, leading to a total of 230 overtreated cases.

Thus, despite a higher risk of overdiagnosis in prostate cancer, the use of conservative treatment for localized prostate cancer and the use of universally invasive treatment in breast cancer have led to very similar magnitudes of overtreatment in prostate cancer and breast cancer screening. Of course, not all overdiagnosed prostate cancer cases are localized disease, and so the estimate of prostate cancer overdiagnosis is overestimated. However, it is likely that the majority of overdiagnosed prostate cancer cases are localized due to longer lead times. Overall, about \( \frac{1}{4} \) of screen-detected cases in both prostate cancer and breast cancer can be expected to be overtreated.

Many issues with study quality affect the accuracy of the overdiagnosis estimates, including length of follow up, assuming an exponential distribution for lead time, and contamination in the control groups. Long term follow up of the ERSPC and PLCO randomized controlled trials of prostate cancer screening will provide better estimates of the true overdiagnosis of prostate cancer due to screening. However, because it has been shown that
breast cancer screening reduces mortality, it is unlikely that better RCT estimates of overdiagnosis due to screening will be forthcoming.

**Conclusion**

Overdiagnosis and overtreatment represent important harms in both prostate cancer screening and breast cancer screening. About \( \frac{1}{4} \) of screen-detected cases can be expected to be overtreated with invasive therapy. Thus, treatment should be used judiciously in screen-detected patients at low risk of disease progression, particularly older patients with lower stage disease in prostate cancer screening and older patients with DCIS in breast cancer screening. Active surveillance of patients with low risk prostate cancer can be used to minimize the risks of overtreatment of this population. In addition, careful consideration should be given to the risks and benefits of surgical treatment for elderly women with DCIS, particularly those over age 70.
<table>
<thead>
<tr>
<th>Study</th>
<th>Study Design</th>
<th>Definition and Expression of Overdiagnosis</th>
<th>Sample Characteristics</th>
<th>Background Studies, Validation, and Parameter Estimation</th>
<th>Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telesca 2007</td>
<td>Stochastic Simulation Model</td>
<td>OD expressed as the fraction of cases detected by PSA screening that, in the absence of the test, would not have been diagnosed within the individuals' lifetimes</td>
<td>Hypothetical cohort of 1 million men between 50-84 yo in the US</td>
<td>Incidence data from US SEER database 1973 to 2000. PSA screening rates from retrospective analysis of SEER database, Medicare claims, and NHIS. Detection rates estimated from previous studies. Mortality from NCHS Vital Stats.</td>
<td>Based on current rates of PSA screening from the Medicare/SEER claims database using PSA &gt; 4.</td>
</tr>
<tr>
<td>Hamashima 2006</td>
<td>Service Screening Model</td>
<td>No explicit definition O/E ratio calculated using: E = I X (P/100000) X ST X SE I = incidence P = population ST = sojourn time (an estimate of lead time) SE = sensitivity Expressed as the percentage increase in prostate cancer diagnosis in the screened population as compared to the unscreened population</td>
<td>2061 male volunteers over 50 enrolled from Tokyo 2/04 - 1/05. Men with previous diagnosis of cancer or self report of precancerous disease were excluded.</td>
<td>Incidence estimated from 11 Cancer Registries in Japan. Sojourn time estimated as 5-15 yrs from 8 previous studies. It is unclear how sensitivity was estimated.</td>
<td>Serum PSA using a cutoff of 2.7 ng/mL</td>
</tr>
<tr>
<td>Parker 2006</td>
<td>Competing-risks hazard model</td>
<td>Overdetection = the detection of cases by screening that otherwise would not have been detected Expressed as the proportion of screen-detected cases of prostate cancer that is overdiagnosed.</td>
<td>Hypothetical cohort of 1 million men with prostate cancer</td>
<td>Survival data from Albertsen 1998 cohort study: 767 men aged 55 to 74 years with clinically localized prostate cancer diagnosed between 1971 and 1984 were treated conservatively for up to 10-20 years. Overdiagnosis estimates validated by previous Draisma estimates.</td>
<td>Biennial PSA screening with cutoff of 3</td>
</tr>
<tr>
<td>Study</td>
<td>Study Design</td>
<td>Definition and Expression of Overdiagnosis</td>
<td>Sample Characteristics</td>
<td>Background Studies, Validation, and Parameter Estimation</td>
<td>Intervention</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------</td>
<td>-----------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Tsodikov 2006\textsuperscript{33}</td>
<td>Mathematical model taken directly from population data (SEER database).</td>
<td>Expressed as the proportion of prostate cancers identified through screening that would have never been detected in the absence of screening.</td>
<td>Incidence data from 350,000 cases of prostate cancer obtained from SEER database using age interval 50-85 from 1973-2000.</td>
<td>Risk of death from other causes derived from the Human Mortality Database. Age at tumor onset modeled by parametric distributions (Weibull, Gamma, MVK). Sojourn time modeled as a Weibull distribution dependent on age and secular trends in detection. PSA screening rates modeled from SEER database. Compare observed SEER incidence by year and age with model predicted incidence for validation.</td>
<td>PSA; available data does not distinguish b/t screening and dx PSA. Tests performed within 3 mos of diagnosis were assumed to be diagnostic. No cutoff specified, but 4 is likely. Assumes screening is a random point process in the population, thus does not model a specific screening regimen</td>
</tr>
<tr>
<td>Ciatto 2005\textsuperscript{34}</td>
<td>Service Screening Model</td>
<td>Detection of latent non-aggressive cancers that will never become clinically evident Expressed as the % increase in prostate cancer diagnosis in the screened population as compared to the unscreened population</td>
<td>6,890 volunteer men aged 60-74 invited in 2 cohorts from a random sample of Italian National Health Service registered GP offices 1991-1994. Excl: Pts with disabling illnesses, unlikely to attend invitation, or w/ PCA.</td>
<td>Expected incidences were taken from the Tuscany Cancer registry age-specific incidence rates. No validation.</td>
<td>2 biennial screening rounds of PSA (cutoff &gt; 10), DRE + TRUS</td>
</tr>
<tr>
<td>Davidov 2004\textsuperscript{35}</td>
<td>Mathematical Model</td>
<td>When a screening exam detects a disease that would have otherwise been undetected in a person's lifetime because the individual would have died of other causes prior to clinical onset</td>
<td>No population of interest stated, but SEER database used for estimation of incidence.</td>
<td>All-cause Mortality from Period Life Table by the SSA describing mortality of US males in 1997. Sojourn time estimates of 5-15 yrs are from 8 previous studies. Incidence from SEER database used in model. Sensitivity estimates of seem arbitrary. No validation.</td>
<td>screening every 5 years with PSA (assume data uses cutoff of 4) in 50-60 yos, 50-70 yos, and 50-80 yos</td>
</tr>
</tbody>
</table>
Table 6: Prostate Cancer Study Characteristics (Cont’d)

<table>
<thead>
<tr>
<th>Study</th>
<th>Study Design</th>
<th>Definition and Expression of Overdiagnosis</th>
<th>Sample Characteristics</th>
<th>Background Studies, Validation, and Parameter Estimation</th>
<th>Intervention</th>
</tr>
</thead>
</table>
| Draisma 2003³⁸ | Markov Model validated by RCT and baseline incidence data | Overdetection = screen-detected cancers that would not have been diagnosed in the absence of screening. Expressed overdiagnosis as: 
1. the relative increase, caused by screening, of the number of men with a cancer diagnosed during their lifetime 
2. the fraction of irrelevant cancers of all detected cancers | Hypothetical cohort of 1 million men aged 55 at the start of screening | Dutch male life table used to estimate time of death from other causes CI (0.19) and stage-specific sensitivities (0.64, 0.91, 0.97 for localized, regional, and distant stages) estimated by the model to fit ERSPC results. Validated in 3 ways: 1. ERSPC Rotterdam; ²⁶³ 42376 men aged 55-74 in Rotterdam pop. registry who were invited to participate and accepted. These men were randomly assigned to control or screening arms (21166 control, 21210 screen). 6 years of follow up in this study. 2. Baseline incidence of prostate cancer in The Netherlands in 1991 (prescreening) 3. Baseline stage distribution of clinically diagnosed cancers in the Rotterdam Cancer Registry for 1992 and 1993 | 19970 men in screen arm received first screening test. First 9766 men received DRE, TRUS, and PSA. Biopsy if DRE or TRUS abnormal or for PSA > 4. 10204 men in first round and all 3545 who had second test received PSA. Sextant biopsy of these men for PSA > 3. Modeled single screening at 55, 60, 65, 70, 75; annual screening at 55-67 and 55-75, and q4 year screening at 55-67 and 55-75.
Table 6: Prostate Cancer Study Characteristics (Cont’d)

<table>
<thead>
<tr>
<th>Study</th>
<th>Study Design</th>
<th>Definition and Expression of Overdiagnosis</th>
<th>Sample Characteristics</th>
<th>Background Studies, Validation, and Parameter Estimation</th>
<th>Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etzioni 2002&lt;sup&gt;11&lt;/sup&gt;</td>
<td>Stochastic Simulation Model</td>
<td>Overdiagnosis = the detection, through screening, of disease that would never have been diagnosed in the absence of such screening Expressed as the fraction of cases detected by PSA screening that, in the absence of the test, would not have been diagnosed within the individuals' lifetimes (the probability of dying of other causes during the lead time)</td>
<td>Hypothetical cohort of 2 million men aged 60-84 in 1988</td>
<td>Age distribution and age-specific mortality rates from census data. PSA testing rates and cancer detection rates for 1988-1998 from US SEER and Medicare claims Databases for men 60-84.&lt;sup&gt;164&lt;/sup&gt;,&lt;sup&gt;165&lt;/sup&gt; Lead times of 3, 5, and 7 years were used from previous studies.&lt;sup&gt;42&lt;/sup&gt;,&lt;sup&gt;158&lt;/sup&gt;,&lt;sup&gt;159&lt;/sup&gt; To estimate the secular trend, they used trends in TURP detected cancers up to 1988 extrapolated out to 1998 in absence of PSA. 10 years of follow up in this data. Not much validation in this model other than SEER incidence.</td>
<td>Modeled 1988-1998 amount of screening using PSA with a cutoff of 4</td>
</tr>
<tr>
<td>Study</td>
<td>Study Design</td>
<td>Definition and Expression of Overdiagnosis</td>
<td>Sample Characteristics</td>
<td>Background Studies, Validation, and Parameter Estimation</td>
<td>Intervention</td>
</tr>
<tr>
<td>------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------</td>
<td>------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Etzioni 1999&lt;sup&gt;43&lt;/sup&gt;</td>
<td>Markov Model</td>
<td>Overdiagnosis = cases of prostate cancer detected by screening that would have never been diagnosed in the absence of screening Expressed as: the number of patients with screen-detected prostate cancer without clinical presentation before death / all patients with stage A1 disease</td>
<td>Hypothetical cohort of 200000 men 30 to 95</td>
<td>All cause mortality, asymptomatic incidence, stage transition rates from previous Markov Model&lt;sup&gt;165&lt;/sup&gt; Death from other causes from US life tables from NCHS. Data from Pre-PSA era used for probability of clinical presentation and stage distribution&lt;sup&gt;167&lt;/sup&gt; PSA distribution from previous study&lt;sup&gt;147&lt;/sup&gt; PCA death rates without screening are from SEER 1973-1987 data. Validation: Modeled age-specific incidence and survival compared to SEER 1984-1988. Modeled sojourn time of 10-11 yrs compared to known estimates&lt;sup&gt;162&lt;/sup&gt; Stage distribution validated against pvs study&lt;sup&gt;154&lt;/sup&gt; Detection rate and positive predictive value compared to pvs cohort study&lt;sup&gt;155&lt;/sup&gt; Sensitivity and specificity validated against retrospective cohort&lt;sup&gt;159&lt;/sup&gt;</td>
<td>5 Interventions Modeled, including annual PSA screening with a cutoff of 4, biennial PSA screening with a cutoff of 4, q5 year PSA screening with a cutoff of 4, annual PSA screening with an age-specific cutoff, and biennial PSA screening with an age-specific cutoff</td>
</tr>
<tr>
<td>Zappa 1998&lt;sup&gt;45&lt;/sup&gt;</td>
<td>Service Screen Model using Monte-Carlo sampling for simulation</td>
<td>Overdiagnosis = Expressed as the proportional excess of cancers detected by screening with respect to that expected in its absence</td>
<td>2 Hypothetical cohorts of 10,000 men each aged 60 or 65 at screen initiation</td>
<td>Model Used detection rates and sensitivities from Florence Pilot Studies: 2740 males living in Florence aged 60-74 years from 1992 to 1995&lt;sup&gt;168&lt;/sup&gt; screened twice over 4 years of follow up</td>
<td>biennial PSA for 5 consecutive rounds</td>
</tr>
<tr>
<td>Study</td>
<td>Reported Results</td>
<td>Overdiagnosed Cases / 1000 Pts Screened</td>
<td>Overdiagnosed Cases / 1000 Cases Detected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Telesca 2007²³ | Overdiagnosis in whites = 22.7%  
Overdiagnosis in blacks = 34.4%                                                      | Cannot calculate without more information, such as the actual number of overdiagnosed cases in whites and blacks | Whites = 227 / 1000 cases  
Blacks = 344 / 1000 cases                |
| Hamashima 2006²⁸ | O/E ratio = 3.43 (p = 0.0022) for ST = 10, SE = 70%  
O/E ratio = 5.36 – 16.07 when ST = 5 and SE = 30-90%.  
O/E ratio > 1 for all ST = 15 yr analyses | 17 excess cancers / 2061 men screened = 8.2 overdiagnosed cases / 1000 patients screened                   | 17 excess cancers / 24 detected cancers = 708 overdiagnosed cases / 1000 cases detected |
| Parker 2006²⁰  | Overdetection rates ranged widely depending on age group and Gleason score  
55-59 yo's with Gleason > 7 had 7.3% OD  
70-74 yo's with Gleason < 7 had 76.6% OD  
Sensitivity ranges were extremely wide for all estimates | Used simulated cohort of 1 million men  
Impossible to calculate this rate without detection rates within each cohort group for biennial screening (not available) |                                            |
| Tsodikov 2006²³ | % overdiagnosis ranges between 58% and 30% depending on the age of the patient; overall risk of overdiagnosis for men entering the age risk zone for prostate cancer = 25%  
Figure 7: overdiagnosis increases from ~10% at the age of 55 to ~70% at the age of 95 | Impossible to calculate this rate without detection rates within each age cohort | Estimated based on graph:  
Age | Overdiagnosis Rate |                                          |
| Ciatto 2005²⁴  | SIR = 1.66% for screened individuals, and 1.26 for all individuals (possible selection bias/external validity issue here). This is over 9 years of follow-up. After 5 years, SIR = 1.17 in invited, screened pts and 1.1 in all patients. | Over 9 years f/u:  
Observed – Expected / # screened = 45.5 / 2664 = 17 cases / 1000 pts screened  
After 5 years f/u:  
(38 – 32.4) / 2664 = 2 cases / 1000 pts screened (fewer pts likely screened 5 yrs after initiation) | Over 9 years f/u:  
Observed – expected / observed = 45.5 / 114 = 399 OD cases / 1000 cases  
After 5 years f/u  
(38 – 32.4) / 38 = 147 OD cases / 1000 cases |
<table>
<thead>
<tr>
<th>Study</th>
<th>Reported Results</th>
<th>Overdiagnosed Cases / 1000 Pts Screened</th>
<th>Overdiagnosed Cases / 1000 Cases Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Davidov 2004²⁵</td>
<td>Overall range = 8.48% to 53.6% depending on lead time, sensitivity, and exam schedule. For 50 – 75 yo men with 10 yr lead time and 0.9 sensitivity, OD = 30% (33% if screened annually)</td>
<td>Cannot be calculated from this model</td>
<td>For Sn = 0.7</td>
</tr>
<tr>
<td></td>
<td>MST = mean sojourn time</td>
<td></td>
<td>For 50-75 q3yr protocol, 10 yr MST, Sn = 0.9, OD = 300 / 1000 cases, or 330 / 1000 cases if annual screening occurs</td>
</tr>
<tr>
<td>D raisma 2003²⁶</td>
<td>Single screening at: 55 = 4 irrelevant cases / 1000 men screened = 27% of detection; 6% increased lifetime risk of PCA dx 60 = 12/1000 = 38%, 18% 65 = 24/1000 = 47%; 38% 70 = 34/1000 = 53%; 54% 75 = 30/1000 = 56%; 47% Annual screening from 55-67 = 51 irrelevant cancers / 1000 men screened = 50% of detection; 80% increased lifetime risk Annual screening from age 55 to 75 = 79/1000 = 56%; 124% 4yr screening from 55-67 = 41/1000 = 48%; 65% 4yr screening from 55-75 = 66/1000 = 54%; 105%</td>
<td>Single Screening at: 55 = 4 OD cases / 1000 men screened 60 = 12 OD cases / 1000 men screened 65 = 24 OD cases / 1000 men screened 70 = 34 OD cases / 1000 men screened 75 = 30 OD cases / 1000 men screened Annual Screening: 55-67 = 51 OD cases / 1000 men screened 55-75 = 79 OD cases / 1000 men screened Screening every 4 years: 55-67 = 41 OD cases / 1000 men screened 55-75 = 66 OD cases / 1000 men screened</td>
<td>Single Screening at: 55 = 270 OD cases / 1000 cases 60 = 380 OD cases / 1000 cases 65 = 470 OD cases / 1000 cases 70 = 530 OD cases / 1000 cases 75 = 560 OD cases / 1000 cases Annual Screening: 55-67 = 500 OD cases / 1000 cases 55-75 = 560 OD cases / 1000 cases Screening every 4 years: 55-67 = 480 OD cases / 1000 cases 55-75 = 540 OD cases / 1000 cases</td>
</tr>
</tbody>
</table>
Table 7: Prostate Cancer Study Results (Cont'd)

<table>
<thead>
<tr>
<th>Study</th>
<th>Reported Results</th>
<th>Overdiagnosed Cases / 1000 Pts Screened</th>
<th>Overdiagnosed Cases / 1000 Cases Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etzioni 2002&lt;sup&gt;41&lt;/sup&gt;</td>
<td>Overall rates of overdiagnosis = 29% for whites and 44% for blacks (best-fitting lead time for blacks was 7 years, for whites it was 5 years). Therefore 29% and 44% of screen-detected cases would be overdiagnosed in whites and blacks, respectively. For lead time of 5 years, OD = 28.8% for whites and 32.6% for blacks. For lead time of 7 years, OD = 39.5% for whites and 43.8% for blacks.</td>
<td>Cohort of 2 million screened—detection rates different depending on the pt age, year of screening, and race. Cannot be calculated without explicit numbers of overdiagnosed cases projected to occur in each race cohort per 2 million screened.</td>
<td>Assuming High “p” value (conservative estimate of OD). LT OD Rate (W,B).</td>
</tr>
</tbody>
</table>
| Etzioni 1999<sup>43</sup> | Annual PSA > 4 = 24.7% of stage A1 cancers were overdiagnosed. Biannual PSA > 4 = 23.1% overdiagnosed. Annual age specific = 16.8% overdiagnosed. Biannual age specific = 15.4% overdiagnosed. 5-year PSA > 4 = 20.5% overdiagnosed. | OD Rate X 80,000 stage A1 cancers / 200,000 persons screened x 1000 = rate / 1000 pts screened. Annual PSA > 4 = 98.8 OD cases / 1000 patients screened. Biannual PSA > 4 = 92.4 OD cases / 1000 pts screened. Annual age specific = 67.2 OD cases / 1000 pts screened. Biannual age specific = 61.6 OD cases / 1000 pts screened. 5-year PSA > 4 = 82 OD cases / 1000 pts screened. | Annual PSA>4.  
-247 OD cases / 1000 stage A1 cases. Biannual PSA>4.  
-231 OD cases / 1000 cases. Annual age-specific PSA cutoffs.  
-168 OD cases / 1000 cases. Biannual age-specific PSA cutoffs.  
-154 OD cases / 1000 cases. q5 yr PSA>4.  
-205 OD cases / 1000 cases. |
| Zappa 1998<sup>45</sup> | For Constant incidence, OD = 51% for age 60; 93% for age 65. If 2% yearly increment in incidence, OD = 25% for age 60; 65% for age 65. | Assumption=100% attendance for 10,000 men screened by biennial PSA x 5 rounds. (Expected cancers with screen – expected cancers w/o screen) / 10,000 screened x 1000 = OD rate. Baseline Inc. Age OD Rate. |
|                         | Assumption=100% attendance for 10,000 men screened by biennial PSA x 5 rounds. (Expected cancers with screen – expected cancers w/o screen) / 10,000 screened x 1000 = OD rate. Baseline Inc. Age OD Rate. | (Expected cancers with screen – expected cancers w/o screen) / expected cancers with screen x 1000 = OD rate. Baseline Inc. Age OD Rate. | (Expected cancers with screen – expected cancers w/o screen) / expected cancers with screen x 1000 = OD rate. Baseline Inc. Age OD Rate. |

<table>
<thead>
<tr>
<th>Study</th>
<th>Reported Results</th>
<th>Overdiagnosed Cases / 1000 Pts Screened</th>
<th>Overdiagnosed Cases / 1000 Cases Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etzioni 2002&lt;sup&gt;41&lt;/sup&gt;</td>
<td>Overall rates of overdiagnosis = 29% for whites and 44% for blacks (best-fitting lead time for blacks was 7 years, for whites it was 5 years). Therefore 29% and 44% of screen-detected cases would be overdiagnosed in whites and blacks, respectively. For lead time of 5 years, OD = 28.8% for whites and 32.6% for blacks. For lead time of 7 years, OD = 39.5% for whites and 43.8% for blacks.</td>
<td>Cohort of 2 million screened—detection rates different depending on the pt age, year of screening, and race. Cannot be calculated without explicit numbers of overdiagnosed cases projected to occur in each race cohort per 2 million screened.</td>
<td>Assuming High “p” value (conservative estimate of OD). LT OD Rate (W,B).</td>
</tr>
</tbody>
</table>
| Etzioni 1999<sup>43</sup> | Annual PSA > 4 = 24.7% of stage A1 cancers were overdiagnosed. Biannual PSA > 4 = 23.1% overdiagnosed. Annual age specific = 16.8% overdiagnosed. Biannual age specific = 15.4% overdiagnosed. 5-year PSA > 4 = 20.5% overdiagnosed. | OD Rate X 80,000 stage A1 cancers / 200,000 persons screened x 1000 = rate / 1000 pts screened. Annual PSA > 4 = 98.8 OD cases / 1000 patients screened. Biannual PSA > 4 = 92.4 OD cases / 1000 pts screened. Annual age specific = 67.2 OD cases / 1000 pts screened. Biannual age specific = 61.6 OD cases / 1000 pts screened. 5-year PSA > 4 = 82 OD cases / 1000 pts screened. | Annual PSA>4.  
-247 OD cases / 1000 stage A1 cases. Biannual PSA>4.  
-231 OD cases / 1000 cases. Annual age-specific PSA cutoffs.  
-168 OD cases / 1000 cases. Biannual age-specific PSA cutoffs.  
-154 OD cases / 1000 cases. q5 yr PSA>4.  
-205 OD cases / 1000 cases. |
<p>| Zappa 1998&lt;sup&gt;45&lt;/sup&gt; | For Constant incidence, OD = 51% for age 60; 93% for age 65. If 2% yearly increment in incidence, OD = 25% for age 60; 65% for age 65. | Assumption=100% attendance for 10,000 men screened by biennial PSA x 5 rounds. (Expected cancers with screen – expected cancers w/o screen) / 10,000 screened x 1000 = OD rate. Baseline Inc. Age OD Rate. | (Expected cancers with screen – expected cancers w/o screen) / expected cancers with screen x 1000 = OD rate. Baseline Inc. Age OD Rate. |</p>
<table>
<thead>
<tr>
<th>Study and Overall Quality</th>
<th>Quality of Study on Which Model Based or Validated By</th>
<th>Completeness of Model (overall completeness, sensitivity analyses)</th>
<th>Similarity of Comparison Groups and Potential for Contamination</th>
<th>Questionable Assumptions Made by the Model</th>
<th>Length of Follow up</th>
<th>Generalizability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zappa 1998&lt;sup&gt;45&lt;/sup&gt;</td>
<td>Fair. Service screen. Only had 4 years of follow up.</td>
<td>Fair. Includes measures of detection rate and sensitivity, but not competing causes of death.</td>
<td>Fair. Authors state that screening was negligible in 1990-1991, but offer no data to support this conclusion.</td>
<td>Assumes either constant incidence or 2% increase in incidence per year. Assumes detection rates at all incidence screens are the same as at the second screen.</td>
<td>Fair. 14 years is a long enough period to most likely cover most lead times.</td>
<td>Fair. Population based and biologically plausible, but only based on ~3000 pts.</td>
</tr>
<tr>
<td>Etzioni 1999&lt;sup&gt;43&lt;/sup&gt;</td>
<td>Fair. Used Cowen 1994 Markov of prostate cancer progression. Incidence validated by comparing model age-specific incidence to SEER 1984-1988. No validation by comparison to prospective study.</td>
<td>Good. Includes measures of natural history, clinical presentation, PSA growth curves, screening strategies, and survival. Performed multiple sensitivity analyses to address some of their assumptions.</td>
<td>No comparison group necessary for this model.</td>
<td>Big assumption is that all of the parameters that they include in the model were calculated from similar enough populations to fit together properly. Also assume no overdiagnosis in stage B, C, or D cancers. Multiple other assumptions about the values of specific parameters in the model.</td>
<td>Good. Patients modeled until age 95 or death.</td>
<td>Fair. Population-based and biologically plausible, but only models white men.</td>
</tr>
<tr>
<td>Etzioni 2002&lt;sup&gt;31&lt;/sup&gt;</td>
<td>Fair. Model primarily used parameters from SEER and census data. Lead time was estimated at 3-7 years from 3 previous studies noted above (1 nested case control, 1 retrospective cohort). Model fit to SEER incidence data from 1988 to 1998.</td>
<td>Fair. Sensitivity analyses varying lead times, p (the proportion of PSA-associated cases whose PSA was screening, not diagnostic), and secular trend help address assumptions. PSA sensitivity not addressed in the model.</td>
<td>No comparison group for this model.</td>
<td>Big assumption is that the lead time parameters are generalizable enough to others to be used together. Assume that lead times were 3,5,or 7 years with gamma distributions. Assumes that all PSA tests within 3 months of dx were screening. Assumptions about detection rates and secular trend.</td>
<td>Good. Patients modeled until death. Incidence data only available for 10 years, though.</td>
<td>Good. Population based, biologically plausible, models white and black men separately.</td>
</tr>
</tbody>
</table>
### Table 8: Quality of Included Prostate Cancer Studies (Cont’d)

<table>
<thead>
<tr>
<th>Study and Overall Quality</th>
<th>Quality of Study on Which Model Based or Validated By</th>
<th>Completeness of Model (overall completeness, sensitivity analyses)</th>
<th>Similarity of Comparison Groups and Potential for Contamination</th>
<th>Questionable Assumptions Made by the Model</th>
<th>Length of Follow up</th>
<th>Generalizability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Draisma 2003^38, Quality: Good</td>
<td>Good. Model validated by data from high-quality ERSPC Rotterdam RCT. Model also validated by prescreening incidence of prostate cancer in The Netherlands and stage distribution at baseline.</td>
<td>Good. Model parameters included cumulative incidence, transition probabilities and dwelling times for various stages and grades, Weibull distributions, and test sensitivities. Includes 3 alternative models (one of which assumes lead time=exponential distribution)</td>
<td>Fair. Model estimated contamination of 20 tests / 1000 man-years. Included a parameter in model to correct for this.</td>
<td>Assumes lead time occurs in Weibull distribution. Assumes that the multiple interventions used in the ERSPC Rotterdam are similar enough that results are meaningful. Assumes that all cause mortality in patients with screen-detected prostate cancer is similar to the population as a whole.</td>
<td>All patients modeled until death. ERSPC Rotterdam had follow-up of only 6 years, though.</td>
<td>Fair. Heavily reliant on population of Rotterdam ERSPC and the 1991 population in The Netherlands. Only 50% of registry volunteered to participate in Rotterdam study; these volunteers are different from the rest of the population. Exclusion criteria not overly strict.</td>
</tr>
<tr>
<td>Davidov 2004^35, Quality: Poor</td>
<td>Poor. Study uses SEER data 1993-1997 for incidence. No validation of model performed</td>
<td>Fair. Includes sensitivity and sojourn time data. Does not include any estimation of secular trend of incidence increase or PSA testing rates. Sensitivity analyses for different mean lead times, screening schedules, and sensitivities. Mathematical design is difficult to understand.</td>
<td>No comparison group necessary in this model.</td>
<td>Assumes sojourn time follows an exponential distribution. Assume sensitivities equal 0.3, 0.7, or 0.9. Assume mortality of patients with prostate cancer is identical to that of US males in 1997.</td>
<td>All patients modeled until death. Only 5 years of incidence data used from SEER database.</td>
<td>Fair. Population-based from the SEER database, but lack of validation compromises this as well.</td>
</tr>
<tr>
<td>Study and Overall Quality</td>
<td>Quality of Study on Which Model Based or Validated By</td>
<td>Completeness of Model (overall completeness, sensitivity analyses)</td>
<td>Similarity of Comparison Groups and Potential for Contamination</td>
<td>Questionable Assumptions Made by the Model</td>
<td>Length of Follow up</td>
<td>Generalizability</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------------------------------------------------</td>
<td>------------------------------------------------------------------</td>
<td>----------------------------------------------------------</td>
<td>-----------------------------------------</td>
<td>-------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Ciatto 2005 Fair Quality: Fair</td>
<td>Fair. Based on 2 service screen populations in Florence between 1991 and 1994 (~7000 men).</td>
<td>Fair to poor. Does not include death rates due to other comorbidities or secular trends in incidence. No sensitivity analyses performed.</td>
<td>Fair to poor. Volunteers are likely different from population in Tuscany Cancer Registry. Screened cohort is part of the reference population (3%). Authors state that opportunistic screening by PSA was almost absent in the district when the pilots were carried out (no data to back this up).</td>
<td>Assumes that all excess incidence is a result of overdiagnosis; does not take into account the possibility that some could be earlier diagnosis of cancers that would have presented clinically. Assumes that 9 years of follow up is long enough to overcome most of the lead times for screen-detected cancer cases.</td>
<td>9 years (subjects linked with regional Tuscany Tumor Registry and Regional Mortality Registry)</td>
<td>Poor. Patients volunteered for screening. Florence population may not be generalizable to more diverse populations. Age range only 60-74. Pts with disabling illnesses and “those unlikely to attend invitation” excluded.</td>
</tr>
<tr>
<td>Tsodikov 2006 Fair Quality: Fair</td>
<td>Fair. Model fit to incidence data from SEER database 1973-2000. SEER/Medicare linkage used to estimate PSA testing rates. Mean sojourn time from published estimates. No validation by epidemiologic studies.</td>
<td>Fair. Model takes many factors into account, including lead time, comorbidities, age at presentation, PSA screening frequencies and detection rates. Sensitivity analyses performed for different age of onset distributions. Mathematical model is difficult to understand</td>
<td>No comparison group used for this model.</td>
<td>Model assumes 100% sensitive PSA test. Assumes age at tumor onset is a parametric distribution (Weibull, MKV, or Gamma). Assumes sojourn time follows a Weibull distribution. Assumes all tests performed within 3 months of diagnosis were diagnostic.</td>
<td>All patients modeled until death. 28 years of incidence data used.</td>
<td>Fair. Validation by population level data, but not epidemiologic studies.</td>
</tr>
<tr>
<td>Study and Overall Quality</td>
<td>Quality of Study on Which Model Based or Validated By</td>
<td>Completeness of Model (overall completeness, sensitivity analyses)</td>
<td>Similarity of Comparison Groups and Potential for Contamination</td>
<td>Questionable Assumptions Made by the Model</td>
<td>Length of Follow up</td>
<td>Generalizability</td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------------------------------------------</td>
<td>-----------------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>------------------------------------------</td>
<td>--------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Parker 2006&lt;sup&gt;10&lt;/sup&gt;</td>
<td>Fair. Model based on survival data from retrospective cohort study. Draisma 2003 model used to estimate overdagnosis.</td>
<td>Good. See Draisma 2003.</td>
<td>Fair. See Draisma 2003.</td>
<td>See Draisma 2003. For this study’s purposes, the authors assumed biennial screening with 100 percent attendance.</td>
<td>All patients modeled until death.</td>
<td>Fair. See Draisma 2003. In addition, survival data used had exclusion criteria excluding patients with metastases, survival less than 6 mos., and cancers</td>
</tr>
<tr>
<td>Hamashima 2006&lt;sup&gt;28&lt;/sup&gt;</td>
<td>Poor. Based on service screen population in Japan from 2004 to 2005.</td>
<td>Fair. Does not include parameter that estimates comorbid death rates. Does perform sensitivity analysis for lead time, but not for sensitivity. Sensitivity estimation is arbitrary.</td>
<td>Poor. Not clear how similar the volunteers for screening are to the reference population from 11 Japanese cancer registries. Potential for contamination is high because authors did not assess for previous PSA screening.</td>
<td>Assumes that all excess incidence is a result of overdiagnosis; does not take into account the possibility that some could be earlier diagnosis of cancers that would have presented clinically. Authors admit that follow up is insufficient to correctly take lead time into account.</td>
<td>Only 1 year of follow up</td>
<td>Fair to poor. Patients are volunteers in Tokyo who were almost all (~70%) 55-70. Exclusion of pts with self reported precancerous disease may remove patients with comorbidities. Relatively small overall numbers of patients.</td>
</tr>
<tr>
<td>Study</td>
<td>Study Design</td>
<td>Definition and Expression of Overdiagnosis</td>
<td>Sample Characteristics</td>
<td>Validation, Analysis, and Parameter Estimation</td>
<td>Intervention</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Duffy 2005</td>
<td>Markov Model fit to incidence data from 2 RCT's, the Swedish 2 County Trial and the Gothenburg Trial (Day 1984 Model)</td>
<td>Overdiagnosis = the diagnosis of cancer as a result of screening, usually histologically confined, that would not have arisen clinically during the lifetime of the host had screening not taken place Expressed as the % of screen-detected breast cancer cases that were overdiagnosed</td>
<td>Swedish Two-County: 77,080 women 40-74 yo randomized to invitation to screening and 55,985 randomized to no invitation to screening between 1977 and 1984. Gothenburg: 21,650 women 39-59 yo randomized to invitation to screening and 29,961 randomized to no invitation between 1982-1989.</td>
<td>Observed incidence from RCT data For expected incidence, fit Poisson distributions to the 2 studies to estimate incidence of preclinical but screen detectable progressive cancers, incidence of overdiagnosed cancers, lead time, and sensitivity using the model of Day and colleagues</td>
<td>S2C—single-view mammography every 2yrs for 40-49 yo and every 3yrs for 50-74 yos GT—2V mammography at first screening with subsequent views dependent on breast density. Interval q18 mos</td>
<td></td>
</tr>
<tr>
<td>Hamashima 2006</td>
<td>Service Screening Model</td>
<td>No explicit definition Expressed as the percentage increase in cancer diagnosis in the screened population vs. unscreened population</td>
<td>1725 female volunteers over 40 enrolled from Tokyo 2/04 – 1/05</td>
<td>Incidence estimated from 11 Cancer Registries in Japan. Sojourn time (4-5 years) and sensitivity (70-80%) estimated from previous studies</td>
<td>2V mammography, US, and physical exam</td>
<td></td>
</tr>
<tr>
<td>de Koning 2006</td>
<td>Markov Model fit to Dutch service screening incidence data and also validated by HIP trial data (Day 1984 Model)</td>
<td>Overdiagnosis = the detection of breast cancers by screening that would otherwise never have been clinically diagnosed but are now treated Expressed as the percent of total breast cancer incidence that is overdiagnosed</td>
<td>Nijmegen Service Screen Women 50-69 invited to screening from 1990 to 1997, then in 1999 women 70-74 invited. The number of women screened is unclear.</td>
<td>Observed incidence data from service screen source of data. For expected incidence, fit Poisson distributions to the 2 studies to estimate incidence of preclinical but screen detectable progressive cancers, incidence of OD cancers, lead time, and sensitivity using the model of Day and colleagues</td>
<td>Biennial mammography screening (intervention is not described in detail)</td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Study Design</td>
<td>Definition and Expression of Overdiagnosis</td>
<td>Sample Characteristics</td>
<td>Validation, Analysis, and Parameter Estimation</td>
<td>Intervention</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
</tbody>
</table>
Expressed as the percentage increase in breast cancer diagnosis in the screened population as compared to the unscreened population | Based their estimate of overdiagnosis on 2 trials they judged to be adequately randomized  
Malmo 1976:  
42,484 women aged 45-69 years randomized by birth year cohort (21,242 screened, 21,240 control)  
Canada 1980:  
89,931 women aged 40-59 randomized individually after accepting invitation | Calculated relative risk of overdiagnosis by dividing incidence in screened patients by incidence in unscreened patients | Malmo 1976:  
2V mammography in 1-2nd screening rounds, 1V or 2V later every 18-24 mos. (5-6 screens)  
Canada 1980:  
Annual 2V mammography (4-5 screens) |
| Olsen 2006    | Markov Model fit to incidence data from Copenhagen service screening program (Day 1984 Model) | Overdiagnosis = the diagnosis of breast cancers that without screening would not have emerged clinically in the woman's lifetime  
Expressed as % of screen-detected cancers at 1st screen, subsequent screens, and % of all cancers diagnosed | 35,123 women aged 50-69 years beginning in 1991 screened at least once. | Observed incidence from the Denmark service screen data.  
For expected incidence, fit Poisson distributions to the 2 studies to estimate incidence of preclinical but screen detectable progressive cancers, incidence of overdiagnosed cancers, lead time, and sensitivity using the model of Day and colleagues | Biennial mammography screening |
<table>
<thead>
<tr>
<th>Study</th>
<th>Study Design</th>
<th>Definition and Expression of Overdiagnosis</th>
<th>Sample Characteristics</th>
<th>Validation, Analysis, and Parameter Estimation</th>
<th>Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paci 2006</td>
<td>Service screening model “corrected for lead time”</td>
<td>Overdiagnosis = detection at screening of breast cancers that would not have been diagnosed in the absence of screening over a subject’s lifetime</td>
<td>13,519 prescreening and 13,999 screen-detected breast cancers diagnosed b/t 1986-2001 in women 40-79 yo in 6 cities in North/Central Italy.</td>
<td>Observed incidence from the service screening data. Expected incidence modeled with 2 step Poisson regression analysis based on prescreening age-, calendar time-, and area-specific incidence rates. Correction for lead time performed by assuming an exponential distribution for lead time and calculating the probability that each case would have been identified clinically each year after detection. Summing these probabilities = the # of screen-detected cases that would have arisen clinically each year in the absence of screening. Overdiagnosis thus is the ratio of actual screen detected cases to screen-detected cases that would have arisen clinically. MST’s estimated from previous literature as 3.7 and 4.2 years for women 50-59 and 60-74 yo</td>
<td>Unclear</td>
</tr>
<tr>
<td>Zackrisson</td>
<td>RCT</td>
<td>Overdiagnosis = the detection of cases that would never have come to clinical attention without screening Expressed as the % excess cumulative incidence in the screened group as compared to the control group</td>
<td>42,283 women born in 1908-1932 (45-69 yo) randomized to invitation to screening or control in 1976. 21,088 invited and 21,195 controls. Birth year cohorts randomized separately. Screening was offered to 45-54 yos at end of randomization, but not to 55-69 yos.</td>
<td>Calculate relative risk by dividing incidence in screened group by incidence in unscreened group</td>
<td>mammography q18-24 mos (see Malmo trial)</td>
</tr>
<tr>
<td>Study</td>
<td>Study Design</td>
<td>Definition and Expression of Overdiagnosis</td>
<td>Sample Characteristics</td>
<td>Validation, Analysis, and Parameter Estimation</td>
<td>Intervention</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Paci 2004[95]</td>
<td>Service Screening Model with correction for lead time</td>
<td>Overdiagnosis = the detection by screening of lesions that would not have arisen in the host’s lifetimes</td>
<td>60,000 women aged 50-69 in Florence invited for screening from 1990 – 1999. 60% compliance (36,000 women screened)</td>
<td>Correct for lead time by calculating the probability that a screen-detected case would have surfaced clinically after the end of the study period, assuming a mean sojourn time of 3.42 years and an exponential distribution of sojourn times. The sum of these probabilities was subtracted from the observed number of cases to correct for lead time. Expected # of cases determined by applying age specific incidence rates from 1985-1989 to the age distribution of the study population.</td>
<td>2V mammography every 2 yrs from 1990-1999</td>
</tr>
<tr>
<td>Zahl 2004[102]</td>
<td>Service Screening Model</td>
<td>Overdiagnosis = detection by screening of low malignancy lesions that otherwise would not be detected in a patient’s lifetime</td>
<td>Norway Service Screen: 165,000 women from 4 Norwegian counties invited to screening every 2 yrs starting in 1996 – 2001 (75% attendance = 120,000 screens) Sweden Service Screen: 900,000 women 50-69 invited to screening at least once screened from 1986-1996 (75% attendance = 675,000 screens)</td>
<td>Poisson regression used to estimate % change in age-specific incidence rates after introduction of screening. Authors exclude DCIS in analysis.</td>
<td>biennial mammography</td>
</tr>
<tr>
<td>Study</td>
<td>Study Design</td>
<td>Definition and Expression of Overdiagnosis</td>
<td>Sample Characteristics</td>
<td>Validation, Analysis, and Parameter Estimation</td>
<td>Intervention</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------</td>
<td>-------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Peeters   | Service Screening Model | Overdiagnosis = A histologically established diagnosis of invasive or intraductal breast cancer that would never have developed into a clinically manifest tumor during the patient’s normal life expectancy if no screening examination had been carried out  
Expressed as the percentage increase in cancer diagnosis in the screened population as compared to the unscreened population | 30,700 women aged 35 and older living in Nijmegen, The Netherlands screened. Incidence in this population was compared to incidence in Arnhem, a neighboring city of similar population that was unscreened. | Nijmegen incidence / Arnhem incidence = RR | 6 screening rounds of biennial mammography (1975-1986) |
<table>
<thead>
<tr>
<th>Study</th>
<th>Reported Results</th>
<th>Overdiagnosed Cases / 1000 Pts Screened</th>
<th>Overdiagnosed Cases / 1000 Cases Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duffy 2005</td>
<td>Overall incidence rates equalized in yr 8-9 in Swedish trial after control group screened. Increase in incidence of DCIS balanced by deficit in incidence of invasive cancers. Sweden upper limit of DCIS OD=15% of DCIS cases, 1% of all tumors after exclusion of 1st screen cases. From model: 3.1% OD at 1st screen (0.1-10.9) = 3.8 overdiagnosed cases/1000 pts screened. -0.3% at incidence screens. Gothenburg upper limit of DCIS OD=18% of DCIS cases, 2% of all tumors after exclusion of 1st screen cases. -4.2% OD at 1st screen (0.0-28.8) = 3.4 overdiagnosed cases/1000 pts screened. -0.3% at incidence screens. Overall, less than 5% of cases at prevalence screen and less than 1% of cases at incidence screen overdiagnosed.</td>
<td>Swedish Two County Trial Data: 3.8 OD cases / 1000 pts screened. Gothenburg Trial Data: 3.4 OD cases / 1000 pts screened</td>
<td>Swedish Two County Trial: 31 OD cases / 1000 screen detected cancers at the 1st screen. 3 OD cases / 1000 screen detected cancers at each subsequent screen. Gothenburg Trial: 42 OD cases / 1000 screen-detected cancers at the 1st screen. 3 OD cases / 1000 screen-detected cancers at each subsequent screen.</td>
</tr>
<tr>
<td>Hamashima 2006</td>
<td>O/E ratio for incidence of breast cancer with screening = 2.41</td>
<td>9 overdiagnosed cases / 1725 patients screened = 5 cases / 1000 pts screened</td>
<td>9 overdiagnosed cases / 15 cases = 600 OD cases / 1000 screen-detected cancers</td>
</tr>
<tr>
<td>de Koning 2006</td>
<td>3% of the total incidence of breast cancer = overdiagnosis, corresponding to 8% of screen-detected cancers</td>
<td>Cannot be determined from the information given</td>
<td>80 overdiagnosed cases / 1000 screen-detected cancers</td>
</tr>
<tr>
<td>Gotzsche 2006</td>
<td>About 30% overdiagnosis could be expected with mammography screening based on the results of RCT's - RR = 1.30 for Canada 1980a (40-49 yo) - RR = 1.26 for Canada 1980b (50-59 yo) - RR = 1.32 for Malmo 1976</td>
<td>Cannot be determined from the information given</td>
<td>230 OD cases / 1000 cancers diagnosed</td>
</tr>
<tr>
<td>Olsen 2006</td>
<td>7.8% of screen-detected cancers overdiagnosed at the first screen and 0.5% at the second screen. 4.8% of total cancer diagnosis is overdiagnosed. When in situ cases are excluded, rates decrease to 7.3%, which indicates ~ 10% overdiagnosis of DCIS.</td>
<td>Overall incidence = 14 OD cases / 1000 patients screened</td>
<td>At the first screen: 78 overdiagnosed cases / 1000 screen-detected cancers. At the second screen: 5 overdetermined cases / 1000 screen-detected cancers. Slight increases in overdiagnosis rates (to 86 / 1000 at 1st screen and 6 / 1000 at 2nd screen) when sensitivity of 80% used.</td>
</tr>
<tr>
<td>Study</td>
<td>Reported Results</td>
<td>Overdiagnosed Cases / 1000 Pts Screened</td>
<td>Overdiagnosed Cases / 1000 Cases Detected</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Paci 2006&lt;sup&gt;90&lt;/sup&gt;</td>
<td>Excess ratio of 4.6% after correction for lead time for both in situ and invasive cases. For invasive cases only, excess incidence of 3.2%.</td>
<td>Cannot be calculated without knowing the total number of patients screened</td>
<td>obs (corrected) – expected / observed 7906 – 7555.5 / 10,294 = 34 OD cases / 1000 cancers diagnosed (not just screen-detected)</td>
</tr>
<tr>
<td>Zackrisson 2006&lt;sup&gt;91&lt;/sup&gt;</td>
<td>The rate of overdiagnosis 15 years after follow up was 10% in women 55-69 who were not screened after the conclusion of follow up.</td>
<td>Cannot be calculated without knowing total numbers of patients screened</td>
<td>intervention – control / intervention = (3.09 -2.81) / 3.09 = 91 OD cases / 1000 cancers diagnosed (not just screen-detected)</td>
</tr>
<tr>
<td>Paci 2004&lt;sup&gt;99&lt;/sup&gt;</td>
<td>Overdiagnosis is less than 5% when invasive and in situ cancers are considered together.</td>
<td>Cannot calculate this number without the actual number of patients screened.</td>
<td>Invasive + in situ cancers: less than 50 OD cases / 1000 cases of breast cancer. Invasive only: 22 OD cases / 1000 cases of breast cancer (in registry)</td>
</tr>
<tr>
<td>Zahl 2004&lt;sup&gt;102&lt;/sup&gt;</td>
<td>One third of all invasive breast cancers in 50-69 year olds are overdiagnosed. Incidence of breast cancer in Norway increased by 54% with screening and incidence in Sweden increased by 45% with screening.</td>
<td>Cannot be calculated without the actual numbers of patients screened.</td>
<td>Norway: 351 OD cases / 1000 cases. Sweden: 310 OD cases / 1000 cases.</td>
</tr>
<tr>
<td>Anttila 2002&lt;sup&gt;107&lt;/sup&gt;</td>
<td>In 1935-39 birth cohort: RR of breast cancer in screened vs reference group= 1.18 = 18% overdiagnosis</td>
<td>obs – expected / pts screened</td>
<td>obs – expected / observed From Figure 3: 4.7 - 4 / 4.7 = 150 OD cases / 1000 cases</td>
</tr>
<tr>
<td>Peeters 1989&lt;sup&gt;112&lt;/sup&gt;</td>
<td>Excess of 11% cases in Nijmegen for the period 1975 to 1986. For given 4 year time periods: 1975-1978 = 1.30 1979-1982 = 1.03 1983-1986 = 1.01 Declining rate ratios suggest a low risk for overdiagnosis</td>
<td>Cannot be calculated</td>
<td>11% overdiagnosis = obs – exp / obs = 1.11 – 1 / 1.11 = 99 OD cases / 1000 cases</td>
</tr>
<tr>
<td>Study and Overall Quality</td>
<td>Quality of Study on Which Model Based or Validated By</td>
<td>Completeness of Model (overall completeness, sensitivity analyses)</td>
<td>Similarity of Comparison Groups and Potential for Contamination</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------------------------------------------</td>
<td>-------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>Duffy 2005&lt;sup&gt;7&lt;/sup&gt;</td>
<td>Fair. 2 sub optimally randomized RCT's (Gotzsche citation)</td>
<td>Fair. Good completeness. Would be better if sensitivity analyses had been performed to address some of the assumptions.</td>
<td>Good. Expected incidence calculated from the same service screening data as the observed incidence.</td>
</tr>
<tr>
<td>Hamashima 2006&lt;sup&gt;28&lt;/sup&gt;</td>
<td>Poor. Based on service screen population in Japan from 2004 to 2005.</td>
<td>Poor. Does not include parameter that estimates comorbid death rates. Does not report results of sensitivity analyses for lead time/sensitivity.</td>
<td>Poor. 18.5% of women had mammography within previous year.</td>
</tr>
<tr>
<td>Study and Overall Quality</td>
<td>Quality of Study on Which Model Based or Validated By</td>
<td>Completeness of Model (overall completeness, sensitivity analyses)</td>
<td>Similarity of Comparison Groups and Potential for Contamination</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------------------------------------------------</td>
<td>-----------------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>de Koning 2006&lt;sup&gt;83&lt;/sup&gt;</td>
<td>Fair. Nijmegen service screen data = observed incidence.</td>
<td>Fair. Includes parameters such as sensitivity, lead time, age and stage. Does not perform sensitivity analyses for these parameters.</td>
<td>Expected incidence is modeled from Dutch service screening data as well, thus the potential for contamination is low.</td>
</tr>
<tr>
<td>Gotzsche 2006&lt;sup&gt;84&lt;/sup&gt;</td>
<td>Excellent. 7 RCT’s: 2 high quality, 5 suboptimally randomized.</td>
<td>Simplistic design. Does not attempt to model the effect of lead time on excess incidence.</td>
<td>Moderate potential for contamination. Random sample of 500 women in Malmo RCT control group showed 24% underwent mammography during the study.</td>
</tr>
<tr>
<td>Study and Overall Quality</td>
<td>Quality of Study Based or Validated By</td>
<td>Completeness of Model (overall completeness, sensitivity analyses)</td>
<td>Similarity of Comparison Groups and Potential for Contamination</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>Olsen 2006&lt;sup&gt;78&lt;/sup&gt;</td>
<td>Fair. Copenhagen service screen data used.</td>
<td>Good. Good completeness. Sensitivity analyses were performed to address the sensitivity assumption, with very little change in overdiagnosis rates.</td>
<td>Good. Expected incidence calculated from the same service screening data as the observed incidence.</td>
</tr>
<tr>
<td>Paci 2006&lt;sup&gt;79&lt;/sup&gt;</td>
<td>Fair. Northern Italian service screen population.</td>
<td>Fair. Models incremental cases (# of screen-detected cases in a given year) and decremental cases (# of screen-detected cases that would have arisen clinically in a given year). Includes a correction of lead time and a sensitivity analysis of different mean sojourn times that showed minimal change in OD. Would be stronger if sensitivity analyses using different lead time distributions had been performed.</td>
<td>Fair. Populations are the same, but authors do not state how much mammography occurred prescreening in the 6 regions of interest.</td>
</tr>
<tr>
<td>Study and Overall Quality</td>
<td>Quality of Study Based on Which Model Validated By</td>
<td>Completeness of Model (overall completeness, sensitivity analyses)</td>
<td>Similarity of Comparison Groups and Potential for Contamination</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>-------------------------------------------------------------</td>
</tr>
<tr>
<td>Zackrisson 2006&lt;sup&gt;91&lt;/sup&gt; Quality: Good</td>
<td>Fair to Good. Based on high quality (Malmo) RCT data.</td>
<td>Fair. Would be better if sensitivity analyses had been performed with various levels of contamination of the control group.</td>
<td>Potential for screening of 55-69 year olds after the end of the trial is high.</td>
</tr>
<tr>
<td>Paci 2004&lt;sup&gt;99&lt;/sup&gt; Quality: Fair</td>
<td>Fair. Based on large Florence service screen population.</td>
<td>Fair to good. Performed sensitivity analysis using different mean sojourn time estimates with estimates ranging from 3-7% for invasive and in situ cancers</td>
<td>High potential for contamination; the authors do not give any data showing low screening rates in 1983-1989.</td>
</tr>
<tr>
<td>Zahl 2004&lt;sup&gt;102&lt;/sup&gt; Quality: Fair</td>
<td>Fair. Based on 2 large service screen populations.</td>
<td>Fair. Would be better if sensitivity analyses had been performed that calculated OD for different baseline incidence rates.</td>
<td>High potential for contamination, because the authors don’t give any.</td>
</tr>
<tr>
<td>Study and Overall Quality</td>
<td>Quality of Study on Which Model Based or Validated By</td>
<td>Completeness of Model (overall completeness, sensitivity analyses)</td>
<td>Similarity of Comparison Groups and Potential for Contamination</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------------------------------------------------</td>
<td>--------------------------------------------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>Anttila 2002[107]</td>
<td>Fair. Based on service screen population in Helsinki from 1986-1997.</td>
<td>Fair. Would be better if the model reported overdiagnosis rates for all patients instead of just the 1935-39 birth cohort.</td>
<td>Authors do not mention the potential for contamination in the group of patients used to calculate the expected incidence.</td>
</tr>
<tr>
<td>Peeters 1989[112]</td>
<td>Fair. Based on service screen population in Nijmegen, The Netherlands.</td>
<td>Fair. Would be better if model compared age-specific incidence rates in Arnhem and Nijmegen.</td>
<td>Poor. Similar incidence rates for breast cancer pre-screening in Nijmegen and Arnhem indicate that the populations are similar. The authors do not give us an estimate of the prevalence of mammography in Arnhem, though, and we don’t know how similar the populations really are.</td>
</tr>
</tbody>
</table>
References


55. Carroll PR. Early stage prostate cancer--do we have a problem with over-detection, overtreatment or both? *J Urol.* 2005;173:1061-1062.


84. Gotzsche PC. Ramifications of screening for breast cancer: overdiagnosis in the Malmo trial was considerably underestimated. *BMJ.* 2006;332:727.


163. de Koning HJ, Auvinen A, Berenguer Sanchez A, et al. Large-scale randomized prostate cancer screening trials: program performances in the European Randomized Screening
for Prostate Cancer trial and the Prostate, Lung, Colorectal and Ovary cancer trial. *Int J Cancer.* 2002;97:237-244.


## Appendix 1

### Prostate Cancer Study Selection

<table>
<thead>
<tr>
<th>Study</th>
<th>Study Design</th>
<th>Inclusion/Exclusion Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pelzer 2007(^\text{23})</td>
<td>Case Series</td>
<td>Excluded: Pathologic definition of overdiagnosis</td>
</tr>
<tr>
<td>Graif 2007(^\text{24})</td>
<td>Case Series</td>
<td>Excluded: Pathologic definition of overdiagnosis</td>
</tr>
<tr>
<td>Telesca 2007(^\text{25})</td>
<td>Model fit to SEER registry data</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Postma 2007(^\text{26})</td>
<td>RCT</td>
<td>Excluded: Does not have an explicit outcome of % overdiagnosis (only has % minimal cancers, but we don’t know how many of these patients were overdiagnosed)</td>
</tr>
<tr>
<td>Delongchamps 2006(^\text{27})</td>
<td>Narrative Review</td>
<td>Excluded: Narrative Review (no search strategy)</td>
</tr>
<tr>
<td>Hamashima 2006(^\text{28})</td>
<td>Service Screening Model</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Miller 2006(^\text{29})</td>
<td>Retrospective cohort</td>
<td>Excluded: Does not focus on screening PSA-detected cases only. Uses a more inclusive definition of overtreatment that encompasses all unnecessary treatment of indolent cases.</td>
</tr>
<tr>
<td>Parker 2006(^\text{30})</td>
<td>Competing-risks hazard model</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Pashayan 2006(^\text{31})</td>
<td>Retrospective Cohort</td>
<td>Excluded: Uses diagnostic PSA positive cases rather than screen positive cases</td>
</tr>
<tr>
<td>Roemeling 2006(^\text{32})</td>
<td>RCT</td>
<td>Excluded: Does not calculate % overdiagnosis as an outcome; speculates that increased metastasis-free survival in screened patients may be due to overdiagnosis</td>
</tr>
<tr>
<td>Tsodikov 2006(^\text{33})</td>
<td>Mathematical model</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Ciatto 2005(^\text{34})</td>
<td>Service Screening Model</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Davidov 2004(^\text{35})</td>
<td>Mathematical Model</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Kwiatkowski 2004(^\text{36})</td>
<td>RCT</td>
<td>Excluded: Does not calculate % overdiagnosis as an outcome; states that further information from ERSPC is needed to assess the degree of overdiagnosis</td>
</tr>
<tr>
<td>Tomblom 2004(^\text{37})</td>
<td>Prospective cohort study</td>
<td>Excluded: Does not calculate % overdiagnosis as an outcome; states that the risk for overdetection in screening programs that require longer follow-up could not be assessed from this study.</td>
</tr>
<tr>
<td>Draisma 2003(^\text{38})</td>
<td>Markov Model validated by RCT and baseline incidence data</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Draisma 2003(^\text{39}) (MISCAN)</td>
<td>Same study as above with a few extra studies reviewed in Discussion</td>
<td>Excluded: Same as other Draisma 2003</td>
</tr>
<tr>
<td>Auvin 2002(^\text{40})</td>
<td>Model based on RCT</td>
<td>Excluded: Does not include overdiagnosis, overdetection, or overtreatment as an outcome. Primary outcome is lead time.</td>
</tr>
<tr>
<td>Etzioni 2002(^\text{41})</td>
<td>Computer model validated by SEER data</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Hugosson 2000(^\text{42})</td>
<td>Nested Case-Control</td>
<td>Excluded: Does not include an estimate of overdiagnosis as an outcome</td>
</tr>
<tr>
<td>Etzioni 1999(^\text{43})</td>
<td>Markov Model</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>McGregor 1998(^\text{44})</td>
<td>Modeling</td>
<td>Excluded: Defines OD as the detection of cancer that, left untreated, would not cause death. This is a less inclusive definition than the definition of interest.</td>
</tr>
<tr>
<td>Zappa 1998(^\text{45})</td>
<td>Model using Monte-Carlo sampling</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Gelabert Mas 1997(^\text{46})</td>
<td>Spanish</td>
<td>Excluded: Spanish</td>
</tr>
<tr>
<td>Study</td>
<td>Study Design</td>
<td>Inclusion/Exclusion Status</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Benoit 1995</td>
<td>Narrative Review</td>
<td>Excluded: No systematic search strategy</td>
</tr>
<tr>
<td>Kramer 1993</td>
<td>Narrative Review</td>
<td>Excluded: No search strategy; overdiagnosis not an explicit outcome</td>
</tr>
<tr>
<td>Slaughter 2002</td>
<td>Narrative Review</td>
<td>Excluded: No search strategy; overdiagnosis not an explicit outcome</td>
</tr>
<tr>
<td>Lotan 2006</td>
<td>Correspondence</td>
<td>Excluded: Correspondence</td>
</tr>
<tr>
<td>Ciatto 2006</td>
<td>Service Screening Pilots</td>
<td>Excluded: Overdiagnosis is not an outcome</td>
</tr>
<tr>
<td>Graif 2006</td>
<td>Narrative Review</td>
<td>Excluded: Narrative Review</td>
</tr>
<tr>
<td>Pashayan 2006</td>
<td>Correspondence</td>
<td>Excluded: Correspondence</td>
</tr>
<tr>
<td>Ross 2005</td>
<td>Nested Case Control</td>
<td>Excluded: Overdiagnosis is not an explicit outcome; overtreatment not mentioned either</td>
</tr>
<tr>
<td>Pinsky 2004</td>
<td>Convolution Model</td>
<td>Excluded: Used Lung Cancer Screening Trial rather than Prostate Cancer Screening trial to fit model</td>
</tr>
<tr>
<td>Sennfalt 2004</td>
<td>RCT</td>
<td>Excluded: Focuses on costs; will mention in discussion section</td>
</tr>
<tr>
<td>Ciatto 2003</td>
<td>RCT methods paper</td>
<td>Excluded: Overdiagnosis not an outcome</td>
</tr>
<tr>
<td>Ciatto 2003</td>
<td>RCT methods paper</td>
<td>Excluded: Overdiagnosis not an outcome</td>
</tr>
<tr>
<td>Draisma 2003</td>
<td>Abstract of study above</td>
<td>Excluded: Just an abstract</td>
</tr>
</tbody>
</table>

Prostate Cancer Study Selection (Cont'd)
# Appendix 2

## Breast Cancer Study Selection

<table>
<thead>
<tr>
<th>Study</th>
<th>Study Design</th>
<th>Inclusion/Exclusion Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duffy 2005&lt;sup&gt;5&lt;/sup&gt;</td>
<td>Markov Model based on 2 RCT's, the Swedish 2 County Trial and the Gothenburg Trial</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Hamashima 2006&lt;sup&gt;35&lt;/sup&gt;</td>
<td>Service Screening Model</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Benoit 1995&lt;sup&gt;44&lt;/sup&gt;</td>
<td>Narrative Review</td>
<td>Excluded: No systematic search strategy performed</td>
</tr>
<tr>
<td>de Koning 2006&lt;sup&gt;83&lt;/sup&gt;</td>
<td>Markov Model based on service screening data</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Duffy 2006&lt;sup&gt;84&lt;/sup&gt;</td>
<td>Narrative Review</td>
<td>Excluded: Narrative review</td>
</tr>
<tr>
<td>Gotzsche 2006&lt;sup&gt;85&lt;/sup&gt;</td>
<td>Correspondence</td>
<td>Excluded: Correspondence re: Zackrisson 2006</td>
</tr>
<tr>
<td>Gotzsche 2006&lt;sup&gt;86&lt;/sup&gt;</td>
<td>Systematic Review</td>
<td>INCLUDED** Also need to include RCT's</td>
</tr>
<tr>
<td>Moller 2006&lt;sup&gt;87&lt;/sup&gt;</td>
<td>Correspondence</td>
<td>Excluded: Correspondence re: Zackrisson 2006</td>
</tr>
<tr>
<td>Olsen 2006&lt;sup&gt;88&lt;/sup&gt;</td>
<td>Markov Model based on service screening data</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Paci 2006&lt;sup&gt;89&lt;/sup&gt;</td>
<td>Service Screening Model Corrected for lead time</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Svendsen 2006&lt;sup&gt;90&lt;/sup&gt;</td>
<td>Service Screening Model</td>
<td>Excluded: No explicit numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Zackrisson 2006&lt;sup&gt;91&lt;/sup&gt;</td>
<td>Follow up study of Malmo RCT</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Zahl 2006&lt;sup&gt;92&lt;/sup&gt;</td>
<td>Correspondence</td>
<td>Excluded: Correspondence re: Zackrisson 2006</td>
</tr>
<tr>
<td>Duffy 2005&lt;sup&gt;93&lt;/sup&gt;</td>
<td>Narrative Review</td>
<td>Excluded: Narrative review that reports results of Paci 2004 and Yen 2003</td>
</tr>
<tr>
<td>Kerlikowske 2005&lt;sup&gt;94&lt;/sup&gt;</td>
<td>Retrospective Cohort</td>
<td>Excluded: No explicit numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Moss 2005&lt;sup&gt;95&lt;/sup&gt;</td>
<td>Narrative Review of RCT data</td>
<td>Excluded: No search strategy</td>
</tr>
<tr>
<td>Paci 2005&lt;sup&gt;96&lt;/sup&gt;</td>
<td>Narrative Review of service screens</td>
<td>Excluded: No search strategy</td>
</tr>
<tr>
<td>McCann 2004&lt;sup&gt;97&lt;/sup&gt;</td>
<td>Service Screening Model</td>
<td>Excluded: No explicit numerical estimate of overdiagnosis (actually assume no OD in model)</td>
</tr>
<tr>
<td>Olsen 2004&lt;sup&gt;98&lt;/sup&gt;</td>
<td>Correspondence</td>
<td>Excluded: Correspondence</td>
</tr>
<tr>
<td>Paci 2004&lt;sup&gt;99&lt;/sup&gt;</td>
<td>Service Screening Model</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Zahl 2004&lt;sup&gt;100&lt;/sup&gt;</td>
<td>Correspondence</td>
<td>Excluded: Correspondence re: Olsen 2003</td>
</tr>
<tr>
<td>Zahl 2004&lt;sup&gt;101&lt;/sup&gt;</td>
<td>Correspondence</td>
<td>Excluded: Correspondence</td>
</tr>
<tr>
<td>Zahl 2004&lt;sup&gt;102&lt;/sup&gt;</td>
<td>Service Screening Model</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Game 2003&lt;sup&gt;103&lt;/sup&gt;</td>
<td>Correspondence</td>
<td>Excluded: Correspondence</td>
</tr>
<tr>
<td>Mandelblatt 2003&lt;sup&gt;104&lt;/sup&gt;</td>
<td>Systematic Review</td>
<td>Excluded: No explicit numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Olsen 2003&lt;sup&gt;105&lt;/sup&gt;</td>
<td>Service Screening Model</td>
<td>Excluded: No explicit numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Yen 2003&lt;sup&gt;106&lt;/sup&gt;</td>
<td>Markov Model based on Swedish Two County RCT and service screen data</td>
<td>Excluded: Inappropriate definition of overdiagnosis (Overdiagnosis = cases of DCIS that are non-progressive Non-progressive (DCIS&lt;sub&gt;0&lt;/sub&gt;) = DCIS which could not have progressed to invasive disease if left untreated)</td>
</tr>
<tr>
<td>Anttila 2002&lt;sup&gt;107&lt;/sup&gt;</td>
<td>Service Screening Model</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Tange 2002&lt;sup&gt;108&lt;/sup&gt;</td>
<td>?</td>
<td>Excluded: Not in English</td>
</tr>
<tr>
<td>Hakama 1995&lt;sup&gt;109&lt;/sup&gt;</td>
<td>Comparison of pathology of screen-detected breast cancers</td>
<td>Excluded: No explicit numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Study</td>
<td>Study Design</td>
<td>Inclusion/Exclusion Status</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------------------------------</td>
<td>-----------------------------------------------------------------</td>
</tr>
<tr>
<td>Jatoi 1995</td>
<td>Narrative Review</td>
<td>Excluded: Narrative Review</td>
</tr>
<tr>
<td>Roselli Del Turco 1993</td>
<td>Narrative Review</td>
<td>Excluded: Narrative Review</td>
</tr>
<tr>
<td>Peeters 1989</td>
<td>Service Screening Model</td>
<td>INCLUDED</td>
</tr>
<tr>
<td>Lundgren 1982</td>
<td>Single arm prospective screening</td>
<td>Excluded: No explicit numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Groot 2006</td>
<td>Cost effectiveness model of breast cancer</td>
<td>Excluded: No explicit numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Shen 2005</td>
<td>Cost-effectiveness study</td>
<td>Excluded: No explicit numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Kosters 2003</td>
<td>Systematic Review</td>
<td>Excluded: No explicit numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Leivo 1999</td>
<td>Cost-effectiveness study</td>
<td>Excluded: No explicit numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Austoker 1994</td>
<td>Narrative Review</td>
<td>Excluded: Narrative Review</td>
</tr>
<tr>
<td>Seppanen 2006</td>
<td>Service Screening Model</td>
<td>Excluded: No explicit definition or numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Jonsson 2005</td>
<td>Service Screening Model</td>
<td>Excluded: No explicit definition or numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Davidov 2004</td>
<td>Mathematical Model</td>
<td>Excluded: No explicit numerical estimate of breast cancer overdiagnosis</td>
</tr>
<tr>
<td>Douek 2004</td>
<td>Abstract only</td>
<td>Excluded: Abstract only</td>
</tr>
<tr>
<td>Fracheboud 2004</td>
<td>Service screening model</td>
<td>Excluded: No explicit definition or numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Hofvind 2004</td>
<td>Service Screening</td>
<td>Excluded: No explicit numerical estimate of overdiagnosis</td>
</tr>
<tr>
<td>Olsen 2004</td>
<td>Correspondence</td>
<td>Excluded: Correspondence re: Zahl 2004</td>
</tr>
<tr>
<td>Zahl 2004</td>
<td>Abstract only</td>
<td>Excluded: Abstract only</td>
</tr>
<tr>
<td>Paci 2004</td>
<td>Abstract only</td>
<td>Excluded: Abstract only</td>
</tr>
<tr>
<td>Gotzsche 2003</td>
<td>Abstract only</td>
<td>Excluded: Abstract only</td>
</tr>
<tr>
<td>Cherevani 1998</td>
<td>Abstract only</td>
<td>Excluded: Abstract only</td>
</tr>
<tr>
<td>Chamberlain 1990</td>
<td>Abstract only</td>
<td>Excluded: Abstract only</td>
</tr>
</tbody>
</table>