Systematic Review of Cancer Screening Literature for Updating American Cancer Society Breast Cancer Screening Guidelines

Prepared for:
American Cancer Society, Inc.
250 Williams Street
Atlanta, GA 30303

Contract No. 13214

Prepared by:
Duke Evidence Synthesis Group
Durham, NC

Investigators:
Laura Havrilesky, M.D.
Jennifer M Gierisch, Ph.D.
Patricia Moorman, Ph.D.
Douglas McCrory, M.D.
Sujata Ghatle, M.D.
John Williams, M.D.
Ranee Chatterjee Montgomery, M.D.
Matt Crowley, M.D.
Andrzej Kosinski, Ph.D.
Lars Grimm, M.D.
Brittany Davidson, M.D.
Amy S. Kendrick, R.N., M.S.N.
Megan Chobot, M.S.L.S.
Rebecca Gray, D.Phil.
Gillian Sanders, Ph.D.
Evan Myers, M.D., M.P.H.

FINAL REPORT: December 5, 2014
# Contents

**Introduction** .......................................................................................................................... 1 
  Background and Objectives ........................................................................................................ 1 
  Approach to Benefits and Harms ............................................................................................ 1 
  Key Questions ............................................................................................................................ 2 
  PICOTS for Key Questions ........................................................................................................ 3 
  Analytic Framework .................................................................................................................. 9 

**Methods** .................................................................................................................................. 11 
  Topic Refinement and Review Protocol .................................................................................. 11 
  Literature Search Strategy ....................................................................................................... 11 
    Search Strategy ...................................................................................................................... 11 
    Inclusion and Exclusion Criteria ........................................................................................... 11 
    Study Selection ..................................................................................................................... 13 
  Evidence Synthesis .................................................................................................................. 15 
    Qualitative Synthesis ........................................................................................................... 15 
    Quantitative Synthesis ......................................................................................................... 15 
  Grading the Overall Strength of the Body of Evidence Using GRADE ................................ 17 
  Peer Review .............................................................................................................................. 19 

**Results** .................................................................................................................................... 19 
  Results of Literature Searches ............................................................................................... 19 
  Key Question 1 ....................................................................................................................... 21 
    Summary ............................................................................................................................... 21 
    Description of Included Studies .......................................................................................... 23 
    Detailed Synthesis ............................................................................................................... 26 
    Harm-benefit Trade-offs ...................................................................................................... 101 
  Key Question 2 ....................................................................................................................... 130 
    Summary ............................................................................................................................... 130 
    Description of Included Studies .......................................................................................... 131 
    Detailed Synthesis ............................................................................................................... 133 
    Harm-benefit Trade-offs: False Positives per Death Prevented ........................................... 144 
  Key Question 3 ....................................................................................................................... 145 
    Summary ............................................................................................................................... 145 
    Description of Included Studies .......................................................................................... 146 
    Detailed Synthesis ............................................................................................................... 146 
    Harm-benefit Trade-offs ...................................................................................................... 150 
  Key Question 4a ..................................................................................................................... 150 
    Summary ............................................................................................................................... 150 
    Description of Included Studies .......................................................................................... 151 
    Detailed Synthesis ............................................................................................................... 152 
  Key Question 4b ..................................................................................................................... 155 
    Summary ............................................................................................................................... 155 
    Description of Included Studies .......................................................................................... 155 
    Detailed Synthesis ............................................................................................................... 156 
  Key Question 5a ..................................................................................................................... 157
Summary ............................................................................................................. 157
Description of Included Studies ........................................................................ 157
Detailed Synthesis ............................................................................................... 157
Key Question 5b .................................................................................................. 158
Summary ............................................................................................................. 158
Harm-benefit Trade-offs: High-risk Women ....................................................... 158

**Screening for Breast Cancer: Overall Discussion** ........................................ 161
Limitations of the Review .................................................................................. 161
Limitations of Breast Cancer Screening ............................................................ 162
Key Findings for Critical Outcomes .................................................................... 163
  - Breast Cancer Mortality ................................................................................ 163
  - Life Expectancy .............................................................................................. 164
  - Overdiagnosis ............................................................................................... 164
  - False Positives .............................................................................................. 165
  - Quality-adjusted Life Expectancy ................................................................. 166
Harm-benefit Trade-offs High-risk Women ......................................................... 166

**References** ..................................................................................................... 168

**Tables**
Table 1. Inclusion and Exclusion Criteria .......................................................... 11
Table 2. Grading the Quality of Evidence for Specific Outcomes at the Individual Study Level .14
Table 3. Rating the Quality of the Body of Evidence using GRADE ...................... 18
Table 4. Summary of RCTs of Mammography (Adapted from UK Independent Panel and Cochrane Reviews) .................................................... 24
Table 5. Pooled Estimates of Breast Cancer Mortality Reduction from Screening Based on European Observational Studies ........................................ 27
Table 6. Individual Cohort Study Estimates of Breast Cancer Mortality Reduction ... 28
Table 7. Individual Case-Control Study Estimates of Breast Cancer Mortality Reduction ... 31
Table 11. Effect of Mammography on Breast Cancer Mortality by Age in RCTs ......... 43
Table 12. Effect of Mammography on Breast Cancer Mortality by Age, Observational Studies ..................................................................... 44
Table 13. Estimated Absolute Effect of Age Group on Breast Cancer Mortality Reduction, by Estimated Relative Reduction Attributable to Screening .......... 48
Table 14. Effect of Mammography on Breast Cancer Mortality by Age and Screening Interval (Canadian Task Force) ......................................................... 49
Table 15. Estimated Gains in Life Expectancy with Biennial and Annual Mammography Screening by Age to Start Screening (Assuming Screening Stops after Age 69) ...........................................55
Table 16. Estimated Gains in Life Expectancy with Biennial and Annual Mammography Screening by Age to Stop Screening (Assuming Screening Starts at Age 50) ...........................................55
Table 17. Effect of Screening Interval on Gains in Life Expectancy by Age of Starting Screening (Assuming Screening Stops after Age 69) ...........................................................................................................56
Table 18. Effect of Screening Interval on Gains in Life Expectancy by Age of Stopping Screening (Assuming Screening Starts at Age 50) ...........................................................................................................56
Table 19. Published Estimates of Overdiagnosis ..................................................................................................................62
Table 20. Percent Change in Age- and Stage-Specific Incidence of Breast Cancer, 1992-2011, SEER ..............................................................................................................................................................72
Table 21. Annual Percent Change in Age- and Stage-Specific Incidence of Breast Cancer, 1992-2011, SEER ..............................................................................................................................................................73
Table 22. Across-Country Variation in the Proportion of DCIS among all Screen-Detected Cancers in Women 50-69, ............................................................................................................................................73
Table 23. Screen-detected and Non-screen-detected DCIS among Women in the BCSC ..........77
Table 24. Studies of “Natural History” of Untreated DCIS ........................................................................................................81
Table 25. Potential Proportion of Screen-detected Lesions that Represent Overdiagnosis under Different Estimates of DCIS Progression and of the Proportion of Small Node-negative Tumors that would not Become Clinically Apparent without Screening, by Age ........................................84
Table 26. Estimated 10-year Cumulative Probability (95% CI) of False Positive Recall in the BCSC by Age, Breast Density, and HRT Status ........................................................................................................87
Table 27. Estimated 10-year Cumulative Probability (95% CI) of False Positive Biopsy in the BCSC by Age, Breast Density, and HRT Status ........................................................................................................90
Table 28. Estimated 10-year Cumulative Probability (95% CI) of a False Positive Biopsy in the BCSC by Radiologist and Patient Risk Level .........................................................................................................................91
Table 29. Estimated 10-year and Lifetime False Positive Recall Probability by Screening Interval and Age of Starting Screening (Assumes Screening Stops after Age 74), Assuming Independence of False Positive Results at Each Examination, Based on BCSC Estimates .........................................................................................93
Table 30. Estimated 10-year and Lifetime False Positive Biopsy Probability by Screening Interval and Age of Starting Screening (Assumes Screening Stops after Age 74), Assuming Independence of False Positive Results at Each Examination, Based on BCSC Estimates ........................................93
Table 31. Utility Weights Used to Estimate QALYs in CISNET and UCSF BCSC Models ......100
Table 32. Incremental False Positive Recalls and Biopsies per Breast Cancer Death Prevented, by Age to Start Screening and Screening Interval (Assuming Screening Stops after Age 69), Calculated from CISNET “Exemplar Model” Results .................................................................................................104
Table 33. Incremental False Positives per Death Prevented with Different Strategies for Use of Digital Mammography (Median Estimates Across 5 CISNET Models for Each Outcome) ......108
Table 34. Model-Estimated Cumulative Probability of Breast Cancer Death by Screening Strategy and Mortality Reduction Estimation (Cumulative Probability in Absence of Screening 3.2%) ........................................................................................................................................110
Table 35. Estimated Overdiagnoses per Breast Cancer Death Prevented among 60- to 69-year-old Invited for Screening, Florence, Italy, 1991-2007 (Adapted from Puliti, 2012) ..........................121
Table 36. Effect of Mammography on Breast Cancer Mortality by Age and Screening Interval (Canadian Task Force) ........................................................................................................................................133
Table 37. Estimated Lifetime Cancer Deaths Prevented per 100,000 by Screening Interval, Stratified by Age at Starting Screening ................................................................. 135
Table 38. Estimated Lifetime Cancer Deaths Prevented per 100,000 by Screening Interval, Stratified by Age at Stopping Screening ................................................................. 135
Table 39. Effect of Screening Interval on Gains in Life Expectancy by Age of Starting Screening ................................................................. 136
Table 40. Effect of Screening Interval on Gains in Life Expectancy by Age of Stopping Screening ................................................................. 137
Table 41. Effects of Screening Interval on Proportion of DCIS vs. Invasive by Menopausal Status and BMI .............................................................................. 138
Table 42. Estimated Effect of Screening Interval on False Positives and False Positive Biopsies by Age of Starting Screening (Assuming Screening Stops after Age 69) ................................................................. 141
Table 43. Estimated Effect of Screening Interval on False Positives and False Positive Biopsies by Age of Stopping Screening (Assuming Screening Starts at Age 50) ................................................................. 141
Table 44. Cumulative Total False Positives and False Positive Biopsies by Interval and Age to Start (Assumes Screening Stops after Age 74) ................................................................. 142
Table 45. Threshold Relative Risks where Screening of 40- to 49-year-olds Results in Equivalent Harm-benefit Ratio to Biennial Screening of 50- to 74-year-olds, by Interval, Measure of Harm-benefit, and Mammography Method ................................................................. 159
Table 46. Outcomes of Annual Mammography and Annual Mammography plus MRI in BRCA1 and BRCA2 Carriers .............................................................................. 160

Figures
Figure 1. Analytic framework .................................................................................... 10
Figure 2. Literature flow diagram ............................................................................. 20
Figure 3. Estimated Relative Reduction (with 95% CI or Range) in Breast Cancer Mortality Associated with Mammography Screening Compared to No Screening, by Study Design among Pooled Studies .............................................................................. 35
Figure 4. Estimated Cumulative Lifetime Number of Breast Cancer Deaths Prevented by Age to Start Screening (Assuming Screening Ends after Age 69) and Screening Interval .............................................................................. 46
Figure 5. Estimated Cumulative Lifetime Number of Breast Cancer Deaths Prevented by Age to Stop Screening and Screening Interval (Assuming Screening Starts at Age 50) .............................................................................. 47
Figure 6. Effect of Age and Comorbidity on Reduction in Breast Cancer Mortality by Continuing to Screen to Given Age (from Data in Lansdorp-Vogelaar, 2014) .............................................................................. 48
Figure 7. Ratio of Cumulative Probability of Death from Breast Cancer to Death from other Causes by Age and Year Post-Diagnosis, SEER 2002-2010 .............................................................................. 54
Figure 8. Age-specific (A) and Cumulative (B) Incidence of In Situ Breast Cancers, Invasive Breast Cancers <2 cm with No Nodes or Distant Metastases, and All Other Invasive Breast Cancers, SEER, 2000-2010 .............................................................................. 76
Figure 9. Distribution of Breast Cancer Diagnoses by Age .............................................................................. 77
Figure 10. Estimated Age-specific Incidence of In Situ, T1N0M0 Invasive Breast Cancer, and All Other Breast Cancers in Unscreeneed (A) and Screened (B) Women .............................................................................. 78
Figure 11. Estimated Distribution of Diagnoses by Age in Unscreeneed (A) and Screened (B) Women .............................................................................. 80
Figure 12. Trends in Incidence of Invasive Cervical, Colorectal, and Breast Cancer, and In Situ Breast Cancer, SEER, 1973-2011 .............................................................................. 83
Figure 13. Estimated Number of (A) Total False Positives and (B) False Positive Biopsies by Age to Start Screening (Assuming Screening Ends after Age 69) and Screening Interval ..........95
Figure 14. Estimated Number of (A) Total False Positives and (B) False Positive Biopsies by Age to Start Screening (Assuming Screening Ends after Age 69) and Screening Interval ..........96
Figure 15. False Positive Biopsies and Breast Cancer Deaths Prevented, by Age to Start Screening and Screening Interval (Assuming Screening Stops at Age 69) ...........................................105
Figure 16. False Positive Biopsies and Breast Cancer Deaths Prevented, by Age to Stop Screening and Screening Interval (Assuming Screening Stops at Age 50) .............................................106
Figure 17. False Positive Biopsies and Deaths Prevented by Age to Start Screening (A and C) and Age to Stop Screening (B and D, Biennial (Solid Line) vs. Annual (Dotted Line) Screening ..........................................................107
Figure 18. Harm-benefit Acceptability Curves for False Positive Biopsies (A and B) and Total False Positives (C and D) by Age to Start Screening and Mortality Reduction .........................112
Figure 19. Harm-benefit Acceptability Curves for False Positive Biopsies by Age to Stop Screening and Mortality Reduction ..................................................................................114
Figure 20. Reported “Willingness to Pay” in Terms of False Positives per Death Prevented ....116
Figure 21. Harm-benefit Acceptability Curve for Overdiagnoses and Breast Cancer Deaths Prevented for Women 60-69 Years Old in Florence, Italy (Derived from Puliti, 2012), “Base Case” Estimates.........................................................................................................................122
Figure 22. Harm-benefit Acceptability Curve for Overdiagnoses and Breast Cancer Deaths Prevented for Women 60-69 Years Old in Florence, Italy (Derived from Puliti, 2012), “Sensitivity Analysis” Estimates .................................................................................................................................................122
Figure 23. Harm-benefit Acceptability Curves: Overdiagnosed Cases of DCIS per Breast Cancer Death Prevented by Relative Risk of DCIS Among Screened Women and Probability of Progression of DCIS to Cancer in the Absence of Treatment, Relative Mortality with Screening 0.62 (95% CI, 0.56 to 0.69) ..........................................................................................................................125
Figure 24. Harm-benefit Acceptability Curves: Overdiagnosed Cases of DCIS per Breast Cancer Death Prevented by Relative Risk of DCIS Among Screened Women and Probability of Progression of DCIS to Cancer in the Absence of Treatment, Relative Mortality with Screening 0.80 (95% CI 0.73 to 0.89) ..........................................................................................................................126
Figure 25. Women’s Threshold for Treatment of DCIS According to Chance of Becoming Invasive .................................................................................................................................129

Appendixes
Appendix A. Exact Search Strings
Appendix B. Data Abstraction Elements
Appendix C. Modeling Methods
Appendix D. List of Included Studies by Key Question
Appendix E. List of Excluded Studies
Appendix F. Key to Included Primary and Companion Articles
Appendix G. Study Characteristics Tables
Appendix H. GRADE Summary Tables
Introduction

Background and Objectives

The overall goal of this project is to support the American Cancer Society (ACS) Guidelines Development Group (GDG) in the development of evidence-based breast cancer screening guidelines that meet the criteria outlined by the Institute of Medicine (IOM) 2011 report, “Clinical Practice Guidelines We Can Trust.” This support includes:

- Systematic review of the scientific literature;
- Synthesis of the evidence using appropriate methods, including both qualitative summaries and quantitative approaches such as meta-analysis and decision analysis;
- Rating the quality of the evidence using criteria developed by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group;
- Summarizing the review, synthesis, and quality rating for the GDG, with an emphasis on presenting the results in a format that will enable the GDG to translate the evidence into guidelines using GRADE; and
- Summarizing the review, synthesis, and quality rating for the public and scientific community with a manuscript to a peer-reviewed journal that describes the methodology and key findings of the systematic review.

Approach to Benefits and Harms

In an “ideal” setting (assuming perfect adherence on the part of patients and clinicians, no resource constraints, etc.), the relative benefits and harms of screening for any cancer are based on four basic considerations:

1) Benefits:
   - What is the probability that screening will detect a potentially fatal cancer earlier in its natural history prior to onset of symptoms, and what is the probability that earlier detection leads better health outcomes (reduced mortality, potentially reduced morbidity) than managing a cancer that presents through clinical signs or symptoms?

2) Harms:
   - What is the probability that a given screening test will result in a suspicious finding requiring additional work-up but not resulting in a cancer diagnosis?
   - What is the probability that false positive test results will lead to worse health outcomes compared to no screening?
   - What is the probability of harms associated with detecting and treating an unsuspected cancer or cancer precursor with a given screening test that would otherwise not have become clinically apparent during a patient’s lifetime (overdiagnosis)?

3) Benefits and harms from screen-detected cancers and cancer precursors:
   - The probability of a previously unknown precursor or invasive breast cancer being present at the time of the screening test (prevalence at the time of screening) is a function of:
     - Age (in all women);
### Presence of risk factors
- Presence of risk factors (including family history, use of hormone replacement therapy, or known genetic predisposition);
- Sensitivity of previous screening test and time since previous screening test;
- Sensitivity of a given test (mammography, clinical breast exam [CBE], magnetic resonance imaging [MRI], etc.) for detecting breast cancer precursors (e.g., ductal carcinoma in situ [DCIS]) and invasive cancer.

The relative probability of death and morbidity due to breast cancer, and of morbidity due to breast cancer treatment, in women with cancers and cancer precursors detected through screening compared to women with cancers diagnosed through clinical signs and symptoms is a function of:
- Effectiveness of treatment in women with screen-detected vs. clinically diagnosed cancers;
- Adverse outcomes of treatment in women with screen-detected cancer precursors, screen-detected cancers, and clinically diagnosed cancers;
- Competing risks for death (in turn a function of age and comorbid conditions);
- The probability of a cancer precursor progressing to invasive cancer.

### 4) Harms from false positives:
The probability of a previously unknown precursor/invasive breast cancer (the lower this probability, the higher the probability of a false positive result) is a function of:
- Age;
- Other risk factors;
- The type and time since any previous screening test;
- The specificity of a given test;
- The health outcomes related to a false positive diagnosis.

Within this framework, the trade-off between benefits and harms resulting from different possible recommendations for breast cancer screening varies primarily based on the probability of cancer/cancer precursors (driven by factors such as age, presence of other risk factors, and screening intervals) and the test characteristics of sensitivity and specificity.

### Key Questions
With input from the ACS and the GDG, we revised the Key Questions (KQs) specified in the original Request for Proposals (RFP) using the general approach of specifying the Populations, Interventions, Comparisons, Outcomes, Timings of outcomes, and Settings (PICOTS) of interest for each KQ (see the next section for details of PICOTS for each KQ). The first three KQs focus on average-risk women; the remaining four questions (KQs 4 and 5 are each split into two parts) focus on women with an increased risk of breast cancer.

KQs were:

- **KQ 1**: What are the relative benefits, limitations, and harms associated with mammography screening compared to no screening in average-risk women ages 40 and older, and how do they vary by age, screening interval, and prior screening history?
• **KQ 2:** In average-risk women who are screened with mammography, what are the relative benefits, limitations, and harms associated with annual, biennial, triennial, or other screening interval, and how do they vary by age?

• **KQ 3:** What are the benefits, limitations, and harms associated with clinical breast examination (CBE) among average-risk women 40 years and older compared to no CBE, and how do they vary by age, interval, and participation rates in mammography screening?

• **KQ 4a:** Among women with an increased risk of breast cancer due to factors known PRIOR to the onset of screening (e.g., family history, BRCA mutation carrier, history of chest irradiation), what are the relative benefits, limitations, and harms associated with different screening modalities compared to no screening (i.e., what ages to start and stop screening) and to each other?

• **KQ 4b:** Among women with an increased risk of breast cancer due to factors identified AS THE RESULT OF screening or diagnosis (e.g., prior diagnosis of proliferative lesions), what are the benefits, limitations, and harms associated with different screening modalities compared to no screening, and to each other?

• **KQ 5a:** Among women with an increased risk of breast cancer due to factors known PRIOR to the onset of screening (e.g., family history, BRCA mutation carrier, history of chest irradiation), what are the relative benefits, limitations, and harms associated with different screening modalities at different intervals, and how do these vary by age?

• **KQ 5b:** Among women with an increased risk of breast cancer due to factors identified AS THE RESULT OF screening or diagnosis (e.g., prior diagnosis of proliferative lesions,) what are the benefits, limitations, and harms associated with different screening modalities at different intervals, and how do these vary by age?

**PICOTS for Key Questions**

In this section, we outline the PICOTS of interest for each KQ.

**KQ 1:** *What are the relative benefits, limitations, and harms associated with mammography screening compared to no screening in average-risk women ages 40 and older, and how do they vary by age, screening interval, and prior screening history?*

**Population:** Women aged 40 and older, who do NOT have a history of:
- Known susceptibility gene mutation (e.g., BRCA1/BRCA2);
- History of previous breast cancer or DCIS;
- Family history of breast cancer (define in terms of number, degree of relation);
- Lobular neoplasia;
- Previous abnormal pathology (proliferative lesions);
- Previous chest irradiation.
Subgroups of interest include:

- **Age:**
  - 40 and older with no upper limit
  - Subgroups by 5-year increments as possible (to get at data that may be hidden in larger 10-year breakdowns)
  - Consider upper age cutoff for highest age group using a range of cut points (e.g., over 65, 70, 75, 80, 85, and recognizing that 5-year interval data may be sparse for older age groups)

- **Race/ethnicity:**
  - White, non-Hispanic and White, Hispanic
  - Black/African-American, non-Hispanic and Black/African-American, Hispanic
  - Asian-Pacific Islander
  - Native American/Alaska
  - Other, Hispanic and non-Hispanic

- **Comorbidities:**
  - Presence or absence of potentially fatal comorbid conditions (e.g., other cancers, chronic heart disease, diabetes) and interaction with age on competing risk of non-breast cancer mortality

**Interventions:**

- Plain film mammography
- Digital mammography
  - Digital direct radiography (DR)
  - Computed radiography (CR)

Note: We did not abstract studies that directly compared two different methods of performing mammography. For each included study of mammography, we recorded important aspects of the method used that might affect test performance (plain film vs. digital, one- vs. two-view, single vs. double reader, computer aided vs. unaided) and used these data to rate the study in terms of direct applicability to current U.S. practice.

**Comparisons:**

- No mammography vs. mammography (plain film or digital) at any screening interval
- Repeat comparison for identified subgroups as defined above

**Outcomes:**

- **Critical:**
  - Breast cancer mortality (breast cancer deaths prevented by screening)
  - Life expectancy (life-years gained by screening)
  - Quality of life (quality-adjusted life-years gained by screening)
  - Overdiagnosis (screen-detected cancers that would not have led to symptomatic breast cancer if undetected by screening)
  - Overtreatment (cancer therapies—surgery, radiation, chemotherapy—performed for screen-detected cancers that would not have led to symptomatic breast cancer if undetected by screening)
o False positive results, stratified as:
  – Repeat examination on same day as positive screening result
  – Additional imaging performed subsequent to screening visit
  – Biopsy resulting in normal diagnosis
• Important but not critical:
  o Stage distribution at diagnosis
  o Emotional impact (anxiety, depression, etc.) of positive results (true and false positives)
• Limited importance (Note: Since, by definition, these outcomes should not be considered in formulating strength of recommendations under GRADE, relevant articles on these outcomes were flagged at the time of screening, but were not abstracted or rated for quality.)
  o Reassurance from true negatives
  o False reassurance from false negatives
  o Secondary effects of test results on health resource utilization, both breast cancer related and non-breast cancer related

Timing of outcomes:
• Immediate (up to 12 weeks after screening)
• Short-term (within 12 weeks to 18 months of screening)
• Longer-term (greater than 18 months after screening)
  o Time intervals for longer-term follow-up were reported specifically as reported in the original study or categorized in systematic reviews.

Settings:
• Screening program
• Opportunistic screening
• Presence/absence of infrastructure to insure adequate follow-up of test results

**KQ 2: In average-risk women who are screened with mammography, what are the relative benefits, limitations, and harms associated with annual, biennial, triennial, or other screening interval, and how do they vary by age?**

PICOTS identical to KQ 1, except:

Comparisons:
• Mammography (digital or plain film) at intervals of:
  o 1 year
  o 2 years
  o 3 years
  o Alternative intervals (e.g., 18 months)
KQ 3: What are the benefits, limitations, and harms associated with clinical breast examination among average-risk women 40 years and older compared to no CBE, and how do they vary by age, interval, and participation rates in mammography screening?

Population: Women aged 40 and older, who do NOT have a history of:
- Known susceptibility gene mutation (e.g., BRCA1/BRCA2);
- History of previous breast cancer or DCIS;
- Family history of breast cancer (need to define in terms of number, degree of relation);
- Lobular neoplasia;
- Previous abnormal pathology (proliferative lesions);
- Previous chest irradiation.

Subgroups of interest include:
- Age:
  - 40 and older with no upper limit
  - Premenopausal vs. postmenopausal (definition of menopause may vary between studies)
  - 5-year age increments, with stopping age varying from 70 up
- Race/ethnicity:
  - White, non-Hispanic and White, Hispanic
  - Black/African-American, non-Hispanic and Black/African-American, Hispanic
  - Asian-Pacific Islander
  - Native American/Alaska
  - Other, Hispanic and non-Hispanic
- Comorbidities:
  - Presence or absence of potentially fatal co-morbid conditions (e.g., other cancers, chronic heart disease, diabetes)
- Adherence to mammography recommendations, characterized as:
  - Ever screened versus never screened
  - Time since last screen

Interventions:
- Clinical breast exam (CBE)

Comparisons:
- CBE (at 1-, 2-, 3-year intervals) vs. no CBE (and no other screening)
- CBE (at 1-, 2-, 3-year intervals) + mammography (at different intervals) vs. mammography alone

Outcomes:
- Same as listed above (KQ 1)

Timing of outcomes:
- Same as listed above (KQ 1)
- Data may not support as discrete an analysis of interval as in mammography
Setting:
- Type of provider (family physician, nurse practitioner, obstetrician/gynecologist, etc.)

**Important note on KQs 4 and 5:** Because our initial review found limited evidence on breast cancer mortality for KQs 4 and 5, we included stage distribution of tumors detected through screening as an alternate critical outcome for these KQs after discussion with the GDG.

*KQ 4a:* Among women with an increased risk of breast cancer due to factors known PRIOR to the onset of screening (e.g., family history, BRCA mutation carrier, history of chest irradiation), what are the relative benefits, limitations, and harms associated with different screening modalities compared to no screening (i.e., what ages to start and stop screening) and to each other?

**Population:**
- Women ages 40 and older with:
  - Known susceptibility gene mutation (e.g., BRCA1/BRCA2);
  - Family history of breast cancer (need to define in terms of number, degree, etc.):
    - Unknown BRCA1/BRCA2 status
    - Test negative BRCA1/BRCA2
  - Previous chest irradiation;

Subgroups of interest include:
- Same as KQ 1 and 2, above—vary by age and/or menopausal status, race/ethnicity, comorbidity

**Interventions:**
- Plain film mammography
- Digital mammography
- CBE
- MRI
- Ultrasound
- Tomosynthesis

**Comparisons:**
- Varying age at starting and age of stopping, and varying order of tests (e.g., mammography followed by MRI followed by mammography)

**Outcomes:**
- All outcomes listed above (KQ 1), *plus*
- Stage distribution of tumors detected through screening (added as an alternate critical outcome in lieu of data on mortality)

**Timing of outcomes:**
- Same as listed above (KQ 1)
Settings:
- Same as listed above (KQ 1)

**KQ 4b:** Among women with an increased risk of breast cancer due to factors identified as the result of screening or diagnosis (e.g., prior diagnosis of proliferative lesions), what are the benefits, limitations, and harms associated with different screening modalities compared to no screening, and to each other?

PICOTS identical to KQ 4a, except:

**Population:**
- Women ages 40 and older with:
  - Lobular neoplasia
  - Previous abnormal pathology (proliferative lesions)

Subgroups of interest include:
- Same as KQs 1 and 2, above—vary by age and/or menopausal status, race/ethnicity, comorbidity

**KQ 5a:** Among women with an increased risk of breast cancer due to factors known prior to the onset of screening (e.g., family history, BRCA mutation carrier, history of chest irradiation), what are the relative benefits, limitations, and harms associated with different screening modalities at different intervals, and how do these vary by age?

PICOTS identical to KQ 4a, except:

**Comparisons:**
- All screening modalities, at intervals of:
  - 1 year
  - 2 years
  - 3 years
  - Alternative intervals (e.g., 18 months)

**KQ 5b:** Among women with an increased risk of breast cancer due to factors identified as the result of screening or diagnosis (e.g., prior diagnosis of proliferative lesions), what are the benefits, limitations, and harms associated with different screening modalities at different intervals, and how do these vary by age?

PICOTS identical to KQ 4b, except:

**Comparisons:**
- All screening modalities, at intervals of:
  - 1 year
  - 2 years
  - 3 years
  - Alternative intervals (e.g., 18 months)
Analytic Framework

Figure 1 depicts the analytic framework for this project.
Figure 1. Analytic framework

Abbreviations: ACS=American Cancer Society; BrCA=breast cancer; KQ=Key Question
Methods

Topic Refinement and Review Protocol

Through a series of conference calls with ACS staff and the GDG, we revised the KQs, PICOTS, and protocol from those originally specified in the RFP and proposal.

Literature Search Strategy

Search Strategy

To identify relevant published literature, we searched PubMed® (March 6, 2014), CINAHL® (September 10, 2013), and PsycINFO® (September 10, 2013). No lower date limit was used for RCTs; for observational studies, we searched for all citations published from January 1, 2000, on. An experienced search librarian advised on all searches. Exact search strings are included in Appendix A. We also checked to ensure that our search results captured all studies included in four key systematic reviews of RCTs and three key systematic reviews of observational studies, particularly for studies reporting mortality. All citations were imported into an electronic database (EndNote® X4; Thomson Reuters, Philadelphia, PA).

Inclusion and Exclusion Criteria

The criteria used to screen articles for inclusion/exclusion at both the title-and-abstract and full-text screening stages are detailed in Table 1.

Table 1. Inclusion and Exclusion Criteria

<table>
<thead>
<tr>
<th>Study Characteristic</th>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
</table>
| Population           | For radiographic studies:  
                        • Women aged 40 and older  
                        • Without known risk factors  
                        • With known risk factors (breast cancer susceptibility gene carrier, previous chest irradiation, family history, previous DCIS or lobular neoplasia, previous abnormal pathology)  
                        For CBE:  
                        • Women aged 40 and older, with and without risk factors listed above |  
                        • Nonhuman subjects  
                        • Male subjects  
                        • Previous invasive breast cancer |
| Interventions        |  
                        • No screening  
                        • Mammography (film and digital)  
                        • CBE  
                        • MRI  
                        • Ultrasound  
                        • Tomosynthesis | Screening modalities other than those listed |
<table>
<thead>
<tr>
<th>Study Characteristic</th>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
</table>
| **Comparators**      | • No screening vs. mammography, CBE, or other modality  
                      • Comparisons between screening methods (e.g., mammography vs. CBE, or mammography vs. MRI)  
                      • Different intervals  
                      • Different outcomes (e.g., studies that compare patient preferences or utilities for different outcomes relative to breast cancer screening) | No comparisons or outcomes of interest between:  
• Screening vs. no screening (any method)  
• Different methods (e.g., mammography vs. MRI, or digital vs. plain film mammography)  
• Different intervals (any method)  
• Different intermediate screening outcomes (same or different methods)—e.g., depression scores after false positive vs. true negative results |
| **Outcomes**         | • Breast cancer mortality  
                      • Life expectancy (life-years gained by screening)  
                      • Quality of life (quality-adjusted life-years gained by screening)  
                      • Overdiagnosis (screen-detected cancers that would not have led to symptomatic breast cancer if undetected by screening)  
                      • Overtreatment (cancer therapies—surgery, radiation, chemotherapy—performed for screen-detected cancers that would not have led to symptomatic breast cancer if undetected by screening)  
                      • False positive results  
                      • Stage distribution at diagnosis  
                      • Emotional impact (anxiety, depression, etc.) of positive results (true and false positives)  
                      • Recall rates  
                      • Sensitivity and specificity (only if a 2x2 table can be completed)  
                      • Patient preferences as measured using validated quality-of-life measures, utilities using accepted methods such as standard gamble or time-trade-off; stated preferences measured by conjoint analysis; revealed preference studies; etc. | • Outcomes not listed  
• Economic outcomes only |
<p>| <strong>Timing of outcomes</strong> | • Studies of any duration | • None |
| <strong>Setting</strong>          | • All settings where screening is provided | • None |</p>
<table>
<thead>
<tr>
<th>Study Characteristic</th>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
</table>
| Study design         | • Controlled studies (RCTs, cohort studies, case-control studies), pooled patient-level meta-analyses, systematic reviews, and study-level meta-analyses  
• Modeling/simulation studies that meet other inclusion criteria (modeling may be the only way to generate estimates of long-term effects of screening in many settings)  
• Observational studies (prospective and retrospective cohort studies, case-control studies, or cross-sectional studies) published since 2000 with an n ≥ 1000 for average-risk women, or n ≥ 100 for high-risk populations | • Not a research study (e.g., editorial, non-systematic review, letter to the editor)  
• Exploratory/pilot study  
Note: Although we did not formally abstract non-systematic reviews, many of these included substantial discussions of important methodological issues. We used these to help inform our review, grading, and discussion of the evidence. |
| Publication type     | • English language only  
• Peer-reviewed articles | • Non-English articles  
• Abstracts only |

Abbreviations: CBE=clinical breast exam; DCIS=ductal carcinoma in situ; MRI=magnetic resonance imaging; RCTs=randomized controlled trials

**Study Selection**

Using the prespecified inclusion and exclusion criteria described in Table 1, two investigators independently reviewed titles and abstracts for potential relevance to the KQs. Articles included by either reviewer underwent full-text screening. At the full-text review stage, paired researchers independently reviewed the articles and indicated a decision to “include” or “exclude” the article for data abstraction. When the two reviewers arrived at different decisions about whether to include or exclude an article, they reconciled the difference through review and discussion, or through a third-party arbitrator if needed. Full-text articles meeting our eligibility criteria were included for data abstraction. We confirmed that we had included all of the studies included in four key recent systematic reviews, particularly for studies reporting mortality. All screening decisions were made and tracked in a DistillerSR database (Evidence Partners Inc., Manotick, ON, Canada).

**Data Extraction**

The research team created data abstraction forms and evidence table templates for abstracting data for each KQ. Based on clinical and methodological expertise, a pair of investigators was assigned to abstract data from each eligible article. One investigator abstracted the data, and the second reviewed the completed abstraction form alongside the original article to check for accuracy and completeness. Disagreements were resolved by consensus, or by obtaining a third reviewer’s opinion if consensus could not be reached. To aid in both reproducibility and standardization of data collection, researchers received data abstraction instructions directly on each form created specifically for this project within the DistillerSR database.

We designed the data abstraction forms to collect the data required to evaluate the specified eligibility criteria for inclusion in this review, to facilitate both data reporting and formal synthesis (e.g., for studies of test characteristics, abstractors may fill in summary estimates of sensitivity, specificity, and predictive values for reporting, as well as 2x2 tables to facilitate potential meta-analysis). Before the data abstraction form templates were used, they were pilot-tested with a sample of included articles to ensure that all relevant data elements were captured.
and that there was consistency/reproducibility between abstractors. Forms were revised as necessary before full abstraction of all included articles. Appendix B provides a detailed listing of the elements included in the data abstraction forms.

We also developed forms and provided instructions for grading the quality of evidence for specific outcomes at the individual study level. We used the GRADE methodology for rating individual study limitations (risk of bias), using a four-point scale from very low to high quality, with randomized controlled trials (RCTs) starting with a high quality rating and observational studies starting with a moderate quality rating, with specific study limitations lowering the rating (Table 2). This will facilitate translation of the review results into a format that will enable the GDG to efficiently review the quality of the evidence and formulate guideline recommendations.

Table 2. Grading the Quality of Evidence for Specific Outcomes at the Individual Study Level

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Initial Quality Rating</th>
<th>Factors Lowering Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCT</td>
<td>High</td>
<td>• Lack of allocation concealment&lt;br&gt;• Lack of blinding&lt;br&gt;• Incomplete accounting of patients and outcome events&lt;br&gt;• Selective outcome reporting bias&lt;br&gt;• Stopping early for benefit&lt;br&gt;• Use of unvalidated outcome measures (e.g., patient-reported outcome)&lt;br&gt;• Carryover effects in cross-over trials&lt;br&gt;• Recruitment bias in cluster randomized trials</td>
</tr>
<tr>
<td>Observational study</td>
<td>Moderate</td>
<td>• Failure to develop and apply appropriate eligibility criteria (inclusion of control population)&lt;br&gt;• Flawed measurement of both exposure and outcome&lt;br&gt;• Failure to adequately control confounding&lt;br&gt;• Incomplete follow-up</td>
</tr>
<tr>
<td>Modeling study</td>
<td>Moderate</td>
<td>• Failure to specify model structure&lt;br&gt;• Failure to identify data sources for parameters&lt;br&gt;• Failure to describe methods of imputation for unmeasurable parameters (such as time to progression for undiagnosed cancers)&lt;br&gt;• Failure to describe and justify key assumptions&lt;br&gt;• Failure to perform sensitivity analyses&lt;br&gt;• If probabilistic analyses performed, failure to describe distributions used, or use of inappropriate distributions (e.g., normal distributions for parameters bounded by 0)</td>
</tr>
</tbody>
</table>

Abbreviation: RCT = randomized controlled trial

For studies that reported on more than one relevant outcome, we performed separate quality ratings for each outcome (i.e., it is possible for a study to be of “Moderate” quality for one outcome but “Low” or “High” for another).

Forms were developed in DistillerSR to record final individual study quality ratings, as well as the specific limitations resulting in any downgrading. For grading the quality of the body of evidence across each KQ outcome, we generated tables using the recommended GRADE format.

Modeling studies are, by definition, indirect evidence. Therefore, even the highest quality modeling study can be, at best, only moderate quality evidence. We rated individual modeling studies using the recently published recommendations of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR).
Evidence Synthesis

Qualitative Synthesis

For all critical outcomes, we discuss results and methodological limitations of included studies, note qualitative patterns or inconsistencies, and discuss common themes and potential explanations for observed patterns or inconsistencies. We identified papers meeting criteria that were relevant to outcomes rated as important by the GDG, but did not abstract them or grade their quality since, under GRADE, they are not directly factored into decisions about recommendations or strength of recommendations. As noted above, because our initial review found limited evidence on breast cancer mortality for KQs 4 and 5, the GDG elected to formally review the evidence for stage distribution of tumors detected through screening for these questions—that is, the GDG chose to treat stage distribution as an alternate critical outcome for KQs 4 and 5.

Quantitative Synthesis

We considered three forms of quantitative data synthesis for this review, based on the results of the literature review and input from ACS and the GDG:

1) Meta-analysis: Meta-analytic results of outcomes may be particularly helpful for GRADE quality rating regarding the precision of estimates of outcomes. Factors that we usually consider in deciding on the utility of meta-analysis are statistical power, conceptual homogeneity across studies, and the feasibility of generating a summary estimate. To perform meta-analyses we use Comprehensive Meta-Analysis v 2.0 (Englewood, NJ: Biostat, Inc), typically using random-effects models. We evaluate heterogeneity both visually and quantitatively, and perform relevant sensitivity analyses (e.g., by study design).

Four high-quality systematic reviews/meta-analyses published within the past 4 years have synthesized the available data, particularly for breast cancer mortality, and have reported roughly similar results.\(^3,4,6,11\) Given the size of the literature, we planned to rely on these reviews, after confirming that they met appropriate methodological standards and used inclusion/exclusion criteria similar to ours. We abstracted data from the most recent article reporting results from each of the key RCTs. Our plan was to abstract additional individual articles only if they were not included in the four key reviews. We planned to conduct our own meta-analyses only if any additional literature (a) was substantially different in results from previous studies, or (b) would substantively improve our ability to grade the quality of evidence for a particular outcome (because it would substantially improve the precision of the estimate of effect on harm or benefit).

We did not identify any updated evidence from the studies included in this review, or new evidence from other studies, that would be likely to substantially change either the direction of effect or the precision of estimates. We also did not identify any new evidence for outcomes that were not amenable to quantitative synthesis in previous reviews (such as overdiagnosis). In our judgment, additional meta-analysis will not substantially help the GDG resolve uncertainties about the evidence.

2) Estimating absolute effects for the U.S. population: The majority of the available literature on screening outcomes, particularly mortality and overdiagnosis, comes from studies conducted outside the U.S. These studies, both alone and when combined in meta-analyses, provide estimates of the relative effect of different screening strategies on outcomes, and, in
some cases, there are estimates of the absolute effect as well. While differences between study settings may affect the magnitude of the relative effect, the more important issue for the purposes of developing guidelines for U.S. women is that estimates of the absolute effect may not be applicable. For example, the absolute difference in breast cancer death attributable to screening is dependent on the incidence of cancer in an unscreened population (which may vary depending on differences in the distribution of cancer risk factors, as well as variations in the likelihood that a woman with a cancer at a given stage will present with symptoms leading to detection and classification as an incident case) and in mortality from cancer at a given point in its natural history (which may vary based on differences in access to care, quality of care, or differences in competing risks of mortality). As we will discuss in the Results, there is also substantial variability between countries in outcomes such as false positives or the diagnosis of in situ cancers (which may contribute to overdiagnosis). Given the large differences between the European countries where the majority of the evidence on screening outcomes was generated and the U.S. in terms of both population characteristics and the health system, estimates of the absolute effect for any outcome provided by European studies may be substantially higher or lower than in the U.S.

In the absence of population-based data on outcomes among screened and unscreened women in the U.S., estimating the absolute effects requires use of either sophisticated mathematical models or cruder approaches requiring a range of simplifying assumptions. Where available, we report on estimates from models reported in the literature. We also used a simpler approach to generate estimates of age-specific incidence, incidence-based mortality, and 15-year survival for breast cancer in U.S. women using SEER*Stat software. (We acknowledge that, as with using non-U.S. data, these results may also under- or overestimate the “true” absolute effects of screening; however, the estimates in this case are derived from observed U.S.-specific data). Age-specific results were also stratified into in situ lesions, invasive cancers <2 cm in diameter with no nodal involvement or distant metastases (T1N0M0), and all other invasive cancers. Given these estimates, literature-based estimates of the relative effect of different screening strategies on the outcome (e.g., relative reduction in breast cancer mortality), and estimates of the prevalence of screening from the National Health Interview Survey, we then calculated event probabilities for screened and unscreened U.S. women. For example, overall breast cancer mortality is the weighted average of mortality among screened women (where pScreened is the proportion of women screened):

\[ \text{Mortality}_{\text{Overall}} = \text{Mortality}_{\text{Screened}} \times p_{\text{Screened}} + \text{Mortality}_{\text{Unscreened}} \times (1 - p_{\text{Screened}}) \]

Since

\[ \text{Mortality}_{\text{Screened}} = \text{Mortality}_{\text{Unscreened}} \times \text{RelativeMortality}_{\text{Screened}} \]

mortality in unscreened women can be calculated as:

\[ \text{Mortality}_{\text{Unscreened}} = \frac{\text{Mortality}_{\text{Overall}}}{(\text{RelativeMortality}_{\text{Screened}} \times p_{\text{Screened}}) + (1 - p_{\text{Screened}})} \]

and mortality in screened women can be estimated by multiplying mortality in unscreened women by the relative reduction attributable to screening.

More details are provided in the individual sections under Results, and in Appendix C.
3) Harm-benefit trade-offs: Simulation models can be especially useful for synthesizing data from a variety of sources, comparing interventions and outcomes that may not be feasible to compare even with observational study designs, and estimating the impact of specific parameters on outcomes. Probabilistic models may be particularly useful as tools for visualizing the effect of uncertainty about harm-benefit trade-offs on the strength and direction of recommendations using GRADE.

Much of the recent controversy about breast cancer screening revolves around whether the benefit of screening is outweighed by potential harms, particularly in certain populations (e.g., Gregory, 2010). The review of the available evidence and the estimates of absolute effects in the U.S. population provide our estimates for mortality reduction and other critical outcomes, but they do not provide direct estimates of harm-benefit ratios (for example, overdiagnoses per breast cancer death prevented) or estimates of uncertainty around these ratios resulting from uncertainty in the estimates of the numerator and denominator (for example, given a point estimate and 95% CIs for overdiagnoses and breast cancer mortality, what is the 95% CI of the harm-benefit ratio?). However, even an estimate of a particular harm-benefit ratio with a 95% CI is not helpful for making decisions if there is no consensus on what a maximal acceptable ratio should be, or if there is likely to be variability among different GDG panel members, patients, providers, and other stakeholders.

In order to provide these estimates, we developed simple models to estimate the joint probabilities of critical outcomes (in particular, breast cancer mortality, overdiagnosis, and false positives) using the age-specific SEER data, and parameter estimates from the literature to generate harm-benefit acceptability curves, which depict the likelihood that a given strategy will be above or below a given harm-benefit ratio; this approach is derived from economic analysis, where the optimal strategy may vary based on “willingness-to-pay” for a given outcome. Again, details are provided in individual sections under Results and in Appendix C.

Grading the Overall Strength of the Body of Evidence Using GRADE

We graded the overall quality of the body of evidence for each outcome per KQ based on the specific criteria outlined by GRADE (Table 3). There is no explicit “formula” for grading strength of evidence when data are available from both RCTs and observational studies, particularly when, as is the case with breast cancer screening, there are differences in the magnitude of effect across different study designs, and where factors other than study internal validity/risk of bias, such as secular trends in incidence, screening technology, and treatment effectiveness may influence the applicability of the evidence to the population of interest. For each outcome per KQ, we provide our assessment of the overall strength of evidence across all included study designs by assessing four domains: risk of bias, consistency, directness, and precision. An additional domain considered was strength of association (magnitude of effect). For risk of bias, we considered basic (e.g., RCT) and detailed study design (e.g., evidence of imbalance between intervention and control groups). We used results from meta-analyses when evaluating consistency (forest plots, tests for heterogeneity), precision (confidence intervals), and strength of association (weighted mean difference). These domains were considered qualitatively, and a summary rating of high, moderate, low, or very low strength of evidence was assigned after discussion by two investigators. This four-level rating scale consists of the following definitions:
• High—We are very confident that the true effect lies close to that of the estimate of the effect. (Alternative: Further research is very unlikely to change our confidence on the estimate of effect.)

• Moderate—We are moderately confident in the effect estimate; the true effect is likely to be close to the estimate of effect, but there is a possibility that it is substantially different. (Alternative: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.)

• Low—Our confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect. (Alternative: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.)

• Very low—We have very little confidence in the effect estimate; the true effect is likely to be substantially different from the estimate of effect. (Alternative: Evidence on an outcome is absent or too weak, sparse, or inconsistent to estimate an effect.)

GRADE also does not provide explicit guidance on how to weight modeling studies. Even the most sophisticated modeling study will be limited by the strength of the evidence available for the most important parameters. In general, because modeling is often most useful for addressing questions where direct evidence is difficult to obtain (comparing a large number of different screening intervals and starting and stopping ages), and because many models require assumptions or imputed values in order to be tractable, there will almost always be residual uncertainty about the results of modeling studies. Therefore, we assumed that modeling studies themselves could be no higher than moderate quality. As part of the total body of evidence, modeling studies raised quality if they contributed to improved consistency of results (e.g., if model-based estimates of mortality reduction were consistent with observational studies that were not used to provide inputs into the model).

Table 3. Rating the Quality of the Body of Evidence using GRADE

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Initial Quality</th>
<th>Lower Quality If</th>
<th>Raise Quality If</th>
<th>Quality of Body of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized trials</td>
<td>High (four plus: ⊕⊕⊕⊕)</td>
<td>Risk of bias:</td>
<td>Large effect:</td>
<td>High (four plus: ⊕⊕⊕⊕)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Serious</td>
<td>+1 Large</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very serious</td>
<td>+2 Very large</td>
<td></td>
</tr>
<tr>
<td>Observational studies</td>
<td>Moderate (three plus: ⊕⊕⊕)</td>
<td>Inconsistency:</td>
<td>Dose response:</td>
<td>Moderate (three plus: ⊕⊕⊕)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Serious</td>
<td>+1 Evidence of a gradient</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very serious</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Indirectness:</td>
<td>All plausible residual confounding:</td>
<td>Low (two plus: ⊕⊕)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Serious</td>
<td>+1 Would reduce a demonstrated effect</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very serious</td>
<td>+1 Would suggest a spurious effect if no effect was observed</td>
<td>Very low (one plus: ⊕)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imprecision:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study Design</td>
<td>Initial Quality</td>
<td>Lower Quality If</td>
<td>Raise Quality If</td>
<td>Quality of Body of Evidence</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Publication bias:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Likely</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very likely</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviation: GRADE=Grading of Recommendations Assessment, Development and Evaluation

**Peer Review**

The peer review process is our principal external quality-monitoring device. After incorporation of initial feedback from the ACS and the GDG, we prepared a revised draft for peer review by external reviewers selected by the ACS. After all comments on this second draft report were received, the ACS and GDG consolidated and prioritized the comments by theme and per report sections. The resulting comments list was reviewed by the Duke Investigator team, and a call was held with the ACS and GDG to discuss plans for revising the report. A table detailing responses to all comments from the prioritized list has been submitted to the ACS/GDG along with this final report.

**Results**

In what follows, we begin by describing the results of our literature searches. The remainder of the chapter is organized by Key Question (KQ). Under each KQ, we begin by listing the key points of the findings (including GRADE strength-of-evidence assessments), followed by a brief description of included studies and a detailed synthesis of the evidence.

**Results of Literature Searches**

Figure 2 depicts the flow of articles through the literature search and screening process. Searches of PubMed, CINAHL, and PsycINFO yielded 10,200 unique citations. Sixty-six more citations were identified through manual searching/referral from investigators, for a total of 10,266 citations. After applying inclusion/exclusion criteria at the title-and-abstract level, 2197 full-text articles were retrieved and screened. Of these, 2037 were excluded at the full-text screening stage, leaving 160 articles for data abstraction. These 160 articles described 93 unique studies. The relationship of studies to the KQs is as follows: 71 studies relevant to KQ 1, 9 studies relevant to KQ 2, 7 studies relevant to KQ 3, 11 studies relevant to KQ 4, and 1 study relevant to KQ 5 (some studies were relevant to more than one KQ). Further details on the studies included for each KQ are provided in the relevant results sections, below.

Appendix D provides a detailed listing of included articles by KQ. Appendix E provides a complete list of articles excluded at the full-text screening stage, with reasons for exclusion. Appendix F provides a “study key” table listing the primary and companion publications for the 93 included studies. Appendix G summarizes important study characteristics for all included studies. Finally, Appendix H provides GRADE summary tables for the critical outcomes evaluated under each KQ.
Some studies were relevant to more than one KQ.

Abbreviation: KQ=Key Question
Key Question 1

What are the relative benefits, limitations, and harms associated with mammography screening compared to no screening in average-risk women ages 40 and older, and how do they vary by age, screening interval, and prior screening history?

Summary

Key Points: Outcomes

Breast Cancer Mortality:
- **Overall effectiveness:** Screening is consistently associated with a reduction in breast cancer mortality across a range of study designs, from trend studies through RCTs.
- **Precision of effect estimate:** There is considerable variability in the estimates of the magnitude of effect across different study designs, although there is less within a given study design.
- Our assessment of the quality of evidence for a reduction in overall breast cancer mortality with the use of mammographic screening is **HIGH.**
- However, because we are uncertain about the magnitude of the expected mortality reduction in future U.S. populations based on the considerations listed above, the overall quality of evidence for the magnitude of breast cancer mortality reduction with the use of mammographic screening is **MODERATE.**
- The available evidence demonstrates a reduction in mortality with screening of women between the ages of 40 and 49, but the quality of evidence for the magnitude of effect is **MODERATE.**
- There are very limited data on screening effectiveness in women older than 70. On average, women diagnosed with breast cancer after age 75 are more likely to die from other causes than from breast cancer, but modeling studies suggest there may be some older women who may benefit from screening based on life expectancy and co-morbidities. We judge the quality of evidence as **LOW.**

Life Expectancy:
- Life expectancy gains from screening are relatively larger at younger ages, and, at those younger ages, are larger with annual than with biennial screening.
- Because estimates of life expectancy gains from screening are by definition indirect, and there is considerable uncertainty about some of the value of several parameters important for estimating these gains (in particular the magnitude of mortality reduction associated with screening at different ages and different intervals), we judge the quality of evidence for the magnitude of the effect of screening on life expectancy to be **LOW.**

Overdiagnosis:
- Estimates of the proportion of screen detected cancers that are overdiagnosed vary widely, ranging from 0 to 50%.
- The magnitude of the estimate varies depending on the definition of overdiagnosis, the denominator used, the method of analysis, the population studied, whether ductal carcinoma in situ (DCIS) is included, and assumptions about the behavior of DCIS. As
with breast cancer mortality reduction, we judge the quality of evidence for the existence of some overdiagnosis to be **HIGH**; however, given the wide range of estimates, the lack of directness (from observational studies in non-U.S. settings, and from model-based estimates), and the uncertainty about the natural history of DCIS and small localized invasive cancers, we judge the quality of evidence on the magnitude of overdiagnosis to be **LOW**.

**False Positives:**
- As with any imperfect test, screening with mammography results in false positive results, some of which result in invasive procedures such as biopsies.
- False positive results have measurable emotional impact, which may be long-lasting in some women (see discussion under Quality-adjusted Life Expectancy).
- Although the per-screen likelihood of a false positive is lower with shorter screening intervals, the cumulative probability of a false positive result increases with more frequent screening.
- False positive probability is affected by breast density (decreased with mostly fatty tissue, increased with extremely dense tissue), family history (increased), and the availability of prior films (decreased). There is also considerable variability between radiologists and facilities.
- We judge the quality of evidence that false positives results are more common with more frequent screening as **HIGH** based on consistency across study designs and settings. Quality of evidence for estimates of the magnitude of the cumulative false positive rate over 10 years in the U.S. is **MODERATE**; there is much greater uncertainty about lifetime probabilities, with evidence quality limited to modeling extrapolations, for overall **LOW** quality evidence.

**Quality-adjusted Life Expectancy:**
- The utility measures used for estimating quality-adjusted life expectancy in U.S. model-based studies are limited by either derivation from non-U.S. populations, who may have quite different preferences, or by lack of any patient or general population-based estimate. In addition, assumptions about the duration of the impact of relevant states are not empirically supported.
- Despite these limitations, common events that have small and short effects on utilities (screening visits themselves, false positive results) consistently have a substantial effect on overall quality-adjusted life expectancy at the population level, which decreases with frequency of screening and the probability of false positive results; the magnitude of this decrease is effected by the magnitude of the disutility.
- Quality-adjusted life expectancy is decreased by overdiagnosis, which is intuitive. Since overdiagnosed cancers would, by definition, not lead to a breast cancer death, patients experience the disutility of diagnosis and treatment with no gain in life expectancy. The impact of overdiagnosis on quality-adjusted life expectancy is dependent not only on the estimate of the rate of overdiagnosis, but also the magnitude and duration of the disutility of treatment of overdiagnosed cancers (including DCIS), the age at which the diagnosis occurs, and, critically, the ratio of overdiagnoses to cancer deaths prevented: if this ratio is substantially above 1.0 and the diagnoses occur at a substantially younger age than the prevented deaths, then it is possible that some screening strategies might result in a net
decrease in quality-adjusted life expectancy compared to no screening. Identifying this threshold ratio should be an important priority for future modeling studies.

- Although the qualitative effects of these parameters on quality-adjusted life expectancy are plausible and consistent, we judge the quality of evidence for the effect of screening on quality-adjusted life expectancy to be LOW, based on the inherent uncertainties in the underlying estimation of life expectancy, the critical uncertainty about the rate of overdiagnosis, and the limitations of the available utility weights.

**Key Points: Balance of Benefits and Harms**

- Estimates of total false positives per breast cancer death prevented from various sources range from approximately 150 to 1500, depending on estimates of mortality reduction, test specificity, age, screening interval, and whether total false positives for the population versus false positives per patient are used as the denominator. Evidence on patient preferences is limited and of LOW quality.

- Estimates of overdiagnosis per breast cancer death prevented are also dependent on mortality reduction and age, but are even more affected by uncertainty about the proportion of cancers that are overdiagnosed. Given that the U.S. has higher rates of DCIS diagnosis than other countries with breast cancer screening, uncertainty about the natural history of DCIS is a major contributor to uncertainty about the relative contribution of DCIS to overdiagnosis, and assumptions about the probability of progression of DCIS.

**Description of Included Studies**

**Studies**

We identified four recent systematic reviews/meta-analyses of RCTs. Three of these were specifically performed to inform screening guidelines, one of which (Canadian Task Force) used GRADE for formulating recommendations.

Our independent searching identified 8 RCTs, 2 of which had separate components, for a total of 10 studies. All 10 studies were included or discussed in the above systematic reviews (some reviews also separate the Swedish Two County trials into 2 separate papers, for a total of 11 RCTs). Table 4 briefly summarizes the characteristics of these studies.
Table 4. Summary of RCTs of Mammography (Adapted from UK Independent Panel and Cochrane Reviews)

<table>
<thead>
<tr>
<th>PICOTS Element</th>
<th>HIP***</th>
<th>Malmo I</th>
<th>Malmo II</th>
<th>Swedish Two-County†‡</th>
<th>Edinburgh**</th>
<th>Canada I§,25</th>
<th>Canada II§,26</th>
<th>Stockholm***</th>
<th>Goteborg*</th>
<th>UK Age****</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total N</td>
<td>62,000</td>
<td>42,482</td>
<td>6,780</td>
<td>133,065</td>
<td>54,656</td>
<td>10,105</td>
<td>9,765</td>
<td>18,939</td>
<td>6,932</td>
<td>Individual</td>
</tr>
<tr>
<td>Age</td>
<td>40-64</td>
<td>45-69</td>
<td>43-49</td>
<td>38-75</td>
<td>45-64</td>
<td>40-49</td>
<td>50-59</td>
<td>39-65</td>
<td>39-59</td>
<td>39-41</td>
</tr>
<tr>
<td>Attendance Rate†</td>
<td>65%</td>
<td>74%</td>
<td>NR</td>
<td>85%</td>
<td>65%</td>
<td>88%</td>
<td>88%</td>
<td>82%</td>
<td>85%</td>
<td>81%</td>
</tr>
<tr>
<td>Interventions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interval</td>
<td>12 mo</td>
<td>18-24 mo</td>
<td>18-24 mo</td>
<td>24-33 mo</td>
<td>24 mo</td>
<td>12 mo</td>
<td>12 mo</td>
<td>24-28 mo</td>
<td>18 mo</td>
<td>12 mo</td>
</tr>
<tr>
<td>N Screening Rounds</td>
<td>4</td>
<td>6-8</td>
<td>608</td>
<td>2-4</td>
<td>4-5</td>
<td>4-5</td>
<td>405</td>
<td>2</td>
<td>4-5</td>
<td>8-10</td>
</tr>
<tr>
<td>N Viewings</td>
<td>2</td>
<td>2 then 1 or 2</td>
<td>2 then 1 or 2</td>
<td>1</td>
<td>2 then 1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2 then 1</td>
<td>2 then 1</td>
</tr>
<tr>
<td>Other interventions</td>
<td>CBE</td>
<td>–</td>
<td>–</td>
<td>SBE</td>
<td>CBE</td>
<td>SBE after initial CBE</td>
<td>CBE+SBE</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Comparator</td>
<td>No Screening</td>
<td>No Screening</td>
<td>No Screening</td>
<td>No Screening</td>
<td>No Screening</td>
<td>SBE after initial CBE</td>
<td>CBE+SBE</td>
<td>No Screening</td>
<td>No Screening</td>
<td>No Screening</td>
</tr>
<tr>
<td>Timing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration of Screening</td>
<td>5 years</td>
<td>12 years</td>
<td>12 years</td>
<td>7 years</td>
<td>6 years</td>
<td>5 years</td>
<td>5 years</td>
<td>4 years</td>
<td>7 years</td>
<td>8 years</td>
</tr>
<tr>
<td>Setting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>U.S.</td>
<td>Non-U.S.</td>
<td>Non-U.S.</td>
<td>Non-U.S.</td>
<td>Non-U.S.</td>
<td>Non-U.S.</td>
<td>Non-U.S.</td>
<td>Non-U.S.</td>
<td>Non-U.S.</td>
<td>Non-U.S.</td>
</tr>
<tr>
<td>Mortality Relative Risk*</td>
<td>0.83 (95% CI 0.61 to 1.07)</td>
<td>0.81 (95% CI 0.61 to 1.07)</td>
<td>Excluded from meta-analyses (no long term follow-up)</td>
<td>0.58 (95% CI 0.45 to 0.76)</td>
<td>0.76 (95% CI 0.61 to 0.95)</td>
<td>Excluded from meta-analyses (imbalances between groups)</td>
<td>0.97 (95% CI 0.87 to 1.06)</td>
<td>0.73 (95% CI 0.73 to 1.13)</td>
<td>0.97 (95% CI 0.78 to 1.16)</td>
<td>0.75 (95% CI 0.58 to 0.98)</td>
</tr>
</tbody>
</table>

*Estimate used in meta-analysis. †Definition (mean across all screens, cumulative across all screens, first screen, etc) variable across studies.

Abbreviations: CBE=clinical breast exam; HIP=Health Insurance Plan of New York; N=number (of); NR=not reported; PICOTS=Populations, Interventions, Comparisons, Outcomes, Timing of outcomes, and Settings; RCT=randomized controlled trial; SBE=self-breast exam
We identified three systematic reviews of observational studies in European populations. Our independent searching identified 63 observational studies relevant to KQ 1 (13 case-control studies, 49 cohort studies, 1 modeling study). The non-randomized prospective cohort studies leveraged screening programs that were begun by regions, with non-screened regions serving as controls.

One U.S.-based study from the Cancer Intervention and Surveillance Modeling Network (CISNET) was a collaboration between seven independent mathematical modeling groups. After agreeing to a common set of base-case parameters for the models, the groups used population-based data on age, period, and cohort-specific incidence, mortality, screening and treatment patterns, and survival to simulate incidence of and mortality from breast cancer in the U.S. from 1975-2000 under four scenarios: no screening or improved treatment, screening only, improved treatment only, and both screening and improved treatment. By comparing the estimated incidence and mortality under each scenario to the observed incidence and mortality, the investigators were able to estimate the relative contribution of screening and improved treatment on the observed decline in breast cancer mortality over this period. As part of this exercise, each group generated an estimate of breast cancer mortality reduction attributable to screening.

The CISNET collaborators also used these models to generate estimates of breast cancer mortality, life expectancy, quality-adjusted life expectancy, and false positive tests under different scenarios of age of starting screening, stopping screening, and screening interval; these estimates were used to support the 2009 update of the U.S. Preventive Services Task Force (USPSTF) screening recommendations.

Population

These studies included women from as low as 39 years of age to as high as 79 years of age, but mostly included 50-69, with several studies aimed specifically at the 40-49 or 45-49 year age group, and one at the 70- to 74-year-old age group. All of these studies focused on screening women at average risk, but varied in the approach to eliminate women who had risk factors from the participant pool; little information was collected or reported about risk factors such as family history of breast cancer, chest irradiation, or known gene mutations.

Of the RCTs, one was performed in the U.S., one in Canada, two in the UK, and four in Sweden. The cohort studies included 7 U.S., 2 Canadian, 2 Japanese, 1 Australian, and 34 European (10 Sweden, 7 Italy, 6 Norway, 5 Denmark, 3 Finland, 1 German, 1 Netherlands, 1 Spain, 1 UK/Sweden, 1 Austria/Finland/Sweden, and 1 Norway/Sweden) studies. The case-control studies included 2 U.S., 2 Australian, and 9 European (5 Netherlands, 2 UK, 1 Iceland, 1 Italy) studies. The one modeling study was from the Netherlands. None reported racial or ethnic characteristics of the study populations, but the geographical distribution suggests that all of the study populations are majority White non-Hispanic.

Interventions

Interventions studied included screen film mammography using either single- or double-views; some used double-view at first screening with single-view at subsequent screens. Many studies employed two readers. Mammography was most often offered as part of an organized screening program rather than opportunistic screening. The screening interval ranged from 1 to 2 years, with most studies striving for 2-year screening intervals. In one study, screening was offered less frequently than planned.
Studies also varied in the use of ancillary techniques for breast cancer detection such as clinical breast exam (CBE) and breast self-examination (BSE).

Outcomes

We identified 43 studies comparing the effect of mammography versus no screening on breast cancer mortality (8 RCTs, 15-22 13 case-control studies, 31-43 and 22 cohort studies 44-65).

We identified 20 studies that estimated overdiagnosis (2 RCTs, 15, 16 17 cohort studies, 45, 47, 48, 66-79 and 1 modeling study 80). Age at mammography screening varied widely among the 20 studies, with biennial screening intervals in the majority. Estimation of the rate of overdiagnosis was made by comparison of breast cancer incidence between screened and unscreened cohorts. Further details on populations, screening interval, and method for estimating incidence in the unscreened population for individual studies are provided in Appendix Table G-1.

Observational studies of overdiagnosis require adjustments for both breast cancer risk differences between screening and control populations and for increased incidence due to lead time in screening cohorts. In most studies, adjustments for breast cancer risk were made for age-, temporal-, and/or geographic-based variations. 47, 66, 68, 71, 73, 74, 77-79, 81, 82 Lead time adjustment methods included observation for a compensatory drop in breast cancer incidence following the end age of a screening program using prolonged follow up of at least 5 years, 47, 66, 68, 73, 74, 81 the inclusion of a prevalence screen at the end of the study period in a non-screened population with observation of its effect on incidence, 16, 71 and the exclusion of years of prevalence screening from screening cohorts. 82

In all we identified 18 studies reporting false positive rates. Sixteen studies looked at subsequent visit repeat examination rates (“recall”), 3 RCTs 17, 18, 21 and 15 observational studies. 49, 83-94 Six observational studies reported false positive biopsies. 48, 49, 83, 87, 92, 95

Most of the studies provided only “base rates” for false positives, 17, 18, 21, 49, 85, 88, 93 three studies analyzed differences in false positive rates by modalities, 84, 89, 95 and two by age. 84, 96 Two studies provided data on the effect of age, screening interval, breast density, first versus subsequent examination, and availability of previous films. 87, 92

Characteristics of the included studies are summarized in Appendix Table G-1. GRADE summary tables for the outcomes described below are provided in Appendix H.

Detailed Synthesis

Breast Cancer Mortality

Effect of Screening on Breast Cancer Mortality across All Ages

Study Results

Systematic Reviews of RCTs

All of the meta-analyses excluded the Edinburgh 20 and Malmo II 18 studies. The Edinburgh trial had substantial differences in baseline socioeconomic characteristics between groups, and the Malmo II study, never fully reported, also had evidence of imbalance between groups. Pooled estimates for breast cancer mortality after 13 years of follow-up were similar for the two meta-analyses using random-effects models (UK Independent Panel, 11 relative risk [RR] 0.80;
95% CI, 0.73 to 0.89; and Canadian Task Force,\(^6\) RR 0.82; 95% CI, 0.74 to 0.94), and for the Cochrane analysis,\(^3\) which used a fixed-effect model (RR 0.81; 95% CI, 0.74 to 0.87). (The USPSTF review\(^4\) did not present results across all ages.) None of the reviews found significant heterogeneity or evidence of publication bias.

**Systematic Reviews of Observational Studies**

Broeders et al.\(^7\) reviewed published studies based on data from European screening programs and synthesized results by study design (Table 5). For ease of reading, we use “relative risk” throughout the report to refer to both a true relative risk/risk ratio (the incidence of an outcome among those exposed divided by the incidence in those unexposed) and to odds ratios (the odds of exposure among those with the outcome of interest divided by the odds of exposure among those without the outcome, in a case-control study), since, in most cases, the odds ratio is a reasonable estimate of the relative risk.

**Table 5. Pooled Estimates of Breast Cancer Mortality Reduction from Screening Based on European Observational Studies\(^7\)**

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Relative Risk for Breast Cancer Mortality (95% CI)</th>
<th>Number of Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trend studies (before and after introduction of screening)</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Cohort studies (incidence-based mortality, screening vs. no screening)</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Invited to screen</td>
<td>0.75 (0.68 to 0.81)</td>
<td></td>
</tr>
<tr>
<td>Accepted screening</td>
<td>0.82 (0.76 to 0.89)</td>
<td></td>
</tr>
<tr>
<td>Case-control studies</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.46 (0.40 to 0.54)</td>
<td></td>
</tr>
<tr>
<td>Adjusted for self-selection</td>
<td>0.52 (0.42 to 0.65)</td>
<td></td>
</tr>
<tr>
<td>Invited</td>
<td>0.69 (0.57 to 0.83)</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviation: CI=confidence interval

Key points from the systematic review include:

- For both incidence-based mortality (cohort) and case-control studies, mortality reductions were greater (RRs lower) when the comparison was between women accepting versus not accepting screening than when the comparison was between women invited versus not invited to screen. RRs for both study designs were lower than the pooled estimates for the RCTs, although confidence intervals overlap when “invited to screen” was the exposure, as in the RCTs. For example, pooled RR estimates for women invited versus not invited to screen were 0.80 (95% CI, 0.73 to 0.89) in the UK Independent Panel meta-analysis of RCTs, 0.75 (0.68 to 0.81) in the meta-analysis of incidence-based mortality studies, and 0.69 (0.57 to 0.83) in the meta-analysis of case-control studies.

- Estimated mortality reductions were greater with case-control studies than with cohort studies.

**Individual Observational Studies**

Table 6 shows results for individual cohort studies, including those published subsequent to the Broeders systematic review,\(^7\) stratified by estimates based on either invitation to screening or attendance at screening. The table also indicates whether the study adjusted for self-selection bias (factors associated with attendance at screening that might also contribute to breast cancer mortality) and the method used for this adjustment.
Table 6. Individual Cohort Study Estimates of Breast Cancer Mortality Reduction

<table>
<thead>
<tr>
<th>Study; Country</th>
<th>Population</th>
<th>Comparator/Study Dates</th>
<th>Breast Cancer Mortality</th>
<th>Method for Adjusting for Selection Bias for Screened vs. Unscreened Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Age</td>
<td>RR (95% CI)</td>
<td></td>
</tr>
<tr>
<td><strong>U.S.-based Study</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schonberg, 2009” U.S.</td>
<td>2011</td>
<td>&gt;80</td>
<td>No screening</td>
<td>1994-2006</td>
</tr>
<tr>
<td><strong>Non-U.S.-based Studies (by Population Age [Lower Bound])</strong></td>
<td></td>
<td></td>
<td>RR (95% CI)</td>
<td></td>
</tr>
<tr>
<td>Tabar, 2001” Sweden</td>
<td>1,939,348 person years</td>
<td>20-69</td>
<td>No screening</td>
<td>1968-1996</td>
</tr>
<tr>
<td>Hellquist, 2011” Sweden</td>
<td>7,261,415 person-years</td>
<td>40-49</td>
<td>No screening</td>
<td>1986-2005</td>
</tr>
<tr>
<td>Jonsson, 2003” Sweden</td>
<td>43,749</td>
<td>40-64</td>
<td>No screening (counties with vs. without Group I, vs. all of Sweden Group II)</td>
<td>1977-1998</td>
</tr>
<tr>
<td>Jonsson, 2007” Sweden</td>
<td>185,000</td>
<td>40-74</td>
<td>No screening</td>
<td>1989-2001</td>
</tr>
<tr>
<td>Hakama, 1997” Finland*</td>
<td>158,755</td>
<td>48-60</td>
<td>No screening</td>
<td>1987-1991</td>
</tr>
<tr>
<td>Jonsson, 2000” Sweden</td>
<td>439,431</td>
<td>&lt;50</td>
<td>No screening</td>
<td>1987-1996</td>
</tr>
<tr>
<td>Kalager, 2010” Norway</td>
<td>462,306</td>
<td>50-69</td>
<td>No screening</td>
<td>1996-2005</td>
</tr>
<tr>
<td>Study; Country</td>
<td>Population</td>
<td>Age</td>
<td>Comparator/Study Dates</td>
<td>Breast Cancer Mortality</td>
</tr>
<tr>
<td>---------------</td>
<td>------------</td>
<td>-----</td>
<td>-----------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Sarkeala, 2008 * Finland*</td>
<td>361,848</td>
<td>50-69</td>
<td>No screening (historical and contemporaneous)</td>
<td>1992-2003</td>
</tr>
<tr>
<td>Paci, 2002 * Italy*</td>
<td>60,000</td>
<td>50-69</td>
<td>No screening</td>
<td>1990-1999</td>
</tr>
<tr>
<td>Olsen, 2005 * Denmark*</td>
<td>NR</td>
<td>50-71</td>
<td>No screening (historical and contemporaneous)</td>
<td>1991-2001</td>
</tr>
<tr>
<td>Puliti, 2012 * Italy</td>
<td>51,096</td>
<td>50-74</td>
<td>No screening (attendance vs. non-attendance)</td>
<td>1991-2008</td>
</tr>
<tr>
<td>Weedon-Fekjaer, 2014 * Norway</td>
<td>15,193,034 person-years</td>
<td>50-79</td>
<td>No invitation</td>
<td>1986-2009</td>
</tr>
<tr>
<td>Parvinen, 2006 * Finland</td>
<td>1,980,026</td>
<td>55-69</td>
<td>Ages 55-59 (Tampere) Ages 55-69 (Turku) No screening (pre-screening, non-screening areas)</td>
<td>1987-2001</td>
</tr>
<tr>
<td>Swedish Organised Service Screening Evaluation Group 2006 * Sweden</td>
<td>1,108,610</td>
<td>&lt;70</td>
<td>No screening (projected based on Poisson regression of pre-screening trends)</td>
<td>2001</td>
</tr>
<tr>
<td>Jonsson, 2003 * Sweden</td>
<td>125,438</td>
<td>70-74</td>
<td>No screening</td>
<td>1986-1998</td>
</tr>
<tr>
<td>Duffy, 2002 * Sweden</td>
<td>7.5 million</td>
<td>NR</td>
<td>Mammography 2 yr</td>
<td>1958-1998</td>
</tr>
</tbody>
</table>

* Included in systematic review. **

Abbreviations: CI=confidence interval; N=number of participants; NR=not reported; RCTs=randomized controlled trials; RR=relative risk; SES=socioeconomic status
Key points from the cohort studies include:

- RR estimates are generally lower (mortality reduction greater) than those observed with the RCTs. The point estimate for the meta-analysis of cohort studies using invitation to screening as the population of interest (0.75) is similar to the lower bound of the 95% CI for the meta-analyses of the RCTs using the same population (women invited to screening) (lower bounds ranged from 0.73 to 0.74).

- The majority of the studies were in the context of organized, rather than opportunistic screening. There are no direct large population-based U.S. studies.

- Mammography technology and standards are closer to current standards than in the RCTs.

- Mortality reductions were consistently greater when the analysis compared screened versus unscreened women rather than women who are invited versus not invited to screen.

- Adjustment for self-selection bias was not consistently performed across all studies.

Table 7 shows results for individual case-control studies, including those published subsequent to the Broeders systematic review, with and without adjustment for self-selection bias.
<table>
<thead>
<tr>
<th>Study; Country</th>
<th>Population</th>
<th>Comparator/Study Dates</th>
<th>Breast Cancer Mortality</th>
<th>Method for Adjusting for Selection Bias for Screened vs. Unscreened Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U.S.-based Study</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norman, 2007** U.S.</td>
<td>4569</td>
<td>40-64</td>
<td>1994-2005</td>
<td>40-49 years: 0.89 (0.65, 1.23) 50-64 years: 0.47 (0.35, 0.63) Premenopausal: 0.74 (0.53, 1.04) Postmenopausal: 0.45 (0.33, 0.62) Conditional logistic regression, with age, race, menopausal status, BMI, family history, education, parity, smoking, alcohol, oral contraception, hormone replacement, income</td>
</tr>
<tr>
<td>Elmore, 2005** U.S.</td>
<td>3852</td>
<td>40-65</td>
<td>1983-1998</td>
<td>Average risk: 40-65 years: 0.86 (0.71, 1.04) 40-49 years: 0.80 (0.62, 1.01) 50-65 years: 1.02 (0.74, 1.39) High risk: 40-65 years: 1.05 (0.80, 1.39) 40-49 years: 1.03 (0.69, 1.52) 50-65 years: 1.13 (0.70, 1.69) Logistic regression, adjusted for race, comorbidity, age at first birth</td>
</tr>
<tr>
<td>Study; Country</td>
<td>Population</td>
<td>Comparator/Study Dates</td>
<td>Non-U.S.-based Studies (in Ascending Order by Population Age [Lower Bound])</td>
<td>Breast Cancer Mortality</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------</td>
<td>------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Age</td>
<td>RR (95% CI) Unadjusted</td>
<td>RR (95% CI) Adjusted for Screening Bias</td>
</tr>
<tr>
<td>Van Schoor, 2010</td>
<td>1632</td>
<td>40-69</td>
<td>No Screening 1975-1990</td>
<td>RR (95% CI) Adjusted for Screening Bias</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>40-49: 0.50 (0.30, 0.82) 0.54 (0.35, 0.85) 60-69: 0.65 (0.38, 1.13)</td>
<td></td>
</tr>
<tr>
<td>Broeders, 2002 **</td>
<td>930</td>
<td>40-79+</td>
<td>No screening 1975-1997</td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td></td>
<td></td>
<td>40-49: 0.84 (0.30, 2.29) 0.65 (0.30, 1.42) 60-69: 0.63 (0.31, 1.28) 70-79: 0.70 (0.32, 1.54) 79 and older: 1.11 (0.19, 6.39)</td>
<td></td>
</tr>
<tr>
<td>Gabe, 2007 ***</td>
<td>1128</td>
<td>43-83</td>
<td>No screening 1987-2002</td>
<td>0.59 (0.39, 0.84) 0.65 (0.39, 1.09)</td>
</tr>
<tr>
<td>Iceland</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roder, 2008 **</td>
<td>1964</td>
<td>45-80</td>
<td>Any screening prior to death, timing of screening relative to death, frequency of screening, No screening 1994-2005</td>
<td>All women: 0.59 (0.47, 0.74) Age at diagnosis: &lt;50 years: 0.53 (0.40, 0.70) 50-69 years: 0.43 (0.25, 0.73) ≥70 years: 0.41 (0.40, 0.65)</td>
</tr>
<tr>
<td>Australia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study; Country</td>
<td>Population</td>
<td>Comparator/Study Dates</td>
<td>Breast Cancer Mortality</td>
<td>Method for Adjusting for Selection Bias for Screened vs. Unscreened Analysis</td>
</tr>
<tr>
<td>---------------</td>
<td>------------</td>
<td>------------------------</td>
<td>-------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Nickson, 2012** Australia</td>
<td>4077</td>
<td>50-69</td>
<td>Mammography (Controls) 2 years 1995-2006</td>
<td>Unadjusted RR (95% CI)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adjusted for Screening Bias</td>
</tr>
<tr>
<td>van Schoor, 2011*** Netherlands</td>
<td>1410</td>
<td>50-69</td>
<td>No screening 1975-2008</td>
<td>Unadjusted RR (95% CI)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adjusted for Screening Bias</td>
</tr>
<tr>
<td>Allgood, 2008****</td>
<td>852</td>
<td>50-70</td>
<td>No screening 1995-NR</td>
<td>Unadjusted RR (95% CI)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adjusted for Screening Bias</td>
</tr>
<tr>
<td>Puliti, 2008***** Italy</td>
<td>8750</td>
<td>50-74</td>
<td>No screening 1988-2002</td>
<td>Unadjusted RR (95% CI)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adjusted for Screening Bias</td>
</tr>
<tr>
<td>Fielder, 2004**** UK</td>
<td>1136</td>
<td>50-74</td>
<td>No screening 1991-2001</td>
<td>Unadjusted RR (95% CI)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adjusted for Screening Bias</td>
</tr>
<tr>
<td>Otto, 2012**** Netherlands</td>
<td>4494</td>
<td>50-75</td>
<td>No screening 1990-2003</td>
<td>Unadjusted RR (95% CI)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adjusted for Screening Bias</td>
</tr>
<tr>
<td>Paap, 2010**** Netherlands</td>
<td>236</td>
<td>50-75</td>
<td>No screening 1995-2005</td>
<td>Unadjusted RR (95% CI)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adjusted for Screening Bias</td>
</tr>
</tbody>
</table>

* Included in Broeders systematic review.

Abbreviations: BMI=body mass index; CI=confidence interval; N=number of participants; RR=relative risk; SES=socioeconomic status
Key points from the case-control studies include:

- As noted above, estimates for mortality reduction were lowest for this study design. The point estimate in the Broeders systematic review for case-control studies adjusted for self-selection bias was lower (RR 0.52; 95% CI, 0.42 to 0.65) than the estimate from cohort studies of screening versus no screening (RR 0.62; 95% CI, 0.56 to 0.69), although there was considerable similarity in terms of study time, populations, screening methodology, etc.

- For the most part, estimates were higher after adjustment for self-selection bias, although two studies from the Netherlands reported lower estimates AFTER adjustment (implying a lower risk of breast cancer mortality among women invited but not attending screening compared to uninvited women).

- Again, the majority of studies are from non-U.S. settings, with inconsistent results across the two U.S. studies.

**Model-based Estimates**

The median estimated reduction in observed breast cancer mortality attributable to screening in the U.S. from 1975-2000 for all 7 models was 15% (equivalent to a RR of 0.85), with a range from 7% (RR 0.93) to 23% (RR 0.77).

**Effects of Study Characteristics on Estimates**

The majority of reviewed studies, both individually and in meta-analyses, found a reduction in breast cancer mortality associated with mammography screening—the differences lie primarily in the magnitude of the reduction. In this section, we discuss factors that may contribute to these differences.

Figure 3 depicts summary estimates with either 95% CIs or ranges for systematic reviews of European observational studies, RCTs in total (similar across all reviews) and simulation model-derived estimates from the U.S.
Figure 3. Estimated Relative Reduction (with 95% CI or Range) in Breast Cancer Mortality Associated with Mammography Screening Compared to No Screening, by Study Design among Pooled Studies

Figure 3 Key:
- Trend=European studies of mortality before and after introduction of screening, range\(^7\)
- CC-screened=European case-control studies, exposure=screened, mean (95% CI)\(^7\)
- CC-screened/adjusted=European case-control studies, exposure=screened, adjusted for potential bias, mean (95% CI)\(^7\)
- CC-invited=European case-control studies, exposure=invitation to screening\(^7\)
- CISNET=Model-based estimate of relative reduction in mortality attributable to screening in U.S., median (range)\(^29\)
- Incidence-screened=European incidence-based mortality cohorts, exposure=screened, mean (95% CI)\(^7\)
- Incidence-invited=European incidence-based mortality cohorts, exposure=invitation to screening, mean (95% CI)\(^7\)
- RCT--All=All RCTs included in meta-analyses\(^5\)

__________________
Key points when comparing estimates across study designs in Figure 3 include:

- Estimated breast cancer mortality reduction increased in parallel with the inherent risk of bias in study design—reduction estimates were lowest in RCTs, then increased as risk of bias increased from cohort studies to case-control studies.

- Considerations for RCTs:
  - Within a given study design, mortality reduction was greater when the exposure was defined as screening attendance, rather than invitation to screening—this may contribute to some of the difference between the observational studies and the RCTs, as the latter primarily used invitation to screening as the intervention. Studies using invitation to screening as the intervention of interest provide evidence for the efficacy or effectiveness of a screening program, which inherently incorporates both the “technical” aspects of screening (sensitivity and specificity, appropriate follow-up and treatment), as well as the effectiveness of the screening program itself in getting women to accept invitations. In the setting of the U.S., where the lack of a formal screening program means that the potential effectiveness of screening is based on whether an individual woman attends screening, this implies that the estimates from the RCTs may underestimate mortality reduction among those women who actually attend screening in the U.S.
  - Mammographic technology and methods differed across the RCTs, and are substantially different from current practice. To the extent that current practice is more sensitive, this means that the RCTs underestimate the potential mortality reduction from screening relative to current practice.
  - On the other hand, the RCTs were largely performed in an era when treatment for more advanced invasive breast cancer was less effective. If treatment of more advanced disease was less effective, screening and detection of earlier stage disease should lead to a greater survival benefit, and therefore the RCTs may overestimate the potential mortality reduction relative to current practice. Even if the relative estimate is still relevant, the absolute estimate will be smaller if the difference in survival between screen-detected and non-screen-detected tumors is smaller than in previous eras.
  - Sample sizes are generally smaller in the RCTs. For populations or subgroups where mortality is lower in the short term (especially younger women), the number of deaths observed during follow-up may not be sufficient to demonstrate reduction in mortality at traditional levels of statistical significance. Cohort studies with sufficient follow-up would have greater power to detect both short- and longer term differences in breast cancer death, assuming adequate control of potential confounding. The potential ability of case-control designs to address this issue is partly dependent on whether a smaller, potentially non-significant reduction in mortality in younger women is due to inadequate power to detect relatively uncommon short-term deaths or lack of sufficient follow-up to detect deaths prevented further in the future—the definition of “exposure” with regards to timing of screening relative to breast cancer death is critical here.

- Considerations for Observational Studies—Trend Studies:
  - None of the direct trend studies adjusted for secular trends in treatment effectiveness. Estimates of the relative contribution of screening and improved treatment vary. For example, significant improvements in breast cancer mortality occurred in both
screened and unscreened age groups after introduction of the Norwegian screening program, attributed to broad-based efforts to coordinate diagnostic and treatment services for breast cancer patients, with an estimate that approximately a third of the mortality reduction was due to screening. In the U.S, the CISNET modelers estimated that the contribution of screening to the observed reduction in U.S. mortality in the years 1975 to 2000 was approximately equivalent to the contribution of more effective treatments, although with a wide range of estimates (median contribution of screening 46%, range 23% to 65%, with 6 of 7 models at 53% or lower). Although reassuring in terms of consistency with other study designs, the difficulty of disaggregating the effects of screening and treatment limits the utility of trend studies for estimating the magnitude of a mortality reduction attributable to screening.

- Other issues with trend studies are the inability to directly measure exposure to screening, the inability to distinguish deaths occurring after the introduction of screening that were attributable to cancers diagnosed prior to the introduction of screening (a problem analogous to the use of crude age-specific mortality, as discussed in more detail below), and variation in the length of observation after the introduction of screening.

- Considerations for Observational Studies—Cohorts and Case-Control Studies:
  - The European observational studies represent more contemporary screening and treatment practices compared to the RCTs, so, in terms of test performance and treatment outcomes, their results may be more applicable to the U.S. Given opportunistic screening in the U.S., analyses based on attendance at screening rather than invitation to screening may be more appropriate in terms of estimation of the relative impact of screening on mortality among U.S. women who undergo screening, although estimates of the absolute impact on number of deaths prevented are not directly applicable to the U.S.
  - Many of the observational studies using attendance at screening adjusted for potential selection bias using a method described by Duffy et al. The observed relative risk (RR) or odds ratio (OR) in mortality between those attending screening compared to those not attending screening is adjusted based on the observed RR of death in women not attending screening compared to women who were not invited to screening (because of participation in a randomized trial, or temporal or geographic variation in implementation of organized screening). This method is relatively simple to implement, providing an estimate of the relevant RR parameter is available, and has the potential to address a wide range of confounders, some of which may not be observed or measurable (as suggested by one of the examples in the paper, which found a greater effect using the RR method than one which adjusted for specific potential confounders).
    - As the authors note, a key assumption of this method is that “…the relative mortality of non-compliers compared with a population not invited for screening is the same in the programme in question as was observed in the previously published randomized trials.” For the most part, adjusting for selection bias using this method resulted in slightly lower estimates of mortality reduction, although, as noted above, in two studies mortality reduction was greater after adjustment for selection bias.
The use of alternatives to RCTs to evaluate comparative effectiveness is currently the subject of much methodological research, particularly to help resolve questions where the feasibility of further RCTs is limited (as it certainly is with breast cancer screening), where there are questions about the generalizability of RCT results, or where there may be potentially important heterogeneity in treatment effects between subgroups of patients. In particular, because of the significant potential for selection bias on the part of both patients and providers in choosing specific interventions, the use of methods such as propensity scores (which can be useful when all confounders are known and measured) and instrumental variables (which can potentially account for unmeasured confounding) is growing. However, there is inconsistency in the degree to which the use of these newer methods produces results that agree with the results of RCTs: a recent review noted that there was substantial inconsistency in the degree to which observational studies using propensity scores agreed with RCTs, and “[e]ven more concerning than the overall lack of agreement across designs is the absence of a clear pattern that could be used to predict the level of agreement in specific cases….

Without a better method of predicting observational study reliability, numerous well-known discordant examples—such as stem cell transplantation for breast cancer or hormone therapy for coronary heart disease prevention—can be mentioned to discredit all observational analyses, even in situations where concordance is highly probable.”

For the most part, adjustment for self-selection bias results in lower mortality reduction in both case-control and cohort studies, but the point estimate is still higher in cohort studies than in case-control studies. Given that the risk of bias is generally considered lower with cohort designs, estimates based on case-control studies will inherently have a higher degree of uncertainty.

Thus, although it is highly plausible that (a) relative mortality reduction among women attending screening compared to women not attending screening is greater than the reduction observed when the comparison is based on invitation to screening, and therefore the RR is lower than the point estimate RRs from the RCTs, (b) improvements in screening technology may also improve mortality reduction relative to the reductions observed in the RCTs, and (c) the most commonly used method for adjusting for selection bias in observational studies provides less biased estimates when appropriate parameters are available, there is still some uncertainty about whether estimates based on observational studies, particularly case-control studies, are free enough of bias to serve as the primary estimate of relative mortality reduction.

As noted above, fewer barriers to appropriate care after an abnormal screening result in European studies may lead to an overestimation of the potential mortality reduction from screening in the U.S. We discuss the potential impact of differences in the post-screening process (including time to diagnosis, receipt of therapy, and adherence to therapy) in more detail below in considering the directness of European evidence to estimates of the potential mortality reduction from screening in the U.S.
Considerations for Computational Models:

- The CISNET results for effectiveness are derived from models based on estimates of test sensitivity and specificity, attendance at screening after the introduction of mammography in the U.S., and stage-specific survival after detection, as well as estimates and assumptions about underlying disease natural history. Although there is no formal screening program in the U.S., these results are analogous to an exposure based on “invitation to screen.” The median RR estimate (0.85) is similar but slightly higher than estimates from European cohort studies or RCTs based on invitation to screen (0.80 to 0.82). There is also a wide range in estimates between models (0.77 to 0.93). Some of this likely reflects inherent differences between models; there may also be differences in post-screening behaviors and access to care, with barriers to receiving appropriate treatment after a screen-detected abnormality a significant issue in the U.S. compared to most European countries contributing to some of the differences.

Estimated Absolute Effects of Screening on Breast Cancer Mortality in the U.S.

Using SEER age-specific incidence-based breast cancer mortality for cases diagnosed between 1992-2010, we estimated the absolute reduction in breast cancer deaths over a 15-year time period using a range of estimated relative reductions from the literature, from 0.60 (the approximate point estimate for the European cohort studies) to 0.90 (a point slightly higher than the upper 95% confidence bound for the RCT meta-analyses and slightly below the upper bound of the CISNET model-based estimate for the U.S.). Total 15-year incidence-based breast cancer mortality was calculated separately for ages 40-49, 50-59, 60-69, and 70-84. As described in Appendix C, estimates of breast cancer-specific mortality were obtained from SEER for single-year age groups for 15 years after diagnosis, based on age at diagnosis. For example, for women at age 40 at diagnosis, estimates were obtained for the proportion dying within 1 year of diagnosis, 2 years, and so on, up to 15 years after diagnosis; similar estimates were obtained for 41-year-olds, 42-year-olds, etc. Estimates for earlier years post-diagnosis will be more precise, because there will be more women. In addition, the mix of treatments received will be more variable—women in the first few years after diagnosis represent the full range of treatments used between 1992 and 2010, while experience in later years post-diagnosis will be over-represented by the treatments used during the earlier part of that time span (e.g., mortality up to 10 years post-diagnosis includes women whose initial diagnosis was made between 1992 and 1999, while mortality for the first 5 years after diagnosis includes women who were diagnosed between 1992 and 2005).

The overall mortality for the given 10-year age groups was derived by multiplying the estimated single-year age 15-year mortality by the proportion of women in each 1-year age interval within each age group based on 2010 U.S. Census estimates. Particularly for older age groups, this results in an average mortality for the age group that is slightly “skewed” towards women at the younger end of the age range. For example, of all women 60-69 years old, 11.7% are 60 years old, while 7.6% are 69. Because 15-year incidence-based mortality declines from 488 per 100,000 for 60-year-olds to 570 per 100,000 for 69-year-olds, this results in a slightly lower cumulative mortality for the age group than if all ages were equally represented within the age group (unweighted cumulative mortality for the age group of 425 per 100,000 versus age-adjusted cumulative mortality of 422 per 100,000).

To estimate the potential reduction in mortality attributable to screening, we defined “screened” as having received a mammogram within the past 2 years, as reported in the National
Health Interview Survey (NHIS), compared to “no screening,” which included women who had been screened at a longer interval, as well as women who had never been screened. This may overestimate the number needed to screen compared to no screening relative to annual screening, since the number of deaths prevented will be greater with annual screening (although the absolute number of deaths prevented with annual screening compared to biennial screening will be smaller, and thus the NNS higher, when comparing the two screening intervals to each other rather than each to no screening). Although screening recommendations vary between groups for some ages, all recommendations are for screening annually or biennially every 2 years between ages 50 and 70, with many organizations recommending annual or biennial screening beginning at age 40—thus, this definition is somewhat analogous to an “accepted invitation” definition in the setting of a formal screening program.

For these estimates, we used 65% as the estimate of “screening” prevalence—although there is some variation across age groups (with rates up to 75% for women ages 50-64), rates across all age groups have consistently been reported as approximately 65% in the NHIS since 1995 (Tables 8-10). The NHIS estimates were also used by the CISNET investigators. These results are similar to other population-based estimates; for example, reported rates of mammography within the previous 2 years among respondents to the 2005 Medical Expenditure Panel Survey (MEPS) were 66.9% for 40- to 49-year-olds, 75.9% among 50- to 64-year-olds, and 67.6% among those 65 and older, with 10% of all respondents 40 and older reporting never having had a mammogram.

<table>
<thead>
<tr>
<th>Relative Reduction</th>
<th>15-year Cumulative Deaths per 100,000</th>
<th>NNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screened</td>
<td>Unscreened</td>
<td>Absolute Difference</td>
</tr>
<tr>
<td>40%</td>
<td>199.2</td>
<td>332.0</td>
</tr>
<tr>
<td>35%</td>
<td>206.7</td>
<td>318.1</td>
</tr>
<tr>
<td>30%</td>
<td>213.6</td>
<td>305.2</td>
</tr>
<tr>
<td>25%</td>
<td>220.0</td>
<td>293.4</td>
</tr>
<tr>
<td>20%</td>
<td>225.9</td>
<td>282.4</td>
</tr>
<tr>
<td>15%</td>
<td>231.4</td>
<td>272.2</td>
</tr>
<tr>
<td>10%</td>
<td>236.5</td>
<td>262.8</td>
</tr>
</tbody>
</table>

Abbreviations: NNS=number needed to screen; SEER=Surveillance, Epidemiology, and End Results

<table>
<thead>
<tr>
<th>Relative Reduction</th>
<th>15-year Cumulative Deaths per 100,000</th>
<th>NNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screened</td>
<td>Unscreened</td>
<td>Absolute Difference</td>
</tr>
<tr>
<td>40%</td>
<td>324.6</td>
<td>541.0</td>
</tr>
<tr>
<td>35%</td>
<td>336.9</td>
<td>518.2</td>
</tr>
<tr>
<td>30%</td>
<td>348.1</td>
<td>497.3</td>
</tr>
<tr>
<td>25%</td>
<td>358.5</td>
<td>478.0</td>
</tr>
<tr>
<td>20%</td>
<td>368.1</td>
<td>460.2</td>
</tr>
<tr>
<td>15%</td>
<td>377.1</td>
<td>443.6</td>
</tr>
<tr>
<td>10%</td>
<td>385.4</td>
<td>428.2</td>
</tr>
</tbody>
</table>

Abbreviations: NNS=number needed to screen; SEER=Surveillance, Epidemiology, and End Results
### Table 10. Estimated 15-year Cumulative Breast Cancer Mortality among Screened and Unscreened Women Aged 60-69 Years Based on SEER Incidence-based Mortality, 1992-2010, Assuming 65% Prevalence of at Least Biennial Screening, by Relative Mortality Reduction

<table>
<thead>
<tr>
<th>Relative Reduction</th>
<th>15-year Cumulative Deaths per 100,000</th>
<th>NNS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Screened</td>
<td>Unscreened</td>
</tr>
<tr>
<td>40%</td>
<td>422.2</td>
<td>703.6</td>
</tr>
<tr>
<td>35%</td>
<td>336.9</td>
<td>518.2</td>
</tr>
<tr>
<td>30%</td>
<td>452.8</td>
<td>646.8</td>
</tr>
<tr>
<td>25%</td>
<td>466.3</td>
<td>621.7</td>
</tr>
<tr>
<td>20%</td>
<td>478.8</td>
<td>598.5</td>
</tr>
<tr>
<td>15%</td>
<td>490.4</td>
<td>576.9</td>
</tr>
<tr>
<td>10%</td>
<td>501.2</td>
<td>556.9</td>
</tr>
</tbody>
</table>

Abbreviations: NNS=number needed to screen; SEER=Surveillance, Epidemiology, and End Results

Key points to consider with these estimates include:

- Within a given estimate of relative breast cancer mortality reduction, the estimated number needed to screen (NNS) to prevent one breast cancer death approximates the NNS based on RCT estimates. For example, the estimated NNS for women 40- to 49-year-olds in the UK Age Trial for 10 years of follow-up and 7-9 years of screening was 2512 at relative reduction of 17% (RR 0.83) (NNS 2315 when restricted to deaths within 10 years among women with the potential for 10 years of follow-up), which is within the range of these estimates for a reduction of 15% (NNS 2469) and 20% (NNS 1781) with 15 years of screening and follow-up.

- These estimates are quite similar to those generated based on a life table/Markov model using age-specific incidence, age-specific disease-specific survival, and competing risks of death (see Appendix C). This is not surprising, since incidence-based mortality is a function of age-specific incidence and post-diagnosis survival. For simplicity, we assumed the mortality reduction attributable to screening occurred immediately. This will tend to overestimate the magnitude of benefit at 15 years, since mortality reductions in the RCTs were generally not observed until 2-3 years after the start of screening, although this overestimate may be compensated for because mortality reductions occur after the 15-year window, leading to a decrease in the overestimate.

- Absolute effectiveness was much more sensitive to the estimate of relative reduction in mortality than it was to the estimate of the proportion of women who were unscreened or underscreened. For example, holding the proportion screened constant at 65%, increasing the estimate of effectiveness from a 30% reduction (RR 0.7) to a 40% reduction (RR 0.6) in 40- to 49-year-olds increases the absolute difference in breast cancer mortality attributable to screening from 91.6 per 100,000 to 132.8 per 100,000 (decreasing NNS from 1092 to 753). However, if survey respondents over-report their frequency of screening, holding mortality reduction to 40% and changing the estimated proportion of screened women from 65% to 50% results in absolute mortality difference decreasing from 132.8 per 100,000 to 122.8 per 100,000 (increasing NNS from 753 to 814). In other words, over-reporting of screening behavior actually leads to an underestimation of the absolute difference—the estimated absolute difference between screened and unscreened increases as the proportion of unscreened decreases. To illustrate, the estimated cumulative 15-year overall mortality for 40- to 49-year-olds is 246 per 100,000. If this represented the overall mortality in a population where everyone was screened, with a
mortality reduction of 40% (RR 0.6) from screening, then the estimated mortality in unscreened women would be:

\[ Mortality_{Unscreened} = \frac{246}{(0.6 \times 1) + (1 - 1)} = 410 \]

resulting in an absolute difference of 410-246=164 per 100,000, for a NNS of 610.

If the observed mortality represented only unscreened women, then the estimated mortality in screened women is 0.6*246, or 148, for an absolute difference of 98, and a NNS of 1020. In other words, if the NHIS estimates of the proportion of women who are unscreened or underscreened are too low, then the estimates in the table of absolute differences are too high, and estimates of NNS too low.

- These estimates assume that the breast cancer mortality risk for women who are never screened is the same as for women who are screened at least once but at some interval greater than every 1-2 years. If this is not the case, then the absolute mortality reduction will be lower than these estimates.

- Because screening may prevent breast cancers deaths that would otherwise occur later than 15 years from the start of screening, truncating the mortality estimates at 15 years post-diagnosis may underestimate the mortality reduction over a longer time horizon, and thus overestimate the NNS. On the other hand, estimates of the likely experience of women diagnosed with breast cancer in the present or in the near future over the next 15 years are also marked by substantial uncertainty because of potential changes in treatment effectiveness, as well as in competing risks from other cause mortality.

- As noted in the discussion of the CISNET estimates, the absolute decrease in mortality attributable to screening is dependent on the underlying incidence of breast cancer that is not attributable to screening; in other words, changes in breast cancer incidence are a function of both changes in detection (through screening) and changes in the underlying natural history of breast carcinogenesis (because of changes in the prevalence of exposure to specific causes or effect modifiers). In addition to a significant decrease in the use of hormonal replacement therapy, changes in the prevalence (or timing) of other potentially relevant exposures, including age at menarche and menopause, age at first pregnancy, breast feeding, patterns of use of oral contraceptives, and obesity, may all affect the underlying biological development (or timing of development) of breast cancer. Estimates of the likelihood of outcomes 10 or more years in the future after implementation of different screening strategies now are based on current evidence about both breast cancer incidence and treatment effectiveness, which is inherently uncertain.

- Lifetime risk of cancer death from the age at which screening might start is a useful metric for comparing strategies, and estimates of this risk under different screening strategies are necessary for generating estimates of the impact of screening on life expectancy and quality-adjusted life expectancy. However, providing information on the benefits (and harms) of screening over a shorter time horizon is also reasonable (and there is no reason that provision of information about lifetime risk precludes providing information about shorter term outcomes) for a number of reasons:
  - As previously mentioned, there is inherent uncertainty about both the underlying risk of breast cancer and the likely outcomes of treatment the longer the time
horizon becomes; breast cancer incidence, treatment, and outcomes for women above age 60 may well be very different for women now in their 40s compared to present treatment. Explicit acknowledgement that future evidence may change the assessment of the balance of benefits and harms for any given screening recommendation may facilitate acceptance of revised recommendations from patients, clinicians, and other stakeholders.

- Lifetime estimates necessarily rely on model-based extrapolations, which have a moderate to high degree of uncertainty in both underlying assumptions and estimates of key parameters.
- In general, people place a higher value on outcomes that occur in the near future compared to the distant future (temporal discounting), there is individual variation in the degree to which future outcomes are discounted, and these time preferences can affect patient decision-making about health behaviors, including screening.\textsuperscript{102-105} All other things being equal, cancer deaths prevented in the near term are more “valuable” than cancer deaths prevented 30 or more years in the future, especially if the likelihood of harms occurs sooner.

**Effect of Screening on Breast Cancer Mortality at Different Ages**

**Study Results**

**Systematic Reviews of RCTs**

Again, all reviews excluded Edinburgh\textsuperscript{20} and Malmo II.\textsuperscript{18} Table 11 presents results for subgroups by age. Screening in women younger than 50 consistently reduces breast cancer mortality by approximately 15%. Results for women 50 years and older showed a slightly greater relative reduction, with most of this decrease attributable to a larger effect in women 60-69 years old. Data on women 70-74 are limited to the Swedish Two-County trial, with differences in the direction of effect variable based on methods for case classification.

**Table 11. Effect of Mammography on Breast Cancer Mortality by Age in RCTs**

<table>
<thead>
<tr>
<th>Review</th>
<th>RR (95% CI)</th>
<th>Included Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Under 50 years</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USPSTF\textsuperscript{*}</td>
<td>0.85 (0.75 to 0.96)</td>
<td>Malmo I, Canada I, Goteborg, HIP, UK Age, Tw o-County, Stockholm</td>
</tr>
<tr>
<td>Canadian Task Force\textsuperscript{*}</td>
<td>0.85 (0.76 to 0.96)</td>
<td>Malmo I, Canada I, Goteborg, HIP, UK Age, Tw o-County, Stockholm</td>
</tr>
<tr>
<td>Cochrane\textsuperscript{*}</td>
<td>0.84 (0.73 to 0.96)</td>
<td>Malmo I, Canada I, Goteborg, HIP, UK Age, Tw o-County, Stockholm</td>
</tr>
<tr>
<td><strong>50 years and older (all categories)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canadian Task Force\textsuperscript{*}</td>
<td>0.79 (0.68 to 0.90)</td>
<td>Malmo I, Canada II, Goteborg, HIP, Tw o-County, Stockholm</td>
</tr>
<tr>
<td>Cochrane\textsuperscript{*}</td>
<td>0.77 (0.69 to 0.86)</td>
<td>Malmo I, Canada II, Goteborg, HIP, Tw o-County, Stockholm</td>
</tr>
<tr>
<td>USPSTF\textsuperscript{*}</td>
<td>0.86 (0.75 to 0.99)</td>
<td>Canada II, Malmo I, Goteborg, Tw o-County, Stockholm</td>
</tr>
<tr>
<td>Canadian Task Force\textsuperscript{*}</td>
<td>0.82 (0.68 to 0.98)</td>
<td>Canada II, Malmo I, Goteborg, Tw o-County, Stockholm</td>
</tr>
<tr>
<td>UK Independent Panel\textsuperscript{*}</td>
<td>0.80 (0.73,0.89)</td>
<td>Canada II, Malmo I, Goteborg, Tw o-County, Stockholm, HIP</td>
</tr>
<tr>
<td><strong>60-69 years</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USPSTF\textsuperscript{*}</td>
<td>0.68 (0.54 to 0.87)</td>
<td>Malmo I, Goteborg</td>
</tr>
<tr>
<td>Review</td>
<td>RR (95% CI)</td>
<td>Included Studies</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Canadian Task Force*</td>
<td>0.69 (0.57 to 0.83)</td>
<td>Malmo, Goteborg, Two-County, Stockholm, HIP</td>
</tr>
<tr>
<td>USPSTF</td>
<td>1.12 (0.73 to 1.72)</td>
<td>Two-County</td>
</tr>
<tr>
<td>Canadian Task Force*</td>
<td>0.68 (0.45 to 1.00)</td>
<td>Two-County</td>
</tr>
</tbody>
</table>

Abbreviations: CI=confidence interval; HIP=Health Insurance Plan; RR=relative risk; USPSTF= U.S. Preventive Services Task Force

**Observational Studies**

Table 12 presents the results of included observational studies which provided separate estimates for mortality reduction from screening by age group.

**Table 12. Effect of Mammography on Breast Cancer Mortality by Age, Observational Studies**

<table>
<thead>
<tr>
<th>Study; Country</th>
<th>Age</th>
<th>Study Dates</th>
<th>RR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age &lt;50 Years</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cohort Studies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hellquist, 2011**</td>
<td>40-49</td>
<td>1986-2005</td>
<td>0.74 (0.66, 0.83)</td>
</tr>
<tr>
<td>Jonsson, 2007**</td>
<td>40-49</td>
<td>1989-2001</td>
<td>0.62 (0.42, 0.91)</td>
</tr>
<tr>
<td>Jonsson, 2000**</td>
<td>&lt;50</td>
<td>1987-1996</td>
<td>0.91 (0.72, 1.15)</td>
</tr>
<tr>
<td><strong>Case-Control Studies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roder, 2008**</td>
<td>&lt;50</td>
<td>1994-2005</td>
<td></td>
</tr>
<tr>
<td>Norman, 2007**</td>
<td>40-49</td>
<td>1994-1998</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elmore, 2005**</td>
<td>40-49</td>
<td>1983-1988</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Age 50-69 Years</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cohort Studies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puliti, 2012**</td>
<td>50-69</td>
<td>1991-2008</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jonsson, 2007**</td>
<td>50-69</td>
<td>1989-2001</td>
<td>0.80 (0.64, 1.0)</td>
</tr>
<tr>
<td>Parvinen, 2006**</td>
<td>55-69</td>
<td>1987-2001</td>
<td>55-59 years: 0.73 (0.45, 1.19)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>60-64 years: 0.64 (0.36, 1.14)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>65-69 years: 0.53 (0.28, 0.99)</td>
</tr>
<tr>
<td>Study; Country</td>
<td>Age</td>
<td>Study Dates</td>
<td>RR (95% CI)</td>
</tr>
<tr>
<td>---------------</td>
<td>-----</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Invited to Screen (Cohort) or Unadjusted (Case Control)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Case-Control Studies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norman, 2007**</td>
<td>50-64</td>
<td>1994-1998</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elmore, 2005**</td>
<td>50-65</td>
<td>1983-1988</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hoder, 2008***</td>
<td>50-69</td>
<td>1994-2005</td>
<td>–</td>
</tr>
<tr>
<td><strong>Age ≥70 Years</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schonberg, 2009***</td>
<td>&gt;80</td>
<td>1994-2006</td>
<td>–</td>
</tr>
<tr>
<td>Jonsson, 2007**</td>
<td>70-74</td>
<td>1989-2001</td>
<td>0.97 (0.62, 1.52)</td>
</tr>
<tr>
<td><strong>Case-Control Studies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hoder, 2008**</td>
<td>≥70</td>
<td>1994-2005</td>
<td>–</td>
</tr>
</tbody>
</table>

Abbreviations: CI=confidence interval; OR=odds ratio; RR=relative risk

**Key points include:**
- Confidence intervals for mortality reductions in women under 50 included 1.0 in two U.S.-based case-control studies: RR 0.80 (95% CI, 0.62 to 1.01)\(^{41}\) and 0.89 (0.65 to 1.23),\(^{40}\) and in one Swedish cohort study (RR 0.91; 95% CI, 0.72 to 1.15).\(^{63}\) However, reductions were larger and statistically significant in more recent Swedish cohort studies: RR 0.62 (95% CI, 0.42 to 0.91)\(^{53}\) and 0.71 (0.62 to 0.80),\(^{46}\) and an Australian case-control study: RR 0.53 (95% CI, 0.40 to 0.70).\(^{38}\)
- The point estimate for mortality reduction in one U.S. case-control study was lower when stratified by menopausal status (RR 0.74; 95% CI, 0.43 to 1.04) than by age (0.89; 95% CI, 0.65 to 1.23).\(^{40}\)
- Mortality reductions in the 50- to 69-year-old age group were consistent with those in the overall observational studies described above, which reflects the fact that this age group is most commonly targeted in the organized screening programs which provide the bulk of the observational evidence.

**Model-based Estimates**

Because there is limited direct evidence on outcomes in the U.S., we also summarize results from the CISNET collaborative modeling group on estimates for the effect of age to stop and start screening, by interval, on mortality outcomes.\(^{30}\) For these analyses, each modeling group used estimates of mammographic sensitivity and specificity (adjusted for age, first vs. subsequent screens, and screening interval), from data from the Breast Cancer Screening
Consortium [BCSC]). Based on underlying models of the natural history of breast cancer in the absence of screening, screening results in changes in stage distribution based on test sensitivity; the mortality effect of screening is estimated based on differences in stage-specific survival obtained from SEER. Differences between models primarily arise based on different assumptions about natural history, since the estimates used for screening outcomes come from the same sources. The figures illustrate estimated numbers of cancer deaths prevented per 100,000 for the U.S. from an “exemplar” model from the CISNET collaboration for varying age at starting from 40-60, stopping after age 69 (Figure 4) and for varying age at stopping from after 69-84, starting at age 50 (Figure 5), for annual and biennial screening. Estimates for the other models were reported to be similar, although there is substantial variability between models for other reported outcomes. Because of the inherent uncertainty in both the inputs used for the models, as well as differences in model structures, the primary value of these analyses is to identify qualitative trends. Note that extending the age to stop screening results in greater incremental gains in cancer deaths prevented than lowering the age to stop screening. We also estimate NNS for each comparison (no screening vs. screening at specified ages to start and stop) by dividing 100,000 by the estimated number of deaths prevented.

Figure 4. Estimated Cumulative Lifetime Number of Breast Cancer Deaths Prevented by Age to Start Screening (Assuming Screening Ends after Age 69) and Screening Interval

![Figure 4](image-url)
Figure 5. Estimated Cumulative Lifetime Number of Breast Cancer Deaths Prevented by Age to Stop Screening and Screening Interval (Assuming Screening Starts at Age 50)\textsuperscript{30}

Note that extending the age to stop screening results in greater incremental gains in cancer deaths prevented (steeper slope between ages) than lowering the age to stop for both annual and biennial screening.

The CISNET collaborators used an age-period-cohort model to estimate incidence and stage distribution of breast cancer in the absence of screening, “…consider[ing] the effect of age, temporal trends in risk by cohort, and time period. Because we do not have data on future incidence of breast cancer, we extrapolate forward assuming that future age-specific incidence increases as women age, as observed in 2000.”\textsuperscript{30} However, breast cancer incidence declined significantly after 2002 in the U.S. and many other developed countries, a decline at least partially attributable to the decline in use of hormone replacement therapy after the publication of the Women’s Health Initiative (WHI) results.\textsuperscript{106-118} To the extent that model predictions of future age-specific breast cancer incidence (and thus potential mortality in the absence of screening) were informed by the increasing use of hormone replacement therapy prior to the WHI, breast cancer incidence, and consequently mortality, may be overestimated in these versions of the CISNET models, as may be the potential absolute benefits of screening at any given estimate of relative mortality reduction.

An additional analysis using two of the CISNET models estimated the joint effects of age and comorbidity on mortality prevention from screening in the elderly.\textsuperscript{119} Figure 6 illustrates the estimated number of deaths prevented by screening for each age, stratified by comorbidity level (none, mild, moderate, and severe). Not surprisingly, the mortality reduction is affected by competing risks of death, both through age (prevented deaths decrease with increasing age) and the presence of comorbidities which increase the age-specific probability of death from other causes (the distance between the lines at any given age). In addition, there is a joint effect of these two sources of competing risk (the comorbidity-specific lines converge with advancing age).
Estimated Absolute Effects of Screening in the U.S.

Table 13 shows estimates for 15-year cumulative breast cancer mortality by age group, stratified by the estimate of relative reduction used (note that the estimates for each age group and level of mortality reduction are identical to those in Tables 8-10, above, presented to highlight the effect of age rather than mortality reduction). Because SEER collapses estimates for women over age 85, similar estimates are not available for women aged 70-79, or 80-84.

Table 13. Estimated Absolute Effect of Age Group on Breast Cancer Mortality Reduction, by Estimated Relative Reduction Attributable to Screening

<table>
<thead>
<tr>
<th>Relative Reduction</th>
<th>Age</th>
<th>15-year Cumulative Deaths per 100,000</th>
<th>NNS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Screened</td>
<td>Unscreened</td>
</tr>
<tr>
<td>40%</td>
<td>40-49</td>
<td>199.2</td>
<td>332.0</td>
</tr>
<tr>
<td></td>
<td>50-59</td>
<td>324.6</td>
<td>541.0</td>
</tr>
<tr>
<td></td>
<td>60-69</td>
<td>422.2</td>
<td>703.6</td>
</tr>
<tr>
<td>30%</td>
<td>40-49</td>
<td>213.6</td>
<td>305.2</td>
</tr>
<tr>
<td></td>
<td>50-59</td>
<td>348.1</td>
<td>497.3</td>
</tr>
<tr>
<td></td>
<td>60-69</td>
<td>452.8</td>
<td>646.8</td>
</tr>
<tr>
<td>20%</td>
<td>40-49</td>
<td>225.9</td>
<td>282.4</td>
</tr>
<tr>
<td></td>
<td>50-59</td>
<td>368.1</td>
<td>460.2</td>
</tr>
<tr>
<td></td>
<td>60-69</td>
<td>478.8</td>
<td>598.5</td>
</tr>
<tr>
<td>10%</td>
<td>40-49</td>
<td>236.5</td>
<td>262.8</td>
</tr>
<tr>
<td></td>
<td>50-59</td>
<td>385.4</td>
<td>428.2</td>
</tr>
<tr>
<td></td>
<td>60-69</td>
<td>501.2</td>
<td>556.9</td>
</tr>
</tbody>
</table>

Abbreviation: NNS=number needed to screen

Note that these estimates are for the NNS for each 10-year age group, while the estimates based on the CISNET models are lifetime (beginning at age 40); thus, the two sets of estimates are not directly comparable. If the mortality reduction for each age group is added and compared to the CISNET estimates for the 40-69 age interval, estimates for the NNS over the entire 30-
year period are reasonably close at higher levels of mortality reduction, especially given that the CISNET estimates are for lifetime mortality and, because of the limitations of the SEER data, 15-year estimates of mortality for women diagnosed after age 70 are not included in our estimates. For example, the total estimate of number of deaths prevented at a 40% mortality reduction over 15 years for all three age groups in Table 13 is (132.8 + 216.4 +281.4), or 630.6 per 100,000, for a NNS of 159, which is reasonably close to the lifetime estimate of biennial screening from the CISNET model for ages 40-69 of 610 per 100,000,\textsuperscript{30} for a NNS of 164.

Effect of Screening Interval on Breast Cancer Mortality

Study Results

Systematic Reviews of RCTs

Table 14 depicts the results of the Canadian Task Force meta-analysis of mortality reduction, stratified by age and screening interval. In women under 50, only intervals of less than 24 months are associated with a significant reduction in mortality. Note that these are not direct comparisons within a given study population.

<table>
<thead>
<tr>
<th>Age Screening Interval</th>
<th>RR (95% CI)</th>
<th>Included Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 50 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;24-month interval</td>
<td>0.82 (0.72, 0.94)</td>
<td>HIP, Canada I, Malmo, Goteborg, UK Age</td>
</tr>
<tr>
<td>≥24-month interval</td>
<td>1.04 (0.72, 1.50)</td>
<td>Two County, Stockholm</td>
</tr>
<tr>
<td>50-69 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;24-month interval</td>
<td>0.86 (0.75, 0.98)</td>
<td>HIP, Canada II, Malmo, Goteborg</td>
</tr>
<tr>
<td>≥24-month interval</td>
<td>0.67 (0.51, 0.88)</td>
<td>Two County, Stockholm</td>
</tr>
</tbody>
</table>

Abbreviations: CI=confidence interval; HIP=Health Insurance Plan; RR=relative risk

Observational Studies

We did not identify direct evidence of the effect of screening interval on breast cancer mortality reduction in the observational studies.

Model-based Estimates

Figures 4 and 5, above, illustrate the joint effects of screening interval and age to stop and start on mortality reduction from the “exemplar” model from the CISNET analysis for the USPSTF,\textsuperscript{30} by age at starting screening (stopping after 69) and age at stopping (starting at age 40). The estimated effect of increasing screening frequency from biennial to annual (the distance between the two lines in the figures) increases as the age to begin screening is lowered; the effect is somewhat smaller for raising the age to stop screening.

Effect of Prior Screening History on Reduction in Breast Cancer Mortality

We did not identify any studies meeting our criteria that reported on the effect of prior screening history on the effectiveness of mammography in reducing breast cancer mortality.

49
Discussion/Conclusions: Screening and Breast Cancer Mortality

Overall Effect of Screening on Breast Cancer Mortality

- **Direction of effect**: Screening is consistently associated with a reduction in breast cancer mortality across a range of study designs, from trend studies through RCTs.
- **Precision of effect estimate**: There is considerable variability in the estimates of the magnitude of effect across different study designs, although there is less within a given study design. Uncertainty about the point estimate is affected by:
  - Risk of bias: The magnitude of mortality reduction is correlated with the inherent risk of bias in study design and conduct.
  - Directness of evidence: The applicability of the evidence to the current and future U.S. population is diminished by:
    - **Timing**: The majority of the RCT evidence comes from an era when both mammography practice and treatment options for women with breast cancer differed from current U.S. practice. These differences could both underestimate (because of improved screening methods) and overestimate (because of improved outcomes even for women with more advanced cancers) screening effectiveness.
    - **Differences in design of screening programs**: Both the RCTs and most of the relevant observational studies took place within formal screening programs, as opposed to the opportunistic screening of the U.S. Within each study type, mortality reduction was greater when the comparison to “no screening” was women attending screening than it was when the intervention group was women invited to screening. The overall effectiveness of screening is a function of:
      - The ability of the screening method to detect cancer earlier in its natural history among women who are screened
      - The proportion of eligible women who are screened—in other words, the effectiveness of the screening program, or policies to increase screening uptake under opportunistic screening, in creating incentives and removing barriers to screening
      - The proportion of women with abnormal screening results who receive appropriate diagnosis and treatment
      - Settings where there are fewer barriers to screening than the U.S. will result in greater reductions in mortality.
    - **Differences in health systems**: The majority of the highest quality evidence, both RCT and observational, comes from settings where barriers to post-screening diagnosis and treatment are considerably lower than in the U.S. In order to reduce mortality, screening results need to be translated into appropriate diagnostic and therapeutic interventions. If a substantial proportion of women with abnormal screening results do not receive appropriate therapy, then the potential for mortality reduction will not be achieved. Although a large proportion of differences in breast cancer mortality observed between African-American and white women in the U.S. are attributable to differences in access to screening, some of the differences are also attributable to differences in post-screening care,
including time to diagnosis and receipt of treatment, differences in types of treatment received, and adherence to adjuvant treatment regimens.\textsuperscript{120-125}

- **Secular trends in breast cancer incidence, treatment effectiveness, and competing risks of mortality:** Even the most sophisticated model for predicting outcomes of different screening policies is dependent on assumptions about factors that may be influenced by secular trends. In particular, the reduction in breast cancer incidence observed in 2003-2010 associated with decreased use of hormone replacement therapy means that projections based on pre-WHI trends in incidence may overestimate mortality by overestimating incidence. To further increase uncertainty, these changes may affect different breast cancer subtypes differently—hormone replacement therapy may have primarily affected the risk of lobular carcinomas compared to ductal carcinomas.\textsuperscript{126-128} To the extent that changes in the distribution of different types of breast cancer could affect post-detection mortality, estimates of expected mortality with and without screening, or with different screening strategies, would be affected. (In addition, these changes could affect estimates of overdiagnosis that include DCIS, since DCIS is assumed to be a precursor only for invasive cancers with a ductal histology.) Potential differences in overall and type-specific incidence, as well as treatment effectiveness and competing risks, across different geographic regions will also increase uncertainty when using results generated within one population to infer likely outcomes in another.

- **Unmeasured differences in tumor biology:** There is evidence that screen-detected breast cancers may be biologically different from clinically detected cancer, even within a given stage—screen-detected cancers have a better prognosis than non-screen detected cancers, even after adjustment for stage.\textsuperscript{129} There is also evidence that the distribution of different biological types (for example, triple negative breast cancers) varies across racial/ethnic groups.\textsuperscript{130-133} For some of these types, the probability of metastasis, particularly distant metastasis, may not be as well-correlated with tumor size.\textsuperscript{133-135} If mortality reduction from screening differs across breast cancer subtypes, the differences in the distribution of those subtypes across populations could affect the applicability of the relative reduction in mortality to one population from an estimate generated in another (for example, estimates generated from studies with predominantly white subjects might not be applicable to those where black women are a substantially larger proportion of the population). These differences would also effect estimation of the absolute effect on mortality.

  - Our assessment of the quality of evidence for a reduction in overall breast cancer mortality reduction with the use of mammographic screening is **HIGH**.
  - However, because we are uncertain about the magnitude of the expected mortality reduction in future U.S. populations based on the considerations listed above, the overall quality of evidence for the quantitative estimate of breast cancer mortality reduction with the use of mammographic screening is **MODERATE**.
Effect of Age of Starting Screening on Breast Cancer Mortality

- Because breast cancer incidence is lower in younger women, and survival higher, even large studies have limited power to detect differences in mortality, particularly within a short time horizon.
- However, pooled RCT data suggest a mortality reduction of approximately 15% (RR 0.85) in women under 50. Notably, the studies that provide the basis for this estimate are the most recent and closest to current mammography practice.
  - There is some evidence that intermittent screening preferentially detects slower growing cancers.\textsuperscript{136} Cancers occurring in patients at the lower end of a particular cancer’s “typical” age-specific incidence represent more aggressive tumors, and are less amenable to screening. This is supported by evidence which suggests that the proportion of screen-detected breast cancers with biological markers of good prognosis increases with age.\textsuperscript{137}
  - The sensitivity of mammography is reduced in younger women,\textsuperscript{138} largely because of increased breast density.
- Some of the ambiguity about effectiveness in younger women may be the result of heterogeneity in factors affecting tumor biology and/or mammographic sensitivity. In particular, there is significant individual variation in time to menopause—only 30% of U.S. women have undergone menopause by age 50 (with a median age of 52).\textsuperscript{139} Evidence from RCTs suggests that mortality reduction is lower in 50- to 59-year-olds compared to 60- to 69-year-olds, and some of this may be attributable to many women in the 50-54 age group still being pre- or perimenopausal. Therefore, some of the effectiveness of mammography may be dependent not so much on an arbitrary age, but on where a given woman is in the menopausal transition. Later age at menopause may contribute to an increased risk both through decreased mammographic sensitivity and through effects of continued exposure to estrogen and progesterone on tumor biology. Although age is clearly the simplest marker for patients, clinicians, and policy makers to consider, ultimately other strategies might prove more effective; for example, anti-mullerian hormone (AMH), produced by the ovary, is a very sensitive predictor of age at menopause,\textsuperscript{140} and could potentially be evaluated as part of risk-based screening strategies.
- Screening effectiveness in younger women may be more susceptible to screening interval. We discuss this in more detail under KQ 2, below.
- Given higher sensitivity with roughly equivalent specificity of digital mammography compared to plain film in younger women and women with dense breasts,\textsuperscript{141,142} performance for younger women now may be better compared to the data from RCTs, all of which were based on plain film studies.
- The combination of a lower incidence of breast cancer, better survival, and lower relative mortality reduction means that the absolute reduction in breast cancer mortality associated with screening is lower in younger women, particularly women under 50 (or, more likely, premenopausal women), compared to older women.
- As with screening overall, our assessment of the quality of evidence for a mortality reduction with mammographic screening in women under 50 is HIGH, based on low risk of bias and consistency. However, the same issues related to directness for the purposes of U.S. recommendations apply, and there is a fair degree of imprecision, particularly for
estimates of absolute effectiveness, so we reduce the overall quality of evidence to MODERATE.

Effect of Age of Stopping Screening on Breast Cancer Mortality

- There is very limited direct evidence on the effectiveness of screening in reducing breast cancer mortality in women 70 years and older.
- Both incidence of breast cancer and mortality from breast cancer increase with age, and model-based estimates suggest greater reductions in breast cancer mortality from increasing the age of stopping screening than decreasing the age of starting screening (with opposite effects on life expectancy, as discussed below).
- We did not identify any direct evidence meeting our inclusion criteria on the effect of prior screening history on the effectiveness of mammographic screening. For some cancers (notably cervical cancer), a history of negative screening results over a period of time has been used as a criterion for withdrawing women from screening. However, although the strategy is based on direct evidence, the likely biological mechanism behind the evidence is the natural history of cervical cancer—the majority of women are infected with oncogenic human papillomavirus as adolescents or in their 20s, and, if a persistent infection has not progressed to cancer by age 50 or 60, most evidence suggests it is unlikely to do so. Because the biology and natural history of breast cancer are quite different, there may not be a similar phenomenon of, “If you haven’t gotten it by now, you probably won’t get it.”
- The effect of competing risks on breast cancer mortality increases with age. SEER separates post-diagnosis mortality by cancer-specific and other causes—as age at diagnosis increases, the risk of death from other causes increases dramatically (Figure 7). At age 70, a woman newly diagnosed with breast cancer is 1.4 times more likely to die from breast cancer than other causes within the first year after diagnosis, but, by 4 years post-diagnosis, she is more likely to die from other causes. Women 75 years and older are more likely to die from other causes after a breast cancer diagnosis than they are from breast cancer. It is important to note that these estimates are based on cause-specific mortality after breast cancer diagnosis in women whose cancer was detected by screening AND those whose cancer was detected through other means, and includes all stages.
Figure 7. Ratio of Cumulative Probability of Death from Breast Cancer to Death from other Causes by Age and Year Post-Diagnosis, SEER 2002-2010

- We assess the quality of evidence for breast cancer mortality reduction with the use of mammographic screening in women 70 years and older as LOW.

**Life Expectancy**
Life expectancy is defined as the average (mean) survival time at a given age. In theory, life expectancy can be directly calculated if all participants in an RCT or cohort study are followed until death. However, more typically, the effect of screening on life expectancy is indirectly estimated based on modeling, and this is the approach adopted here.

- Total life expectancy is estimated based on the annual probability of death, stratified by, at least, age, and frequently sex and race/ethnicity. The probability of death from the condition of interest is subtracted to obtain an estimate of the annual probability of death from all other causes.
- The effects of different strategies for screening and treatment on the probability of death from breast cancer are then modeled.
- The difference between cumulative life expectancy under assumptions of no screening and different screening strategies is then expressed as life-years gained from the intervention.
- The gains in life expectancy for a given strategy can be compared either to a common baseline of no screening, or to other strategies (incremental life-years gained).

**Effect of Screening on Life Expectancy across All Ages**
Model results for the CISNET collaboration were presented stratified by ages to stop and start screening.

**Effect of Screening on Life Expectancy at Different Ages**
Because life expectancy is highly correlated with age, the estimated effect of screening on life expectancy is highly sensitive to the ages at which the prevented breast cancer deaths would
have occurred. Tables 15 and 16 illustrate CISNET estimates of life expectancy gains in terms of both the level of the overall population (life-years per 100,000 women) and an individual woman (because the life expectancy gains at this level are well less than 1 year per woman, results have been converted to days). Not surprisingly, differences are greater from extending the age to start screening to earlier ages than from extending the age to stop screening to older ages (since younger women have a lower risk of death from other causes and have a greater potential number of years of life saved by preventing a breast cancer death). As noted above, this is the opposite of the effect of age on breast cancer mortality reduction—the estimated number of breast cancer deaths is more affected by extending screening to older ages.

**Table 15. Estimated Gains in Life Expectancy with Biennial and Annual Mammography Screening by Age to Start Screening (Assuming Screening Stops after Age 69)**

<table>
<thead>
<tr>
<th>Age to Start Screening</th>
<th>Life-years Gained per 100,000 Women</th>
<th>Days Gained per Woman</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Compared to No Screening</td>
<td>Compared to 5 Years later Age to Start</td>
</tr>
<tr>
<td><strong>Biennial</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>5200</td>
<td>19.0</td>
</tr>
<tr>
<td>55</td>
<td>8000</td>
<td>29.2</td>
</tr>
<tr>
<td>50</td>
<td>9900</td>
<td>36.1</td>
</tr>
<tr>
<td>45</td>
<td>11,600</td>
<td>42.3</td>
</tr>
<tr>
<td>40</td>
<td>12,000</td>
<td>43.8</td>
</tr>
<tr>
<td><strong>Annual</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>6900</td>
<td>25.2</td>
</tr>
<tr>
<td>55</td>
<td>10,200</td>
<td>37.2</td>
</tr>
<tr>
<td>50</td>
<td>13,200</td>
<td>48.2</td>
</tr>
<tr>
<td>45</td>
<td>15,200</td>
<td>55.5</td>
</tr>
<tr>
<td>40</td>
<td>16,400</td>
<td>59.9</td>
</tr>
</tbody>
</table>

**Table 16. Estimated Gains in Life Expectancy with Biennial and Annual Mammography Screening by Age to Stop Screening (Assuming Screening Starts at Age 50)**

<table>
<thead>
<tr>
<th>Age to Stop Screening</th>
<th>Life-years Gained per 100,000 Women</th>
<th>Days Gained per Woman</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Compared to No Screening</td>
<td>Compared to 5 Years Earlier Age to Stop</td>
</tr>
<tr>
<td><strong>Biennial</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>9900</td>
<td>36.1</td>
</tr>
<tr>
<td>74</td>
<td>12,100</td>
<td>44.2</td>
</tr>
<tr>
<td>79</td>
<td>13,000</td>
<td>47.5</td>
</tr>
<tr>
<td>84</td>
<td>13,800</td>
<td>50.4</td>
</tr>
<tr>
<td><strong>Annual</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>13,200</td>
<td>48.2</td>
</tr>
<tr>
<td>74</td>
<td>15,600</td>
<td>56.9</td>
</tr>
<tr>
<td>79</td>
<td>17,000</td>
<td>62.1</td>
</tr>
<tr>
<td>84</td>
<td>17,800</td>
<td>65.0</td>
</tr>
</tbody>
</table>

As age to start screening decreases, the relative gains in life expectancy are greater at a fixed age to stop than they are when extending the age to stop screening at a fixed age to start screening; this is true for both annual and biennial screening intervals. These results are expected, given the larger potential gains in life expectancy at younger ages.

**Effect of Screening at Different Intervals on Life Expectancy**

Tables 17 and 18 present the same CISNET model estimates, stratified by screening interval within a given age to start and stop screening.
Table 17. Effect of Screening Interval on Gains in Life Expectancy by Age of Starting Screening (Assuming Screening Stops after Age 69)\(^{10}\)

<table>
<thead>
<tr>
<th>Age to Start Screening</th>
<th>Interval</th>
<th>Life-years Gained per 100,000 Women</th>
<th>Days Gained per Woman</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Compared to No Screening</td>
<td>Compared to Biennial</td>
</tr>
<tr>
<td>60</td>
<td>Biennial</td>
<td>52</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>69</td>
<td>17</td>
</tr>
<tr>
<td>55</td>
<td>Biennial</td>
<td>80</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>102</td>
<td>22</td>
</tr>
<tr>
<td>50</td>
<td>Biennial</td>
<td>99</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>132</td>
<td>33</td>
</tr>
<tr>
<td>45</td>
<td>Biennial</td>
<td>115</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>152</td>
<td>36</td>
</tr>
<tr>
<td>40</td>
<td>Biennial</td>
<td>120</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>164</td>
<td>44</td>
</tr>
</tbody>
</table>

The estimated impact of shortening screening interval on life expectancy is greater when younger women are included in the screening group (for example, 16.1 days increase for annual compared to biennial for women screened age 40–69, compared to 12.0 days increase for annual compared to biennial for women screened from age 50–69). In contrast, the relative gains at any given stopping age from 69 through 74 are smaller. Again, this is expected given the differences in life expectancy as women age.

We emphasize that these are point estimates based on only one of the CISNET models; there is undoubtedly uncertainty even within this model. These estimates are also dependent on underlying assumptions about incidence and mortality of breast cancer without screening, as well as on the test characteristics of mammography (although the inputs for mammography performance used by the CISNET groups are derived from U.S. data).

### Discussion/Conclusions: Effect of Screening on Life Expectancy

- Life expectancy is not synonymous with all-cause mortality. Depending on when death occurs, it is possible to have identical all-cause mortality (the probability of death from any cause over a given time point) and large differences in life expectancy. Pooled estimates of all-cause mortality in the RCTs show no effect of screening on all-cause mortality, with relative risks very close to 1.00 (although a reduction in all-cause mortality was observed in the Swedish Two-County Trial.\(^{143}\) This is, in one sense, reassuring, since it makes it unlikely that mammographic screening and follow-up
treatment substantially increase the overall risk of death (for example, from the consequences of chemotherapy) within the follow-up period of the trials compared to no screening. However, there is debate about whether all-cause mortality should be a primary outcome in evaluations of screening effectiveness. Because breast cancer is relatively uncommon compared to other causes of death, even very large trials are unlikely to be sufficiently powered to detect a difference in all-cause mortality.

- There is no direct evidence on the effect of screening on life expectancy. Model-based estimates of gains compared to no screening for U.S. women are in the range of 19-65 days, depending on age and screening interval. These estimates are qualitatively similar to other analyses of the impact of breast cancer screening on life expectancy, and are smaller than estimates for other interventions derived using similar methods. For example, the estimate for 10-year biennial mammography beginning at age 50 was 0.8 months compared to 2-2.5 months for colorectal cancer screening and 3.1-3.2 months for cervical cancer screening in a frequently cited paper from 1998 comparing estimates from the contemporary literature. As a caveat about the dependence of model-based estimates on the quality of the available data, we note that the same paper estimated a life expectancy gain of 13 months for the use of estrogen-only hormonal replacement therapy in women who had had a hysterectomy, based on a model using the available pre-WHI observational data.

- Life expectancy estimates are typically derived by using cross-sectional data on age-specific mortality and survival to project the experience of hypothetical cohorts. Because both the incidence and mortality from competing risks may change within and between birth cohorts through changes in exposures, risk modifiers, or treatment effectiveness, these estimates always have some inherent uncertainty, particularly for longer time horizons. In the case of breast cancer, where incidence may be decreasing in part through reduction in exposure to hormone replacement therapy, this means that gains in life expectancy for future cohorts may be different.

- Life expectancy gains from screening are relatively larger at younger ages, and, at those younger ages, are larger with annual compared to biennial screening. This is the opposite of the effect of age and screening on breast cancer mortality. The magnitude of harm-benefit trade-offs will likely vary depending on whether the measure of benefit is breast cancer deaths prevented or life-years gained.

- Reducing breast cancer mortality should increase life expectancy, and, since we judge the quality of the evidence that screening reduces mortality HIGH, we judge the quality of the evidence that screening will increase life expectancy as HIGH as well, despite the lack of direct evidence. However, because (a) estimates of life expectancy gains from screening are by definition indirect, (b) there is considerable uncertainty about the estimates of several screening-specific parameters important for estimating these gains (in particular the magnitude of mortality reduction associated with screening at different ages and different intervals), and (c) there is considerable uncertainty about the impact of secular trends on key parameters (such as exposure to exogenous hormones, treatment effectiveness, and competing risk mortality), we judge the quality of evidence for the magnitude of the effect of screening on life expectancy in the U.S. to be LOW.
Overdiagnosis/Overtreatment

Overdiagnosis, defined as the diagnosis of cases of breast cancer through screening that would otherwise not have been detected, either because of very slow growth or because of death from other causes prior to the breast cancer becoming symptomatic, is a clear potential consequence of screening, but the optimal methods for defining and estimating the extent of overdiagnosis with a specific screening strategy in a specific population are not at all obvious. Most studies included in our review found evidence of some degree of overdiagnosis, but the results varied widely depending on how overdiagnosis was defined, how the estimate was generated, and the study setting.

The methodological complexities of estimating overdiagnosis have been reviewed in detail by others.\textsuperscript{80,147,148} Because the question of how estimates of the amount of overdiagnosis associated with different screening strategies should be weighed in formulating recommendations about breast cancer screening is perhaps even more controversial than the question of how much screening reduces breast cancer mortality, we briefly review the key methodological issues discussed in these reviews, following the structure of the most recent paper by Etzioni and colleagues.\textsuperscript{147} Specific methodological issues that contribute to the wide range of overdiagnosis estimates include:

Variation in method of measurement across studies: Definitions used by different investigators identified by de Gelder and colleagues\textsuperscript{80} included:

- Relative increase over a lifetime (ages 0-100 years), defined as the difference between excess cases (defined as a model-based estimate of the number of cancers detected with screening and the predicted numbers without screening) and deficit cases in age groups after screening (defined as the difference between predicted numbers of cancers without screening and modeled numbers with screening), divided by the predicted number of cancers in women aged 0-100 without screening.
- Relative increase during and after screening only, where the numerator is the same, but the denominator is the predicted number of cases without screening over the same age range (age to start screening until death).
- Relative increase during screening, where the numerator is the same, but the denominator is the estimated number of cases only until the end of screening.
- Proportion of all diagnosed cancers (screen detected and interval cancers) that are overdiagnosed (same numerator).
- Proportion of all screen-detected cancers that are overdiagnosed (same numerator, denominator is only screen-detected cancers).
- Relative risk of breast cancer for women of screening age versus predicted number in women of same age without screening, possibly adjusted for lead time (excess incidence).
- Relative risk of breast cancer in women of screening age with screening versus predicted number of cancers with screening if no overdiagnosis takes place.

Applying these different definitions to a microsimulation model of the Dutch population, de Gelder reported a 3.2-fold difference, ranging from 2.8% (when estimated over a lifetime) to 8.9% (when estimated as a proportion of all screen-detected cancers). Estimates also varied based on timing of the estimation (lower when the screening program reached “steady-state”) and with longer follow-up after the end of screening (because of lead time).

Variation in methods for estimating incidence in the absence of screening: As with estimates of mortality reduction, choices for control groups include (a) women randomized to no invitation
to screening within the RCTs, (b) concurrent controls from observational studies where screening was introduced in some regions of a country prior to others, (c) estimates based on projecting observed incidence during a time period preceding the introduction of screening (typically based on Poisson regression), or (d) estimates based on models of the underlying natural history of breast cancer in the absence of screening.

**Variation in population-specific natural history in the absence of screening:** There are a number of potential differences in exposures or practices between populations that can affect the incidence of breast cancer without screening. These include differences in factors that may affect the development and rate of progression of breast cancer, such as fertility patterns, use of breast feeding, use of exogenous female hormones, competing risks for mortality, etc. These can also include differences in factors which do not affect the natural history per se, but which can affect the timing at which a given cancer is detected and becomes “incident” — such as differences in access to diagnostic services, or cultural differences in willingness to seek medical attention. The degree to which these other factors are different between the control population, whether historical or concurrent, and the screened population may lead to an over- or underestimation of the predicted incidence in the screened population in the counterfactual scenario of no screening. This will in turn lead to a biased estimate of the degree of overdiagnosis.

**Variation in differences in diagnostic intensity across populations:** Factors such as frequency of screening, thresholds for recommending biopsy, adherence to recommendations for screening and diagnosis (on the part of both patients and providers), and variability in diagnostic criteria (for both screening and diagnostic tests) can affect the estimate of overdiagnosis (for example, as discussed below, there is substantial unexplained variability in the detection of DCIS across screening programs — to the extent that non-progressive DCIS contributes to overdiagnosis, this variability will lead to variability in the estimate of overdiagnosis across populations).

**Variation in methods used to estimating overdiagnosis:** Etzioni and colleagues describe two basic approaches:

- **Excess incidence:** The difference between incidence with screening and incidence without screening. Issues with this approach include:
  - Inclusion of cases during early implementation/dissemination of screening will bias estimates of overdiagnosis upward, since extra cases in the early years will include both cases that would never progress to symptomatic cancer and prevalent asymptomatic progressive cancers detected through screening.
  - Limiting the ages at risk for incident cancer to those eligible for screening will bias estimates of overdiagnosis upward by not capturing the expected “compensatory drop” in incidence resulting from earlier detection of a given progressive tumor.
  - Methods for estimating the incidence of cancer among the screened population in the counterfactual scenario where that specific population had not been screened (note that this is a basic issue with observational research for any outcome, including mortality). Approaches include projections based on trends in observed incidence in the specific population prior to the implementation, or changes in the distribution of known risk factors for breast cancer across time or space. Another issue here is adjustment for lead time, which is dependent on both the accuracy of the estimate of lead time, and the assumption that the population from which the lead time estimate was derived was similar to the population in terms of factors affecting lead time (including age, the distribution of different subtypes of breast cancer).
cancer, prevalence of risk factors, and prevalence of non-biological factors affecting time to diagnosis).

- **Lead time:** This approach uses “modeling techniques to infer the lead time and the corresponding fraction of cases overdiagnosed from the pattern of excess incidence under screening.” Issues include:
  - An estimate of the incidence in the counterfactual unscreened scenario is also required. This may be based on an underlying model of the natural history of breast cancer, or fitting estimates of lead time and overdiagnosis to observed incidence with screening. Again, even if the parameter estimates (including those which are ultimately unmeasurable and can only be imputed, such as rates of biological disease progression in the absence of screening) are accurate for a given population, they may over- or underestimate expected incidence without screening in a different population.
  - Model structure and assumptions are also critical and can affect results. For example, simulated estimates of overdiagnosis in breast cancer screening varied greatly based on assumptions about the proportion of overdiagnosed cases that represent true non-progressive lesions versus those that would be progressive but never become symptomatic because of competing mortality risks. Etzioni and colleagues imputed lead times for early stage invasive breast cancers in the U.S. under the assumption that all in situ cases were overdiagnosed, under different scenarios for the shape of the distribution of mean lead time, and found that lead times consistent with a reported 30% overall overdiagnosis rate estimated by Bleyer and Welch were significantly higher than estimates of lead time derived from the Swedish Two-County Trial. However, the underlying assumption was that all screen-detected cancers would ultimately progress; as the authors noted, if this assumption is incorrect, then the estimates of the lead time distribution will also be incorrect, since non-progressive cancers have an infinite lead time. This can lead to an underestimate of the overdiagnosis probability.

**Overall Estimates of Overdiagnosis**

Given these considerations, and the diversity of approaches reported, we agree with Etzioni and colleagues that “[o]ur examination of variation in study features and methods leads us to wonder whether it is possible to compare and integrate results across published studies of overdiagnosis.” In this section, we expand on the examples discussed in the Etzioni review and summarize qualitative findings.

**RCTs**

Based on reported follow-up in the seven main RCTs, the Cochrane review estimated an increase in incidence in invited versus uninvited women of 29% (95% CI, 23% to 35%).

The UK Independent Panel limited their meta-analysis to the three trials where screening was not offered to the control groups at the immediate end of the trial (Malmo I and the two Canadian trials) in order to avoid the effect of prevalent cases detected when screening was offered to control participants at the end of the study period in the other trials. The analysis generated two estimates, which differed in the denominator used. The first, favored by the Panel for estimating population impact, expressed overdiagnosis as the proportion of all cancers diagnosed over the entire follow-up period for women invited for screening (10.7%; 95% CI, 9.3 to 12.2%). The
second, for estimating individual risk of overdiagnosis, expressed it as the proportion of all cancers diagnosed during the screening period in women invited for screening (19.0%; 95% CI, 15.2 to 22.7%). Individual study estimates were higher for women under 50 (Canada I) than for women 50 and older (Canada II).

**Observational Studies and Longer Term Follow-up of RCTs**

A pooled analysis of 13 studies reporting 16 estimates of overdiagnosis from 7 European countries (the Netherlands, Italy, Norway, Sweden, Denmark, UK, and Spain) found crude estimates ranging from 0 to 54%.\(^9\) After adjustment for breast cancer risk and lead time, estimates were reduced to 1% to 10%.

Findings from relevant systematic reviews of RCTs\(^3,11\) and additional individual studies identified in our review\(^16,45,47,66-80,82,152\) are summarized in Table 19.
<table>
<thead>
<tr>
<th>Study; Country; Dates; Population Age; Screening Interval</th>
<th>Methods</th>
<th>Overdiagnosis Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Definition of Overdiagnosis</td>
<td>Estimated Incidence without Screening</td>
</tr>
<tr>
<td><strong>Systematic Reviews of RCTs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cochrane, 2013³</td>
<td>Excess cases/cases observed without screening</td>
<td>Comparison of incidence in invited vs. uninvited women in all RCTs</td>
</tr>
<tr>
<td>Sweden and Canada</td>
<td>Variable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Excess cases/cases over entire follow-up period in women invited for screening</td>
<td>Comparison of incidence in invited vs. uninvited women in Malmo I, Canada I, Canada II RCTs</td>
</tr>
<tr>
<td></td>
<td>• Excess cases/cases diagnosed during screening period in women invited for screening</td>
<td></td>
</tr>
<tr>
<td>UK Independent Panel, 2013¹</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sweden and Canada</td>
<td>Variable</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study; Country; Dates; Population Age; Screening Interval</td>
<td>Definition of Overdiagnosis</td>
<td>Estimated Incidence without Screening</td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
<td>----------------------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td><strong>U.S.-based Study</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bleyer, 2012 U.S.</td>
<td>Excess cases/total detected cases over time period</td>
<td>Base case assumed incidence unchanged from 1976-1978; separate analyses assumed constant increase in incidence, and constant increase plus highest incidence of late-stage disease</td>
</tr>
<tr>
<td>United States</td>
<td>Excess defined as difference between increase in DCIS and Stage I vs. decrease in Stage II-IV</td>
<td></td>
</tr>
<tr>
<td>1976-2008</td>
<td>40 and older</td>
<td></td>
</tr>
<tr>
<td>Opportunistic (annual-biennial)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Non-U.S.-based Studies (by Country)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morrell, 2010 Australia</td>
<td>Excess cases/expected cases without screening</td>
<td>Modeled from prescreening age-specific incidence (adjusted for changes in prevalence of HRT, obesity, and nulliparity)</td>
</tr>
<tr>
<td>Australia</td>
<td></td>
<td>Lead time estimates of 2.5 and 5 years</td>
</tr>
<tr>
<td>1999-2001</td>
<td>50-69</td>
<td></td>
</tr>
<tr>
<td>Biennial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study; Country; Dates; Population Age; Screening Interval</td>
<td>Definition of Overdiagnosis</td>
<td>Estimated Incidence without Screening</td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
<td>----------------------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>Coldman, 2013&lt;sup&gt;66&lt;/sup&gt; Canada 1970-2009 40-49 (annual) 40-79 (biennial)</td>
<td>Excess cases/expected cases without screening</td>
<td>Excess incidence based on Poisson regression of trends in 1970-1979.  - Observed vs. predicted cumulative incidence in women screened vs. unscreened or stopped screening  - Population observed vs. expected</td>
</tr>
<tr>
<td>Njor, 2013&lt;sup&gt;67&lt;/sup&gt; Denmark 1991-2009 56-69 (Copenhagen) 59-69 (Funen) Biennial</td>
<td>Excess cases/cases expected without screening</td>
<td>3 controls:  - Historical controls for regions with screening,  - National controls from regions without screening, and  - Historical national controls</td>
</tr>
<tr>
<td>Study; Country; Dates; Population Age; Screening Interval</td>
<td>Methods</td>
<td>Overdiagnosis Estimate</td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
<td>---------</td>
<td>------------------------</td>
</tr>
<tr>
<td><strong>Jørgensen, 2009</strong>&lt;sup&gt;73&lt;/sup&gt; Denmark 1991-2003 50-69 Biennial</td>
<td>Cases overdiagnosed/all diagnosed cancers Poisson regression based on time trends, varied implementation of program, pre-screening era geographic variation Excess incidence</td>
<td>Invasive Cancer Only 4.8% both rounds Invasive Cancer + DCIS 33% (no CIs given)</td>
</tr>
<tr>
<td><strong>Olsen, 2006</strong>&lt;sup&gt;73&lt;/sup&gt; Denmark (Copenhagen) 1991-1995 50-69 Biennial</td>
<td>Cases overdiagnosed/ detected cases Multistate model, based on observed screen-detected and interval cancers in the first 2 screening rounds Lead time</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; screen: 7.3% (95% CI, 0.3% to 26.5%) 2&lt;sup&gt;nd&lt;/sup&gt; screen: 0.5% (95% CI, 0.02% to 2.1%)</td>
</tr>
<tr>
<td><strong>Puliti, 2012</strong>&lt;sup&gt;73&lt;/sup&gt; Italy 1991-2009 60-69 Biennial</td>
<td>Excess cases/cases observed in unscreened women (non-attenders) Estimated from non-attenders in screening program Excess incidence</td>
<td>5% (95% CI, -7% to 18%) Increased to 10% by excluding 34 non-attenders with breast cancer diagnosis within 6 months of index date 10% (95% CI, -2% to 23%)</td>
</tr>
<tr>
<td>Study: Country; Date; Population Age; Screening Interval</td>
<td>Methods</td>
<td>Overdiagnosis Estimate</td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
<td>---------</td>
<td>------------------------</td>
</tr>
<tr>
<td><strong>Definition of Overdiagnosis</strong></td>
<td><strong>Estimated Incidence without Screening</strong></td>
<td><strong>Methodological Approach</strong></td>
</tr>
<tr>
<td>Puliti, 2009*</td>
<td>Excess cases/cases expected without screening</td>
<td>Excess incidence</td>
</tr>
<tr>
<td>Italy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990-2005</td>
<td>Biennial</td>
<td></td>
</tr>
<tr>
<td>60-69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paci, 2006**</td>
<td>Cases overdiagnosed/ cases expected without screening</td>
<td>Estimated based on prescreening incidence, published estimates of mean sojourn time</td>
</tr>
<tr>
<td>Italy (5 regions)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1986-2001</td>
<td>Biennial</td>
<td></td>
</tr>
<tr>
<td>50-74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paci, 2004***</td>
<td>Cases overdiagnosed/ cases expected without screening</td>
<td>Estimated based on prescreening incidence, published estimates of mean sojourn time</td>
</tr>
<tr>
<td>Italy (Florence)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990-1999</td>
<td>Biennial</td>
<td></td>
</tr>
<tr>
<td>50-69</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study; Country; Dates; Population Age; Screening Interval</td>
<td>Methods</td>
<td>Overdiagnosis Estimate</td>
</tr>
<tr>
<td>-----------------------------------------------------------</td>
<td>---------</td>
<td>------------------------</td>
</tr>
<tr>
<td>De Gelder, 2011\textsuperscript{80} Netherlands 1990-2006 40-69 Biennial</td>
<td>Microsimulation model of natural history of breast cancer, screening characteristics</td>
<td>Lead time</td>
</tr>
<tr>
<td>Lund, 2013\textsuperscript{77} Norway 2002-2010 52-79 Biennial</td>
<td>Observed incidence in women in prospective cohort who self-reported no mammograms, compared to women with at least one screening mammogram not national program</td>
<td>Excess incidence</td>
</tr>
<tr>
<td>Hofvind, 2012\textsuperscript{78} Norway 1996-2007 50-69 Biennial</td>
<td>Non-participants among women invited to participate in national program</td>
<td>Excess incidence</td>
</tr>
<tr>
<td>Study; Country; Dates; Population Age; Screening Interval</td>
<td>Definition of Overdiagnosis</td>
<td>Estimated Incidence without Screening</td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
<td>-----------------------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>Kalager, 2012 Norway 1996-2005 50-69 Biennial</td>
<td>Excess cases/expected cases without screening</td>
<td>Women in counties with current screening compared to • Current counties without screening • Prescreening incidence in counties with current screening • Prescreening incidence in counties without current screening Lead time estimates of 2 and 5 years</td>
</tr>
<tr>
<td>Zahl, 2004 Norway and Sweden 1971-2001 30 and older Biennial</td>
<td>Excess cases/cases expected without screening</td>
<td>Estimated incidence without screening based on Poisson regression of prescreening trends, with comparison to counties without screening (Norway)</td>
</tr>
<tr>
<td>Study; Country; Dates; Population Age; Screening Interval</td>
<td>Methods</td>
<td>Overdiagnosis Estimate</td>
</tr>
<tr>
<td>--------------------------------------------------------</td>
<td>---------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Hellquist, 2012\textsuperscript{22} Sweden 1986-2005</td>
<td>Cases overdiagnosed/ cases expected without screening</td>
<td>Observed incidence in women not invited to screen, adjusted for differences in incidence by county prior to screening (1970-1985), lead time</td>
</tr>
<tr>
<td>Yen, 2012\textsuperscript{25} Sweden 1977-2005</td>
<td>Excess cases/cases expected without screening</td>
<td>Cumulative incidence in women randomized to no invitation (but invited to screening at end of trial)</td>
</tr>
<tr>
<td>Zahl, 2011\textsuperscript{23} Sweden 1986-2000</td>
<td>Excess cases/cases expected without screening</td>
<td>Cumulative 6-year incidence in cohort of women 40-69 prior to invitation to screening</td>
</tr>
<tr>
<td>Study; Country; Dates; Population Age; Screening Interval</td>
<td>Methods</td>
<td>Overdiagnosis Estimate</td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
<td>---------</td>
<td>------------------------</td>
</tr>
<tr>
<td><strong>Definition of Overdiagnosis</strong></td>
<td><strong>Estimated Incidence without Screening</strong></td>
<td><strong>Methodological Approach</strong></td>
</tr>
<tr>
<td>Excess cases/cases expected without screening</td>
<td>Estimated incidence without screening based on Poisson regression of prescreening trends</td>
<td>Excess incidence</td>
</tr>
<tr>
<td>Jonsson, 2005(^{77}) Sweden 1985-2000 40-69 Biennial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases overdiagnosed/women screened</td>
<td>Poisson regression based on observed ages-specific incidence trends from 1974-1988 relative to incidence in women under age 45 (note: would not capture effects of HRT)</td>
<td>Excess incidence</td>
</tr>
<tr>
<td>Duffy, 2010(^{78}) United Kingdom 1989-2004 50-70 Triennial</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: BMI=body mass index; CI(s)=confidence interval(s); DCIS=ductal carcinoma in situ; HRT=hormone replacement therapy; OR=odds ratio; RCTs=randomized controlled trials; RR=relative risk
Major qualitative results include:

- Estimates based on the excess incidence approach are consistently higher (often substantially higher) than estimates based on the lead time approach.

- Estimates based on the excess incidence approach are lower when (a) follow-up time is increased, and/or (b) adjustments are made for lead time. For example, the estimate used by the UK Independent Panel was almost twice as high (19.0% vs. 10.7%) when the follow-up time was restricted to the screening period than when follow-up was extended over a longer period.\(^\text{11}\)

- Estimates under both methods are higher when DCIS and other in situ lesions are included, although the magnitude of the increase is variable. Given the variation in rates of DCIS diagnosis discussed below, this has implications for the generalizability of overdiagnosis estimates across populations.

- Estimates are higher when the analysis is based on comparing women attending screening versus not attending than when the analysis is based on women invited to screen versus not invited; for example, this was seen within the same study among women in British Columbia (increase from 6.7% to 17.3%),\(^\text{66}\) and in two separate Italian studies using different units of analysis (increase from 1%\(^\text{74}\) to 10%\(^\text{45}\)). Although this is expected, and is similar to the effect of changing the comparison groups on mortality reduction described above, it has substantial implications for estimating, even qualitatively, the harm-benefit ratio in terms of the number of expected overdiagnoses per breast cancer death prevented with screening: if the estimates for overdiagnosis are derived from studies that are substantially different in population, comparison groups, length of follow-up, screening strategies, etc., from those used for the mortality estimate, then the resulting ratio will be biased. For example, for the purposes of estimating overdiagnoses per breast cancer death prevented for an individual woman, estimates for both overdiagnosis and mortality reduction derived from studies comparing screened to unscreened women, rather than invited to uninvited, should be used (assuming that any biases are in the same general direction, and that methods for adjusting for those biases are appropriate). However, if the estimate of overdiagnosis is based on studies using population (invited vs. uninvited) estimates, but the mortality reduction estimate is based on screening attendance, then the ratio will be biased downwards.

- Another issue with studies that compare screening attenders to non-attenders is that there may be differences in factors affecting cancer incidence between the two groups that lead to biased estimates. For example, if family history contributes to screening attendance, then the expected incidence would be higher among attenders, leading to an overestimate of overdiagnosis. On the other hand, if attenders have a higher degree of concern about cancer, or lower threshold for seeking medical care, than non-attenders, then it is possible that a cancer would have been detected earlier in its course even without screening. This would overestimate the gain in lead time associated with screening, with subsequent implications for adjusted incidence estimates. Although it is plausible that many of the same contributors to self-selection bias in studies of mortality among screened and unscreened women contribute to biased estimates of incidence, the quantitative estimate of the bias may be different.

- Depending on the size of the overall study population, relatively small changes in the number of observed cases can change the estimate; in one of the Italian studies,\(^\text{45}\) the overdiagnosis percentage increased from 10% to 15% when 34 women with a diagnosis
of breast cancer within 6 months of their invitation to screen were excluded (because of the possibility that these represented cases that were already in the process of diagnosis and therefore not screen-detectable).

There is only one analysis based on observed U.S. data. Estimates were derived using the excess incidence approach, where the excess was defined as the difference between the increase in DCIS and local cancers and the decrease in regional and distant disease, under a base-case assumption of constant age-specific incidence from 1976-1978. In sensitivity analysis, this was varied assuming a constant rate of increase based on the observed annual increase (0.25%/year) in women under 40 years of age; a “best case” scenario for screening doubled the annual increase to 0.5%/year and used the highest observed incidence of regional and distant disease (113 per 100,000 in 1985) as the baseline. Adjustment for the potential effect of hormone replacement therapy on increases in incidence was performed by truncating incidence in 1990-2005 (the years of increasing hormone therapy use) to that of 2006-2008, the most recent data used in the analysis. Estimates for invasive cancer alone were 20% (no CIs given), and 31% when DCIS was included. This decreased to 28% and 22% under the different scenarios about background changes in incidence and incidence of late-stage disease.

In addition to the general limitations of using historical data to estimate current incidence in the absence of screening noted above, using changes in stage distribution as a proxy for overdiagnosis has limitations, particularly in the context of SEER data, where the lack of an indicator for whether or not a specific tumor was detected through screening is a major limitation.

For example, there have been changes in staging definitions, as well as in the type and sensitivity of procedures and technologies for staging over time. Tables 20 and 21 show the overall percent change in age-specific incidence by stage within SEER (Table 20) and the annual percentage change (Table 21) using the adjusted 6th edition American Joint Committee on Cancer (AJCC) breast cancer staging, between 1992 and 2011 (estimated using SEER*Stat software). The goal of cancer staging is to provide prognostic information, and stages may not necessarily represent the natural history of all cancers of a given type (for example, it is certainly possible to have distant metastases without an intermediary step of positive axillary lymph nodes), and, within breast cancer, micrometastases to the regional lymph nodes with a small primary tumor are included in Stage I (Stage IB).

Table 20. Percent Change in Age- and Stage-Specific Incidence of Breast Cancer, 1992-2011, SEER

<table>
<thead>
<tr>
<th>Stage</th>
<th>40-44</th>
<th>45-49</th>
<th>50-54</th>
<th>55-59</th>
<th>60-64</th>
<th>65-69</th>
<th>70-74</th>
<th>75-79</th>
<th>80-84</th>
<th>85+</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>75.3</td>
<td>34.2</td>
<td>45.9</td>
<td>50.2</td>
<td>69.1</td>
<td>89.3</td>
<td>77.3</td>
<td>83.8</td>
<td>64.2</td>
<td>25.5</td>
</tr>
<tr>
<td>I</td>
<td>24.6</td>
<td>6.4</td>
<td>0.1</td>
<td>-2.2</td>
<td>10.1</td>
<td>12</td>
<td>12.7</td>
<td>6.5</td>
<td>10.7</td>
<td>35.7</td>
</tr>
<tr>
<td>II A</td>
<td>15.2</td>
<td>3.1</td>
<td>14</td>
<td>-2.7</td>
<td>16.4</td>
<td>26.7</td>
<td>13.9</td>
<td>-2.3</td>
<td>7.8</td>
<td>-1</td>
</tr>
<tr>
<td>II B</td>
<td>38.3</td>
<td>54.7</td>
<td>50.9</td>
<td>32.1</td>
<td>38.9</td>
<td>32.2</td>
<td>28.9</td>
<td>27.8</td>
<td>51.2</td>
<td>9.3</td>
</tr>
<tr>
<td>II NOS</td>
<td>-92.4</td>
<td>-93.2</td>
<td>-89.5</td>
<td>-97.1</td>
<td>-95.1</td>
<td>-89</td>
<td>-91</td>
<td>-94.2</td>
<td>-86.3</td>
<td>-94.3</td>
</tr>
<tr>
<td>II A</td>
<td>2.1</td>
<td>1.9</td>
<td>-3</td>
<td>-37</td>
<td>-19.1</td>
<td>-9.9</td>
<td>-6.6</td>
<td>-15.8</td>
<td>-45.2</td>
<td>-14.5</td>
</tr>
<tr>
<td>II B</td>
<td>-11.6</td>
<td>-2.7</td>
<td>-9.9</td>
<td>32.4</td>
<td>11.4</td>
<td>-1.7</td>
<td>19</td>
<td>-2.4</td>
<td>-33</td>
<td>0.5</td>
</tr>
<tr>
<td>II C</td>
<td>-26.2</td>
<td>-34.5</td>
<td>-46.4</td>
<td>-45.3</td>
<td>-29.1</td>
<td>-41.7</td>
<td>-4.5</td>
<td>-39.5</td>
<td>-6.6</td>
<td>16.2</td>
</tr>
<tr>
<td>IV</td>
<td>30.7</td>
<td>27.6</td>
<td>15.2</td>
<td>49.7</td>
<td>17.5</td>
<td>6.9</td>
<td>5.8</td>
<td>9.7</td>
<td>4.5</td>
<td>-6.6</td>
</tr>
<tr>
<td>Unstaged</td>
<td>-59.3</td>
<td>-67.4</td>
<td>-65.4</td>
<td>-64.1</td>
<td>-64.5</td>
<td>-71</td>
<td>-70.2</td>
<td>-70.8</td>
<td>-66.3</td>
<td>-67.4</td>
</tr>
</tbody>
</table>

Abbreviations: NOS=not otherwise specified; SEER= Surveillance, Epidemiology, and End Results
Table 21. Annual Percent Change in Age- and Stage-Specific Incidence of Breast Cancer, 1992-2011, SEER

<table>
<thead>
<tr>
<th>Stage</th>
<th>Age</th>
<th>40-44</th>
<th>45-49</th>
<th>50-54</th>
<th>55-59</th>
<th>60-64</th>
<th>65-69</th>
<th>70-74</th>
<th>75-79</th>
<th>80-84</th>
<th>85+</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.4*</td>
<td>2.1*</td>
<td>1.8*</td>
<td>1.2*</td>
<td>2.3*</td>
<td>2.7*</td>
<td>2.3*</td>
<td>2.8*</td>
<td>2.5*</td>
<td>2.4*</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>1.0*</td>
<td>0.5*</td>
<td>-0.5*</td>
<td>-0.8</td>
<td>0.1</td>
<td>0.7*</td>
<td>0.2</td>
<td>0.2</td>
<td>0.6*</td>
<td>1.4*</td>
<td></td>
</tr>
<tr>
<td>IIA</td>
<td>1.0*</td>
<td>0.4*</td>
<td>-0.2</td>
<td>-0.3</td>
<td>0.8*</td>
<td>0.9*</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.7*</td>
<td></td>
</tr>
<tr>
<td>IIB</td>
<td>2.1*</td>
<td>1.6*</td>
<td>1.4*</td>
<td>0.8</td>
<td>2.1*</td>
<td>1.7*</td>
<td>1.5*</td>
<td>1.6*</td>
<td>1.3*</td>
<td>1.8*</td>
<td></td>
</tr>
<tr>
<td>IIINOS</td>
<td>-12.7*</td>
<td>-13.9*</td>
<td>-13.7*</td>
<td>-14.4*</td>
<td>-12.9*</td>
<td>-11.0*</td>
<td>-11.4*</td>
<td>-11.4*</td>
<td>-9.0*</td>
<td>-8.4*</td>
<td></td>
</tr>
<tr>
<td>IIIA</td>
<td>0.5</td>
<td>-0.2</td>
<td>-0.7</td>
<td>-1.3</td>
<td>-0.9</td>
<td>-0.5</td>
<td>-1.2*</td>
<td>-0.9</td>
<td>-1.4*</td>
<td>-1.0*</td>
<td></td>
</tr>
<tr>
<td>IIIB</td>
<td>1.5</td>
<td>-0.3</td>
<td>0.3</td>
<td>1.3</td>
<td>0.4</td>
<td>0.2</td>
<td>-0.1</td>
<td>-0.2</td>
<td>-1.2*</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>IIIC</td>
<td>-2.2*</td>
<td>-3.0*</td>
<td>-2.8*</td>
<td>-2.8*</td>
<td>-1.8*</td>
<td>-1.7*</td>
<td>-1.0*</td>
<td>-2.0*</td>
<td>-0.5</td>
<td>2.4*</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>1.9</td>
<td>1.5*</td>
<td>0.4</td>
<td>1.1*</td>
<td>0.7</td>
<td>0.6</td>
<td>-0.1</td>
<td>0</td>
<td>0.5</td>
<td>-0.2</td>
<td></td>
</tr>
<tr>
<td>Unstaged</td>
<td>-6.3*</td>
<td>-6.6*</td>
<td>-7.4*</td>
<td>-7.3*</td>
<td>-6.6*</td>
<td>-6.6*</td>
<td>-7.7*</td>
<td>-8.1*</td>
<td>-7.7*</td>
<td>-5.9*</td>
<td></td>
</tr>
</tbody>
</table>

*Significantly different than no change, p<0.05.

Abbreviations: NOS=not otherwise specified; SEER= Surveillance, Epidemiology, and End Results

Some of the largest statistically significant changes are seen in the Stage III NOS (not otherwise specified) and Unstaged categories across all ages. This creates substantial difficulty for interpreting changes in stage-specific incidence, or changes in the distribution of stages within a given age group, over time, since the distribution of disease severity, especially within the unstaged group, is likely to be different than the distribution within staged groups.

Another limitation is that changes in the distribution of disease severity within a given stage may not be captured. Tumor size within stage is decreased among women who are screened, consistent with an effect of screening. Under a classic stochastic model of cancer growth, tumor volume is directly related to the number of cells, and, based on chance alone, the probability of a given tumor accumulating enough mutations to develop the ability to metastasize should be correlated with size—thus, all things being equal, a smaller tumor within a given stage should be less likely to have metastasized at the time of detection, and survival will be improved (although the extent to which this translates into improved mortality, versus the effects of lead time, may vary). Therefore, changes in the size distribution within stages that have implications for decreased mortality may not be captured by analysis of stage-specific trends, particularly the relatively imprecise local/regional/distant classification. However, it must also be noted that there is evidence that tumor size alone is not universally predictive of outcome, based on a combination of observational, laboratory, and modeling studies, and that certain cancers may be small yet biologically quite aggressive (and vice versa), perhaps because tumor metastatic potential is derived from specific stem cells. This phenomenon would be consistent with the lack of change in the incidence of stage IV disease—small tumors with a high biological predisposition to metastasis may have already spread at the time of detection through screening.

The lack of other estimates of overdiagnosis for the U.S. is particularly problematic given wide variation in the detection of DCIS between countries; Table 22 depicts rates for women 50-69 in select countries participating in the International Cancer Screening Network.

Table 22. Across-Country Variation in the Proportion of DCIS among all Screen-Detected Cancers in Women 50-69

<table>
<thead>
<tr>
<th>Country</th>
<th>All Screens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age-Standardized Incidence/1000</td>
</tr>
<tr>
<td>Subsequent Screens</td>
<td></td>
</tr>
</tbody>
</table>

73
Invasive Cancer
DCIS
Age-Standardized DCIS/1000
Invasive Cancer Subsequent/All Invasive Cancer

<table>
<thead>
<tr>
<th>Country</th>
<th>Invasive Cancer</th>
<th>DCIS</th>
<th>Age-Standardized DCIS/1000</th>
<th>Invasive Cancer Subsequent/All Invasive Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>3.19</td>
<td>1.00</td>
<td>24%</td>
<td>0.98</td>
</tr>
<tr>
<td>Denmark</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Copenhagen</td>
<td>6.65</td>
<td>1.55</td>
<td>19%</td>
<td>1.38</td>
</tr>
<tr>
<td>Fyn</td>
<td>5.83</td>
<td>0.64</td>
<td>10%</td>
<td>0.62</td>
</tr>
<tr>
<td>Norway</td>
<td>4.30</td>
<td>0.93</td>
<td>18%</td>
<td>0.86</td>
</tr>
<tr>
<td>Netherlands</td>
<td>4.06</td>
<td>0.80</td>
<td>16%</td>
<td>0.76</td>
</tr>
<tr>
<td>Italy</td>
<td>3.98</td>
<td>0.72</td>
<td>15%</td>
<td>–</td>
</tr>
<tr>
<td>Finland</td>
<td>4.81</td>
<td>0.45</td>
<td>9%</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Abbreviation: DCIS = ductal carcinoma in situ

Data from the U.S. are from the BCSC. Key points include:

- Rates and proportions vary widely across countries and are not correlated with invasive cancer rates. The U.S., Denmark, and Luxembourg (not shown) are all outliers, with higher DCIS detection relative to invasive cancer.
- DCIS detection rates are not related to the ratio of cancers detected at subsequent screens to all invasive cancers, a measure used as a simple measure of screening program performance, where a ratio of 1.5 is approximately equivalent to a program sensitivity of 75%. The lower number for the U.S. is partly a function of a substantial number of women receiving annual screening (since there is less time between screens, the number of screen-detectable cancers at each round will be smaller at a given rate of tumor growth).
- For the U.S., in particular, DCIS detection rates do not decrease substantially with subsequent screens.
- Table 22 illustrates data for women aged 50 and older. Given the SEER data below showing high rates of DCIS diagnosis in younger women (Table 23 and Figures 8-9), the overall proportion of DCIS in the U.S. among all women screened might be substantially higher than 24%, given active screening among women 40-49.

This variability across countries is also seen across centers in the U.S.—in the BCSC registry, DCIS detection rates at individual centers varied from 14.6-23.8% overall, and from 18-30% in women 40-49 years old. This variability means that, even if all other parameters are equivalent, estimates of the contribution of DCIS to overdiagnosis derived from screening programs in other countries may not apply to the U.S. Perhaps more importantly from the perspective of an individual woman in the U.S., it means that there is likely to be substantial uncertainty about estimations of her individual risk of having an overdiagnosed (and overtreated) cancer because of the variability in DCIS detection rates across centers that may be independent of variation in risk.

Model-based Estimates for the U.S.

The CISNET collaborators reported that five of the six models estimated overdiagnosis rates, but did not show the actual estimates. A study from one of the CISNET groups that estimated the retrospective cost-effectiveness of screening based on patterns observed in the U.S. reported that the incidence of cancer was approximately 25% higher with screening than without, but did not provide any additional details.
Other model-based studies, either from the CISNET group\textsuperscript{157,158} or the Breast Cancer Surveillance Consortium\textsuperscript{159} comment on the qualitative effect of overdiagnosis on quality-adjusted life expectancy, but do not provide specific estimates of the probability of overdiagnosis.

**Effect of Age on Estimates of Overdiagnosis**

Qualitatively, the risk of overdiagnosis among the CISNET models increased with age, with the increase accelerating because of competing risks of mortality.\textsuperscript{30} Overdiagnosis was higher for DCIS than for invasive cancer, with more overdiagnoses due to DCIS in younger women, but because of the competing risk of mortality, extending screening beyond age 69 had a greater effect on overdiagnosis than starting screening earlier. Estimates were reported to be sensitive to whether a given model included DCIS in the underlying history, and to assumptions about the behavior of DCIS and small localized cancers, but again, the quantitative effects of these assumptions was not presented.

**Effect of Screening Intervals on Overdiagnosis**

In the CISNET models, biennial screening strategies reduced overdiagnosis compared to annual strategies, “…but by much less than one half.”\textsuperscript{30}

**Estimated Absolute Effect Size in the U.S. Population**

Given the extreme uncertainty about the magnitude of overdiagnosis, we are unable to make a direct estimate of the absolute number of women in the U.S. who are overdiagnosed as the result of breast cancer screening. However, it is possible to estimate the size of the potential “pool” of tumors which could be overdiagnosed through screening under certain assumptions about which tumors are most likely to be overdiagnosed. Different estimates of the proportion of overdiagnosis can then be applied to this pool to provide a range of plausible estimates.

Here, we assume that overdiagnosed breast cancers are drawn from in situ lesions (primarily DCIS) and small (<2 cm) invasive cancers without involvement of regional lymph nodes or distant metastases (T1N0M0 under the TNM staging system). It is possible that more advanced cancers could also represent overdiagnosis (for example, in the setting of older women with high near- or intermediate-term risk of death from another cause because of age and/or comorbidities).

Figure 8 shows the age-specific and cumulative incidence of in situ, T1N0M0 invasive cancers, and all other invasive cancers from SEER 2000-2010.
As Figure 9 illustrates, the combination of in situ and T1N0M0 tumors is at least 50% at all ages, with in situ being somewhat more common at younger ages. The proportion of cancers that are in situ estimated here is quite similar to that in a recent international comparison across
screening programs, where the U.S. had the highest proportion (24%) of DCIS lesions among all breast cancers.\textsuperscript{154}

**Figure 9. Distribution of Breast Cancer Diagnoses by Age**

To estimate the distribution of these diagnoses in screened and unscreened women, we take an approach similar to the one used to estimate breast cancer mortality reduction, assuming:

- 65\% of women are screened at least biennially.
- Based on data from the BCSC, the relative risk of an in situ diagnosis with screening varies with age, with a relative risk (RR) ranging from 7.0 at age 40 to 4.9 at age 70;\textsuperscript{155} the overall age-adjusted RR was 6 (0.78 per 1000 screen-detected vs. 0.13 per 1000 non-screen-detected; Table 23).

**Table 23. Screen-detected and Non-screen-detected DCIS among Women in the BCSC\***

<table>
<thead>
<tr>
<th>Age</th>
<th>DCIS Rate per 1000 Mammograms (95% CI)</th>
<th>RR (Calculated from Mean Incidence)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Screen-detected</td>
<td>Non-screen-detected</td>
</tr>
<tr>
<td>40-49</td>
<td>0.56 (95% CI, 0.41 to 0.70)</td>
<td>0.08 (95% CI, 0.02 to 0.13)</td>
</tr>
<tr>
<td>50-59</td>
<td>0.68 (95% CI, 0.52 to 0.85)</td>
<td>0.09 (95% CI, 0.03 to 0.05)</td>
</tr>
<tr>
<td>60-69</td>
<td>1.03 (95% CI, 0.83 to 1.23)</td>
<td>0.19 (95% CI, 0.11 to 0.28)</td>
</tr>
<tr>
<td>70-84</td>
<td>1.07 (95% CI, 0.87 to 1.27)</td>
<td>0.22 (95% CI, 0.13 to 0.31)</td>
</tr>
</tbody>
</table>

*Adapted from Table 4 in Ernster et al.\textsuperscript{153}

Abbreviations: BCSC=Breast Cancer Surveillance Consortium; CI=confidence interval; DCIS=ductal carcinoma in situ; RR=relative risk

- As a sensitivity analysis, we used a RR of 3.0 across all ages for screened versus invited but not screened women in the Norwegian screening program.\textsuperscript{70}
- The RR of having a tumor <2 cm with no nodes is 1.5 with screening, and for having nodes 1.25 with no screening, based on a recent systematic review.\textsuperscript{160}
- Note that, for incidence, this simply partitions the age-specific incidences into screened and unscreened, without an explicit adjustment for lead time, or assumptions about the
effect of diagnosing and treating of DCIS on incidence of future invasive cancers (of all stages).

Figure 10 shows the estimated age-specific incidence of in situ, T1N0M0, and all other invasive cancers in screened and unscreened women based on these parameters, while Figure 11 shows the age-specific distribution of each diagnosis.

Figure 10. Estimated Age-specific Incidence of In Situ, T1N0M0 Invasive Breast Cancer, and All Other Breast Cancers in Unscreened (A) and Screened (B) Women

(A) Unscreened Women

(B) Screened Women
Figure 11. Estimated Distribution of Diagnoses by Age in Unscreened (A) and Screened (B) Women

(A) Unscreened Women

(B) Screened Women

Over half of all diagnosed cancers in screened women are in situ or T1N0M0, with the proportion of in situ lesions ranging from over 30% at age 40 to approximately 16% at age 80, rates consistent with those reported by the BCSC.\textsuperscript{155} In the 2002 BCSC report, 86.0\% of all DCIS lesions were screen-detected. The incidence of DCIS has increased since 1996-1997, the time period analyzed by Ernster and colleagues:\textsuperscript{155} age-adjusted incidence of DCIS in 1996 was 54.3 per 100,000, compared with 71.0 per 100,000 in 2010.\textsuperscript{12} Applying the age-specific relative risks derived from the Ernster paper to current DCIS incidence rates results in an estimated 92\%
of DCIS cases being screen-detected—the relatively small increase in the proportion attributable to screening is plausible given the large increase in incidence (some of which may be attributable to the increase in the proportion of the population in the 50 to 70 age range because of the aging of the baby boom generation).

There is considerable uncertainty about the natural history of DCIS; in particular, the proportion of detected DCIS lesions that would ultimately progress to symptomatic cancer in the absence of screening is unclear. Given that, at least in the U.S., most DCIS lesions are treated using modalities identical to those used to treat early invasive cancers, the advantages of detecting and treating DCIS, compared to a “watchful waiting” approach, are not clear.161 Perhaps even more than with invasive cancers, estimations of the proportion of DCIS that is not progressive (as opposed to potentially progressive cancers that would not become symptomatic because of competing mortality risks) are critical to estimating the risk of overdiagnosis with different screening strategies.

Estimates of the proportion of DCIS lesions that will progress to invasive cancer vary widely. The most direct estimates come from follow-up studies of breast biopsies initially read as normal where DCIS was identified on subsequent review. Table 24, adapted from the review of Erbas and colleagues,162 shows the results from the four available studies.

Table 24. Studies of “Natural History” of Untreated DCIS*

<table>
<thead>
<tr>
<th>Study</th>
<th>No. of Benign Biopsies Examined</th>
<th>No. of Misdiagnosed DCIS (No. for whom Follow-up Available)</th>
<th>No. Subsequently Invasive</th>
<th>Age at Initial Biopsy</th>
<th>Follow-up Period</th>
<th>% Invasive (95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eusebi, 1994163</td>
<td>9520 (histological reassessment on only 9446)</td>
<td>80 (80)</td>
<td>11</td>
<td>24–77 years</td>
<td>1–14 years</td>
<td>0.14 (0.07, 0.23)</td>
</tr>
<tr>
<td>Sanders, 2005†</td>
<td>11,760</td>
<td>28 (28)</td>
<td>11</td>
<td>33–74 years</td>
<td>Median 31 years</td>
<td>0.32 (0.15, 0.49)</td>
</tr>
<tr>
<td>Rosen, 1970165</td>
<td>&gt;8000 reported as benign</td>
<td>30 (15)</td>
<td>8</td>
<td>Not reported</td>
<td>1–24 years</td>
<td>0.53 (0.28, 0.79)</td>
</tr>
<tr>
<td>Collins, 2004166</td>
<td>1877</td>
<td>13 (13)</td>
<td>6</td>
<td>41–63 years</td>
<td>4–18 years</td>
<td>0.46 (0.19, 0.73)</td>
</tr>
</tbody>
</table>

*Adapted from Erbas, 2006.162

†Update of study included in Erbas, 2006.162

Abbreviations: CI=confidence interval; DCIS=ductal carcinoma in situ; No.=Number

In addition to the small numbers and subsequent wide confidence intervals, these studies have additional limitations:
- Less aggressive lesions might have been more likely to have been originally misdiagnosed, leading to an underestimate of progression.
- The process of inflammation and wound healing induced by the biopsy may have affected the natural history of the lesion.
- Loss to follow-up was high in the study by Rosen et al;165 if women with subsequent breast cancer were more likely to have been detected, then the estimate would be biased upward.
As with overdiagnosis in general, an alternative approach is to use mathematical models to impute the proportion of progressive DCIS from observed data. Yen and colleagues used data from prevalent and incident screens in the Swedish Two-County trial and a range of observational data from service screening programs to estimate the proportion of screen-detectable DCIS lesions that would progress to invasive cancer, using a Markov model, and estimated proportions of non-progressive DCIS ranging from 19-46% at the time of the prevalence screen, and 3-21% at the subsequent screen (with 7 out of the 8 estimates of the proportion for the subsequent screen being 7% or less). In addition to issues concerning the validity of assumptions about the appropriateness of an exponential distribution for the transition times, and of progressive versus non-progressive behavior of DCIS being the only source of heterogeneity in transition rates, which were discussed by the authors, there is a more fundamental assumption that is not discussed which could affect the estimates of both the proportion of non-progressive DCIS and lesions and the transition rates.

The Markov model as described in the paper apparently assumes that invasive cancer is necessarily preceded by DCIS. However, DCIS can only be a non-obligate precursor for invasive ductal carcinoma, which, while the most common single type, only accounts for 69% of all invasive cancers in the U.S., with proportions ranging from 75% in 40- to 44-year-olds to 57% in women 85 and older; other histologic types have different pre-invasive states. All of the studies used for the analysis reported total invasive cancer cases, without any description of histologic types. If the model assumes that all observed cancers (a proportion of which will be non-ductal) necessarily pass through a DCIS state in order to become invasive, then the estimated proportion of non-progressive DCIS will necessarily need to be low in order to fit the observed data. However, if 20-30% of invasive cancers never pass through a progressive DCIS state because they are not ductal in origin, then a higher proportion of non-progressive DCIS would be compatible with the observed data. In other words, the apparent structural assumptions of the model lead to a potential overestimation of the proportion of DCIS which must progress in order to fit observed invasive cancer incidence.

Another indirect line of evidence that the proportion of DCIS that is non-progressive may be relatively high is the lack of a clear decrease in the incidence of invasive cancer of any stage as detection and treatment of DCIS has increased. Assuming no major changes in the underlying natural history of the disease in the presence of a detectable preclinical stage, screening should lead to both a shift to earlier stages of invasive disease and a decrease in overall incidence as preclinical lesions are treated and removed from the risk pool. This has been observed with cervical and colorectal cancer, but not with breast cancer, where, despite marked increases in the detection of DCIS, the incidence of invasive disease has increased or not changed (with the exception of a few years after the release of the Women’s Health Initiative results) (Figure 12).
Given the uncertainty about the proportions of both DCIS and invasive lesions that are potentially non-progressive, we can only provide a range of estimates under different assumptions about those proportions. Table 25 presents the potential proportion of overdiagnosed lesions based on a range of estimates of the proportion of DCIS lesions that progress, and the proportion of small node-negative lesions that would not progress given the observed age-specific incidence of each and the assumptions above (65% screened within the
past 2 years across all age groups, RR for DCIS among screened women ranging from 7.0 to 4.86, fixed RR with screening for T1N0M0 tumors); results did not differ substantially using a fixed RR for DCIS with screening of 3.0 (which, given the observed difference in DCIS rates between the U.S. and Norway, is conservative).

Table 25. Potential Proportion of Screen-detected Lesions that Represent Overdiagnosis under Different Estimates of DCIS Progression and of the Proportion of Small Node-negative Tumors that would not Become Clinically Apparent without Screening, by Age

<table>
<thead>
<tr>
<th>Age</th>
<th>20%</th>
<th>50%</th>
<th>80%</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>40-44</td>
<td>6.0%</td>
<td>15.1%</td>
<td>24.1%</td>
<td>1.5%</td>
<td>2.9%</td>
<td>4.4%</td>
<td>5.8%</td>
</tr>
<tr>
<td>45-49</td>
<td>6.1%</td>
<td>15.1%</td>
<td>24.2%</td>
<td>1.6%</td>
<td>3.2%</td>
<td>4.7%</td>
<td>6.4%</td>
</tr>
<tr>
<td>50-54</td>
<td>6.0%</td>
<td>15.0%</td>
<td>24.1%</td>
<td>1.7%</td>
<td>3.4%</td>
<td>5.0%</td>
<td>6.9%</td>
</tr>
<tr>
<td>55-59</td>
<td>5.5%</td>
<td>13.7%</td>
<td>22.0%</td>
<td>1.8%</td>
<td>3.7%</td>
<td>5.5%</td>
<td>7.5%</td>
</tr>
<tr>
<td>60-64</td>
<td>5.1%</td>
<td>12.7%</td>
<td>20.3%</td>
<td>2.0%</td>
<td>4.0%</td>
<td>6.0%</td>
<td>8.0%</td>
</tr>
<tr>
<td>65-69</td>
<td>4.9%</td>
<td>12.4%</td>
<td>19.8%</td>
<td>2.1%</td>
<td>4.2%</td>
<td>6.4%</td>
<td>8.4%</td>
</tr>
<tr>
<td>70-74</td>
<td>4.6%</td>
<td>11.5%</td>
<td>18.4%</td>
<td>2.2%</td>
<td>4.4%</td>
<td>6.6%</td>
<td>8.9%</td>
</tr>
<tr>
<td>75-79</td>
<td>4.2%</td>
<td>10.5%</td>
<td>16.8%</td>
<td>2.2%</td>
<td>4.4%</td>
<td>6.6%</td>
<td>8.9%</td>
</tr>
<tr>
<td>80-94</td>
<td>3.5%</td>
<td>8.8%</td>
<td>14.1%</td>
<td>2.1%</td>
<td>4.3%</td>
<td>6.4%</td>
<td>8.6%</td>
</tr>
</tbody>
</table>

Discussion/Conclusions: Overdiagnosis

- The lack of consensus on the most appropriate methodology for defining and estimating overdiagnosis is a major barrier to comparing published estimates, or to deriving estimates for the U.S. based on relative estimates generated in other settings.
- Variations in the rates of diagnosis of DCIS between screening programs, and variations in assumptions about the natural history of DCIS and its role in the biology of invasive cancer, account for a substantial proportion of the uncertainty about the rates of overdiagnosis from both observational and modeling studies. Rates of DCIS diagnosis in the U.S. are higher than in other countries, meaning that the potential contribution of DCIS to overdiagnosis in the U.S. is substantial, even if no screen-detected invasive cancers are overdiagnosed. Given that current practice is for all women with DCIS to be treated relatively aggressively, and that the diagnosis itself creates considerable confusion and anxiety for many women, this may have a substantial impact on quality of life and quality-adjusted life expectancy, as we discuss below.
- As with breast cancer mortality reduction, we judge the quality of evidence for the existence of overdiagnosis to be HIGH; however, given the wide range of estimates, the lack of directness (from observational studies in non-U.S. settings, and from model-based estimates), and the uncertainty about the natural history of DCIS and small localized invasive cancers, we judge the quality of evidence on the estimate for the quantitative magnitude of overdiagnosis in the U.S. to be LOW. The high incidence of DCIS among screened women, the variability in rates of diagnosis even within countries, and the high degree of uncertainty about the proportion of DCIS that has the potential to progress to symptomatic invasive cancer all contribute to high degrees of uncertainty about the
probability of an overdiagnosed DCIS or invasive lesion for U.S. women under different screening policies at both the population and individual levels.

**False Positives**

By definition, women who are not screened cannot have a false positive result, so we report estimates only for screened women. Although we report on results from other settings, we emphasize those from U.S. population-based data as most relevant to recommendations for U.S. screening practice; in addition, as the results show, there is substantial variability in false positive probabilities for both recall visits and biopsies across and within countries. For non-U.S. studies, our discussion here for the most part focuses on results from recent pooled analyses or systematic reviews; results from individual studies that met our inclusion criteria that are not discussed below are presented in Appendix Table G-1.

**Observational Studies**

**False Positive: Same Day Repeat Examination**

We did not identify any studies that separately reported same day repeat examination false positive rates.

**False Positive: Subsequent Visit Repeat Examination (Recall)**

**Single Screening Visit**

**Non-U.S. Studies**

False positive recalls (defined as screening tests resulting in a repeat examination performed at some future time, with no cancer detected at the subsequent examination) in identified European screening studies ranged from 1.1% to 10.6% per screen among average-risk women.\(^{21,83,84,88,90,91,93}\)

False positive recall rates are consistently higher with a first screen compared to subsequent screens. In a pooled summary of results from 20 screening programs in 17 European countries between 2005 and 2007 (screening ages 50-69, with biennial screens), Hofvind et al.\(^8\) reported recall rates of 9.3% (range 2.2% to 15.6%) for the initial screen and 4.0% (range 1.2% to 10.5%) for subsequent screens. Positive predictive value was 9.6% (range 4.9% to 24.2%) for first screens and 18.6% (range 6.8% to 49.5%) for subsequent screens.

In the UK Age RCT of screening, false positive probability (women aged 39-41) was 4.9% at first screen and 3.2% at subsequent screens.\(^{17}\)

**U.S. Studies**

The best available population-based U.S. data are from the Breast Cancer Surveillance Consortium (BCSC). As with the European data, false positive recall probabilities were higher for first screens than for subsequent screens. For first screens, false positive recall ranged from 16.4% for women aged 40-44 to 19.7% for women aged 55-59; for subsequent screens, proportions ranged from 8.9% for 40- to 44-year-olds to 9.6% for women 65 years old and over. In a multivariate analysis, initial screen false positive probabilities were significantly higher with increasing age: using the probability for women aged 40-44 as the reference, odds ratios (95% CIs) were 1.27 (1.21 to 1.33) for 45- to 49-year-olds, 1.39 (1.31 to 1.47) for 50- to 54-year-olds, and 1.24 (1.15 to 1.36) for 55- to 59-year-olds. In contrast, false positive probabilities for
subsequent examinations were not statistically higher for women older than 40-44 except for women aged 45-49 (OR 1.07; 95% CI, 1.01 to 1.12).

In addition to age, first screen false positive probabilities were significantly increased by:

- A family history of breast cancer (20.5% compared to 17.6% in women without a family history; OR 1.21; 95% CI, 1.13 to 1.30).
- Breast density had a significant effect on initial false positive results. Compared to women with scattered fibroglandular densities (BI-RADS 2), women with heterogeneously dense breasts (BI-RADS 3) had significantly increased false positive recall rates (19.3% compared to 17.8% in women with BI-RADS 2; OR 1.11; 95% CI, 1.06 to 1.16); false positive probabilities were significantly decreased in women with breasts that were either extremely dense (BI-RADS 4; OR 0.85; 95% CI, 0.79 to 0.91) or with a density that of almost entirely fat (BI-RADS 1; OR 0.62; 95% CI, 0.56 to 0.67).
- Time; false positives increased in every time period, from 13.6% prior to 1997 to 20.7% after 2004, with ORs relative to pre-1997 statistically significant for all time periods.
- Current hormone replacement therapy was not significantly associated with an increased false positive probability.

For subsequent screens:

- Family history was not associated with an increased probability of a false positive examination (OR 0.98; 95% CI, 0.91 to 1.05).
- As with first examinations, breast density affected false positive probability. A high fat content decreased false positive probability relative to scattered fibroglandular densities (OR 0.45; 95% CI, 0.40 to 0.50), while probability was increased with both heterogeneously dense (OR 1.40; 95% CI, 1.33 to 1.46) and extremely dense breasts (OR 1.16; 95% CI, 1.08 to 1.25).
- False positive probabilities for subsequent screens also increased significantly with time, from 8.6% prior to 1997 to 11.0% after 2004.
- False positive probabilities with subsequent screens were halved when a comparison mammogram was available, from 15.8% to 8.7% (OR 0.50; 95% CI, 0.45-0.56).
- False positive probability increased as time since last screen increased, from 8.3% for an interval of 9-18 months (reference) to 9.3% for 19-30 months (OR 1.13; 95% CI, 1.08 to 1.19) to 10.7% for intervals longer than 30 months (OR 1.33; 95% CI, 1.26 to 1.40).

For both first and subsequent screens, false positive probabilities within the BCSC are approximately twice as high as the pooled results from the European studies.\(^8\)

**Cumulative False Positive Recall Probability**

**Non-U.S. Studies**

The estimate of cumulative risk of any false positive result (defined as further assessment without a diagnosis of cancer, both recall and biopsy) from a pooled analysis of three European studies over 10 rounds of biennial screening in women aged 50-69 years was 19.7% (CIs not reported).\(^8\)

The estimated cumulative lifetime risk over 13 examinations from ages 50 through 75 in one Dutch program was 7.3% (95% CI, 5.5 to 9.0%) for women with an initial screen between 1997 and 2006, an increase from 4.4% (95% CI, 3.3 to 5.1%) for women with a first screen in 1975.\(^88\)
For women aged 40-49 in the intervention arm of the UK Age trial, 18.1% of women attending at least one screening visit had one or more false positive screens. Observed cumulative probability of at least one false positive result over seven screens was 20.5%, with an estimate of 21.6% based on an assumption of independence of risk at each examination for each woman. Based on observed attendance, the investigators estimated a 28.0% probability of a false positive over 10 screens.17

U.S. Studies

Overall 10-year cumulative risk of a false positive estimated based on a multivariate model that accounted for age, family history, breast density, year of first examination, availability of previous mammograms, individual registry within the BCSC, and a variable for random radiologist within the BCSC varied by screening frequency, but not by age of starting screening. For women with a first screen at age 40, estimated 10-year cumulative risk of a false positive was 61.3% (95% CI, 59.4% to 63.1%) for annual screening versus 41.6% (95% CI, 40.6% to 42.9%) for biennial screening. For women with a first screen at age 50, estimated 10-year cumulative risk was 61.3% (95% CI, 58.0% to 64.7%) for annual screening, and 42.0% (95% CI, 40.4% to 43.7%) for biennial screening.92

In a subsequent analysis where results were presented stratified by age (40-49 years vs. 50-74 years), breast density categories, and use of hormone replacement therapy, cumulative 10-year false positive rates for women 40-49 years were higher than for women 50 and older who were not using hormone replacement therapy (with CIs not overlapping, suggesting a significant difference), but not for women using hormone replacement therapy.87 Table 26 presents these results for women under 50, women 50 and older not using hormone replacement therapy, and women on combination hormone replacement; results for women using estrogen only were similar to those for women using combination therapy.

Table 26. Estimated 10-year Cumulative Probability (95% CI) of False Positive Recall in the BCSC by Age, Breast Density, and HRT Status*

<table>
<thead>
<tr>
<th>Age and Measure</th>
<th>Breast Density</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fatty (BI-RADS 1)</td>
<td>Scattered Fibroglandular Densities (BI-RADS 2)</td>
<td>Heterogeneously Dense (BI-RADS 3)</td>
<td>Extremely Dense (BI-RADS 4)</td>
</tr>
<tr>
<td>Age 40–49:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First mammography</td>
<td>11.2 (10.3–12.2)</td>
<td>17.0 (16.6–17.4)</td>
<td>18.0 (17.6–18.4)</td>
<td>15.1 (14.4–15.8)</td>
</tr>
<tr>
<td>Cumulative probability of false positive after 10 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual</td>
<td>36.3 (34.3–38.3)</td>
<td>60.0 (58.6–61.3)</td>
<td>68.9 (67.6–70.1)</td>
<td>65.5 (64.0–66.9)</td>
</tr>
<tr>
<td>Biennial</td>
<td>21.2 (20.0–22.3)</td>
<td>38.5 (37.8–39.3)</td>
<td>46.3 (45.5–47.1)</td>
<td>43.2 (42.3–44.1)</td>
</tr>
<tr>
<td>Age 50–74 (no HRT):</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First mammography</td>
<td>9.9 (9.1–10.8)</td>
<td>16.5 (16.0–17.1)</td>
<td>19.0 (18.2–19.8)</td>
<td>16.3 (14.3–18.5)</td>
</tr>
<tr>
<td>Cumulative probability of false positive after 10 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual</td>
<td>30.3 (29.3–31.3)</td>
<td>49.8 (49.0–50.6)</td>
<td>60.2 (59.3–61.0)</td>
<td>58.5 (57.1–59.8)</td>
</tr>
<tr>
<td>Biennial</td>
<td>17.4 (16.8–18.0)</td>
<td>30.7 (30.2–31.2)</td>
<td>38.9 (38.3–39.5)</td>
<td>37.5 (36.6–38.4)</td>
</tr>
<tr>
<td>Age 50–74 (combination HRT):</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First mammography</td>
<td>11.1 (8.4–14.6)</td>
<td>18.5 (16.8–20.4)</td>
<td>19.6 (17.5–21.8)</td>
<td>14.7 (10.7–19.7)</td>
</tr>
</tbody>
</table>

87
**Age and Measure**

<table>
<thead>
<tr>
<th>Age and Measure</th>
<th>Fatty (BI-RADS 1)</th>
<th>Scattered Fibroglandular Densities (BI-RADS 2)</th>
<th>Heterogeneously Dense (BI-RADS 3)</th>
<th>Extremely Dense (BI-RADS 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulative probability of false positive after 10 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual</td>
<td>34.4 (32.7–36.2)</td>
<td>58.6 (57.5–59.8)</td>
<td>68.1 (67.0–69.2)</td>
<td>65.8 (64.2–67.4)</td>
</tr>
<tr>
<td>Biennial</td>
<td>19.7 (18.7–20.8)</td>
<td>37.1 (36.3–37.9)</td>
<td>45.3 (44.4–46.2)</td>
<td>43.2 (41.9–44.5)</td>
</tr>
</tbody>
</table>

*From Kerlikowske, 2013.*

**Abbreviations:** BCSC = Breast Cancer Surveillance Consortium; BI-RADS = Breast Imaging Reporting and Data System; CI = confidence interval; HRT = hormone replacement therapy

As with the single examination rates, the cumulative 10-year estimates for the U.S. are substantially higher than the lifetime risk estimates from European screening programs. The BCSC investigators noted that estimating lifetime cumulative probabilities “…require[s] extrapolation beyond the length of observation in the current study.”

We discuss some of the difficulties inherent in this extrapolation after presenting the results for false positives biopsy recommendations.

**False Positive: Biopsy**

An abnormal finding on mammography can result in a recommendation for pathological examination to determine the presence of cancer, with the method for obtaining tissue varying from aspiration using a small-bore needle to a more extensive biopsy requiring local, regional, or general anesthesia. Depending on the study, whether or not a woman who received a recommendation for a biopsy after an abnormal mammogram actually underwent a procedure may not be recorded, and, depending on how these women are included in calculations of sensitivity and specificity of mammography, the false positive rate of the screen itself may be under or over-estimated. For example, if the denominator is all screening mammograms with a recorded referral for biopsy, and the numerator is all women undergoing biopsy after a recommendation who did not have cancer detected, the calculated false positive rate would be lower than the rate using only women actually undergoing biopsy if a substantial number of women either never underwent biopsy, or did not have results included. (The same is also true for false positive recall.) The details of these definitions are variable from study to study; in addition, type of biopsy (needle aspiration versus surgical) is often not provided.

For ease of presentation and reading, we refer to “false positive biopsies” throughout the following section, even though, for some studies, “false positive biopsy recommendations” may be more appropriate, and we do not attempt to distinguish between needle aspiration or surgical biopsy.

**Single Examination**

**Non-U.S. Studies**

In a pooled summary of results from 20 screening programs in 17 European countries between 2005 and 2007 (screening ages 50-69, with biennial screens), Hofvind et al. reported overall biopsy rates of 2.2% (range 0.8% to 3.3%) for the initial screen and 1.1% (range 0.3% to 1.5%) for subsequent screens. The ratio of benign to malignant histology was 0.27 (range 0.18 to 0.66) for first screens and 0.11 (range 0.02 to 0.21) for subsequent screens. In subsequent screens, younger women were less likely to undergo biopsy after referral for further assessment,
but the overall positive predictive value of screening was lower. Of those women who did undergo biopsy, the benign-to-malignant ratio was highest (0.22) in women aged 50-54 years (ratio 0.12 in women 55-59 years, 0.10 in women 60-64 years, and 0.08 in women 65-69 years; p=0.07 for trend). This is consistent with an increasing incidence of breast cancer with age. The reported proportion of women undergoing surgical intervention was 0.19% for first examinations and 0.07% for subsequent examinations; it is unclear from the text whether this included needle biopsies or only incisional biopsies.

**U.S Studies**

As seen with false positive recall, false positive biopsy recommendations were higher with first screens than with subsequent screens, and the probability significantly increased with age for first screens and most age categories for subsequent screens. For first screens, false positive biopsy recommendations ranged from 2.0% for 40- to 44-year-olds to 3.0% for 55- to 59-year-olds; for subsequent screens, proportions were 0.8% for 40- to 44-year-olds to 1.5% for women 65 years old and over. In a multivariate analysis, initial screen false positive probabilities were significantly higher with increasing age: using the probability for women aged 40-44 as the reference, odds ratios (95% CIs) were 1.40 (1.24 to 1.57) for 45- to 49-year-olds, 1.75 (1.53 to 2.00) for 50- to 54-year-olds, and 1.48 (1.23 to 1.79) for 55- to 59-year-olds. In contrast to the false positive recall probability, false positive biopsy recommendations significantly increased with age: compared to 40- to 44-year-olds, odds ratios (95% CIs) were 1.19 (1.02 to 1.39) for 45- to 49-year-olds, 1.33 (1.11 to 1.60) for 50- to 54-year-olds, 1.27 (1.01 to 1.59) for 55- to 59-year-olds, 1.09 (0.79 to 1.50) for 60- to 64-year-olds, and 1.91 (1.15 to 3.16) for women 65 years or older.

In addition to age, first screen false positive biopsy probabilities were significantly increased by:

- A family history of breast cancer (3.3% compared to 2.3% in women without a family history; OR 1.47; 95% CI, 1.25 to 1.72).
- Heterogeneously dense breasts (2.6% compared to 2.3% in women with scattered fibroglandular densities; OR 1.12; 95% CI, 1.01 to 1.24); the probability in women with extremely dense breasts was not significantly different (OR 0.98; 95% CI, 0.83 to 1.16) compared to the reference. False positive probabilities were significantly decreased in women with almost entirely fatty breasts (OR 0.67; 95% CI, 0.54 to 0.85).
- In contrast to false positive recall, false positive biopsy rates did not increase over time (2.2% pre-1997 compared to 2.4% after 2004; OR 1.09; 95% CI, 0.84 to 1.42).

As with false positive recall, current hormone replacement therapy was not significantly associated with an increased false positive biopsy probability.

For subsequent screens:

- Family history was not associated with an increased probability of a false positive examination (OR 0.91; 95% CI, 0.73 to 1.12).
- As with first examinations, breast density affected false positive biopsy probability with subsequent screens as well. A high fat content decreased false positive probability relative to scattered fibroglandular densities (OR 0.53; 95% CI, 0.38 to 0.76), while probability was increased with both heterogeneously dense (OR 1.47; 95% CI, 1.28 to 1.68) and extremely dense breasts (OR 1.57; 95% CI, 1.28 to 1.94).
- Again in contrast with false positive recall rates with subsequent screens, false positive biopsy probability did not change over time, from 0.8% pre-1997 to 0.9% after 2004.
As with false positive recall rates, false positive biopsy rates for subsequent screens were significantly associated with the availability of previous films and screening interval:

- False positive biopsy probabilities with subsequent screens were decreased when a comparison mammogram was available, from 1.3% to 0.9% (OR 0.70; 95% CI, 0.52 to 0.93).
- False positives increased as time since last screen increased from 0.8% for an interval of 9-18 months (reference) to 1.0% for 19-30 months (OR 1.22; 95% CI, 1.05 to 1.41) to 1.3% for intervals longer than 30 months (OR 1.60; 95% CI, 1.37 to 1.86).

**Cumulative False Positive Biopsy Probability**

**Non-U.S. Studies**

Estimated cumulative risk of undergoing a biopsy from a pooled analysis of three European studies over 10 rounds of biennial screening in women aged 50-69 years was 2.9% (CIs not reported).8

**U.S. Studies**

In the BCSC multivariate model, overall 10-year cumulative false positive biopsy rate was again associated with screening interval; although cumulative probabilities were approximately 2% higher for women beginning screening at age 50 compared to age 40, confidence intervals overlapped. For women with a first screen at age 40, estimated 10-year cumulative risk of a false positive biopsy was 7.0% (95% CI, 6.1% to 7.8%) for annual screening versus 4.8% (95% CI, 4.4% to 5.2%) for biennial screening. For women with a first screen at age 50, estimated 10-year cumulative risk was 9.4% (95% CI, 7.4% to 11.5%) for annual screening, and 6.4% (95% CI, 5.6% to 7.2%) for biennial screening.92

In contrast to the results for false positive recall (Table 26), there was no apparent interaction between age and hormone replacement therapy status in estimated cumulative false positive biopsy probability—screening interval and breast density were the major determinants in the stratified analysis (Table 27).87

### Table 27. Estimated 10-year Cumulative Probability (95% CI) of False Positive Biopsy in the BCSC by Age, Breast Density, and HRT Status*

<table>
<thead>
<tr>
<th>Age and Measure</th>
<th>Breast Density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fatty (BI-RADS 1)</td>
</tr>
<tr>
<td>Age 40–49:</td>
<td></td>
</tr>
<tr>
<td>First mammography</td>
<td>1.6 (1.3–2.1)</td>
</tr>
<tr>
<td>Cumulative probability of false positive after 10 years</td>
<td></td>
</tr>
<tr>
<td>Annual</td>
<td>5.5 (4.5–6.7)</td>
</tr>
<tr>
<td>Biennial</td>
<td>2.9 (2.4–3.4)</td>
</tr>
<tr>
<td>Age 50–74 (no HRT):</td>
<td>2.4 (2.0–2.9)</td>
</tr>
<tr>
<td>First mammography</td>
<td></td>
</tr>
<tr>
<td>Cumulative probability of false positive after 10 years</td>
<td></td>
</tr>
<tr>
<td>Annual</td>
<td>5.0 (4.5–5.6)</td>
</tr>
<tr>
<td>Biennial</td>
<td>2.8 (2.5–3.1)</td>
</tr>
</tbody>
</table>
Within the BCSC registry, there was substantial variation depending on radiologist, leading to substantial variability in the estimates of cumulative false positive biopsy probability depending on the interaction between an individual woman’s risk (based on age, breast density, family history, and availability of prior examination) and radiologist variability (Table 28). (Variability was similar for false positive recall, ranging from 29.4% for woman at low risk for a false positive with results consistently read by a radiologist in the 25th percentile for false-positive risk screened annually, to 71.6% for the same woman screened annually with readings by a radiologist at the 75th percentile for false positive risk.)

Table 28. Estimated 10-year Cumulative Probability (95% CI) of a False Positive Biopsy in the BCSC by Radiologist and Patient Risk Level*

<table>
<thead>
<tr>
<th>Overall Risk Group</th>
<th>Age 40 at First Mammogram</th>
<th>Age 50 at First Mammogram</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Annual Screening</td>
<td>Biennial Screening</td>
</tr>
<tr>
<td>Overall</td>
<td>7.0 (6.1–7.8)</td>
<td>4.8 (4.4–5.2)</td>
</tr>
<tr>
<td>Radiologist in 25th percentile for false-positive risk:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woman at low false-positive risk</td>
<td>3.2 (2.4–4.0)</td>
<td>2.4 (1.8–3.0)</td>
</tr>
<tr>
<td>Woman at intermediate false-positive risk</td>
<td>5.0 (4.0–6.0)</td>
<td>3.7 (3.1–4.3)</td>
</tr>
<tr>
<td>Woman at high false-positive risk</td>
<td>6.1 (4.9–7.3)</td>
<td>4.5 (3.7–5.3)</td>
</tr>
<tr>
<td>Woman at very high false-positive risk</td>
<td>7.6 (6.0–9.2)</td>
<td>5.6 (4.6–6.6)</td>
</tr>
<tr>
<td>Radiologist in 50th percentile for false-positive risk:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woman at low false-positive risk</td>
<td>3.5 (2.5–4.5)</td>
<td>2.6 (2.0–3.2)</td>
</tr>
<tr>
<td>Woman at intermediate false-positive risk</td>
<td>5.4 (4.4–6.4)</td>
<td>4.0 (3.2–4.8)</td>
</tr>
<tr>
<td>Woman at high false-positive risk</td>
<td>6.7 (5.3–8.1)</td>
<td>4.9 (4.1–5.7)</td>
</tr>
<tr>
<td>Woman at very high false-positive risk</td>
<td>8.3 (6.5–10.1)</td>
<td>6.1 (4.9–7.3)</td>
</tr>
<tr>
<td>Radiologist in 75th percentile for false-positive risk:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woman at low false-positive risk</td>
<td>4.2 (3.2–5.2)</td>
<td>3.0 (2.2–3.8)</td>
</tr>
<tr>
<td>Woman at intermediate false-positive risk</td>
<td>6.4 (5.2–7.6)</td>
<td>4.7 (3.9–5.5)</td>
</tr>
</tbody>
</table>

*From Kerlikowske, 2013.**
### Overall Risk Group

<table>
<thead>
<tr>
<th>Overall Risk Group</th>
<th>Age 40 at First Mammogram</th>
<th>Age 50 at First Mammogram</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Annual Screening</td>
<td>Biennial Screening</td>
</tr>
<tr>
<td>Woman at high false-positive risk</td>
<td>7.9 (6.3–9.5)</td>
<td>5.8 (4.8–6.8)</td>
</tr>
<tr>
<td>Woman at very high false-positive risk</td>
<td>9.8 (7.8–11.8)</td>
<td>7.2 (5.8–8.6)</td>
</tr>
</tbody>
</table>

*From Hubbard, 2011.*

Abbreviations: BCSC=Breast Cancer Surveillance Consortium; CI=confidence interval

Note: False-positive risk profiles are based on multivariable logistic regression models including age, year of first examination, hormone replacement therapy use, family history of breast cancer, breast density, availability of comparison mammogram, registry, and random radiologist intercepts. Risk profiles have year of first examination in 1997–1999, no hormone replacement therapy, and comparison mammogram available at subsequent screenings. Levels were defined as follows: low = no family history of breast cancer, Breast Imaging Reporting and Data System (BI-RADS) 1 breast density; intermediate = no family history of breast cancer, BI-RADS 2 breast density; high = no family history of breast cancer, BI-RADS 3 breast density; very high = family history of breast cancer, BI-RADS 3 breast density.

### Estimating Lifetime Probabilities in the U.S.

As noted above, there are no direct U.S. population-based estimates of the lifetime cumulative probability of a false positive result, either one resulting in a repeat visit alone or one resulting in a biopsy. Estimates from the CISNET investigators described below are derived from observed sensitivity and specificity estimates from the BCSC applied to underlying mathematical models of breast cancer natural history, and are subject to uncertainty inherent in the validity of those models and the parameters that are used; one advantage of this approach is that it does allow for the impact of competing risks on lifetime probability. The multivariate predictive model used in the papers reporting the results of the BCSC does not extend beyond 10 years but could presumably provide a lifetime estimate under different assumptions about false positive probabilities.

To provide a simple estimate based on the observed BCSC data, we use the approach described by the UK Age trial investigators, which includes an assumption that the probability of a false positive at any given examination is independent of previous examinations (which the BCSC data clearly show is not the case and will overestimate the cumulative probability), and calculate the cumulative risk over *n* screening examinations as:

\[
(1 - \text{Probability}_{\text{False Positive First Exam}}) \times (1 - \text{Probability}_{\text{False Positive Subsequent Exams}})^{n-1}
\]

We also assume that the probability of a false positive biopsy on subsequent exam is not related to age (which will underestimate the cumulative probability), although we do vary it based on screening interval as estimated in Hubbard et al.

These results based on the adjusted estimates for first and subsequent false positive recall (Table 29) and biopsies (Table 30) for annual and biennial screening beginning at age 40, 45, and 50, with cumulative probabilities over both 10 years and to a fixed age of 74.
Table 29. Estimated 10-year and Lifetime False Positive Recall Probability by Screening Interval and Age of Starting Screening (Assumes Screening Stops after Age 74), Assuming Independence of False Positive Results at Each Examination, Based on BCSC Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Annual Screening</th>
<th>Biennial Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Age</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>False positive probability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First screen</td>
<td>16.4%</td>
<td>19.9%</td>
</tr>
<tr>
<td>Subsequent screens</td>
<td>8.3%</td>
<td>8.3%</td>
</tr>
<tr>
<td>Cumulative Probability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 years</td>
<td>61.7%</td>
<td>63.3%</td>
</tr>
<tr>
<td>To age 74</td>
<td>95.2%</td>
<td>92.9%</td>
</tr>
</tbody>
</table>

Table 30. Estimated 10-year and Lifetime False Positive Biopsy Probability by Screening Interval and Age of Starting Screening (Assumes Screening Stops after Age 74), Assuming Independence of False Positive Results at Each Examination, Based on BCSC Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Annual Screening</th>
<th>Biennial Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Age</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>False positive probability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First screen</td>
<td>2.0%</td>
<td>2.8%</td>
</tr>
<tr>
<td>Subsequent screens</td>
<td>0.8%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Cumulative Probability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 years</td>
<td>8.8%</td>
<td>9.6%</td>
</tr>
<tr>
<td>To age 74</td>
<td>24.8%</td>
<td>22.4%</td>
</tr>
</tbody>
</table>

With the assumption of independence, 10-year cumulative estimates are higher than those reported by the BCSC (for example, for false positive biopsies with screening beginning at 40, 8.8% here vs. 7.0% for the BCSC for annual screening and 5.9% vs. 4.8% for biennial screening, with similar differences for screening beginning at 50 and for false positive recall). The variation in estimates of absolute differences is similar, although there is less consistency in whether these estimates are higher or lower than those reported for the BCSC (for annual vs. biennial screening for 40-year-olds, we estimate an absolute difference in 10-year false positive biopsy risk of 1.4%, vs. the BCSC estimate of 2.4%; for 50-year-olds, the difference is 2.9% here vs. 1.6%).

Cumulative risks to age 74 are likely to be an overestimate both because of the independence assumption and the presence of competing risks, although some of this overestimation, particularly for false positive biopsy recommendations, would be attenuated by the increasing risk with age.

The main qualitative results here are:
- Accounting for higher false positive probabilities at the time of the first screen and with longer screening interval reduces differences in the cumulative 10-year probability of both false positive recalls and biopsies associated with varying age to start screening and screening interval.
- However, the cumulative effect of an extra 5 to 10 screens over a lifetime still leads to a greater cumulative risk of at least one false positive recall or biopsy when screening starts at younger ages or occurs at more frequent screening intervals.

This is consistent with the qualitative description provided by Hubbard and colleagues: “Over a lifetime of screening, beginning screening 10 years earlier would result in an additional 10 screening mammograms under annual screening and 5 under biennial screening and the lifetime risk for false-positive mammography results will thereby be increased.”"
CISNET Estimates

Joint Effects of Age of Stopping and Starting Screening, and Interval, on False Positive Recall and Biopsies

Figures 13 and 14 illustrate the expected lifetime number of false positives per 100,000 women from the “exemplar” CISNET model, varying age to start (Figure 13) or stop (Figure 14) screening, by annual or biennial screening interval. The models use estimates of sensitivity and specificity from the BCSC, adjusted for age, screening interval, and first versus subsequent examinations. The CISNET models either use these values directly as input variables, for calibration purposes, or to fit test characteristic estimates from both the BCSC and other sources; the “exemplar” model is the one that calibrates its results to the BCSC estimates.
Figure 13. Estimated Number of (A) Total False Positives and (B) False Positive Biopsies by Age to Start Screening (Assuming Screening Ends after Age 69) and Screening Interval

(A) Total False Positives

(B) False Positive Biopsies
Figure 14. Estimated Number of (A) Total False Positives and (B) False Positive Biopsies by Age to Start Screening (Assuming Screening Ends after Age 69) and Screening Interval

A. Total False Positives

B. False Positive Biopsies

Note that estimated rates of total false positives and false positive biopsies are much more sensitive to age of starting screening than age of stopping screening (the slopes of the lines, which represent the incremental difference between two ages, is steeper for extending to younger ages). The slope is also steeper for annual screening compared to biennial.

The estimates in these tables suggest that screening interval has a greater effect on false positives than age alone, but rates go up much more rapidly with earlier age to start than later age to stop. Although not directly comparable, the results here do illustrate the effect of including competing risks on the lifetime estimate: estimated false positive biopsy rates in this CISNET
model for ages 50 to 74 are 11.0% for annual screening and 6.6% for biennial screening, while our crude estimates, which do not include competing risks, are 19.8% for annual screening and 13.6% for biennial screening (although the absolute difference in lifetime estimates is similar). Estimates for cumulative false positive recall are higher with the CISNET model (and in fact are greater than 100% for almost all scenarios where screening begins at 50 or younger, except for biennial screening beginning at 50), presumably because each false positive is counted (including multiple false positives in the same patient), whereas the estimate used in our crude analysis was based on the probability of “at least one” false positive, which would not count multiple false positive results in the same woman.

Discussion/Conclusions: False Positives

This discussion emphasizes conclusions drawn from the available U.S.-based evidence, primarily from the BCSC; because of the substantially higher rates of false positives (both recall and biopsy) for both first and subsequent screens in the U.S. compared to European studies, the applicability of quantitative estimates derived from studies performed outside the U.S. to estimations of outcomes within the U.S. is extremely limited.

- Screening with currently available mammography inevitably results in false positive results, some of which result in invasive procedures, including biopsies.
- False positive results have measurable emotional impact, which may be long-lasting in some women (see discussion under Quality-adjusted Life Expectancy).
- The likelihood of a false positive result, whether recall or biopsy, is highest at the time of the first screen, but decreases with subsequent screens.
- However, the likelihood of a false positive biopsy recommendation on subsequent examinations increases with age; the effect of this on cumulative probability of a false positive biopsy over extended periods of time is not clear.
- There is substantial direct and indirect evidence that the probability of a false positive result, whether one resulting only in additional radiologic examination or one resulting in a biopsy, is influenced by the radiologist reading the film:
  - Under a Bayesian model of screening and diagnosis, a radiologist refers a patient for further evaluation when the post-mammography probability of breast cancer is above some threshold; the threshold for referral to biopsy is higher than for recall. The post-test probability of cancer is a function of the sensitivity and specificity of the screening test, and the pre-test probability of cancer—the likelihood that a given patient has cancer. The higher the pre-test probability of cancer, the greater the post-test probability at any fixed level of sensitivity and specificity. There are patient-specific factors that increase false positive recalls (family history of breast cancer) and biopsy (family history of breast cancer, age) at both first and subsequent screens. These factors ARE associated with an increased risk of a prevalent cancer and should increase post-test probability (and therefore false positive rates). The fact that they are associated with an increased false positive risk suggests that their effect on radiologists’ threshold for referral is substantially greater than the quantitative association would suggest—in other words, some radiologists may overestimate the pre-test probability of cancer based on these risk factors.
Factors that in effect increase the precision of the radiologists’ estimate of the prior probability of disease (first examination vs. subsequent examination, availability of prior examinations) also reduce the false positive rate.

The consistent association between increasing screening interval and increased per-examination false positive probability (for both recall and biopsy) is consistent both with an increased pre-test probability of disease (a longer interval increasing the chance of a new cancer) and with decreased precision of the estimate (the potential consequences of a false negative reading would be greater with a longer screening interval, so the need for certainty is increased).

The estimated cumulative 10-year risk varies widely when the variability across individual radiologists’ variation is taken into account. Taken together, this evidence means that, even if a more precise estimate of the risk of a false positive recall or biopsy were available based on high quality population-based data, potential variation in who will be interpreting a given screening test means that there is substantial uncertainty about the cumulative risk of a false positive result for an individual woman from this source alone. Given relatively high geographic mobility, high turnover in insurance coverage, and potential turnover regarding which radiologists are covered by which payer, this is particularly the case for women not covered by Medicare.

- False positive probabilities for both recall and biopsy increased substantially in the U.S. from pre-1997 to the period after 2004. If this trend is continuing, then, as with estimations of future cancer incidence and mortality, uncertainty about false positive probabilities increases with time horizons for future predictions (i.e., given the same estimates, predictions about outcomes in 20 years are more uncertain than predications over the next 5-10 years).

- Although the 10-year probability of a false positive recall or biopsy appears similar when screening begins at age 40, 45, or 50 (because of differences in age-specific false positive rates with first examination), the cumulative effect of an additional 5-10 screens means that earlier ages for starting screening will result in higher lifetime false positive risks for any given fixed stopping age. This effect would be attenuated if false positive results decreased with increasing number of previous negative examinations, but there is no evidence to suggest that; the significant association between increasing age and increased false positive biopsy probability with subsequent examinations suggests that the effect of age on pre-test probability may outweigh any effects of a long history of negative examinations. Much of the effect of younger age on false positive probability appears to be related to breast density, rather than age alone.

- Similarly, although the per-screen probability of false positive biopsy or recall decreases with as screening interval shortens, this is not enough to compensate for the cumulative effect of a larger number of lifetime screens on the cumulative risk of false positive biopsy or recall.

- We judge the quality of evidence that, qualitatively, the lifetime risk of a false positive recall or biopsy increases with younger age to start screening or with more frequent screening as HIGH, based on consistency across study designs and settings. For women in the U.S., quality of evidence for estimates of the magnitude of the cumulative false positive rate over a relative short time horizon (up to 10 years) is MODERATE; results are relatively consistent, particularly for absolute differences between different strategies.
However, there is (a) uncertainty about future trends in test performance, (b) substantial methodological limitations associated with estimations of lifetime risk, and (c) substantial variability in false positive rates between radiologists which, subsequently, may lead to potential variation over a woman’s lifetime in the per-examination risk of a false positive as geographic mobility, insurance coverage, and providers covered by that insurance change. Therefore, we judge the uncertainty surrounding the cumulative probability of a false positive recall or biopsy to be high, and the quality of evidence for the magnitude of an individual woman’s lifetime risk associated with different screening strategies to be LOW.

Quality-adjusted Life Expectancy

Quality-adjusted life expectancy (measured in quality-adjusted life-years, or QALYs) is a measure which integrates the effects of different health interventions on both mortality (through estimates of life expectancy) and morbidity (through adjustments for quality-of-life preferences). In theory, quality-adjusted life expectancy captures both benefits and harms in a single measure, facilitating comparisons between strategies; for this reason, it is the recommended standard denominator for use in cost-effectiveness analysis.\(^{168}\)

Quality-adjusted life expectancy is calculated by defining a set of relevant health states—for example, no breast cancer, DCIS, and local, regional, and distant invasive cancer. A weight (utility) is assigned to each state relative to “perfect” health (a value of 1.0) and death (a value of 0), using one or more of a range of standard instruments for capturing relative preferences. The state-specific weight is then applied to the measured or estimated duration of time spent in each state to estimate quality-adjusted life expectancy. If the mean survival with distant invasive breast cancer is 3 years and the utility is 0.6, then the quality adjusted life expectancy is 3\(\times\)0.6, or 1.8 QALYs. “Disutility” is sometimes used to refer to the decrement in utility associated with the health state—if the utility measurement is 0.6, the disutility is 1 minus 0.6, or 0.4.

Utilities can be assessed in the general population using stated preference methods such as the time trade-off or standard gamble, or they can be collected from patients as part of a research protocol using instruments such as the European Quality of Life-5 Dimensions (EQ-5D). If direct measurement with an instrument such as the EQ-5D is used, then quality-adjusted life expectancy could be directly estimated if all subjects are followed to death. More typically, utility values are used in conjunction with models to estimate the expected QALYs associated with different strategies; in this case, quality-adjusted life expectancy is, by definition, an indirect measure, subject to the same limitations as model-based estimates of life expectancy or other outcomes, with the additional need to ensure that the utility weights are appropriate for a given population.

Utility Weights used in Estimates of the Effect of Screening on Quality-adjusted Life Expectancy

Before describing the reported estimated effects of screening on quality-adjusted life expectancy in the CISNET models and in the University of California at San Francisco (UCSF) model based on Breast Cancer Surveillance Consortium data,\(^{159}\) it is worth discussing the utility weights used in these estimates.

Studies from the CISNET collaborators that estimate QALYs\(^{156,157,169}\) use two sources for utility weights. First, age- and sex-specific EQ-5D scores from the 2000 Medical Expenditure Panel Survey (MEPS)\(^{170}\) were used to establish “healthy” QALYs. Weights for screening...
attendance and diagnostic evaluation were obtained from a 1991 Dutch literature review, while weights for in situ, localized, and distant cancer were apparently assigned by the investigators “consistent with treatment-specific quality-of-life weights reported in other studies,” referencing a 2000 review of utility weights across oncology, which noted substantial methodological weaknesses in the utility measures.

Another U.S.-based model from the Breast Cancer Surveillance Consortium used directly measured EQ-5D values, but these values were from a 2007 study of Swedish patients, and the rationale for some of the assumptions about duration of the health state is not clear (for example, a diagnosis of DCIS is associated with a decreased utility only for the first year after diagnosis, which is not consistent with the experience of many patients).

Table 31 presents these utility weights for both sets of models.

Table 31. Utility Weights Used to Estimate QALYs in CISNET and UCSF BCSC Models

<table>
<thead>
<tr>
<th>State</th>
<th>CISNET Utility</th>
<th>1-Utility</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screening</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CISNET</td>
<td>0.994</td>
<td>0.006</td>
<td>1 week</td>
</tr>
<tr>
<td>UCSF</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Work-up of Abnormal Result</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CISNET</td>
<td>0.895</td>
<td>0.105</td>
<td>5 weeks</td>
</tr>
<tr>
<td>UCSF (False positive only)</td>
<td>0.987</td>
<td>0.013</td>
<td>?1 year</td>
</tr>
<tr>
<td>DCIS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CISNET</td>
<td>0.90</td>
<td>0.10</td>
<td>2 years</td>
</tr>
<tr>
<td>UCSF: Year 1</td>
<td>0.904</td>
<td>0.096</td>
<td>1 year</td>
</tr>
<tr>
<td>UCSF: Subsequent years</td>
<td>1</td>
<td>0</td>
<td>?Until Death</td>
</tr>
<tr>
<td>Local Invasive Cancer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CISNET</td>
<td>0.90</td>
<td>0.10</td>
<td>2 years</td>
</tr>
<tr>
<td>UCSF: Year 1</td>
<td>0.846</td>
<td>0.154</td>
<td>1 year</td>
</tr>
<tr>
<td>UCSF: Subsequent years</td>
<td>0.98</td>
<td>0.02</td>
<td>?Until Death</td>
</tr>
<tr>
<td>Regional Invasive Cancer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CISNET</td>
<td>0.75</td>
<td>0.25</td>
<td>2 years</td>
</tr>
<tr>
<td>UCSF: Year 1</td>
<td>0.753</td>
<td>0.247</td>
<td>1 year</td>
</tr>
<tr>
<td>UCSF: Subsequent years</td>
<td>0.905</td>
<td>0.095</td>
<td>?Until Death</td>
</tr>
<tr>
<td>Distant Invasive Cancer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CISNET</td>
<td>0.60</td>
<td>0.40</td>
<td>until death</td>
</tr>
<tr>
<td>UCSF: Year 1</td>
<td>0.753</td>
<td>0.247</td>
<td>1 year</td>
</tr>
<tr>
<td>UCSF: Subsequent years</td>
<td>0.832</td>
<td>0.168</td>
<td>?Until Death</td>
</tr>
</tbody>
</table>

Abbreviations: BCSC=Breast Cancer Surveillance Consortium; CISNET=Cancer Intervention and Surveillance Modeling Network; DCIS=ductal carcinoma in situ; NR=not reported; QALYs=quality-adjusted life-years; UCSF=University of California at San Francisco

Effects of Parameters and Assumptions on Estimates of Quality-Adjusted Life Expectancy

Because of variability in the utility estimates, as well as differences in the models, we focus here on author-reported qualitative effects of different assumptions and parameter values on estimates of quality-adjusted life expectancy. The most important of these were:

- The small disutility associated with undergoing screening had a major effect on QALYs, particularly for more frequent screening strategies.
- The disutility of false positive results had a substantial effect on QALYs, enough to raise the estimated cost/QALY above $100,000 from a base case value of $72,000. This may underestimate the effects, since there is consistent evidence in the literature that some measures of the emotional impact of false positive results may persist for at least a year in a substantial proportion of women, affecting subsequent screening behavior, although
this effect, which may be more cancer-specific, may not be observable with standard measures of generalized anxiety or utility.\textsuperscript{175}

- Quality-adjusted life expectancy was affected by assumptions about overdiagnosis in all models that included overdiagnosis,\textsuperscript{156,157,159,169} although the quantitative relationship between overdiagnosis and QALYs was not presented, only qualitative statements such as, “We found that the harm-benefit ratio QALYs lost/life-year gained was sensitive to the amount of overdiagnosis with an increasing number of QALYs lost with an increasing amount of overdiagnosis.”\textsuperscript{157}

**Discussion/Conclusions: Effects of Screening on Quality-adjusted Life Expectancy**

- The utility measures used for estimating quality-adjusted life expectancy in U.S. model-based studies are limited by either derivation from non-U.S. populations, who may have quite different preferences, or by lack of any patient- or general population-based estimate. In addition, assumptions about the duration of the impact of relevant states are not empirically supported.

- Despite these limitations, common events that have small and short-term effects on utilities still have a major effect on overall quality-adjusted life expectancy, which decreases with frequency of screening and the probability of false positive results; the magnitude of this decrease is affected by the magnitude of the disutility.

- Quality-adjusted life expectancy is decreased by overdiagnosis, which is intuitive. Since overdiagnosed cancers would, by definition, not lead to a breast cancer death, patients experience the disutility of diagnosis and treatment with no gain in life expectancy. The impact of overdiagnosis on quality-adjusted life expectancy is dependent not only on the estimate of the rate of overdiagnosis, but also the magnitude and duration of the disutility of treatment of DCIS or small localized invasive cancer, the age at which the diagnosis occurs, and, critically, the ratio of overdiagnoses to cancer deaths prevented: if this ratio is substantially above 1.0 and women with overdiagnosed cancers are on average younger than the age at “death” for prevented cancer deaths, then it is possible that screening strategies which increase the risk of overdiagnosis relative to reductions in mortality would result in a net decrease in quality-adjusted life expectancy compared to strategies which prevented fewer deaths but also had fewer overdiagnoses.

- Although the qualitative effects of these parameters on quality-adjusted life expectancy are plausible and consistent, we judge the quality of evidence for the magnitude of the effect of different screening strategies on quality-adjusted life expectancy to be LOW, based on the inherent uncertainties in the underlying estimation of life expectancy, the critical uncertainty about the rate of overdiagnosis, and the limitations of the available utility weights.

**Harm-benefit Trade-offs**

Estimating the quantitative trade-off between the benefits of screening and the potential harms to inform recommendations for screening for U.S. women is inherently difficult, due to:

- The inherent uncertainty in the estimate of the relative reduction in mortality attributable to screening for U.S. women, given the considerations discussed above (including
generalizability of results from non-U.S. studies, both randomized and observational, to the U.S. setting).

- The even greater uncertainty about the absolute reduction in mortality expected for a given relative reduction, given the lack of population-based data for estimating breast cancer incidence and mortality in the absence of screening over the next 10-20 years.

- The uncertainty surrounding estimates of overdiagnosis. In particular, the lack of any reliable estimate for the proportion of screen-detected DCIS that would ultimately develop into symptomatic invasive cancer is a major driver of uncertainty about the risk of overdiagnosis associated with screening in the U.S., given that the U.S. has the highest rates of DCIS among countries with active screening.

- Although there is generally less uncertainty about estimates of false positive recall and biopsy with different screening strategies (certainly less than there is for overdiagnosis), the wide range of cumulative risk based on individual women’s risk factors and variability in radiologist thresholds means that estimates at the population-level may not capture the uncertainty for an individual woman.

Since the trade-off between benefits and harms is frequently expressed as a “harm-benefit” ratio (analogous to a cost-benefit ratio---false positive biopsies per breast cancer death prevented, overdiagnoses per breast cancer death prevented), the uncertainty surrounding the estimates of each component in the numerator and the denominator is propagated in the estimate of the ratio. The estimate of the harm-benefit ratio has, or should have, confidence intervals around it that reflect the uncertainty about the quantitative estimates of benefits and harms.

This uncertainty has generally not been systematically discussed or addressed, either in individual studies, in reviews, or in guidelines recommendations. In addition, there is a notable lack of consensus (or even an attempt to develop one) about the definition of an acceptable threshold for a particular trade-off. Guidelines developers have generally not explicitly stated their threshold, or the criteria for identifying such a threshold, at which a recommendation or the strength of recommendation, for or against a specific policy would change.

In this section, we discuss published estimates of these trade-offs for the U.S. population and provide some additional estimates using a range of “simple” approaches. There are limitations to these approaches, as well as to the available evidence, and we fully acknowledge that other approaches could result in different estimates (both for mean ratios and the uncertainty around them). Our purpose in presenting these results is not to provide a definitive analysis, but to illustrate the effects of uncertainty surrounding the individual outcome estimates on the uncertainty in the estimate of the trade-off. We believe that since formal guidelines processes such as GRADE explicitly call for weighing the balance of benefits and harms, exploring the effect of uncertainty about the evidence for individual benefits and harms on the estimate of that balance, as well as the effect of different methodological approaches to generating estimates of the balance, provides useful background.

Our basic approach is explicitly derived from economic analysis. A “harm-benefit” ratio is analogous to a cost-effectiveness ratio: a strategy is preferred relative to another if it results in greater benefit or effectiveness at an acceptable “price” in terms of harms or monetized costs. In cost-effectiveness analysis, the preferred approach to comparing relative costs and effectiveness between two options is the incremental cost-effectiveness ratio (ICER). Given two options, A and B, and assuming Option B is more expensive than Option A, the ICER is defined as:
Option B is preferred if the ICER is at or below the maximum “willingness-to-pay” threshold in terms of dollars per unit of effectiveness gained.

In the context of developing recommendations for breast cancer screening, estimates are needed for the incremental ratio of critical harms (false positives, especially false positive biopsies, and overdiagnoses) and critical benefits (particularly breast cancer deaths prevented) between available options, along with some measure of the uncertainty surrounding this estimate (expressed as the probability that the “true” estimate is below or above that threshold). The question of what that threshold should be, and the degree of certainty required to formulate a specific recommendation, is a judgment which must be made by those developing the recommendations.

In the following sections, we discuss the available evidence for the specific trade-offs of false positives (both recall and biopsy) per breast cancer death prevented, and overdiagnoses per breast cancer death prevented, again with an emphasis on estimates applicable to the U.S. population.

**False Positives per Breast Cancer Death Prevented**

**Model-based Estimates: Ages to Start and Stop Screening and Screening Interval**

The published CISNET estimates of the benefits and harms of different screening strategies used to inform the 2009 USPSTF recommendations present graphs of number of mammograms per death prevented, or per life-year saved, and tables of estimates of the number of expected false positive recalls and biopsies with different strategies compared to no screening, but do not directly provide incremental values. We have presented results for specific outcome graphically in the previous sections. Here, we use the published estimates of the “exemplar model” (Table 4 in Mandelblatt et al.30) to generate incremental harm-benefit ratios for false positives (both total and biopsy only) per breast cancer death prevented. For simplicity, we assume that biennial screening starting at age 50 is an “acceptable” strategy and compare only annual and biennial screening beginning at ages 40, 45, and 50, assuming screening stops after age 74 (the constraints of the data presented in the paper).

Table 32 presents false positive recall, false positive biopsies, and deaths prevented for each strategy in ascending order of false positives (i.e., starting with the least “expensive” alternative). Incremental ratios are calculated in three ways. First, incremental ratios are calculated for each screening option compared to the next least “expensive” option (for example, biennial screening starting at age 45 compared to biennial screening at age 50). Next, options which are “dominated” (more false positives but fewer deaths prevented) are removed, and the incremental ratio recalculated; in this example, because biennial screening beginning at age 40 results in more false positives with fewer deaths prevented than biennial screening beginning at age 45, biennial screening at age 40 is removed, and the incremental ratio between annual screening at 50 and biennial screening at age 45 is calculated. Finally, options can be eliminated through “extended dominance.” The recalculated incremental false positive biopsy ratio between annual screening at age 50 and biennial screening starting at age 45 is 19, while the incremental ratio between biennial screening at age 45 and biennial screening at age 50 is 24. Implicitly, if a ratio
of 24 is acceptable, then a ratio of 19 is acceptable, and a decision maker willing to adopt biennial screening at age 45 at a false positive biopsy per deaths prevented ratio of 24 would also be willing to adopt annual screening at age 50 with a ratio of 19 (more deaths prevented at an “acceptable” cost). After removing biennial screening at age 45, the ratio is recalculated between annual screening beginning at 50 and biennial screening beginning at 50.

Table 32. Incremental False Positive Recalls and Biopsies per Breast Cancer Death Prevented, by Age to Start Screening and Screening Interval (Assuming Screening Stops after Age 69), Calculated from CISNET “Exemplar Model” Results.30 Shaded areas identify strategies eliminated by dominance and extended dominance (see text for explanation).

<table>
<thead>
<tr>
<th>Strategy (Interval, Starting Age)</th>
<th>Outcomes per 100,000 Women</th>
<th>Incremental False Positives/Death Prevented (Compared to Preceding Strategy)</th>
<th>Incremental False Positives/Death Prevented (Eliminating Dominated* Strategies)</th>
<th>Incremental False Positives/Death Prevented (Eliminating Dominated and Extended Dominated† Strategies)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biennial, 50</td>
<td>78,000</td>
<td>5500 540 144 10 144 10 144 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biennial, 45</td>
<td>105,000</td>
<td>7400 620 338 24 338 24 - -</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biennial, 40</td>
<td>125,000</td>
<td>8800 610 -2000 -140 - - - -</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual, 50</td>
<td>135,000</td>
<td>9500 730 83 6 273 19 300 21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual, 45</td>
<td>180,000</td>
<td>12,600 800 643 44 643 44 643 44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual, 40</td>
<td>225,000</td>
<td>15,800 830 1500 107 1500 107 1500 107</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Strategies that have higher false positives but fewer deaths prevented than alternative strategy with fewer false positives.

†Strategies that have an incremental ratio lower than an alternative strategy with fewer false positives.

Figure 15 presents these results graphically for false positive biopsies and breast cancer deaths prevented; the figure for false positive recalls is identical, except for the values on the x-axis. The slope of the lines connecting the included strategies is equivalent to the incremental harm-benefit ratio.
Figure 15. False Positive Biopsies and Breast Cancer Deaths Prevented, by Age to Start Screening and Screening Interval (Assuming Screening Stops at Age 69). Line connects strategies remaining (biennial screening at 50, annual screening at 50, 45, and 40) after elimination through dominance and extended dominance.

Figure 16 presents the results after using the same approach for age to stop screening (assuming screening begins at 50). In this case, the only strategies remaining after eliminating dominated strategies are biennial screening ending at age 84 (false positive recalls per death prevented 118, incremental biopsies per death prevented 8) and annual screening ending at age 84 (incremental false positive recalls per death prevented 188, incremental biopsies 20).
Figure 16. False Positive Biopsies and Breast Cancer Deaths Prevented, by Age to Stop Screening and Screening Interval (Assuming Screening Stops at Age 50). Line connects strategies remaining (biennial screening stopping after age 84 and annual screening stopping after age 84) after elimination through dominance and extended dominance.

Qualitatively, because the estimated number of deaths prevented by extending screening past age 70 is substantially greater than the estimated number of deaths prevented by extending screening to younger ages (because of an absolute smaller number of deaths in younger women), the incremental ratios for extending screening to older women using this specific metric are smaller than the incremental ratios for extending screening to younger women.

This point is illustrated graphically by comparing the slopes of the curves for false positive biopsies as age is extended to younger or older ages (Figure 17A and 17B) to the slopes for deaths prevented (Figure 17C and 17D).

Subsequent to the USPSTF estimates, updated analyses from the CISNET investigators have provided incremental estimates for overall false positives per death prevented (but not false positive biopsies prevented). One analysis, discussed in more detail under Key Question 2, identified thresholds for breast cancer relative risk where screening women under 50 would result in similar harm-benefit ratios to biennial screening for women aged 50-74 (median total false-positive per death prevented ratio compared to no screening across 5 models 146, range 128-151):157

- Biennial screening at age 40 compared to biennial screening at age 50: median 393, range 363-896)
- Annual screening at age 40 compared to annual screening at age 50: median 1030, range 567-1579)

Mixed strategies, such as annual screening from ages 40-49 with biennial screening from ages 50-74 were not evaluated.
Figure 17. False Positive Biopsies and Deaths Prevented by Age to Start Screening (A and C) and Age to Stop Screening (B and D, Biennial (Solid Line) vs. Annual (Dotted Line) Screening. Slopes of lines represent changes in absolute numbers of outcomes with change in age to start or stop; distance between two lines represents difference between annual and biennial screening at any given age.
A more recent analysis retrospectively estimated the cost-effectiveness of introducing digital mammography into the U.S., using biennial film mammography from ages 50 through 74 as the reference case.\textsuperscript{158} Although the analysis did not explicitly estimate harm-benefit ratios, focusing on cost per quality-adjusted life-year as the primary metric, estimates of the median and range for false positives and deaths prevented across the five models were reported (although separate estimates for false positive biopsies were not included) (Table 33).

**Table 33. Incremental False Positives per Death Prevented with Different Strategies for Use of Digital Mammography (Median Estimates Across 5 CISNET Models for Each Outcome)**\textsuperscript{158}

<table>
<thead>
<tr>
<th>Strategy (Technology, Age to Start and Stop, Interval)</th>
<th>Outcomes per 100,000 Women Screened</th>
<th>Incremental False Positives/Death Prevented (Compared to Preceding Strategy)</th>
<th>Incremental False Positives/Death Prevented (Eliminating Dominated* Strategies)</th>
<th>Incremental False Positives/Death Prevented (Eliminating Dominated and Extended Dominated† Strategies)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Film 50-74 Biennial</td>
<td>89,100</td>
<td>580</td>
<td>154</td>
<td>154</td>
</tr>
<tr>
<td>Digital 50-74 Biennial</td>
<td>111,100</td>
<td>680</td>
<td>220</td>
<td>220</td>
</tr>
<tr>
<td>Digital 40-74 Biennial</td>
<td>174,000</td>
<td>760</td>
<td>788</td>
<td>788</td>
</tr>
<tr>
<td>Digital 50-74 Annual</td>
<td>189,400</td>
<td>780</td>
<td>765</td>
<td>765</td>
</tr>
<tr>
<td>Digital 40-49, Annual</td>
<td>222,500</td>
<td>840</td>
<td>552</td>
<td>552</td>
</tr>
<tr>
<td>Digital 40-74 Annual BI-RADS 3, 4* Biennial BI-RADS 1, 2</td>
<td>237,900</td>
<td>900</td>
<td>257</td>
<td>257</td>
</tr>
<tr>
<td>Digital 40-74 Annual</td>
<td>301</td>
<td>980</td>
<td>794</td>
<td>794</td>
</tr>
</tbody>
</table>

*BI-RADS: Breast Imaging Reporting and Data System breast density categories: BI-RADS 1=mostly fatty tissue, BI-RADS 2=scattered areas of fibroglandular density, BI-RADS 3=heterogeneously dense breasts, BI-RADS 4=extremely dense breasts. Film mammographic sensitivity is decreased in women with BI-RADS 3 and 4.

The analysis included tailored strategies to account for reduced sensitivity of screening in either younger women (annual screening for women 40-49 with biennial screening afterwards), or women with denser breasts (annual screening for women 40-74 with Breast Imaging Reporting and Data System [BI-RADS] density categories 3 and 4, biennial screening for women 40-74 with BI-RADS 1 or 2). The results are not directly comparable to the 2009 analysis of film-only strategies, but the qualitative results are similar—the cumulative effects of more frequent screening on false positives increase at a greater rate than the reduction in number of deaths.

Key points about these analyses include:
- Qualitatively, for this specific trade-off, decreasing the interval from biennial to annual, and/or extending screening to younger ages, increases the estimated false positive
probability for both recall and biopsy at a faster rate than the decrease in the number of estimated deaths. Although there is substantial uncertainty about the absolute values, these qualitative results are consistent across a wide range of models using a relatively wide range of approaches.

- If harm-benefit ratios are to be used to assist with decision making, either at the individual level or in formulating recommendations or policies, then an incremental approach identical to the one used in cost-effectiveness analysis should be used, even if only for comparative purposes. There is no reason that the principles of dominance and extended dominance cannot be applied to harm-benefit analysis. As the results in the tables above show, this approach can lead to different ways of thinking about alternative strategies—for example, it is not immediately intuitive that, if the harm-benefit ratio associated with biennial film screening beginning at age 50 is acceptable, then only annual screening at age 50 or younger needs to be considered as an alternative because biennial screening at younger ages is eliminated through extended dominance.

- Because some women may experience more than one false positive result over a lifetime of screening, the cumulative total for a given population typically exceeds the size of the population with longer screening duration, especially with annual screening under an assumption that the probability of a false positive in a given woman with a given set of risk factors for a false positive is independent of a prior history of a false positive result. At the population level, using false positives per death prevented as a measure of one particular harm-benefit trade-off is reasonable. However, at the individual level, the trade-off may be different, depending on the distribution of false positives. For example, the cumulative false probability estimate from the original CISNET estimates for annual screening beginning at age 40 and ending after age 69 is 225,000 per 100,000. Although this is equivalent to a mean number of false positives per woman screened of 2.25, some women will experience no false positives, most only one, and relatively small number multiple false positives.

- Although the results as presented are useful for identifying qualitative trends, they do not capture the inherent uncertainty in the estimates, either within individual models or across all models. The wide range for mean estimates for false positive probabilities and deaths prevented across individual models implies that the harm-benefit ratios may vary—especially when there is lack of consensus about an appropriate threshold for a given harm-benefit, a more complete description of the variability in the estimates would be helpful.

One approach for displaying both the quantitative uncertainty around the harm-benefit ratio and the effect of varying thresholds for a value of the ratio that would change a particular decision is the use of harm-benefit acceptability curves. In the next section, we present some exploratory analyses using this approach.

**Harm-benefit Acceptability Curves**

The following figures represent the results of probabilistic (Monte Carlo) analyses of the simple Markov model described in Appendix C. The model is run as a two-dimensional analysis, drawing from the distributions of key variables, in particular estimates of mortality reduction, overdiagnosis, and false positive probability, and varying other parameters such as age to start screening or stop screening.
We simulated a cohort of U.S. women from age 40 to 100, under a variety of scenarios:

- Screening beginning at ages 40, 45, or 50 and continuing through age 74, or screening beginning at age 50 stopping after ages 74, 79, or 84.
- Mortality reductions attributable to screening of 0.62 (95% CI, 0.56 to 0.59), based on the pooled results of observational studies of incidence-based mortality, and 0.80 (95% CI, 0.73 to 0.89) based on the meta-analysis of RCTs performed for the UK Independent Panel. Within the simulation, the mortality reduction is modeled as a hazard ratio applied to the conditional probability of dying of breast cancer given age at diagnosis during each year after diagnosis (SEER*Stat). The age-specific probability of breast cancer death in the cohort is the sum of the number of deaths occurring among women of that age from breast cancer diagnosed from age 40 through that age, divided by the number of women alive at that age—in other words, the incidence-based age-specific mortality.
- Per-screen false positive rates adjusted for first versus subsequent screen, age at screening (for initial total false positives and biopsies, and subsequent biopsies—age was not a significant predictor of subsequent false positive recalls—and screening interval taken from the BCSC data. For the results shown below, we assume biennial screening.
- For the results shown here, we restricted the pool of women at risk for a false positive only to those who had not previously had a false positive result—this results in an estimate of the proportion of women having one or more false positive results, rather than the total number of false positives. The cumulative probability of either type of false positive outcome can never be above 100% in this case.

For all models, the estimated cumulative probability of breast cancer death from age 40 to 100 was approximately 3.2% (reported estimates in the CISNET models are approximately 3.0%). Estimates for the different screening strategies under different screening effectiveness are shown in Table 34 (as described in Appendix C, mortality reductions attributable to screening continue after screening stops; for women diagnosed after the cessation of screening, there is no mortality benefit, so the risk of breast cancer death is the same, resulting in a slight decrease in overall mortality reduction by extending follow-up over a lifetime).

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Mortality RR 0.62</th>
<th>Mortality RR 0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screen Ages 50-74</td>
<td>1.6%</td>
<td>2.50%</td>
</tr>
<tr>
<td>Screen Ages 45-74</td>
<td>1.8%</td>
<td>2.58%</td>
</tr>
<tr>
<td>Screen Ages 40-74</td>
<td>1.9%</td>
<td>2.83%</td>
</tr>
</tbody>
</table>

The X-axis varies the “acceptable” harm-benefit ratio for a given set of benefits and harms, starting at 0; this is analogous to the “willingness to pay (WTP)” in cost-effectiveness analysis. The Y-axis represents the proportion of simulations where a given option is optimal at a given WTP. If the WTP is 0, then the option with the smallest probability of harms is preferred. As the WTP increases (for example, as the willingness to accept the number of false positives for every breast cancer death prevented increases), the probability that options with higher ratios would be optimal increases. An alternative way to understand the acceptability curves is that the X-axis represents the incremental harm-benefit ratio of one strategy compared to the next least harmful (or “expensive”) strategy, and the Y axis represents the cumulative density function for that
ratio; when only two strategies are being compared, the point on the X-axis where the lines cross at 50% on the Y axis represents the median of the harm-benefit ratio—there is a 50% probability that the “true” incremental ratio is less than that value, and a 50% probability that is greater than that value. For example, if the value on the X axis at 50% on the Y axis is 10, and at 90% the X value is 20, then there is a 10% chance that the “true” ratio is greater than 20. If 20 represents the upper limit of an acceptable threshold, then, based on the evidence and assumptions that went into estimating the ratio, choosing that strategy would result in a 10% chance of making a “wrong” decision.

Figure 18 presents acceptability curves for age to begin screening of 40, 45, and 50 years with a stopping age of 74, at mortality reduction of 0.62 and 0.80, for false positive biopsies and total false positives.
Figure 18. Harm-benefit Acceptability Curves for False Positive Biopsies (A and B) and Total False Positives (C and D) by Age to Start Screening and Mortality Reduction

A. False Positive Biopsies, Mortality Reduction 0.6

B. False Positive Biopsies, Mortality Reduction 0.8

C. Total False Positives, Mortality Reduction 0.6

D. Total False Positives, Mortality Reduction 0.6
Key qualitative results include:

- Ratios are highly dependent on uncertainty surrounding the mortality reduction (the denominator in the ratio).
- For both types of false positive outcomes, screening beginning at age 45 is eliminated by extended dominance, so that the alternatives become screening beginning at 50 versus 40.
- Incorporating age-dependency on the probability of false positives affects strategies—for total false positives, lack of an age effect on the likelihood of subsequent false positives results in elimination of screening starting at age 50 by extended dominance.
- Restricting false positives only to women who have not had a previous false positive within the model results in substantially smaller cumulative risks, especially for total false positives. Lifting the restriction increased mean ratios by approximately 100-150 false positives per death prevented.

Figure 19 illustrates acceptability curves for age to stop screening for each mortality reduction estimate for false positive biopsies.
Figure 19. Harm-benefit Acceptability Curves for False Positive Biopsies by Age to Stop Screening and Mortality Reduction

A. False Positive Biopsies, Mortality Reduction 0.6

B. False Positive Biopsies, Mortality Reduction 0.8
Key qualitative results include:

- Incremental ratios for extending screening beyond age 74 are higher because of a higher overall false positive rate at older ages, and because, as discussed earlier, the risk of competing mortality is very high above age 74, even when accounting for the risk reduction attributable to screening.

Evidence on Patient Preferences for False Positive versus Death Prevented Trade-off

We identified one study that provides explicit evidence on U.S. patient values on the trade-off between false positive results and breast cancer mortality, and another which, while not directly measuring preferences for false positives versus mortality prevention, does provide some evidence on preferences for false positives relative to other aspects of mammography. In 1997, Schwartz and colleagues conducted a national mail survey of 800 randomly selected women (oversampling women 40-69, the potential screening population), asking about understanding about sensitivity and specificity of mammography using a validated visual analogue scale. \(^{176}\) Response rate was 65.6% (n=503), of whom 497 had no history of breast cancer.

Ninety-two percent of women believed that mammography could not cause harm; of those who did, none cited false positives as a harm. Ninety-nine percent believed false positives were possible, with a median estimate of the 10-year probability of a false positive of 20%. There was a high “willingness to pay” in terms of false positives per death prevented—63% were willing to accept a value of 500 or more, with 37% willing to accept 10,000 or more (Figure 20). If anything, a history of a false positive result made women more likely to accept a higher number of false positives, a finding consistent with systematic reviews that find a higher probability of subsequent screening after a false positive, at least in U.S.\(^{174}\) and UK.\(^{177}\) This tolerance for a high false positive/death prevented ratio was not influenced by a substantial overestimation of the benefits of mammography—none of the respondents thought that mammography eliminated the risk of breast cancer death, with most respondents stating a reduction of 30-50% (consistent with contemporary reports on mammography effectiveness).
Limitations of the study primarily involve generalizability to current practice. Respondents had telephones, had agreed to potentially participate in survey research, and had higher income and education levels compared to the total U.S. female population; they were also almost exclusively white. In addition, the ongoing debate over the benefits and harms of mammography during the past 15 years may have led to changes in patient tolerance for false positives.

More recent evidence on patient preferences and outcomes after a false positive result comes from a substudy of the Digital Mammographic Imaging Screening Trial (DMIST). Because this study is prospective and uses standard instruments for measuring anxiety and preferences, we believe it is worth some detailed discussion. Eligible women presenting for screening who agreed to undergo follow-up mammography and provided written consent for participation underwent both digital and screen-film mammography. The substudy consisted of a telephone survey of random samples of women with a positive screening mammogram (any mammogram where additional workup or consultation was recommended, and those with a negative screening mammogram), matched by institution and age. Anxiety was measured using the Spielberger State-Trait Anxiety Inventory (STAI), a validated measure of general anxiety, and the U.S. version of the EuroQol EQ-5D instrument, which consists of five questions related to health (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), with three levels (no problem, some problem, extreme problem). A validated scoring system allows preference weights to assign an overall utility to the current health state. Telephone interviews were conducted after the baseline mammogram and approximately 12 months later. In addition, women were asked to trade off time against false positive results (measured by asking the amount of travel women would undertake to gain access to a new type of mammography that produced fewer false-positives while detecting the same number of cancers), and trade off...
discomfort versus false positives (by asking whether they would prefer a new type of mammogram that was just as sensitive in detecting cancers as current technology that resulted in fewer false positives but required the same amount of breast compression, or a new type that had equivalent sensitivity and specificity but less breast compression).

Approximately 85% of the 1450 eligible women enrolled (1226), with follow-up interviews for 1028 (83.8% of those enrolled). Women with false positive results were significantly younger (44.1% less than 50 compared to 38.6% for women with negative results, p<0.05), but were otherwise similar demographically. At baseline, mean STAI state anxiety was higher among women with a positive mammogram, but EQ-5D scores were similar; at 12 months, there was a significant decrease in STAI state anxiety scores among women with false positive exams. Not surprisingly, of those with false positives, 66.2% had additional imaging (compared to 4.5% after a negative screen), and 14.6 had a biopsy (vs.1.1% after a negative screen). Compared to women with negative exams, 50.6% with false positive results reported moderate or higher levels of anxiety associated with their additional care (vs. 15.6% in the negative group), with 18.8% reporting “a lot” of anxiety, and 4.5% reporting “extreme” anxiety. Similarly, discomfort associated with additional care was more common after false positives (35.2% vs. 14.3%), with 7.9% reporting “a lot” and 4.3% reporting “extreme” discomfort.

At 12 months, women with a false positive mammogram stated they were more likely to use screening in the future than before their result (25.7% vs. 14.2% for those with negative screens), although there was no difference in anticipated anxiety/concern between groups. There were no significant differences between groups in preferences for fewer false positives versus less discomfort during screening (approximately 80% in both groups preferred fewer false positives to less compression while holding sensitivity equivalent), or for “willingness to pay” for fewer false positives (approximately 16% in both groups willing to travel over 4 hours, with 10% willing to stay overnight.)

This large, well-designed study, which used standard assessment tools for measuring generalized anxiety and health preferences in a cohort of women undergoing screening, demonstrated that, although generalized state anxiety was increased after a false positive result, anxiety scores for most women had declined by 1 year after the result. Although this is reassuring, there are several limitations, most of which were mentioned by the study authors (and in an accompanying editorial):

- The STAI is a measure of generalized anxiety. The majority of the literature shows larger and more persistent effects on cancer-specific anxiety, worry, or other quality-of-life domains; in a recent meta-analysis, anxiety was the only generalized domain that showed significant effects. The extent to which cancer-specific concerns affect overall quality of life is unclear. In both this study and others, having a false positive result increases the likelihood of future screening—one mechanism for this may be increased cancer concern prompted by the original false positive result. To the extent that having a false positive may identify someone at higher risk for future breast cancer, this may be a net beneficial outcome, although additional evidence (including use of models that incorporate individual variation in screening behavior) would be helpful.
- The emotional subscale in the EQ-5D does not distinguish between depression and anxiety and has only three levels, so it may not be as sensitive to anxiety-specific changes, especially in the aggregate.
Because of the design of the study, there is no evidence for the duration of the increased anxiety, or the distribution of duration among women (i.e., were some women affected for 6 months or longer).

There are no data presented on whether women who underwent biopsy had higher levels of anxiety, or long lasting anxiety, than women who only had repeat examinations or imaging. Disaggregating the effects of false positive biopsies from repeat examinations is an important consideration for weighing the public health impact of false positives. Intuitively, a false positive biopsy is a “worse” harm than a false positive resulting only in repeat examinations because of the need for an invasive procedure with attendant risks of complications, and, presumably, greater anxiety/worry. In the DMIST substudy, 23.4% of women with a false positive result reported “a lot” or “extreme” anxiety, but only 14.6% of women with a false positive underwent a biopsy. Even if all of the women undergoing biopsy experienced “a lot” or “extreme” anxiety, this still means that an additional 9-10% of women with a false positive resulting in only a repeat examination had an emotional experience (at least as measured using these instruments) similar to the women undergoing biopsy. Given the much larger number of false positive recalls than biopsy, this is a large absolute number of women. In other words, even if the average response to a false positive that does not lead to biopsy is mild and transient, these data are consistent with the possibility that the emotional impact in some women is significant, and that using false positive biopsies alone as a metric for “significant” false positive results may miss clinically meaningful outcomes in a substantial number of women.

Both the study authors and the editorial point out that women participating in a clinical research study may be different from the general population in attitudes about screening, education, comfort with risk, etc., in ways that may affect the applicability of these results to a wider range of women. In this specific study, there is an additional aspect of research participation that may affect generalizability. The primary objective of the DMIST study was to compare diagnostic accuracy (sensitive and specificity) of the two types of mammography. Presumably, since false positive results were part of the primary outcome, the informed consent process included a discussion of the possibility of a false positive result (perhaps even a discussion of the chances of a false positive result), as well as the possible consequences. This discussion was likely much more comprehensive than many women experience given the time constraints of a typical office visit—if participants in the study had a better understanding of the possibility of a false positive result than many women undergoing screening in the community, then the level of anxiety prior to a final determination of no cancer may have been lower, and/or resolution of the anxiety faster, than would be expected in the general population.

Finally, although the study provided evidence that minimizing false positives is important to women, as measured both by their willingness to travel for a procedure that reduced the risk of a false positive and in their preference for a new procedure that reduced false positives over reduction in examination discomfort, both of these questions were asked under the explicit presumption of no decline in the ability of the test to detect early cancers (and reduce mortality). While extremely useful for providing evidence on the impact of false positives on quality-of-life measures (the EQ-5D data in particular is helpful for health economic analyses), the study provides no evidence on whether women
would be willing to accept any increase in mortality (or decrease in test sensitivity) to reduce false positives (increase specificity).

**Discussion/Conclusions: Harm-benefit of False Positives per Death Prevented**

- In the CISNET models, depending on screening interval, age of screening, estimates of mortality reduction, and estimates of false positive probability, the estimated total false positives per breast cancer death prevented at the population level is in the range of 100-200 for different strategies compared to no screening, and 50 to 1500 when screening strategies are compared to each other; rates for false positive biopsies are lower, in the range of 10-100.
- When an incremental approach to comparing the published results is used, dominance or extended dominance eliminates several strategies—if biennial screening at age 50 is used as the reference threshold, extended dominance eliminates biennial screening at younger ages, and the next strategy for consideration is annual screening beginning at age 50.
- Recent evidence on the 12-month impact of false positive results in U.S. women participating in a clinical study suggest that the effect of false positives on generalized state anxiety are resolved within a year for most women, but effects on cancer-specific domains, differential impact of biopsies versus recall alone, or whether a proportion of women were more likely to experience more prolonged or severe anxiety were not reported.
- False positive biopsies are a more “severe” outcome because they carry the risk of complications, are associated with greater pain and discomfort than additional imaging, and, presumably, because patients may associate them with a greater probability of cancer, more severe anxiety consequences. However, there is little available U.S.-based evidence on differences in quality-of-life impact between biopsies and recall examinations; in the DMIST substudy, the proportion of women experiencing “a lot” or “extreme” anxiety was higher by approximately 10% than the proportion of women undergoing biopsy, suggesting that a proportion of women with a false positive resulting in recall alone may experience emotional consequences comparable in severity to women undergoing biopsy.
- Evidence on “willingness-to-pay” for the trade-off of false positives versus cancer death in the U.S. is limited to a single pre-2000 survey. This study suggested that most U.S. women have a very high “willingness-to-pay” for this harm-benefit ratio, with a median value of well above 1000. However, the quality of this evidence is LOW, because of the relatively small sample size and the potential impact of subsequent debate about the benefits and harms of mammography. Although the recent DMIST analysis assessed women’s willingness to trade off reductions in false positives against travel time and discomfort during the test, this was done under the explicit assumption of equivalent sensitivity and thus does not provide any additional evidence for the specific trade-off of false positives (either recall or biopsy) versus test sensitivity (and, by extension, mortality reduction).

**Overdiagnoses per Breast Cancer Death Prevented**

Estimates of overdiagnosis per death prevented have only recently become an outcome of interest, and there are relatively few available estimates; interpretation of these results is subject to all of the uncertainties discussed above, particularly regarding the estimation of overdiagnosis.
Literature-based Estimates

Non-U.S. Estimates

Using estimates of overdiagnosis based on follow-up from the three RCTs where women randomized to no screening were not offered screening at the end of the trial (Malmo I and the two Canadian trials), and estimates of mortality reduction based on the pooled RCTs, the UK Panel estimated approximately three overdiagnoses per death prevented in women screened biennially between the ages of 50 and 70 during screening, with extensive discussion of the high degree of uncertainty resulting from issues of study design, methodology, generalizability, as well as statistical uncertainty.

Duffy and colleagues estimated ratios of overdiagnoses per death prevented over 20 years of biennial screening from 50-70 years of age of 0.49 (based on projections from the incidence screens of the Swedish Two-County Trial), and 0.40 (based on projections of incidence and mortality in the absence of screening in the UK derived from trends prior to the implementation of the national screening program). No direct measure of the precision of these estimates, such as 95% confidence intervals, was provided.

Using an excess incidence (including DCIS) approach for estimating overdiagnosis and observed mortality among women aged 60-69 years attending the Florence, Italy, screening program to women in the same age group (the only group with sufficient follow-up), Puliti and colleagues estimated a ratio of 0.6 overdiagnoses per cancer death prevented, when 34 women with a cancer diagnosis within 6 months of the invitation for screening (who presumably were already being evaluated for cancer at the time of the screening invitation and could not have benefited from screening) were excluded, the reported ratio was 1.0. Confidence intervals were not reported for either estimate. Mortality differences were adjusted for marital and socioeconomic status.

From the confidence intervals reported for the individual components, we can estimate confidence intervals around the ratio, assuming that overdiagnosis and mortality are independent (an assumption that may not be valid—presumably, increasing the ability of the screening test to detect smaller lesions will both decrease mortality and increase the probability of detecting a lesion that would otherwise have gone undetected). For the base case, we used the adjusted confidence intervals reported in the paper; for the sensitivity analysis, where confidence intervals were not reported, we assumed that all 34 cases were in the non-attending group, and that median follow-up was 15 years. Subtracting these 34 cases from the number of incident cases among the non-attenders, and subtracting $34*15 = 510$ person-years of follow-up, we recalculated an unadjusted risk ratio and confidence intervals, with a resulting point estimate for the risk ratio identical to the one reported in the paper (1.15). The number of deaths among this group was not reported, and the authors state that the mortality reduction for 60- to 69-year-olds was “essentially unchanged” at 0.48. For simplicity, we assumed that the width of the confidence interval for the ratio was also unchanged, and simply lowered the upper and lower bounds by 0.01 (see Table 35). We then generated confidence intervals for the ratio by running 10,000 simulations, multiplying the incidence in non-attenders by the estimated relative risk, drawing the value for the relative risk from lognormal distributions characterized by the estimates in Table 35.
Table 35. Estimated Overdiagnoses per Breast Cancer Death Prevented among 60- to 69-year-old Invited for Screening, Florence, Italy, 1991-2007 (Adapted from Puliti, 2012<sup>45</sup>)

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Non-Attenders</th>
<th>Attenders</th>
<th>RR (95% CI) Adjusted for Age, Marital and Socioeconomic Status</th>
<th>Mean Excess Cases or Deaths Prevented (95% CI)</th>
<th>Overdiagnoses/Death Prevented (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Case</td>
<td>–</td>
<td>0.0034</td>
<td>–</td>
<td>–</td>
<td>0.67 (-0.14 to 1.67)</td>
</tr>
<tr>
<td>Incidence</td>
<td>0.0032</td>
<td>0.0034</td>
<td>1.10 (0.98 to 1.23)</td>
<td>0.00032 (-0.00006 to 0.00074)</td>
<td>–</td>
</tr>
<tr>
<td>Mortality</td>
<td>0.00093</td>
<td>0.00040</td>
<td>0.49 (0.38 to 0.64)</td>
<td>0.000474 (0.000335 to 0.000577)</td>
<td>–</td>
</tr>
<tr>
<td>Sensitivity Analysis</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1.05 (0.14 to 2.17)</td>
</tr>
<tr>
<td>Incidence†</td>
<td>0.0030</td>
<td>0.0034</td>
<td>1.15 (1.02 to 1.28)*</td>
<td>0.00045 (0.00006 to 0.00084)</td>
<td>–</td>
</tr>
<tr>
<td>Mortality†</td>
<td>0.0093</td>
<td>0.0040</td>
<td>0.48 (0.37 to 0.63)</td>
<td>0.000484 (0.000344 to 0.000586)</td>
<td>–</td>
</tr>
</tbody>
</table>

<sup>*</sup>Calculated from data provided in paper, RR not adjusted.

<sup>†</sup>Assumption based on description in paper.

Abbreviations: CI=confidence interval; RR=relative risk

In the base case estimates, the lower bound of the ratio is less than 0 because the lower bound of the CI for the relative risk is less than 1.0 (consistent with incidence in unscreened women being higher than in screened women).

Although the confidence intervals around the ratios are useful for illustrating the uncertainty around the estimate, another way to visualize the uncertainty is through the use of a harm-benefit acceptability curve (as we did with the estimates of false positives per death prevented). Figures 21 and 22 illustrate these curves for the data from the Puliti paper.<sup>45</sup>
Figure 21. Harm-benefit Acceptability Curve for Overdiagnoses and Breast Cancer Deaths Prevented for Women 60-69 Years Old in Florence, Italy (Derived from Puliti, 2012 [45]), “Base Case” Estimates. Vertical line indicates 1 overdiagnosis per cancer death prevented.

Figure 22. Harm-benefit Acceptability Curve for Overdiagnoses and Breast Cancer Deaths Prevented for Women 60-69 Years Old in Florence, Italy (Derived from Puliti, 2012 [45]), “Sensitivity Analysis” Estimates. Vertical line indicates 1 overdiagnosis per cancer death prevented.

Again, the “screened” curve in each graph is the cumulative density function of the incremental harm-benefit ratio—in the base case, the mean/median value is 0.67, and there is an approximately 70% probability that the value is less than 1.0. For the “sensitivity analysis” graph (Figure 22), the median value is approximately 1—there is a 50% probability that the true ratio is at least 1.0.
These graphs primarily illustrate the considerable quantitative uncertainty surrounding the harm-benefit trade-off, even within a well-defined cohort using a specific method for estimating overdiagnosis, there may be. Key points include:

- The threshold for “acceptability” is critical. Even with favorable estimates for overdiagnosis and mortality reduction (since the method used for adjusting for self-selection bias may not have accounted for all confounding), there is still a 30% probability that the true overdiagnosis to death prevented ratio is greater than 1.0. Depending on the judgment of patients or policy makers on acceptable trade-offs, a 30% probability may be uncertain enough to affect strength of recommendations.

- Relatively minor methodological issues can affect certainty; removal of a small number of ambiguously classified cases changed the probability of the value being greater than 1.0 from 30% to 50%; if 1.0 were the threshold for acceptability, this would definitely affect strength of recommendation.

- These estimates assume independence of the overdiagnosis and mortality estimates. As noted earlier, it is plausible that there is a correlation—increasing screening sensitivity would lead to both greater mortality reduction and a higher probability of overdiagnosis (the same correlation is also likely for false positives and mortality reduction). Depending on the strength of the correlation, accounting for dependence between the two could lead to wider or narrower confidence intervals and further affect the degree of certainty about the estimate.

As discussed above, even if there were no uncertainty about the generalizability of relative effect estimates from studies in other populations to the U.S., and even if there was consensus about the most appropriate method for estimating overdiagnosis, estimates of the absolute effects for both numerator and denominator are needed for the U.S. population in order to inform U.S. recommendations.

**U.S. Estimates**

Welch and Passow recently estimated a range of overdiagnoses per death prevented for the U.S. of 3-20, depending on the sources used.\(^{182}\) For mortality reduction, the upper bound was based on the 30-year follow-up of the Two-County Trial (31% reduction multiplied by 85% adherence, for a total reduction of 36%), and an arbitrary lower bound of 5% (based on the lack of statistical significance in the Canadian trials). Estimates for 10-year mortality reduction were based on projected 15-year risk of death for 2007-2009 from SEER (based on age-specific mortality, not incidence-based mortality), adjusted for prevalence of screening in the National Health Interview Survey. For overdiagnosis, the lower bound was based on excess incidence estimates from Malmo I, and an upper bound estimate of 33% based on a trend analysis of SEER incidence and mortality, and the Cochrane meta-analysis applied to projected cumulative incidence from SEER and age-specific reported screening rates.

While providing estimates based on U.S. data, the wide range is difficult to interpret. On the one hand, the range does highlight the inherent difficulties in estimating the absolute impact of screening in the U.S. setting of opportunistic screening and lack of data on the screening history of cancer cases in available population-based registries. However, issues include:

- As discussed above and in Appendix C, crude age-specific mortality is not appropriate for estimating the impact of screening using an approach which partitions event rates based on exposure (in this case, to screening) and a relative risk estimate, since deaths
occurring at a given age may represent cases diagnosed prior to the start of the interval of interest. For example, some breast cancer deaths in 52-year-olds represent cases diagnosed prior to age 50, so screening beginning at age 50 would not affect these deaths.

- Both the estimates of relative mortality reduction and overdiagnosis are subject to a very high degree of uncertainty for all of the reasons discussed earlier. Relative mortality reduction attributable to screening could plausibly be greater (because of improved screening sensitivity and differences in estimates based on screened vs. unscreened compared to invited vs. uninvited), and overdiagnosis plausibly lower (because of differences in definitions and methods for estimation).

- Although estimating outcomes over a 15-year time horizon from the onset of screening is reasonable for many reasons (including the need for fewer assumptions about the applicability of current screening and treatment outcomes, cancer incidence in the absence of screening, and competing risks, as well as less dependence on implicit or explicit assumptions about individual preferences for benefits and harms incurred in the near or distant future), a shorter time horizon may lead to overestimation of the overdiagnosis to death prevented ratio, since incidence in the screened group drops after screening stops due to lead time effects, and mortality reductions for cases detected later during the screening period may not be apparent for years after the cessation of screening.

None of the publications from the CISNET group or other recent U.S. modeling studies provided explicit estimates of overdiagnosis or the ratio of overdiagnosis to benefits (deaths prevented, life-years saved). We did not identify any modeling study that explicitly estimated rates of overdiagnosis, or quantified the effect of the substantial uncertainty about overdiagnosis to trade-offs.

**Model-based Estimates**

Given the high degree of uncertainty surrounding the most appropriate method for estimating the probability of overdiagnosis of invasive cancers under different screening strategies, it is extremely difficult to estimate the absolute risk of this component of overdiagnosis for the U.S. However, with data on the overall incidence of DCIS, estimates of the relative risk of DCIS among screened versus unscreened women, estimates of the mortality reduction attributable to screening, and estimates of the prevalence of screening, we can generate estimates of the overall ratio of DCIS to deaths prevented for screened women, and, by varying the proportion of DCIS that would progress if undetected, generate estimates of the ratio of overdiagnosis attributable to DCIS to deaths prevented by screening.

As with false positive biopsy results, we ran two-dimensional Monte Carlo simulations for U.S. women from age 40 to 100, under a variety of scenarios:

- Identical screening strategies and mortality reductions.
- Estimates of DCIS progression probability of 0.2, 0.5, and 0.8. Note that these probabilities are applied to both screen- and non-screen-detected DCIS.
- Relative risks of DCIS of 3.0, based on Norwegian data, and 6.0 (modeled as age-specific relative risks ranging from 4.9 to 7.0) based on BCSC data.

Figure 23 (mortality reduction of 0.62) and Figure 24 (mortality reduction of 0.8) show the results of these analyses.
Figure 23. Harm-benefit Acceptability Curves: Overdiagnosed Cases of DCIS per Breast Cancer Death Prevented by Relative Risk of DCIS Among Screened Women and Probability of Progression of DCIS to Cancer in the Absence of Treatment, Relative Mortality with Screening 0.62 (95% CI, 0.56 to 0.69)
Figure 24. Harm-benefit Acceptability Curves: Overdiagnosed Cases of DCIS per Breast Cancer Death Prevented by Relative Risk of DCIS Among Screened Women and Probability of Progression of DCIS to Cancer in the Absence of Treatment, Relative Mortality with Screening 0.80 (95% CI 0.73 to 0.89)
Key points include:

- Screening beginning at age 45 is eliminated by extended dominance—the incremental ratio of overdoses to deaths prevented comparing screening beginning at 45 is greater than the ratio for screening beginning at age 50 compared to no screening under every scenario.

- Curves for screening ages 40-74 indicate the incremental overdoses attributable to DCIS to cancer deaths prevented compared to screening 50-74.

- Uncertainty about the likelihood of DCIS progression is a large driver of uncertainty about the ratio, indicated by the shifting of the curves to the left (smaller ratios) moving down each column in the graphs (increasing likelihood of progression).

- Uncertainty about the mortality reduction attributable to screening is also a major contributor (curves in Figure 23, with mean mortality reduction of 0.62, are further to the left than curves in Figure 24, with mean mortality reduction of 0.8).

- The impact of uncertainty about the relative risk of DCIS attributable to screening is qualitatively smaller than the effect of the other two main parameters (for any given level of mortality reduction and progression probability, the shift to the right from increased relative risk of DCIS from screening is smaller than the shifts resulting from changes in mortality reduction or progression probability).

- At estimates of DCIS progression of 50% or lower, the probability that the ratio is above 1.0 is close to 100% across all scenarios.

- At the high end of progression probability (80%), and mortality reduction (0.62), the probability that the ratios for either strategy are less than 1.0 is approximately 90%.

Across all combinations of mortality reduction, relative risk of DCIS, and probability of DCIS progression, extending the age for screening always resulted in:

- Elimination of extending screening to age 79 by extended dominance.

- Incremental ratios for extending to screening through age 84 compared to stopping at age 75 that were lower than screening ages 50-74 compared to no screening.

We did not attempt to disaggregate the effects of screen-detection of DCIS on subsequent incidence of invasive cancer or on breast cancer mortality. To the extent that detection of DCIS results in prevention of breast cancer mortality, some of the effect is “baked in”—the observed reduction in mortality is partly attributable to detection and treatment of DCIS, although at least one study has suggested that this contribution is relatively small (5-12%), with the majority of the mortality reduction attributable to shifts to early stage invasive disease.\(^{183}\)

However, because of the high competing risk of mortality, the relative contribution of very early detection of invasive cancers to overdiagnosis is likely to be a more important consideration for older women than for younger women. In addition, there may well be age-specific effects on the probability of DCIS progression that are not captured by simply varying an assumed overall probability of progression. Although we believe that using estimates of the probability of detection of non-progressive DCIS through screening is a reasonable basis for providing plausible ranges of this component of overdiagnosis overall, it is less useful applied to the upper end of possible screening ages.

Given the high degree of uncertainty about any of these estimates, these analyses can only illustrative of the possible range of the overdiagnosis to death prevented trade-off under a variety of reasonable assumptions. Inclusion of invasive cancers which were overdiagnosed would
increase the ratio, but whether this would substantially change the likelihood that a given strategy would exceed an acceptable threshold is not clear (if there is a significant impact, it is likely to be at the upper end of the age range for screening). If there were consensus on the maximum acceptable threshold, further analyses using alternative approaches could be used to help guide strength of recommendations and additional research to resolve key areas of uncertainty.

**Evidence on Patient Preferences for Overdiagnosis versus Death Prevented Trade-Off**

The survey conducted by Schwartz and colleagues also asked about non-progressive lesions. In 1997, only 7% of respondents were aware of the possibility that some lesions might not develop into symptomatic cancer, or, in the case of DCIS, develop into invasive cancer. Sixty percent felt the information would be important for decision making about mammography, with younger women more interested in having the information (71% of women aged 18-39). Subjects were asked to specify a probability of progression to invasive cancer at which they would want to have DCIS treated—40% would wish to be treated if the progression probability was 1%, while 78% would want to be treated at a threshold of 33% (the approximate midpoint of the range used in our analyses (Figure 25). Unfortunately, unlike the evidence on false positives per death prevented, this evidence does not directly inform decisions based on uncertainty about the trade-offs between critical outcomes. Since a substantial proportion of invasive cancers, both screen-detected and clinically detected, will not result in death from breast cancer, a more useful way to frame the question would be as the hypothetical probability of ultimately dying from a potentially detectable DCIS lesion which progresses to invasive cancer. The other limitations of this study listed above in terms of generalizability of respondents and possible secular trends in understanding and preferences about mammography screening are also true for this outcome.
Two recent qualitative studies, one from the UK and one from Australia, explored women’s understanding of overdiagnosis in the context of breast cancer screening. In both studies, investigators found little pre-existing knowledge of overdiagnosis, with most women expressing surprise at the possibility. The concept was initially hard to understand for many participants, but most eventually expressed comprehension. Most women in both groups felt that the information was important to provide to patients, but that issues related to overdiagnosis/overtreatment would not affect their decision to be screened, but might affect their decisions about treatment in the event of a screen-detected cancer. We did not identify any similar recent U.S.-based studies.

### Discussion/Conclusions: Harm-benefit of Overdiagnosis per Death Prevented

- The uncertainty about the true proportion of overdiagnoses among screened women, together with uncertainty about the magnitude of the effect of screening on mortality, precludes estimating the ratio with any degree of precision.
- Probabilistic analyses show that, for DCIS-related overdiagnosis, the likelihood that a given strategy will have an acceptable threshold is primarily driven by the proportion of DCIS that would progress to invasive cancer if undetected and untreated and by the mortality reduction attributable to screening; the relative risk of DCIS from screening has a smaller effect on the incremental overdiagnosis to death prevented ratio. With a high probability of progression (80%) and a high degree of mortality reduction (0.6), the probability that the ratio of overdiagnosis attributable to DCIS to deaths prevented will be
less than 1.0 is high (90% or greater). For other combinations of progression probability and mortality reduction, the ratio is much more likely to be above 1.0.

- Inclusion of overdiagnoses from invasive cancer would increase the probability that the ratio is above 1.0 for all scenarios, but the magnitude of this effect, and thus its impact on whether a given screening strategy was optimal, is uncertain.
- There are very limited data on patient preferences for this trade-off, particularly for the U.S., and no evidence of any formal assessment or discussion of an appropriate threshold for this trade-off from any group making recommendations about breast cancer screening.

**Key Question 2**

In average-risk women who are screened with mammography, what are the relative benefits, limitations, and harms associated with annual, biennial, triennial, or other screening interval, and how do they vary by age?

**Summary**

**Key Points: Outcomes**

**Breast Cancer Mortality:**
- **Direction of Effect:** Direct and indirect evidence suggests some reduction in mortality with more frequent screening (annual vs. biennial) in women under the age of 50, but not in women 50 years and older. We judge the quality of evidence for these effects to be **LOW** because of risk of bias, indirectness, and imprecision. Reduced mortality from more frequent screening in younger women is biologically plausible, since the proportion of cancers that are rapidly progressive may be higher in younger women.
- **Magnitude of Effect:** We judge the quality of evidence for estimating the magnitude of any effect of interval on mortality as **VERY LOW**.

**Life Expectancy:**
- **Direction of Effect:** Model-based estimates suggest improved life expectancy with more frequent screening, especially in younger women, but because these estimates are dependent on empirical data on the effect of interval on mortality, which have a high degree of uncertainty, we judge the quality of evidence to be **LOW**.
- **Magnitude of Effect:** The effects of increasing screening frequency on extending life expectancy are always greater in younger populations, but again, because of the **VERY LOW** quality of the existing evidence to inform the models, the quality of evidence is **VERY LOW**.

**Overdiagnosis:**
- Qualitative descriptions of modeling results suggest an effect of screening interval on overdiagnosis, with overdiagnosis increasing with more frequent screening, but there are no quantitative estimates. Because of the fundamental uncertainties surrounding overdiagnosis discussed under KQ 1, we judge the quality of evidence to be **VERY LOW**.
False Positives:
- **Direction of Effect:** Evidence from observational studies consistently shows a higher lifetime cumulative risk of false positive results and false positive biopsies with more frequent screening. Modeling studies also find higher cumulative false positive rates with more frequent screening; at any given level of test specificity, more frequent screening should result in more false positives. We judge the quality of this evidence for the DIRECTION of effect to be **HIGH**.
- **Magnitude of Effect:** The effect of more frequent screening on false positive rates is higher in settings where test specificity is decreased, such as screening in younger women or women with dense breasts. This finding is consistent, although there is imprecision in the estimates. We judge the quality of evidence to be **MODERATE**.

**Quality-adjusted Life Expectancy:**
- **Direction and Magnitude of Effect:** Modeling studies consistently find that more frequent screening leads to gains in quality-adjusted life expectancy compared to less frequent screening, but the size of the gains is decreased relative to unadjusted life expectancy, especially if disutilities are assigned to screening itself and to false positive results. The potential effects of overdiagnosis are not clear. The incremental gains in quality-adjusted life expectancy are smallest in younger women, again especially when disutilities are assigned to false positives (because of the greater likelihood of false positives in younger women). Because of the inherent uncertainties in the models, particularly for overdiagnosis, and the concerns about the utility weights used, we judge the quality of this evidence to be **LOW**.

**Key Points: Harm-benefit Trade-offs**
- Model-based estimates of incremental false positives per breast cancer death prevented by decreasing screening interval from biennial to annual differ based on whether estimates are derived using total population false positives (including women with multiple false positives) or “at least one” false positive. In both cases, the ratios are well within the range judged to be acceptable by the one U.S.-based study of women’s willingness to accept trade-offs of breast cancer screening (a study which has limitations in terms of its applicability to current recommendations).
- Model-based estimates of false positive biopsy rates per death prevented also increase with screening frequency, but are much lower than for overall false positives; we did not identify any evidence on patient preferences for this specific trade-off.

**Description of Included Studies**

**Studies**
We identified nine studies that evaluated the relative benefits, limitations, and harms associated with annual, biennial, triennial, or other screening interval in average-risk women. All nine were cohort studies: eight were prospective and one retrospective. Six of the prospective studies used data from the same U.S. registry, the Breast Cancer Surveillance Consortium (BCSC). Of the remaining prospective studies, one was a cohort study from Finland where screening interval varied by birth year, and one used data from the Screening Mammography Programme of British Columbia (SMPBC). The
retrospective study was based on a database of the Massachusetts General Hospital Avon Comprehensive Breast Center.\textsuperscript{188}

\section*{Population}
All studies described screening programs for women at average risk of breast cancer. The age groups described ranged from 40-89 years of age, with studies stratifying by age groups of 40-49,\textsuperscript{50,87,187,189,191} 40-59,\textsuperscript{92} 50-74,\textsuperscript{87} 50-79,\textsuperscript{50} 66-74,\textsuperscript{186} and 75-89;\textsuperscript{186} Yankaskas et al.\textsuperscript{190} included ages 40-89, stratified into 5-year age groups (except for 75-89 years). Blanchard et al.\textsuperscript{188} did not report findings stratified by age. In addition to age and menopause status, individual studies reported results stratified by race/ethnicity,\textsuperscript{189} breast density, hormone replacement status,\textsuperscript{87} comorbid conditions in older women (defined using the Charlson comorbidity index),\textsuperscript{186} and body mass index (BMI), stratified as normal (BMI 18.5-24.9), overweight (BMI 25.0-25.9), and obese (BMI $\geq$30.0).\textsuperscript{191}

\section*{Intervention}
All studies evaluated the screening method of two-view screening mammography. One study\textsuperscript{187} described having a second reader for all screening mammograms, while other studies either did not describe their interpretation method or had a single reader. One study randomized participants age 40-49, by their year of birth, to screening intervals of triennial screens and annual screens.\textsuperscript{187} One study compared biennial screening to annual screening,\textsuperscript{50} which occurred as a change in the screening program protocol during the time period studied. Other studies did not clearly define the screening interval as a prescribed program. Rather, one study described cohorts followed over 10, 8, and 5 years and reported results by numbers of screening mammograms that women chose to have over these periods of time.\textsuperscript{188} All but one of the BCSC registry studies described results by screening intervals of 1 year vs. 2 years,\textsuperscript{186,191} or of 1 vs. 2 vs. 3 years, based on the time between the two most recent mammograms,\textsuperscript{87,92,186,189} as well as women’s self-report.\textsuperscript{92} In these studies, “annual” was defined as an interval of 9-15 months, “biennial” as greater than 18-30 months, and “triennial” as greater than 30 to 42 months. Using definitions based on the observed distribution of screening intervals within the BCSC, Yankaskas and colleagues\textsuperscript{190} defined “months since previous mammogram” (MSPM) in intervals of 9-15 months, 16-20 months, 21-27 months, and 28 or more months.

\section*{Outcomes}
Two studies reported breast cancer mortality.\textsuperscript{50,187} Both used cancer registries and vital statistics databases for their respective countries, Finland\textsuperscript{187} and Canada,\textsuperscript{50} to validate their outcome.

One of the BCSC studies\textsuperscript{191} reported on overdiagnosis.

All six BCSC studies evaluated the outcome of false positive screens with recall and with biopsy,\textsuperscript{87,92,186,189-191} while one study evaluated total recall rates and rates of biopsies and negative biopsies.\textsuperscript{188} Definitions of false positive and positive screens were consistent across the studies. All studies used radiologists’ interpretation of mammograms based on the Breast Imaging Reporting and Data System (BI-RADS). A false positive or positive result from a screening mammogram was defined as an initial BI-RADS assessment with 0 (needs additional imaging), 4 (suspicious abnormality), 5 (highly suggestive of malignancy), or 3 (probably benign finding with a recommendation of immediate evaluation). Following this initial assessment, based on further imaging and/or biopsy results, a false positive screen was associated with no
diagnosis of invasive carcinoma or ductal carcinoma in situ (DCIS) within 1 year of the initial positive screen or before the next screening exam, whichever occurred first.

**Timing of Outcomes**

Studies evaluating the outcome of breast cancer mortality followed participants for the longest period of time. The study based on the SMPBC database included women aged 40-79 who were first screened between July 1988 and December 2005, with follow-up regarding data on death completed on December 31, 2005. A Finnish study evaluated breast cancer mortality, following women aged 40-49 who were screened beginning in 1987; these women were followed until age 52 for breast cancer mortality. The cohort studies evaluating the outcome of false positives differed somewhat in duration: 1999-2006, 1996-2008, 1994-2004/2007, and 1985-2002; however, studies were fairly consistent with their definition of a false positive as having no invasive carcinoma or DCIS diagnosed within approximately 1 year after the positive screen.

**Settings**

Six studies describe data from the BCSC, a U.S. mammography registry, with data from screening mammograms done in the U.S., a country that does not have nationally or regionally organized screening program. The data from the UK, Finland, and British Columbia (Canada) are from organized population-based screening programs.

More detailed characteristics of the included studies are summarized in Appendix Table G-2. GRADE summary tables for the outcomes described below are provided in Appendix H.

**Detailed Synthesis**

**Breast Cancer Mortality**

**RCTs**

The Canadian Task Force review indirectly compared the effects of screening interval on breast cancer mortality in women under 50 and 50 years and older from the RCTs (Table 36). An interval of 24 months or less significantly reduced mortality in younger women compared to no screening, but a longer interval did not. Breast cancer mortality was significantly reduced across all intervals compared to no screening for women 50 years old and older. Note that this analysis compared results by interval across studies, rather than within studies.

**Table 36. Effect of Mammography on Breast Cancer Mortality by Age and Screening Interval (Canadian Task Force)**

<table>
<thead>
<tr>
<th>Age Range and Screening Interval</th>
<th>RR (95% CI)</th>
<th>Included Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 50 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;24 months interval</td>
<td>0.82 (0.72 to 0.94)</td>
<td>HIP, Canada I, Malmo, Goteborg, Age</td>
</tr>
<tr>
<td>≥24 month interval</td>
<td>1.04 (0.72 to 1.50)</td>
<td>Two-County, Stockholm</td>
</tr>
<tr>
<td>50-69 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;24 months interval</td>
<td>0.86 (0.75 to 0.98)</td>
<td>HIP, Canada II, Malmo, Goteborg</td>
</tr>
<tr>
<td>≥24 month interval</td>
<td>0.67 (0.51 to 0.88)</td>
<td>Two-County, Stockholm</td>
</tr>
</tbody>
</table>

Abbreviations: CI=confidence interval; HIP=Health Insurance Plan of New York; RR=relative risk
Observational Studies

Two cohort studies describe the outcome of breast cancer mortality in women who underwent screening mammography at different time intervals. One compared triennial screening to annual screening in women age 40-49, while the other compared biennial screening to annual screening in two different cohorts aged 50-79. Neither study showed a difference in breast cancer mortality with these different screening intervals.

One study from Finland invited women aged 40-49 for screening at different time intervals based on their birth year: those born in an even calendar year were invited to annual screening, while those born in an odd calendar year were invited to screening every 3 years. Participants were followed until age 52, for a mean of 12.8 years across all birth cohorts, for an incident breast cancer and for death, either from breast cancer or all causes. With follow-up stopping at age 52, women in their late 40s would have less follow-up time to detect differences in mortality. Compared to the group receiving triennial screens, those participants receiving annual screens had a relative risk (RR) of breast cancer mortality of 1.14 (95% CI, 0.59 to 1.27). However, all-cause mortality was also higher in the annually screened group (RR 1.20; 95% CI, 0.99 to 1.49), with marked differences between the groups in causes of death, suggesting substantial differences between the groups and a high risk for bias.

One study from British Columbia, Canada, describes results for breast cancer mortality among women aged 50-79, comparing two different time periods during which intervals for screening mammography changed. From 1988, when the SMPBC started, through June 1997, all women aged 40-79 were advised to have annual screens. In July 1997, women aged 50-79 were advised to undergo biennial screens, while the recommendations for women aged 40-49 remained unchanged. The breast cancer mortality ratio for women 50-79 who had biennial screening compared to women in the same age group who had annual screening was not significantly increased (1.06; 95% CI, 0.76 to 1.46); despite an increase in the number of screen-detected cases with positive nodes, survival was also not changed. This study also evaluated the change in breast cancer mortality among women aged 40-49, for whom the screening interval did not change, comparing the mortality rates before and after the policy change, and there was no difference in mortality rates (which would be expected).

One limitation of this before/after study design is that changes in treatment effectiveness may play a role in similar mortality rates—in other words, if the mortality advantage of more frequent screening is due to increasing the detection of more rapidly progressive cancers before progression, and changes in available treatments improve mortality in more advanced disease, then one would expect minimal differences in mortality. Another possibility is improved sensitivity of mammography which balances the effect of less frequent screening.

Model-based Estimates

Tables 37 and 38 present estimates of the effect of annual versus biennial screening on breast cancer mortality from the “exemplar” model from the CISNET analysis for the USPSTF, by age at starting screening (stopping after 69) and age at stopping (starting at age 40). We note that results from other models or confidence intervals around the estimates are not presented, but the CISNET analysis paper states that results of other models were consistent with these. Incremental results (number of deaths per 100,000 prevented with annual screening compared to biennial screening) were calculated from the data presented in the table.
Table 37. Estimated Lifetime Cancer Deaths Prevented per 100,000 by Screening Interval, Stratified by Age at Starting Screening. The model simulates a cohort of women with screening starting at the specified age at the specified interval and stopping after age 69.

<table>
<thead>
<tr>
<th>Age to Start Screening</th>
<th>Interval</th>
<th>Cancer Deaths Prevented per 100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Compared to No Screening</td>
</tr>
<tr>
<td>60</td>
<td>Biennial</td>
<td>340</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>460</td>
</tr>
<tr>
<td>55</td>
<td>Biennial</td>
<td>490</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>610</td>
</tr>
<tr>
<td>50</td>
<td>Biennial</td>
<td>540</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>730</td>
</tr>
<tr>
<td>45</td>
<td>Biennial</td>
<td>620</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>800</td>
</tr>
<tr>
<td>40</td>
<td>Biennial</td>
<td>610</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>830</td>
</tr>
</tbody>
</table>

Table 38. Estimated Lifetime Cancer Deaths Prevented per 100,000 by Screening Interval, Stratified by Age at Stopping Screening. The model simulates a cohort of women with screening starting at age 50 at the specified interval and stopping after the specified age through age 100.

<table>
<thead>
<tr>
<th>Age to Stop Screening</th>
<th>Interval</th>
<th>Cancer Deaths Prevented per 100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Compared to No Screening</td>
</tr>
<tr>
<td>69</td>
<td>Biennial</td>
<td>540</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>730</td>
</tr>
<tr>
<td>74</td>
<td>Biennial</td>
<td>750</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>950</td>
</tr>
<tr>
<td>79</td>
<td>Biennial</td>
<td>940</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>1110</td>
</tr>
<tr>
<td>84</td>
<td>Biennial</td>
<td>960</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>1220</td>
</tr>
</tbody>
</table>

Qualitatively:
- The estimated lifetime number of breast cancer deaths prevented by annual screening compared to biennial screening increases as the age to start screening is lowered.
- The estimated lifetime number of breast cancer deaths prevented by annual screening compared to biennial screening increases as the age to stop screening is raised.

Discussion/Conclusions: Screening Interval and Breast Cancer Mortality
- There is limited direct evidence on the effect of screening interval on breast cancer mortality. Indirect evidence from RCTs suggests some benefit from more frequent screening in younger women (Table 36), but this was not observed in the one relevant cohort study reviewed; of note, this study had substantial methodological issues. Model-based estimates suggest there may be greater effect of screening interval on younger women. Given that cancers in younger women are likely to be more aggressive, more frequent screening would in theory be needed to detect faster-growing tumors before they became symptomatic, or had metastasized. In the U.S.-based BCSC registry, stage distribution was significantly improved with annual screening compared to biennial screening in women under 50, particularly for women with dense breasts, but not in women 50 years and older. Since stage distribution is a surrogate for survival (but not necessarily mortality), this finding is consistent with the possibility of a benefit for more frequent screening in younger (or premenopausal) women. An analogy from another
cancer site might be ovarian cancer, where there are no physical barriers to metastasis and the time of progression from local (confined to the ovary) and distant (metastases to other intra-abdominal organs) is likely to be short; model-based analyses suggest that shorter screening intervals are necessary to maximize mortality reduction.\textsuperscript{192,193}

- There is some consistency to the evidence that a more frequent screening interval reduces breast cancer mortality in women 40-49 years; however, there is substantial risk of bias in the observational studies (e.g., younger women who undergo more frequent screening may be at increased risk of breast cancer, or may have other characteristics that affect post-diagnosis mortality, such as better adherence to therapeutic recommendations). On the other hand, the one study which did not show an effect of screening interval\textsuperscript{187} has a high risk of bias because of the likelihood of substantial differences between the groups. Because (a) the evidence in favor of a comparative benefit from annual screening on mortality from the RCTs is indirect, and (b) there is substantial risk of bias against a benefit for annual screening on mortality in the one observational study directly comparing mortality across different intervals in younger women, we judge the quality of the evidence for reduced mortality with annual screening compared to biennial among women 40-49 years as LOW, and evidence for the magnitude of effect as LOW.

- For women 50 and older, the limited evidence suggests no measurable difference in mortality comparing annual to biennial screening, but the only direct evidence is a single study limited by risk of bias. We judge the quality of the evidence for no difference in mortality by screening interval in women over 50 as LOW, and evidence for the magnitude of effect as LOW.

**Life Expectancy**

As noted in the section on KQ 1, life expectancy is rarely, if ever, directly estimated from empiric studies, but is usually estimated from models.

**Model-based Estimates**

Tables 39 and 40 present the same CISNET model estimates, stratified by screening interval within a given age to start and stop screening.

**Table 39. Effect of Screening Interval on Gains in Life Expectancy by Age of Starting Screening.** \textsuperscript{30} The model simulates a cohort of women with screening starting at the specified age at the specified interval and stopping after age 69.

<table>
<thead>
<tr>
<th>Age to Start Screening</th>
<th>Interval</th>
<th>Life-years Gained per 100,000 Women</th>
<th>Days Gained per Woman</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Compared to No Screening</td>
<td>Compared to 5 Years Later Age to Start</td>
</tr>
<tr>
<td>60</td>
<td>Biennial</td>
<td>52</td>
<td>19.0</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>69</td>
<td>25.2</td>
</tr>
<tr>
<td>55</td>
<td>Biennial</td>
<td>80</td>
<td>29.2</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>102</td>
<td>37.2</td>
</tr>
<tr>
<td>50</td>
<td>Biennial</td>
<td>99</td>
<td>36.1</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>132</td>
<td>48.2</td>
</tr>
<tr>
<td>45</td>
<td>Biennial</td>
<td>116</td>
<td>42.3</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>152</td>
<td>55.5</td>
</tr>
<tr>
<td>40</td>
<td>Biennial</td>
<td>120</td>
<td>43.8</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>164</td>
<td>59.9</td>
</tr>
</tbody>
</table>
Table 40. Effect of Screening Interval on Gains in Life Expectancy by Age of Stopping Screening.\textsuperscript{30}

The model simulates a cohort of women with screening starting at age 50 at the specified interval and stopping after the specified age.

<table>
<thead>
<tr>
<th>Age to Stop Screening</th>
<th>Interval</th>
<th>Life-years Gained per 100,000 Women</th>
<th>Days Gained per Woman</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Compared to No Screening</td>
<td>Compared to 5 Years Earlier Age to Stop</td>
</tr>
<tr>
<td>69</td>
<td>Biennial</td>
<td>99</td>
<td>36.1</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>132</td>
<td>33</td>
</tr>
<tr>
<td>74</td>
<td>Biennial</td>
<td>121</td>
<td>44.2</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>156</td>
<td>35</td>
</tr>
<tr>
<td>79</td>
<td>Biennial</td>
<td>130</td>
<td>47.5</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>170</td>
<td>40</td>
</tr>
<tr>
<td>84</td>
<td>Biennial</td>
<td>138</td>
<td>50.4</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>178</td>
<td>40</td>
</tr>
</tbody>
</table>

Qualitatively:
- The estimated gains in life expectancy from increasing screening frequency from biennial to annual screening are greater as the age of beginning screening is lowered (16.1 additional days for annual screening compared to biennial screening beginning at age 40, compared to 12.0 additional days for annual compared to biennial screening when starting screening at age 50).

Discussion/Conclusions: Screening Interval and Life Expectancy
- There is no direct evidence of the impact of screening interval on life expectancy, and model-based estimates are dependent on the reliability of estimates of the effects of interval on mortality at different ages. Since we view the quality of evidence for the effect of screening interval on mortality as LOW, we judge the quality of evidence for the effect of screening interval on life expectancy as VERY LOW.

Overdiagnosis/Overtreatment

RCTs/Observational Studies
- We did not identify any direct estimates of the effect of screening interval on overdiagnosis.
- In an analysis of BCSC data, Dittus et al.\textsuperscript{191} reported on the effects of screening interval on the proportion of detected lesions that were DCIS versus invasive, stratified by menopausal status and BMI. Among premenopausal women, the relative proportion of lesions that were DCIS was higher with biennial screening compared to annual, while the opposite was true among postmenopausal women. This trend was consistent across all BMI classes, although it was only statistically significant for normal weight postmenopausal women (Table 41).
Table 41. Effects of Screening Interval on Proportion of DCIS vs. Invasive by Menopausal Status and BMI

<table>
<thead>
<tr>
<th>Menopausal Status</th>
<th>Normal Weight</th>
<th>Overweight</th>
<th>Obese</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proportion DCIS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Premenopausal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biennial</td>
<td>30.8%</td>
<td>25.5%</td>
<td>24.5%</td>
</tr>
<tr>
<td>Annual</td>
<td>24.8%</td>
<td>21.9%</td>
<td>20.2%</td>
</tr>
<tr>
<td>Odds Ratio (95% CI) for Invasive, Biennial vs. Annual*</td>
<td>0.71 (0.48 to 1.06)</td>
<td>0.70 (0.38 to 1.29)</td>
<td>0.61 (0.29 to 1.24)</td>
</tr>
<tr>
<td>Postmenopausal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion DCIS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biennial</td>
<td>17.5%</td>
<td>16.2%</td>
<td>18.2%</td>
</tr>
<tr>
<td>Annual</td>
<td>25.8%</td>
<td>20.1%</td>
<td>20.7%</td>
</tr>
<tr>
<td>Odds Ratio (95% CI) for Invasive, Biennial vs. Annual*</td>
<td>1.43 (1.02 to 2.02)</td>
<td>1.21 (0.83 to 1.76)</td>
<td>1.43 (0.94 to 2.16)</td>
</tr>
</tbody>
</table>

*Adjusted for registry, race/ethnicity, age, and family history of breast cancer.

Abbreviations: BMI=body mass index; CI=confidence interval; DCIS=ductal carcinoma in situ

Model-based Estimates

The CISNET collaborators reported that biennial screening strategies reduced overdiagnosis compared to annual strategies, “...but by much less than one half.” Details, including whether there was any age effect, were not provided.

Discussion/Conclusions: Screening Interval and Overdiagnosis

- We did not identify any direct evidence of an effect of screening interval on overdiagnosis or overtreatment.
- Model-based studies suggest that screening interval affects overdiagnosis, but do not describe the magnitude of effect, or any age-related differences.
- If screening interval does affect the probability of overdiagnosis, it may vary by age.
  - As discussed above, there are several different definitions of overdiagnosis. For neoplasms where spontaneous regression of mild, potentially premalignant changes is not uncommon (e.g., cervical intraepithelial neoplasia), more frequent screening will be more likely to detect disease that would possibly go away on its own. With breast cancer, the assumption is that some pre-invasive lesions, such as DCIS, or even small invasive cancers will not become symptomatic and/or metastatic, not that they will spontaneously regress. If non-progressive in situ lesions or very slow-growing invasive cancers do not spontaneously regress, screening intervals may not affect the probability of detection through screening unless the cancers become symptomatic, or are detected serendipitously through some other means, especially in younger women.
  - However, in older women, more frequent screening might detect slow-growing in situ or small invasive cancers that would not have become symptomatic before death from another cause.
  - The proportion of cases that are DCIS versus invasive was observed to vary by interval in one analysis of the BCSC registry data, but this was only significant among normal weight, postmenopausal women, where annual screening resulted in a significantly increased proportion of DCIS lesions. A relatively higher
proportion of invasive cancers with annual screening in younger women compared to older women is consistent with the possibility that cancers in younger women are more likely to be rapidly progressive (since rapidly progressive cancers have a greater chance of being detected clinically as screening intervals lengthen). Even if this is the case, however, the implications for inferences about the effect of screening interval on overdiagnosis are unclear because of the fundamental uncertainty about the natural history of DCIS.

- We judge the quality of evidence on both direction and magnitude of the effect of interval on overdiagnosis to be VERY LOW, primarily because of the fundamental uncertainty about measuring overdiagnosis.

**False Positives**

**RCTs**

We did not identify any evidence from RCTs on the effect of screening interval on false positives.

**Observational Studies**

**Recall**

Six cohort studies evaluated the outcome of having a false positive screening mammogram requiring follow-up (recall) with different screening intervals.\(^{87,92,186,189-191}\) All studies defined a false positive similarly, based on BI-RADS scores of screening mammograms which would require follow-up, with no diagnosis of carcinoma or carcinoma in situ within 12 months or before the next screening visit. All studies demonstrated a higher risk or probability of having a false positive with recall of a screening mammogram with shorter screening intervals compared to longer screening intervals.

One study examined the probability of having at least one false positive screening mammogram with recall for women age 40-49 and 50-74 over a 10-year period, by their breast density status, as well as by their hormone therapy status for those aged 50-74.\(^{87}\) This study found that false positives were generally higher for women with extremely dense and heterogeneously dense breasts compared to women with scattered fibroglandular densities and fatty breasts; however, for all four breast density groups, for both women 40-49 and 50-74, probabilities for false positives were higher with screening intervals of 1 year compared to 2 years and 3 years. The probability of having a false positive for those screened yearly was uniformly over twice the probability of those having screens done every 3 years for all groups by age, breast density, and hormone status. Among women with denser breasts, probabilities ranged from approximately 60-69% for women undergoing yearly exams to 28-33% for women undergoing screening every 3 years.\(^{87}\)

Similar to the study above, another study examined the probability of false positive screening mammograms with recall in women aged 40-59 using the same database, the BCSC.\(^{92}\) This study examined recall rates by screening intervals, defined as time since last mammogram. This study also examined the probability of having a false positive screen with recall by the age at which screening was initiated—age 40 or age 50. Again, annual screening was associated with a higher cumulative probability of having a false positive recall than were longer screening intervals.\(^{92}\) Cumulative probabilities over 10 years were higher for women beginning screening at age 50
compared to age 40, presumably because of a higher initial adjusted false positive rate among the
closer counterparts. Of note, the probability that any given individual mammogram would result in a
false positive result increased with increasing duration since the last screen: using a 9-18 month
interval as the reference, the adjusted odds ratio (OR) for a false positive result for an interval of
19-30 months since last screen was 1.13 (95% CI, 1.03 to 1.19), and for greater than 30 months
the OR was 1.33 (1.26 to 1.40). A similar finding of decreased specificity with increasing time
since last mammogram was also observed in an earlier analysis of the BCSC registry.

Another study examined the probability of having a false positive with need of recall among
older women aged 66-89 at time of screening mammogram. Given that it is unclear whether or
not screening for breast cancer is beneficial among women of this age due to other comorbid
conditions, this study stratified results by the women’s Charlson index, a score which takes into
account comorbid conditions that are independent risk factors for death. In this study, over a 10-
year period, the probability of having at least one false positive screening mammogram with
recall was higher for annual screening compared to biennial screening for all women, aged 66-74
and 75-89 years old, with a Charlson score of 0 and with a Charlson score ≥1. For women in
both age groups and with either none or at least one comorbidity, the probability of having at
least one false positive screen with recall over this 10-year period ranged from 47-50% for those
undergoing annual screens to 27-30% for those undergoing screens every 2 years. Similar
results of the effect of screening interval on false positive recall were observed when stratified by
race/ethnicity and BMI.

**Biopsy**

Five cohort studies evaluated the outcome of having a false positive screening mammogram
requiring biopsy with different screening intervals. Similar to the results described for false positive
with recall, in all six studies, shorter screening intervals were associated with a higher probability of false positive screens resulting in biopsies compared to longer screening
intervals.

In the study examining the probability of having at least one false positive screening
mammogram with biopsy for women by age, breast density status, and hormone therapy status
for those aged 50-74, 8-12% of women getting annual screening had a false positive screen
requiring biopsy, compared to 3-7% who had screening at 2- or 3-year intervals. While false
positive results of screening mammograms with biopsies were more common among women
with denser breasts, this pattern of a higher percentage of women getting these results with
shorter screening intervals was consistent for women aged 40-49, women with denser and less
dense breasts, and women on or not on hormone replacement therapy. Similar results were found
among older women with or without comorbidities, across racial/ethnic groups and by
BMI.

Finally, the study of the Massachusetts General Hospital Avon Comprehensive Breast Center
database, similar to the results of false positives with recall, found that those women who had
screening mammograms more frequently during their follow-up period had a higher likelihood of
having a false positive screening requiring biopsy compared to those women who had fewer
screens during their follow-up period. Among the women followed for 10 years, 9-11% of
those women who had 8-10 screens during this time had biopsies performed that did not reveal
cancer, compared to 6-7% of the women who had 1-2 screening mammograms performed during
the same time period.

Overall, in all of these studies, the qualitative relationship between cumulative false positive
biopsy risk (greater with annual compared to biennial screening) was similar to that reported for
any false positive result, although the absolute risk of a false positive biopsy was substantially lower than for any false positive result (with cumulative probabilities of false positive biopsies approximately 5-10% of the cumulative probability of false positive recall at any given screening interval).

**Model-based Estimates of False Positives**

The estimated effect of screening interval on cumulative total false positives and false positive biopsies at a given age to start (Table 42) and stop (Table 43) screening from the same CISNET “exemplar” model are shown below.  

**Table 42. Estimated Effect of Screening Interval on False Positives and False Positive Biopsies by Age of Starting Screening (Assuming Screening Stops after Age 69)**

<table>
<thead>
<tr>
<th>Age to Start Screening</th>
<th>Interval</th>
<th>Total False Positives per 100,000 Women</th>
<th>False Positive Biopsies per 100,000 Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Compared to No Screening</td>
<td>Compared to Biennial</td>
</tr>
<tr>
<td>60</td>
<td>Biennial</td>
<td>34,000</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>60,000</td>
<td>26,000</td>
</tr>
<tr>
<td>55</td>
<td>Biennial</td>
<td>59,000</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>95,000</td>
<td>35,000</td>
</tr>
<tr>
<td>50</td>
<td>Biennial</td>
<td>78,000</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>135,000</td>
<td>57,000</td>
</tr>
<tr>
<td>45</td>
<td>Biennial</td>
<td>105,000</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>180,000</td>
<td>75,000</td>
</tr>
<tr>
<td>40</td>
<td>Biennial</td>
<td>125,000</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>225,000</td>
<td>100,000</td>
</tr>
</tbody>
</table>

**Table 43. Estimated Effect of Screening Interval on False Positives and False Positive Biopsies by Age of Stopping Screening (Assuming Screening Starts at Age 50)**

<table>
<thead>
<tr>
<th>Age to Stop Screening</th>
<th>Interval</th>
<th>Total False Positives per 100,000 Women</th>
<th>False Positive Biopsies per 100,000 Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Compared to No Screening</td>
<td>Compared to Biennial</td>
</tr>
<tr>
<td>69</td>
<td>Biennial</td>
<td>78,000</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>135,000</td>
<td>57,000</td>
</tr>
<tr>
<td>74</td>
<td>Biennial</td>
<td>94,000</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>157,000</td>
<td>63,000</td>
</tr>
<tr>
<td>79</td>
<td>Biennial</td>
<td>102,000</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>174,000</td>
<td>72,000</td>
</tr>
<tr>
<td>84</td>
<td>Biennial</td>
<td>113,000</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>188,000</td>
<td>75,000</td>
</tr>
</tbody>
</table>

Screening interval has a greater effect on false positives than age alone, but rates go up much more rapidly with earlier age to start than later age to stop. Because it is possible for a woman to have more than one false positive, some combinations of ages to start and stop result in estimates of the number of false positives recalls to be greater than the number of women screened. A 1998 estimate of the 10-year cumulative probability of screening in a large health maintenance organization (which did not meet our date criteria for inclusion) reported 23.8% cumulative risk of at least one false positive mammograms, with 4% of women having two or more false positives. The more recent studies included in our review did not report the proportion of women with multiple false positives.
For false positive biopsies, the qualitative pattern was similar, although the estimated probability of a false positive biopsy was substantially less than for a false positive test requiring recall—lifetime estimated probabilities for a false positive biopsy were 90-95% lower for a false positive biopsy result than for any false positive test result at every age to start and stop screening for both annual and biennial screening.

Note that although these estimates are based on the BCSC data and incorporate differences in specificity associated with first versus subsequent screens, age, and screening interval, the estimates are for the total population, rather than for individual women—estimates of total false positives greater than the size of the population represent some women having more than one false positive.

Table 44 presents estimates for the lifetime risk of total false positives and false positive biopsies using the model developed for this report, which also uses the BCSC estimates (false positive rates are higher with first than subsequent screens, higher with older age to start and, for biopsies, with older age in general, and higher with longer screening intervals). False positives are restricted, so that these represent cumulative probabilities of at least one outcome rather than total (when unrestricted, total false positives exceeded 100%, similar to the results with the CISNET models).

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Total False Positives</th>
<th>False Positive Biopsies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biennial, Start Age 50</td>
<td>71.3%</td>
<td>16.1%</td>
</tr>
<tr>
<td>Biennial, Start Age 45</td>
<td>78.4%</td>
<td>18.2%</td>
</tr>
<tr>
<td>Biennial, Start Age 40</td>
<td>82.3%</td>
<td>19.4%</td>
</tr>
<tr>
<td>Annual, Start Age 50</td>
<td>83.6%</td>
<td>22.5%</td>
</tr>
<tr>
<td>Annual, Start Age 45</td>
<td>88.9%</td>
<td>25.3%</td>
</tr>
<tr>
<td>Annual, Start Age 40</td>
<td>92.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>

Discussions/Conclusions: Effect of Screening Interval on False Positives

- Not surprisingly, increasing screening frequency consistently increases the cumulative likelihood of a false positive result in observational studies.
- Evidence from a U.S. community-based registry suggests that the probability of any given mammogram resulting in a false positive result increases as the interval since last screen increases. This may be the result of radiologists lowering their threshold for further evaluation based on both a higher estimate of prior probability given the longer time since last screen, and increased concern about the development of an interval cancer given a longer expected time to next screen. However, even with this higher individual probability of a false positive with longer intervals, the cumulative probability remains higher with shorter intervals for both.
- For total false positive results, the estimated 10-year cumulative probability is higher with annual screening (approximately 61%) compared to biennial screening (approximately 42%) whether women begin screening at age 40 or age 50 based on an analysis of the BCSC data, due to a higher probability of an initial false positive at first examination in older women. The absolute difference in cumulative 10-year false positive biopsy rates is approximately 2% higher with annual screening than with biennial screening at either starting age, and 2% higher starting at 50 compared to starting at 40 (7.0% for annual screening beginning at 40, 9.4% for annual screening beginning at 50,
4.8% for biennial screening beginning at 40, and 6.4% for biennial screening beginning at 50).

- Conversely, the model-based estimated lifetime probability of the effect of screening interval on false positive recall or biopsy increases with an earlier age to start screening. These results are not necessarily inconsistent—it is entirely possible for the cumulative probability of a false positive result to be lower in the 10 years after beginning screening in women aged 40-49 compared to women who begin ages 50-59, but for the lifetime cumulative risk to be higher for women who begin screening at younger ages (i.e., the 35-year cumulative probability compared to the 25-year cumulative probability), because of more opportunities for a false positive to occur. This highlights the inherent uncertainty in estimating quantitative effects beyond the time period for which data are available—when estimates are available only for the 10 year cumulative risk for a given age group, estimating cumulative probabilities over a longer time horizon requires making decisions about whether to apply observed probabilities to longer time periods, which may lead to over- or underestimation.

- Estimates of lifetime risk also vary depending on whether the total number of false positives (which include women who experience more than one) or the number of women experiencing at least one false positive are used in the numerator. The former is a better measure of population burden, while the latter is a better indicator for individual women.

- As discussed in the section on false positives in KQ 1, the variability in false positive risk based on patient characteristics such as breast density, the high degree of variability in false positive rates by radiologists, and the potential effects of geographic mobility and changes in insurance coverage on the availability of prior films (which decreases false positive probability), create additional uncertainty around estimates of the lifetime risk of a false positive for an individual woman.

- Because of its consistency across a variety of studies and patient subgroups, in the setting of opportunistic community practice in the U.S., we judge the strength of evidence that more frequent screening increases the cumulative risk of both false positives test results and false positive biopsies to be HIGH; however, the strength of evidence for the estimate of the magnitude of the effect, for both test results and biopsies, is at best MODERATE for intervals up to 10 years. For longer time horizons, the strength of evidence for the quantitative estimates is LOW, since it based primarily on modeling studies of moderate or low quality (compared to direct evidence—as previously discussed, evidence from the most sophisticated modeling exercise is limited by indirectness and the necessity of unverifiable assumptions about unobserved, often unobservable, events).

- The quality of evidence for a greater cumulative lifetime risk of false positives with a younger age to start screening is MODERATE, but for the quantitative estimate is LOW.

**Quality-adjusted Life Expectancy**

The limitations noted under KQ 1 for estimates of quality-adjusted life expectancy also hold here. We summarize qualitative effects of screening interval that are consistent across all models discussed above.

- Increasing screening frequency results in gains in unadjusted life expectancy, but the incremental gains decrease as screening interval becomes smaller.
• Incremental gains in quality-adjusted life expectancy as screening interval decreases are even smaller, especially if screening itself and false positive results are assigned a disutility. The more often screening occurs, the greater the cumulative impact of these small disutilities on quality-adjusted life expectancy. Because both breast cancer and breast cancer mortality are much less common, and the gains from more frequent screening much smaller, the losses from the minor utilities contribute more to net quality-adjusted life expectancy than the gains from avoiding breast cancer death.

• Although none of the models explicitly quantifies the effect of assumptions about overdiagnosis on quality-adjusted life expectancy, all note that including it decreases estimated QALYs with screening, with a variable effect of screening interval.

Discussion/Conclusions: Effect of Screening Interval on Quality-adjusted Life Expectancy

• Although the qualitative effects of screening interval on quality-adjusted life expectancy are consistent across studies, we judge the quality of evidence to be LOW, based on the inherent uncertainty in the models (especially surrounding overdiagnosis, which may have a substantial impact on quality-adjusted life expectancy), the variability in quantitative estimates derived from the models, and the concerns about the utility weights used raised in KQ 1.

Harm-benefit Trade-offs: False Positives per Death Prevented

Published Estimates

Tables 32 and Figures 15 and 16, above, present the joint effects of screening interval and ages to start and stop screening on total false positives and false positive biopsies per cancer death prevented based on the CISNET analyses. 30 As previously noted, biennial screening at ages 45 or 40 are eliminated by extended dominance—only annual screening strategies are potentially reasonable options as the “acceptable” threshold for the harm-benefit trade-off increases.

Model-based Estimates

Estimates of the joint effect of screening interval on mortality and false positive probability over a lifetime are probably best made using models of the underlying natural history of breast cancer, with test sensitivity and specificity, adjusted for age, screening interval, and potentially other factors such as distribution of breast density used to impute both outcomes—i.e., models such as the CISNET models. The simpler model based on incidence-based mortality we have used for generating alternative estimates for this report can account for the effect of screening interval on false positive outcomes, but without reliable estimates of both individual relative risks for mortality reduction by interval and the proportion of women in the U.S. undergoing annual versus biennial screening, deriving mortality estimates directly is impossible.

Discussion/Conclusions: Harm-benefit Trade-offs

• Based on the CISNET analysis, higher incremental ratios for false positive test results per breast cancer death prevented are seen when screening begins before 50, due to the combined effect of lower mortality and higher false positive rates. The increase in the false positive/deaths prevented ratio between annual and biennial screening beginning at
age 50 and age 40 (approximately 1.5 times higher) is greater than the increase in the false positives/life-year gained ratio (1.3 fold increase) because of the added years of life expectancy.

- The CISNET results were similar for false positive biopsies results (higher ratios as age to start screening was lowered for a fixed stopping age, with much less of an effect as age to stop screening was increased).

- Because these estimates are necessarily based on modeling which uses parameter estimates with a high degree of uncertainty, we judge the quality of evidence for these qualitative effects to be MODERATE, but for the quantitative estimate LOW.

- The false positive test results per death prevented ratios for annual compared to biennial screening at any given age are well within the acceptable range reported in the 1997 survey by Schwartz and colleagues;\textsuperscript{176} we did not identify any similar evidence on an acceptable threshold for false positive biopsies per breast cancer death prevented.

**Key Question 3**

What are the benefits, limitations, and harms associated with clinical breast examination (CBE) among average-risk women 40 years and older compared to no CBE, and how do they vary by age, interval, and participation rates in mammography screening?

**Summary**

**Key Points: Outcomes**

**Breast Cancer Mortality:**

- Direction of effect: The available evidence suggests no effect of CBE alone on breast cancer mortality. This conclusion is based primarily on a single U.S. case-control study, which was graded as moderate quality based on study characteristics. However, this study also found no effect of mammography screening on mortality, which is inconsistent with other studies, particularly other case-control studies. We rate the quality of evidence for this conclusion as VERY LOW. We did not identify any evidence of an incremental mortality benefit of adding CBE to mammography.

- Magnitude of effect: The quality of evidence is VERY LOW, based on imprecision and lack of data on consistency.

**False Positives:**

- Direction of effect: The available evidence suggests that adding CBE to mammography screening increases the false positive rate, based on cohort studies conducted in the U.S., Canada, and Japan, and RCTs conducted in Sudan and India. We rate the quality of evidence for this conclusion as MODERATE based on directness, consistency, and relatively low risk of bias for an observational study.

- Magnitude of effect: In both studies, an estimated 55 false positives were generated for each additional cancer detected. We rate the quality of this evidence as MODERATE based on directness, consistency, and relatively low risk of bias.
Other Critical Outcomes:
- We identified no studies that assessed other critical outcomes for CBE.

Key Points: Harm-benefit Trade-offs
- We did not identify any studies assessing the potential harm-benefit trade-offs of the use of CBE either alone or as an adjunct to mammography or other screening modality.

Description of Included Studies
We identified seven studies (one case-control study, three RCTs, three cohort studies) that evaluated the benefits, limitations, or harms associated with CBE. One U.S.-based case-control study, encompassing two separately published analyses, compared breast cancer mortality after screening with mammography and/or CBE, mammography alone, or CBE alone versus no screening. In addition, an early U.S.-based RCT assessed breast cancer mortality and survival among women randomized to either annual film mammography plus annual CBE or usual care. Five studies assessed the number of false positives, defined as recalls or interventions which led to a benign diagnosis on either follow-up or pathology. Of these, one cluster randomized controlled trial in India compared three rounds of triennial CBE to no screening among healthy women aged 30-69 to determine if CBE alone can reduce the incidence rate of advanced cancers and breast cancer mortality. To date, this study has reported on only one round of CBE screening and assessed the performance characteristics of CBE compared to no screening. Another RCT from Sudan determined the false positive rate in average-risk women screened with CBE. One prospective cohort study from the U.S. reported on the potential contribution of CBE alone or added to mammography compared to mammography alone and assessed performance statistics among women aged 40 and over, while another Japanese study (no age range recorded) compared false positives between three different methods of screening—CBE (in combination with mammography or ultrasound), mammography, and ultrasound. The remaining study was a Canadian retrospective cohort study that estimated the number of false positives in women (age range 50-69) screened with mammography versus mammography and/or CBE. These studies included women of average and high risk but did not stratify false positives by risk status.

More detailed characteristics of the included studies are summarized in Appendix Table G-3. Detailed Synthesis
We classified studies and organized findings by outcome. The low number of studies, and the heterogeneity in design, prohibited quantitative synthesis; therefore, we synthesize findings qualitatively.

Breast Cancer Mortality
Study Results
Two studies (one high and one moderate quality) assessed the impact of CBE on mortality. The first study was the Health Insurance Plan (HIP) RCT. This U.S.-based trial, started in 1963, randomized approximately 62,000 women aged 40 to 64 who were HIP members.
for at least 1 year to annual mammography plus CBE or usual care. Screenings continued annual for 3 years. Mammography and CBE were conducted independently. CBE was conducted by a physician who was usually a surgeon, and mammography was via two-view film. About 67% of women randomized to screening received the initial exam. Even if women disenrolled from HIP, they still continued to receive breast cancer screenings. There were no significant baseline differences between women randomized to screening and control; however, there were significant baseline differences between women in the intervention group who initiated screening versus those who did not (refusers).

By 1975, (through 9 years of follow-up), women randomized to receive three rounds of annual mammography plus CBE experienced 30% fewer breast cancer deaths compared to those in the control group (91 vs. 128 deaths; p<0.01). Although mortality was not reported by mode of detection, case fatality rate over time (essentially, survival) in the control group was 46.7 per 100 breast cancer cases compared to 35.2 per 100 breast cancer cases (p<0.01) in the mammography plus CBE group, controlling for 1 year of lead time bias and 7 years of follow-up. The overall case fatality rate per 100 cases of breast cancer detected at screening was 28.3 at 8 years following diagnosis. When broken down by screening modality, the case fatality rate among breast cancer cases detected at screening was 41.4 per 100 for cancer detected for both mammography and CBE, 14.4 per 100 for mammography only, and 31.8 per 100 for CBE only, consistent with detection of smaller tumors with mammography.

Two analyses of a U.S.-based case-control study, rated moderate quality based on study characteristics, assessed breast cancer mortality associated with the three definitions of screening compared to no screening: mammography and/or CBE, mammography only, and CBE only. The total study population included a combination of average- and high-risk women aged 40-65 who were enrolled in six health plans. Female plan members who died of breast cancer between 1983 and 1998 (n=1351) were matched with cases (n=2501) on age, health plan, and level of breast cancer risk. Elevated risk was defined as a documented history of a previous breast biopsy and family history of breast cancer.

For average-risk women aged 40-65, obtaining a CBE in the previous 3 years resulted in no significant difference in breast cancer mortality compared to no screening (OR 0.94; 95% CI, 0.79 to 1.12). The association between mortality and receipt of CBE in past 3 years was greater for women aged 40 to 65 at increased risk, but the difference still was not statistically significant (OR 0.80; 95% CI, 0.59 to 1.08). Of note, in this study, mammography alone or in combination with CBE was also not associated with a decreased risk of breast cancer death.

In a separately published analysis of data from the same study, the authors identified women who had had a screening CBE within 1 year of breast cancer diagnosis among women who eventually died of breast cancer. Only 105 of 485 had a screening CBE diagnosis of “suspicious” or “indeterminate,” for an estimated sensitivity of 21.6% (95% CI, 18.0 to 25.6%). Sensitivity was significantly decreased when Pap tests were performed at the same visit (suggesting less time was given to the CBE), or for advanced stage cancers.

**Effects of CBE on Mortality at Different Ages**

The HIP trial also reported case fatality effects stratified by age. While there was a statistically significant difference in case fatality/survival rates favoring use of annual mammography plus CBE, age-stratified analysis demonstrated that nearly all of the effect of screening was observed in women aged 50-59 (53.5 vs. 32.1 breast cancer fatalities per 100 breast cancer cases; p<0.01). Above age 59, there was no significant difference in case fatality
rates for screening versus usual care (32.6 vs. 40.5 deaths per 100 cases), and for women aged 40 to 49, the case fatality rates were nearly identical in the two groups (42.0 vs. 40.9 deaths per 100 cases).

The above-cited case-control study also reported effects on mortality stratified by age. Mortality was not significantly reduced with CBE compared to no screening in either women aged 40 to 49 (OR 0.91; 0.73 to 1.13) or aged 50 to 65 (OR 0.98; 0.74 to 1.31).

One potential criticism of the case-control study is that the case definition of “breast cancer” included both DCIS and invasive cancers; since DCIS is much more likely to be detected via mammography than CBE, inclusion of DCIS cases could potentially affect the relative impact of CBE on overall breast cancer mortality. However, DCIS is rarely fatal, and even more rarely listed as the cause of death—for example, the number of death certificates reporting invasive breast cancer as the primary cause of death in the U.S. between 1999 and 2010 was 498,046, while the number listing DCIS as the primary cause of death was 10. Thus, inclusion of DCIS in the case definition seems unlikely to have biased the results against any benefit from CBE. However, it is possible that the relatively short interval for definition of receipt of a CBE (3 years) resulted in an underestimation of any effects of CBE on longer term breast cancer mortality.

Discussion/Conclusions: Breast Cancer Mortality
- There was no evidence of reduced mortality with CBE alone, based on very low quality evidence (single case-control study, pre-2000 cancer deaths, no observed effect of mammography on mortality, wide confidence intervals). Data from the HIP study does not provide any interpretable evidence on either the benefit of CBE alone, or the incremental benefit of adding CBE to mammography on breast cancer mortality. We rate the quality of this evidence as VERY LOW.

Life Expectancy
We identified no studies that assessed this outcome for CBE.

Overdiagnosis
We identified no studies that assessed this outcome for CBE.

False Positives

Overall Estimates of False Positives
Five studies—one high, two moderate, and two low quality—reported on the number of false positives resulting from screening with CBE. All studies included a combination of average- and high-risk women, and none stratified their results by age groups. Across all studies, false positives were defined as any recalls that required further testing with subsequent benign diagnosis on either follow-up or pathology. Although the outcome of interest for these trials was the same, populations, comparators, and who performed CBE (e.g., community health worker, registered nurse) varied greatly; therefore, a meta-analysis was not performed. Overall, false positive rated ranged from 0.9% (compared to no screening) to 8.7% (comparing mammography to CBE + mammography). Key false positive results for each trial are described below.

One high quality U.S.-based prospective cohort study reported on the potential contribution of CBE alone or added to mammography compared to mammography alone in
detecting invasive cancers among 61,688 women aged 40 and over who received at least one breast cancer screening from 1996 to 2000 identified through the Breast Cancer Screening Program from the Group Health Cooperative at Puget Sound. Mammography and CBE were at 1- to 2-year intervals, and two-view mammograms were performed. A registered nurse performed the CBEs. Sensitivity of detecting invasive cancers increased when adding CBE to mammography, but specificity and positive predictive value decreased with the addition of CBE. Using data from the full cohort, we calculated the false positive rates for mammogram alone, mammography plus CBE, and CBE alone; these were 0.89%, 3.0%, and 2.2%, respectively. Sensitivity for detecting cancer was increased by the addition of CBE—for the entire group, 0.4 additional cancers were detected per 1000 women with the addition of CBE, with an extra 20.7 false positives (55 false positives per incremental cancer detected). Positive predictive value for mammography alone was 43.9%, declining to 20.1% with the addition of CBE. Both the increase in sensitivity and decrease in specificity were most pronounced in women with dense breasts.

One Canadian retrospective cohort study, rated moderate quality, compared cancer detection rates and false positive rates between women who received biennial routine breast screening at centers offering mammography alone versus mammography and CBE over a 1-year period between 2002 and 2003 in an organized screening program. Mammography was performed with screen film technique, and all patients were imaged with standard craniocaudal and mediolateral oblique views. All images were interpreted by a single radiologist. CBE was performed by trained and certified nurses at centers offering this service. The cancer detection rate was 5.9 per 1000 women with mammography screening alone and 6.3 per 1000 with screening mammography and CBE. The false positive rate was 6.5% for mammography alone and 8.7% with mammography and CBE. The addition of CBE resulted in an additional 0.4 cancers detected per 1000 women, with a concomitant increase of 22 false positives (or 55 false positives per additional cancer detected). Note the similarity between this study and the U.S.-based study discussed immediately above in incremental false positives per additional cancer detected.

By contrast, a prospective study from Japan, rated low quality, compared sensitivity of screening CBE (in combination with mammography or ultrasound), mammography, and ultrasound in a cohort of 3453 asymptomatic women from 1999 to 2000. All three screening techniques were performed simultaneously, and participants were followed for 2 years in a biennial program. Mammography was performed with a single mediolateral oblique view and interpreted by two radiologists. Ultrasound was performed by a trained technologist using a 7.5 MHz transducer. CBE was performed by surgeons. A total of 530 (15.3%) participants were recalled for additional testing during the study period; 159 (4.6%) after CBE, 279 (8.1%) after mammography, and 165 (4.8%) after ultrasound. During the study and 2-year follow-up period, a total of 13 patients were diagnosed with breast cancer; 11 detected by mammography, 7 by ultrasound, and 2 by CBE, with sensitivities of 61.5%, 53.8%, and 23.1%, respectively, and with false positive rates of 8% for mammography, 5% for ultrasound, and 5% for CBE alone. There were no cases diagnosed on the basis of CBE alone, so the incremental effect of adding CBE to mammography cannot be estimated in this study.

We identified two RCTs of CBE performed by trained community health workers in developing country settings, which, while having low direct applicability to the U.S. setting, provide estimates of false positive rates. One cluster randomized controlled trial of moderate quality in India is comparing three rounds of triennial CBE to no screening among healthy women aged 30 to 69 to determine if CBE alone can reduce the incidence rate of advanced
The first round of screening was initiated in 2006 and completed in 2009. To date, authors have reported only on this one round of screening and only outcomes related to CBE performance statistics. Women aged 30 to 69 were eligible to participate if they had intact breasts and no history of breast cancer. Clusters (n=275) derived from electoral wards were randomized to annual CBE or no screening. In total, 50,366 women in the intervention group had CBE compared to 54,020 in the control group. Trained female community health workers performed the CBE in women’s homes. Women who screened positive were sent to biweekly breast clinics set up by study staff where they were examined by a doctor and sent on for further evaluation, if warranted. Preliminary data from this first round of screening found a false-positive rate of 5.7% (95% CI 5.5% to 5.9%).

A larger RCT from Sudan, rated low quality, enrolled 10,309 women from several villages in Sudan for organized breast cancer screening with CBE from 2010 to 2012. Two counties were randomly assigned by coin toss to receive either the intervention of village women trained to give CBE or the comparator, no training of villagers to give CBE. Participants from villages that had trained volunteers received one screening exam during the 2-year study period. A total of 138 participants were recalled for additional testing. Of these, 20 were lost to follow-up. The remaining 118 (0.9%) had subsequent biopsies, with malignancy diagnosed in 17 (0.16%) and benign changes detected in the remaining 101 participants, resulting in a false positive rate of 0.9%. The control village, consisting of 24,550 women, was not invited to participate in the lay health CBE volunteer program.

**Discussion/Conclusions: False Positive Results**

- Increased false positive rates with addition of CBE to mammography were observed in two large observational studies in the U.S. and Canada, which were consistent in both the direction and magnitude of the observed effect. The cancer detection rate was also improved with addition of CBE. In both studies, an additional 55 false positives occurred for each additional cancer detected. We judge the quality of evidence for increasing false positives by adding CBE to mammography as MODERATE, based on consistency, directness, and precision, with a decrease for risk of bias.

**Quality-adjusted Life Expectancy**

We identified no studies that assessed this outcome for CBE.

**Harm-benefit Trade-offs**

We did not identify any studies assessing the potential harm-benefit trade-offs of the use of CBE either alone or as an adjunct to mammography or other screening modality using the critical outcomes specified by the ACS Guidelines Panel.

**Key Question 4a**

Among women with an increased risk of breast cancer due to factors known PRIOR to the onset of screening (e.g., family history, BRCA mutation carrier, history of chest irradiation), what are the relative benefits, limitations, and harms associated with different screening modalities compared to no screening (i.e., what ages to start and stop screening) and to each other?
Summary

Key Points: Outcomes

Breast Cancer Mortality:
- One case-control, one retrospective cohort, and one prospective cohort reported decreased mortality with screening in high-risk women. The estimate from a UK study may have been too low because of the choice of comparison group, and the confidence intervals for the U.S. study include 1.0 (OR 0.74; 95% CI, 0.53 to 1.03). The retrospective study did not have an unscreened comparison group within its cohort of high-risk women, but rather compared its mortality experience to other cohorts of varying ages and screening histories. Modeling studies suggest that mortality reduction with screening are greater in women at higher risk than in average-risk women.
- Because we judge the quality of evidence for some reduction in breast cancer mortality for average-risk women as HIGH, we also judge the quality of evidence for a breast cancer mortality reduction with screening for women at higher risk as HIGH. However, the quality of evidence for the magnitude of effect is LOW.

Stage Distribution:
- Stage distribution is consistently improved with the use of more sensitive modalities, either MRI compared to mammography, or the combination of MRI and mammography compared to MRI alone. The evidence for the direction of this effect is MODERATE, but for magnitude of effect LOW, making the overall quality of evidence LOW.

False Positives:
- MRI alone, or MRI in addition to mammography, consistently results in more false positives than mammography alone, but, because of imprecision across studies and risk of bias the overall quality of evidence is LOW.
- A number of studies did not report results separately for women at high risk because of genetic or familial predisposition and for women with a prior history of breast cancer and thus did not meet inclusion criteria.

Other Critical Outcomes:
We identified no studies that assessed other critical outcomes for KQ 4a.

Key Points: Harm-benefit Trade-offs
- We discuss the evidence for harm-benefit trade-offs for all high-risk women at the end of the section for KQ 5.

Description of Included Studies
Eight studies were included as relevant to KQ 4a. Two of these—one cohort study and one case-control study—evaluated breast cancer mortality among women at high risk of breast cancer due to a positive family history. Known BRCA1/BRCA2 mutation carriers were included in the cohort study, whereas BRCA1/2 mutation status was not mentioned in the case-control study. Differences in the groups being compared across the two studies precluded
combining the data for meta-analysis. A third study reported outcomes in a cohort of women aged 35-39 with a family history of breast cancer and compared them to other cohorts of younger women (<50 years) with various screening histories.

As noted in the Introduction, because our initial review found limited evidence on breast cancer mortality for KQs 4 and 5, we included stage distribution of tumors detected through screening as an alternate critical outcome for these KQs after discussion with the Guidelines Development Group (GDG). Stage distribution was reported in three prospective studies of high-risk women defined as having a family history of breast cancer or a BRCA1/2 mutation, and two retrospective studies comparing screened high-risk women aged 35-39 to women from other cohorts with various ages and screening histories. Two studies compared the characteristics of tumors detected with MRI screening versus conventional screening, two compared screen-detected tumors in high-risk women to tumors in unscreened women of similar age, and one compared characteristics of tumors diagnosed in screened and unscreened high-risk women. Characteristics of the breast cancers were described by stage, tumor size, and/or nodal status.

False positive outcomes, which are recognized as a limitation of screening with breast MRI, have been examined in several studies of high-risk women. Most of these studies defined high risk on the basis of having a BRCA1/2 mutation, a strong family history of breast cancer, or a personal history of breast cancer. However, one of our a priori screening criteria excluded studies conducted in women with a prior history of breast cancer. Applying this criterion left one study of high-risk women defined by familial or genetic predisposition, and one study of survivors of Hodgkin lymphoma that reported on false positive outcomes.

More detailed characteristics of the included studies are summarized in Appendix Table G-4. GRADE summary tables for the outcomes described below are provided in Appendix H.

**Detailed Synthesis**

**Breast Cancer Mortality**

**RCTs**

We did not identify any RCT evidence for high-risk women.

**Observational Studies**

The effect of screening on breast cancer mortality in women at high risk due to family history was reported in one prospective cohort study, one case-control study, and one retrospective cohort study. The prospective cohort study was conducted in the UK and compared women aged <50 years at high risk for breast cancer (>1 in 6 lifetime risk) who underwent mammographic screening every 12 months versus average-risk women of similar age who were not screened. The screened high-risk women were at significantly lower risk for death from breast cancer (HR 0.24; 95% CI, 0.09 to 0.66). It should be noted, however, that this was not a simple comparison of screened vs. unscreened high-risk women, rather it was a comparison of screened, high-risk women vs. unscreened average-risk women. The use of an unscreened average-risk comparison group rather than an unscreened high-risk group likely resulted in an underestimate of the HR.

The case-control study conducted in six health plans across the U.S., compared the 3-year screening history of women ages 40-65 who died from breast cancer to that of matched control
women without breast cancer. The associations between breast cancer mortality and screening by either clinical breast exam (CBE) or mammography were not statistically significant for either average-risk (OR 0.96; 95% CI, 0.80 to 1.14) or high-risk women (OR 0.74; 95% CI, 0.53 to 1.03; = approximately 95% probability that the OR is below 1.0). Similar trends were observed in younger (ages 40-49) and older (ages 50-65) women.

The retrospective cohort study, conducted in the UK, compared outcomes for women aged 35-59 years with a lifetime cancer risk of ≥17% who had annual mammography screening to outcomes in several other cohorts, including unscreened women <50 years, unscreened women <40 years and women aged 40-49. Among women diagnosed with breast cancer in the various cohorts, the breast cancer mortality was 9% among the 35- to 39-year-old screened cohort compared to 15 to 19% in the comparison cohorts. It is notable that the comparison cohorts differed in age range and time of recruitment and follow-up.

Data from these studies were inadequate to conclude that screening with mammography or a combination of mammography and CBE reduces mortality from breast cancer in high-risk women. None of the studies had a clean comparison of a single screening modality to an unscreened group.

Model-based Estimates

The 2009 USPSTF recommendation against routine screening in 40- to 49-year-olds was based on judgments about the balance of benefits and harms in this group, informed by the analysis of the CISNET collaborators. The USPSTF did recommend biennial screening for women aged 50-74 years, judging the balance of benefits and harms to be favorable. Subsequent to the modeling analysis conducted for the USPSTF, four of the CISNET groups performed additional analyses to identify thresholds of increased risk of breast cancer in 40- to 49-year-olds where the harm-benefit ratio was identical to that for biennial screening for 50- to 74-year-olds, thus justifying a recommendation for screening. The paper did not report specific estimates of reduction in breast cancer mortality among women at higher risk.

Because this paper addressed generic increases in risk rather than specific risk factors, we will discuss the effect of increased risk on overall harm-benefit assessment for KQs 4 and 5 together.

Discussion/Conclusions: Breast Cancer Mortality

- We identified minimal direct evidence on the effect of screening, or more intensive screening regimens, in women at higher than average risk for breast cancer.
- The data we did identify suggested a greater reduction in mortality in high-risk women compared to average women, but all available studies had issues with risk of bias.
- Since the benefits of screening in general should be at least as favorable for high-risk women as they are for women at average risk, the body of evidence for a reduction in mortality with screening compared to no screening in average-risk women should apply to high-risk women as well (quality of evidence HIGH for a qualitative effect, MODERATE for the quantitative estimate,).
- The quality of evidence for any specific modality or screening interval is LOW.
Stage Distribution

Observational Studies
Each of the five studies evaluating this outcome reported a more favorable stage distribution for the more highly screened group. The comparison of high-risk women screened with mammography versus unscreened average-risk women showed significantly smaller tumors (72% vs. 39% <2 cm; p<0.001) and less node involvement (66% vs. 47% node negative; p=0.013) among the screened women.204 Similarly, the comparison of screened and unscreened high-risk women showed less favorable tumor characteristics for the unscreened women (OR for tumor size >15mm 9.72; 95% CI, 1.01-93.61; OR for positive nodes 1.77; 95% CI, 0.36-8.63; OR for Stage II-IV 7.80; 95% CI, 1.18-51.50). The retrospective cohort reported a more favorable stage distribution for the screened women (74% of tumors were <2 cm in the cohort of 35- to 39-year-olds screened with mammography versus 39% and 45% in the two unscreened comparison cohorts, p<0.0001 and p=0.0018). The other two studies, which compared different screening modalities in high-risk women, showed that MRI screening resulted in more favorable tumor characteristics. One reported that 1/9 (11%) cancers diagnosed in the MRI plus mammography group was ≥Stage 2 compared to 6/20 (30%) cancers diagnosed in the mammography alone group.209 Similar findings were reported in the other study, with 85% of cancers being node negative and <2 cm in the MRI group as compared to 54% in the comparison group (p=0.004).208

Discussion/Conclusions: Stage Distribution
Because an additional study in a slightly different population also reported on stage distribution, we discuss the quality of evidence for stage distribution for KQs 4a and 4b together below.

Life Expectancy
Model-based estimates were derived for higher risk women in the CISNET analysis, and are discussed in the context of harm-benefit trade-offs below.

Overdiagnosis/Overtreatment
We did not identify any direct estimates of overdiagnosis/overtreatment in high-risk women in RCTs, observational studies, or model-based estimates. Conceptually, the risk of overdiagnosis should be smaller in women at greater risk of developing breast cancer, particularly at younger ages, but we found no empirical evidence for this.

False Positives

Biopsies: Observational Studies
A prospective study of 1952 women from the Netherlands—of whom 1909 had a familial or genetic predisposition to breast cancer—who were under surveillance for a median of 2.1 years reported a total of 67 biopsies performed in the study group.206 The reported false positive rate for biopsies performed due to mammography findings was 28.0% (7/25) and for biopsies performed due to MRI was 42.9% (24/56). Applying these numbers to the total population, 7/1909 (0.4%) women had false positive biopsies as a result of mammography and 24/1909 (1.25%) women had false positive biopsies as a result of MRI.
In a study of 148 Hodgkin lymphoma survivors in the U.S., 63 biopsies in 45 women were performed during the 3-year study. The false positive biopsy rates for MRI and mammography, respectively, were 13.4% and 5.9% for year 1, 9.0% and 9.0% for year 2, and 2.2% and 7.5% for year 3 (test for trend was not statistically significant).

These studies address high-risk populations that are defined on the basis of different criteria so results should not be combined. Nonetheless, both studies report that MRI screening results in more false positive biopsies than mammograms. The second study suggests that the difference between the modalities is most pronounced when screening is first initiated.

Model-based Estimates

Model-based estimates of false positive results in higher risk women are discussed as part of the integrated presentation of harm-benefit trade-offs below.

Discussion/Conclusions: False Positives

Because additional studies in different high-risk populations address false positives, we discuss the overall quality of evidence for KQs 4a and 4b together below.

Quality-adjusted Life Expectancy

Model-based estimates of the impact of different screening strategies on quality-adjusted life expectancy are discussed below.

Key Question 4b

Among women with an increased risk of breast cancer due to factors identified as the result of screening or diagnosis (e.g., prior diagnosis of proliferative lesions), what are the benefits, limitations, and harms associated with different screening modalities compared to no screening, and to each other?

Summary

Key Points: Outcomes

Stage Distribution:

- Stage distribution was improved by the addition of MRI to mammography compared to MRI alone in two observational studies. We rate the quality of evidence as LOW, based on imprecision and risk of bias.

False Positives:

- Both studies found that the probability of a false positive test increased with the addition of MRI. The quality of evidence was also LOW.

Description of Included Studies

Three studies were included as relevant to KQ 4b. All were retrospective cohort studies that addressed screening outcomes among women who had a prior diagnosis of LCIS or LCIS or AH. All three studies were conducted at the same institution in the U.S., with the most recent study expanding on the results of the prior studies. All data were abstracted from medical
records and compared outcomes for screening with MRI to screening with mammography. Details regarding characteristics of the mammography screening (film/digital, number of views, and number of readers) were not described. The exact age range was not reported; however, 15% of subjects were ≤45 years and 28% were >60 years of age. The outcomes reported varied between the papers with stage at distribution and false positive biopsies each reported in two of the three studies. Because all three studies were observational and there seemed to be a high possibility of bias related to which women received MRI screening, they were judged to be low quality.

More detailed characteristics of the included studies are summarized in Appendix Table G-5. GRADE summary tables for the outcomes described below are provided in Appendix H.

**Detailed Synthesis**

**Stage Distribution**

The stage distributions of cancers diagnosed with MRI and mammography were compared in two studies.\textsuperscript{217,218} The later study\textsuperscript{217} encompasses data from the earlier report.\textsuperscript{218} A trend of earlier diagnosis with MRI was reported, with a smaller median tumor size compared to the conventional screening group (0.5 cm vs. 0.95 cm, \(p=0.09\)). No significant difference in node status was reported (21% in the MRI group versus 24% in the conventional screening group).

**Discussion/Conclusions: Stage Distribution (KQs 4a and 4b)**

- Stage distribution is a surrogate for survival, which may also be a surrogate for mortality (depending on the effectiveness of treatment and the significance of lead time bias for a given cancer).
- Six observational studies suggest that the addition of MRI to other screening modalities improves stage distribution. We judge the quality of evidence to be LOW, primarily because of relatively small sample sizes and resulting imprecision, and some risk of bias inherent in the study designs. The evidence is consistent, however.

**False Positives: Biopsy**

All studies also reported on false positive biopsies. All studies reported a similar proportion of patients diagnosed with breast cancer in the MRI and mammography groups (6.4% vs. 6.1%,\textsuperscript{216} 2.7% vs. 3.6%,\textsuperscript{218} and 13% vs. 13%\textsuperscript{217}). A larger proportion of women in the MRI groups than the mammography groups underwent biopsies (27% vs. 12%\textsuperscript{216} and 25% vs. 11%\textsuperscript{218}). The proportion of women who had a false positive biopsy was higher in the MRI groups than in the mammography groups (22% vs. 9.3\textsuperscript{216}, 22% vs. 7.1\textsuperscript{218} and 36% vs. 13\textsuperscript{217}).

**Discussion/Conclusions: False Positives for KQs 4a and 4b**

- One general point in considering false positive rates is that, for any given level of specificity, the likelihood of a false positive will be reduced in higher risk women, because of the greater prior probability of disease (i.e., improved positive predictive value).
- Three observational studies in different high-risk populations suggest an increase in false positives rates with MRI compared to mammography. We judge the quality of evidence
to be **LOW**, based on issues related to precision and risk of bias. Results were consistent between studies.

**Key Question 5a**

Among women with an increased risk of breast cancer due to factors known PRIOR to the onset of screening (e.g., family history, BRCA mutation carrier, history of chest irradiation), what are the relative benefits, limitations, and harms associated with different screening modalities at different intervals, and how do these vary by age?

**Summary**

**Key Points: Outcomes**

**Stage Distribution:**
- We identified one study with substantial risk of bias that reported more favorable extent of disease at the time of detection (tumors more likely to be less than 20 mm and less likely to have positive lymph nodes) with annual screening compared to biennial screening in women aged 50-69 years with a first-degree relative with a history of breast cancer. We rate the quality of evidence as **LOW**.

**Description of Included Studies**

We identified one cohort study, conducted in Australia, that compared outcomes by screening interval for women with a family history of breast cancer. Mammography was conducted as part of an organized screening program in New South Wales. Four screening sites offered annual mammography screening and four offered biennial screening for women aged 50 to 69 years with a family history of breast cancer in a first-degree relative. BRCA1/BRCA2 mutation status was not addressed in the study. Details on the characteristics of the mammography including film or digital, number of views, and number of readers were not provided. The study was judged to be of low quality, primarily because of risk of bias.

Outcomes reported included the odds ratios (ORs) and 95% CIs for having a tumor size of <20 mm, a well-differentiated tumor, and node-negative cancer, comparing women in the annual screening group to those in the biennial screening group. No other critical outcomes were analyzed in this study.

More detailed characteristics of the included study are summarized in Appendix Table G-6. GRADE summary tables for the outcomes described below are provided in Appendix H.

**Detailed Synthesis**

**Stage Distribution**

Breast cancers diagnosed through annual screening were significantly more likely than cancers diagnosed through biennial screening to be <20 mm (OR 1.91; 95% CI, 1.21 to 3.02) and to be node-negative (OR 1.61; 95% CI, 1.03 to 2.50). Differences in tumor grade were not statistically significant between the screening groups.
**Key Question 5b**

Among women with an increased risk of breast cancer due to factors identified as the result of screening or diagnosis (e.g., prior diagnosis of proliferative lesions), what are the benefits, limitations, and harms associated with different screening modalities at different intervals, and how do these vary by age?

**Summary**

We did not identify any studies that specifically addressed KQ 5b.

**Harm-benefit Trade-offs: High-risk Women**

The benefits and harms of different screening modalities are of particular interest for women at high risk of breast cancer due to a family history of cancer, a mutation in BRCA1 or BRCA2, or medical radiation exposure. There are significant challenges in evaluating the benefits and risks of screening in this population. For both ethical and pragmatic reasons, there have not been—and it is unlikely there will be—any RCTs assessing the efficacy of screening in reducing breast cancer mortality in this population. Thus, judgments on the benefits and risks of different screening modalities must be derived from observational studies and must be based on less critical outcomes than mortality, including stage distribution at diagnosis and false positive biopsies.

Although a relatively large number of studies were identified that examined screening outcomes in high-risk women, an important limitation is that most of them included women with a personal history of breast cancer. An *a priori* decision was made during protocol development, in consultation with ACS and the GDG, to exclude studies of women with a prior diagnosis of breast cancer because they clearly represent a different population both in terms of risk level and the likely outcomes from screening. Unfortunately, the studies that included women with a prior history of breast cancer did not report findings stratified by personal history. Thus there were a very small number of studies available that provided data on mortality, false positives, and stage distribution for women at high risk for breast cancer. Assessments of the benefits of screening in high-risk women are limited by differences between studies in definitions of high risk, short follow-up times, and a limited number of breast cancer diagnoses in most studies.

In this section, we describe modeling studies that provide some additional insight into the outcomes and trade-offs in high-risk women.

**CISNET: Identification of Risk Thresholds**

After review of the evidence, including the CISNET analyses, the USPTF gave a B recommendation for biennial mammographic screening for women aged 50-74, based on “moderate certainty that the net benefit is moderate.” Reasoning that this recommendation established an implicit threshold for “willingness-to-pay,” four of the CISNET groups subsequently performed an analysis to vary the risk of breast cancer to identify thresholds of increased risk where a given strategy for screening 40- to 49-year-old women would meet this threshold. Median estimates across these four CISNET models for biennial screening for 100,000 women aged 50-74 were 630 deaths prevented, 10,900 life-years gained, and 88,300 false positives, for a false positive per death prevented ratio of 146 and false positive per life-year gained of 8.3. In this updated analysis, the investigators included digital mammography as an option.
Table 45 presents the results of this threshold analysis. Key points are:

- The choice of measure of harm and benefit is important. Threshold relative risks were significantly lower when life expectancy was used as the measure of benefit compared to deaths prevented (consistent with the results presented under KQ 1—although the number of deaths prevented by screening younger women is lower, the life expectancy gains are greater).
- Although data were not presented, the authors noted that threshold relative risks increased when quality-adjusted life expectancy, which incorporates the impact of both false positive results and overdiagnosis, was used as the denominator in the harm-benefit ratio.
- Although annual screening is expected to prevent more deaths and result in greater gains in life expectancy in this age group, the increase in false positives is substantially greater, resulting in higher risk thresholds.

Table 45. Threshold Relative Risks where Screening of 40- to 49-year-olds Results in Equivalent Harm-benefit Ratio to Biennial Screening of 50- to 74-year-olds, by Interval, Measure of Harm-benefit, and Mammography Method

<table>
<thead>
<tr>
<th>Interval</th>
<th>Harm-benefit</th>
<th>Mammography Method</th>
<th>Relative Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biennial (compared to no screening of 40- to 49-year-olds)</td>
<td>False positives per death prevented</td>
<td>Film</td>
<td>2.7</td>
</tr>
<tr>
<td></td>
<td>False positives per life-year gained</td>
<td>Digital</td>
<td>3.3</td>
</tr>
<tr>
<td>Annual (compared to biennial screening of 40- to 49-year-olds)</td>
<td>False positives per death prevented</td>
<td>Film</td>
<td>5.1</td>
</tr>
<tr>
<td></td>
<td>False positives per life-year gained</td>
<td>Digital</td>
<td>6.1</td>
</tr>
</tbody>
</table>

The authors note that a systematic review found that women with a first-degree relative with breast cancer had a two-fold or greater risk of breast cancer, which would meet the threshold relative risk, particularly when life-years gained is the measure of benefit. BRCA1 and BRCA2 mutation carriers are also at markedly higher risk: in a recent prospective study, estimated cumulative risk of developing breast cancer by age 70 in a cohort of women with a mean age of 40 at baseline was 60% for BRCA1 carriers and 55% for BRCA2 carriers, approximately 6 times the 8.5% risk between ages 40 and 70 in the general population. A six-fold relative risk would justify annual screening with digital mammography in younger women based on false positives per death prevented alone. However, the risk in BRCA1/2 mutation carriers is so high that increasing sensitivity through more frequent screening and, potentially, adding other modalities might have an even more favorable balance between benefits and harms.

**BRCA1/BRCA2 Mutation Carriers**

Estimating harm-benefit trade-offs in BRCA1/BRCA2 carriers is particularly complex—these women are also at substantially increased risk of ovarian cancer, which has a much poorer prognosis than breast cancer and for which screening is largely ineffective. Strategies for primary prevention of ovarian cancer may affect the underlying risk of breast cancer, either increasing it (oral contraceptives), or decreasing it (risk-reducing salpingoophorectomy).

In 2006, Plevritis and colleagues published a Monte Carlo simulation model evaluating the effectiveness and cost-effectiveness of screening BRCA1/2 mutation carriers with mammography + MRI compared with mammography alone. The model simulated the life
histories of women with BRCA1/2 mutation carriers and incorporated the potential health benefits and harms of strategies of (1) no screening, (2) annual mammography from ages 25 to 69 years, and (3) annual mammography from ages 25 to 69 years plus annual MRI for specific age groups. The accuracy of mammography and breast MRI was estimated based on published data in BRCA1/2 mutation carriers. Breast cancer survival in the absence of screening was based on SEER data. Relevant to our systematic review, the model estimated the proportion of overdiagnosed cases, life expectancy, and breast cancer mortality reduction for women in the different strategies; quality-adjusted life expectancy was derived using utility weights from a time-trade-off survey conducted in 33- to 50-year-old women, including breast cancer patients, women at high risk for breast cancer, and non-high-risk women.\textsuperscript{226} Extensive sensitivity analyses were performed to explore uncertainties in the data and modeling assumptions. One noteworthy feature of this model is that DCIS was not included, largely because the uncertainty about the natural history of DCIS in this population is even greater than it is for non-BRCA mutation carriers.

Predicted outcomes are presented in Table 46. False positive rates were reported to increase from 5\% with mammography alone to 25\% with the addition of MRI, although it is unclear whether this is annually or cumulative over some unspecified interval. Quality-adjusted life expectancy was discounted at a 3\% annual rate (the value of future years was decreased relative to the present, in order to account for people’s preference for their current health over their health status in the future), which prohibits direct comparison to the other outcomes.

Table 46. Outcomes of Annual Mammography and Annual Mammography plus MRI in BRCA1 and BRCA2 Carriers\textsuperscript{225}

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Outcome</th>
<th>Compared to No Screening</th>
<th>Compared to Mammography Alone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mammography Alone</td>
<td>Mammography + MRI</td>
</tr>
<tr>
<td>BRCA1</td>
<td>Life-years gained</td>
<td>0.7</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>Overdiagnosis</td>
<td>1.4%</td>
<td>2.0%</td>
</tr>
<tr>
<td></td>
<td>Relative reduction in breast cancer mortality</td>
<td>14%</td>
<td>38%</td>
</tr>
<tr>
<td>BRCA2</td>
<td>Life-years gained</td>
<td>0.6</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>Overdiagnosis</td>
<td>1.4%</td>
<td>2.2%</td>
</tr>
<tr>
<td></td>
<td>Relative reduction in breast cancer mortality</td>
<td>16%</td>
<td>38%</td>
</tr>
</tbody>
</table>

Abbreviations: BRCA1/2=breast cancer susceptibility gene 1/2; MRI=magnetic resonance imaging

**Discussion/Conclusions: Harm-benefit Trade-offs in High-risk Women**

- In modeling studies, increased risk of cancer substantially improves the harm-benefit ratio of screening overall, or of more sensitive but less specific strategies such as MRI alone, or MRI plus mammography compared to mammography alone.
- Within a given modeling framework, it is possible to identify thresholds of increased risk where the harm-benefit ratio for screening younger high-risk women is equivalent or better to strategies recommended for older average-risk women.
- As noted in the discussion for KQ 1, there is no consensus on an acceptable harm-benefit trade-off for any patient population, or for any combination of benefits and harms.
Screening for Breast Cancer: Overall Discussion

We summarize here some of the potential limitations of our systematic review methodology and the inherent limitations of breast cancer screening given its underlying biology. We then highlight the key findings from our report related to the critical outcomes of breast-cancer mortality, life expectancy, overdiagnosis, false positives, and quality of life. Finally, we discuss our findings in relation to harm-benefit trade-offs and high-risk women as a subgroup of specific interest.

Limitations of the Review

- Our search strategy, inclusion/exclusion criteria, and choice of outcomes were all developed in consultation with the ACS and the Guidelines Development Group (GDG), and we used standard methods as recommended by the Institute of Medicine for systematic review. However, it is always possible that relevant published peer-reviewed evidence was not identified, or that there is reasonable disagreement about whether specific articles should have been included. One of the purposes of multiple reviews is to minimize the chances that our final report will have excluded any crucial evidence that would improve the GDGs confidence in the evidence, and thus influence the strength of recommendation. We reviewed articles identified by the GDG and peer reviewers that were either excluded or missed by our initial and follow-up search, or that were published subsequent to the cut-off date of the follow-up search. Articles that met inclusion criteria, or upon re-review should have been included, were included; in a few cases, articles that otherwise met exclusion criteria (because of cut-off dates for publication or sample size) were included if they provided directly relevant evidence, or if they were part of a systematic review we had included.

- Particularly for breast cancer mortality and overdiagnosis, although qualitative effects are consistent, the quantitative estimates of effect vary widely, depending on study design, when and where the study was performed, and the methods of analysis used to estimate effects. The uncertainty in these estimates in the context of recommendations for U.S. women in 2014 and beyond is exacerbated by trends in factors that may affect the absolute risk of breast cancer (such as the decline in the use of hormone replacement therapy), the absolute risk of dying once diagnosed with breast cancer (such as advances in treatment), and factors that may affect the consequences of overdiagnosis (such as development and validation of markers for prediction of progression in DCIS). In our judgment, there are reasonable arguments for why both relative mortality and overdiagnosis estimates derived from a particular part of the evidence may be too high or too low, particularly in the context of the U.S. population and health system. Therefore, we have presented quantitative estimates across a range of “optimistic” and “pessimistic” assumptions.

- Our quantitative methods are relatively simple compared to the range of other models available, particularly those of the CISNET collaborators. It is quite possible that different estimates would be derived from alternative approaches. However, we are reasonably confident, given the underlying uncertainty, that the relative size of the estimates is reasonable, particularly for the harm-benefit trade-offs. In general, our approach is biased in favor of screening. Specific methodological issues and limitations are discussed in Appendix C.
• There are certainly grounds for reasonable disagreement about judgments of evidence quality and the extent to which those judgments translate into certainty about the quantitative estimate of the probability of specific benefits and harms of screening when applied to the U.S. population. In the absence of direct evidence for the U.S. (in particular, the absence of a link between population-based cancer incidence and mortality data and screening history), we have attempted to generate estimates across a plausible range (and, again, there can be reasonable disagreement about whether the approach used to generating those estimates is optimal). For the purposes of guidelines development under GRADE, the major issue is the likelihood that the plausible range of those estimates, particularly for harm-benefit trade-offs, includes a threshold of acceptability. In the absence of consensus thresholds, that judgment is up to the members of the GDG.

Limitations of Breast Cancer Screening

The primary purpose of breast cancer screening is to reduce mortality from breast cancer through detection of asymptomatic cancers at a stage of development when treatment is more likely to be successful. As a secondary goal, effective treatments of less advanced cancers may involve less morbidity.

The paradigm of successful cancer screening has been the major reduction in both incidence and mortality from cervical cancer in countries where widespread screening has been introduced, and the success of cervical cancer screening has served as an implicit target for screening for other cancers. However, cervical cancer has unique biological characteristics which make it particularly amenable as a target of population-based screening:

• Cervical cancer has a single necessary cause, infection with oncogenic human papillomavirus (HPV).
• There is a relatively narrow window of exposure to HPV associated with the early years of sexual activity—oncogenic HPV incidence and prevalence both decrease substantially by age 30 in most populations.
• There is a long (10-15 years) stage of detectable pre-invasive changes, where treatment has close to 100% likelihood of prevention of invasive disease.
• Most invasive cervical cancers are relatively slow-growing squamous tumor—early spread in most cancers is primarily by direct extension for several years, followed by spread to regional lymph nodes; local and regional treatments with surgery and/or radiation are highly effective.
• Indeed, screening has been less successful in preventing cervical adenocarcinomas, which are harder to detect and which tend to spread more rapidly.

Unfortunately, few other cancers share these characteristics. Particularly for cancers with no identifiable pre-invasive stage and rapid progression to distant metastases, such as ovarian cancer, screening may never be effective at an acceptable frequency of screening.

It is likely that breast cancer lies somewhere in between cervical and ovarian cancers in the potential of screening to reduce mortality.

• The underlying etiology of breast cancer is not as clearly understood as it is for cervical cancer, but clearly exposure to estrogens and progestins, both natural and exogenous, plays a role in the development of breast cancer, and the duration and intensity of this exposure will vary widely among women.
• There is no known obligatory pre-invasive stage; DCIS may well be a precursor in some cases, but not all DCIS will progress, and not all breast cancers are preceded by a detectable in situ lesion. In addition, at least in the U.S., approximately 30% of cancers will be non-ductal, meaning that detection of DCIS will not affect subsequent incidence or mortality.

• The likelihood of metastases may be higher at earlier stages of growth of the primary tumor than they are for squamous cervical cancer, requiring systemic therapy for a larger proportion of cancers. There is a growing body of basic science evidence suggesting that, for some breast cancers, size alone may not be the primary predictor of biological behavior, particularly metastatic behavior; screening with imaging, which is based on the fundamental principle that smaller tumors are less likely to have progressed and have a greater probability of cure, may not be helpful in reducing mortality attributable to these subtypes of cancers.

We believe it is important for policy makers, clinicians, and patients to understand that the fundamental biology of each cancer type (including different subtypes within a particular organ or tissue, and even individual cancers within a specific subtype) is different, and that the success of screening in preventing cancer death may be even more dependent on the cancer itself than on the screening methods used. Screening for breast cancer is highly unlikely ever to be as successful as screening for cervical cancer.

**Key Findings for Critical Outcomes**

**Breast Cancer Mortality**

• We believe the strength of evidence that screening with mammography reduces breast cancer mortality is **HIGH**. However, we are less certain about the magnitude of this reduction. There are features of both the randomized controlled trials (RCTs), which have served as the basis for most recommendations about screening, and the more recent observational evidence, which, when used as the basis for future recommendation, may lead to an over- or underestimation of the impact of screening on breast cancer mortality. To help illustrate the impact of this uncertainty, we have used pooled estimates from both sources. Since judgments about trade-offs may vary depending on these estimates, and additional RCTs are not currently planned, we believe that future observational research should focus on estimating mortality reduction within the U.S., using state-of-the-art methods for causal inference, such as propensity scores, marginal structural models, and/or instrumental variables.

• Evidence is consistent that estimates of mortality reduction are greater when the comparison is between screened and unscreened women than when the comparison is between women invited to screening versus women not invited. This is intuitive, and since the U.S. does not have a formal screening program, estimates based on this comparison may be more applicable. The major methodological concerns here are the potential for unmeasured confounding and the potential effect of differences in post-screening diagnosis and treatment outcomes on applicability of estimates derived from non-U.S. settings. Again, U.S.-based studies using advanced methods would help increase certainty about the quantitative magnitude of mortality reduction.
• The strength of evidence that screening reduces mortality at all ages is HIGH, but, again, there is uncertainty about the magnitude of this effect. Estimated absolute reduction is lower in younger women than in older women, because of a lower overall incidence of breast cancer, but direct evidence for older women is very limited, and registry data strongly suggests that women 75 and older diagnosed with breast cancer are more likely to die from other causes than from breast cancer.

• We have LOW confidence that annual screening reduces mortality in women 40-49 compared to biennial screening, but does not affect mortality in women 50 and older—the evidence is suggestive, and biologically plausible, but the magnitude of effect is relatively small.

• We have LOW confidence in the evidence that CBE does not reduce mortality when added to mammography.

Life Expectancy

• The evidence for life expectancy gains from breast cancer screening is all model-based, and subject to the limitations of both the models themselves and the quality of the data for model parameters. All things being equal, preventing breast cancer deaths should increase life expectancy, and preventing deaths at younger ages should lead to bigger life expectancy gains than preventing deaths in older women. However, we have LOW confidence in the estimates of the size of these gains (primarily because of the uncertainty surrounding the magnitude of mortality reduction).

Overdiagnosis

• Given the frequency of diagnosis of DCIS, and the likelihood that a substantial proportion of DCIS lesions would not have progressed to invasive cancer, we have HIGH confidence that overdiagnosis is a consequence of mammographic screening, but LOW confidence in the magnitude of overdiagnosis, particularly for small localized invasive cancers.

• The extent to which DCIS represents overdiagnosis is ultimately dependent on the proportion of lesions that would eventually progress to symptomatic invasive cancer; uncertainty about this proportion is a major driver of uncertainty about the harm-benefit trade-off of overdiagnoses versus mortality reduction. In addition, because of substantial variability in the rates of DCIS diagnosis both across countries (with the U.S. having the highest rate among countries reporting screening outcomes) and within countries (with substantial variation between centers reporting to the Breast Cancer Screening Consortium), even a better estimate of the probability of progression would still result in substantial uncertainty about the risk of overdiagnosis at the individual patient level.

• Based on the relative incidence of DCIS compared to small localized cancers, and the inherent methodological difficulties in estimating overdiagnosis in invasive cancers, identifying those patients with DCIS who do not need aggressive therapy would likely have a larger impact on the overall estimate of the harms of mammography than a more precise estimate of the proportion of invasive cancers that are overdiagnosed.

• The impact of screening frequency on overdiagnosis is likely to vary by age, both because of differences in competing risks of mortality and likely differences in the
likelihood of progression of small asymptomatic cancers; however, we did not find any direct evidence on this.

- Resolving uncertainty about quantitative estimates of overdiagnosis would be considerably easier if investigators could agree on a common set of methods for this estimation.

### False Positives

- For overall false positives (both those resulting in biopsies and those with only repeat examinations), cumulative 10-year rates are similar whether screening begins at age 40 or at age 50, but are approximately 20% higher with annual screening compared to biennial screening. On a per-screen basis, false positive rates increase with age at first screen, longer screening intervals, family history of breast cancer, and breast density, but decrease with the availability of prior examinations. There is also considerable variation between radiologists. Our confidence in these estimates, derived from observational data of a large population-based registry representing community practice in the U.S., is **MODERATE**.

- For false positive biopsies, cumulative 10-year rates are higher with an older age to start screening (2% difference for age 40 vs. age 50), and with more frequent screening intervals (2% difference for annual vs. biennial screening). Per-screen rates increase with age at subsequent examination, longer screening intervals, family history, and breast density, but also decrease with availability of prior examination; again, there is considerable variation between radiologists, and our confidence in the estimates is **MODERATE**.

- Although the cumulative 10-year rates of false positives are similar or even higher (for biopsies) when women begin screening at age 50 compared to age 40, estimates of the cumulative risk of either type of false positive outcome are consistently higher when screening begins at younger ages (simply because of an increased number of screening examinations). Quantitative estimates of the cumulative lifetime risk are variable, depending on assumptions about the independence of false positive probability, the extent to which individual patient variation is captured, the presence of competing risks, and whether the number of total false positives across the population (which includes women with multiple false positives) or the number of women with at least one false positive is used as the numerator. Unless false positive rates become negligible at some point after extended screening (which seems unlikely, especially given the observed increased risk with age for a false positive result with subsequent screens in the BCSC data), we have **HIGH** confidence that the lifetime probability of a false positive result increases with younger age to start screening, but **LOW** confidence in the quantitative estimate due to the need to rely on models or extrapolations.

- Because the quality-of-life and emotional effects of screening were not critical outcomes, we did not systematically review the evidence on these outcomes in relation to false positive results. A recent meta-analysis suggested that cancer-specific domains are more likely to be affected, and for a longer duration, than generalized measures of anxiety, a finding verified in a recent U.S.-based study which showed only transient effects, on average, on generalized anxiety. Of note in that study, the proportion of women experiencing “a lot” or “extreme” anxiety from a false positive result was 10% higher than the proportion of women undergoing a false positive biopsy, suggesting that the
emotional consequences of a false positive resulting only in a recall examination are similar to those of women undergoing biopsy in some women.

Quality-adjusted Life Expectancy

- We have LOW confidence in estimates of the effect of different screening strategies on quality-adjusted life expectancy, both because of the inherent uncertainties in estimates of life expectancy, and the relative weakness of the utility weights used in the current literature.
- Given the importance of breast cancer screening, obtaining higher quality evidence on patient preferences should be a high research priority.
- Better evidence on utilities is particularly important in helping resolve the importance of overdiagnosis and false positives in the harm-benefit estimation of breast cancer screening. The evidence consistently shows reductions in quality-adjusted life expectancy when false positives are included; if the impact of false positives is longer lasting than currently modeled, as suggested by several systematic reviews, this could further reduce gains in quality-adjusted life expectancy, particularly for alternative strategies where the increase in false positives is much greater than the gains in life expectancy (such as annual compared to biennial screening).
- Depending on the ratio of overdiagnoses to death prevented, the distribution of age at diagnosis for overdiagnosed cancers relative to age at cancer death in unscreened women, and the duration of disutility associated with a cancer diagnosis, it is possible that quality-adjusted life expectancy could be decreased in some screening scenarios relative to no screening. Identifying ranges for these parameters that meet this threshold is an important priority for modelers.

Harm-benefit Trade-offs

- Model-based estimates of total false positives per cancer death prevented are well below the threshold reported in a single 1997 study which has issues with generalizability. There is some uncertainty around these estimates, primarily related to estimates of mortality reduction. The ratio increases with younger age to start screening because of the higher estimated cumulative risk of false positives over a lifetime.
- Overdiagnosis per death prevented ratios, using only the detection of non-progressive DCIS through screening as the definition, are highly dependent on estimates of mortality reduction and DCIS progression, with some variation also dependent on the relative risk of DCIS attributable to screening. The ratio is less than 1.0 at high estimates of DCIS progression (80%) and mortality reduction (62%), but greater than 1.0 when progression probability is 50% or less. Given the LOW confidence in the overdiagnosis estimates, we have LOW confidence in these results.
- There is less evidence on how patients view trade-offs surrounding overdiagnosis and mortality prevention, and none from the U.S.
- Updated evidence on patient preferences for harms/benefit trade-offs should be a high priority.
- In addition, consensus on acceptable thresholds would facilitate both guidelines development (by focusing attention on the evidence needed to achieve certainty about the “true” ratio), and help prioritize future research (by identifying thresholds of effect size
that would either change recommendations or lead to increased strength of recommendations).

**High-risk Women**

- Our confidence that screening reduces mortality in high-risk women is even higher than for average risk women, but we have only *MIXED* confidence in the estimate of the size of the effect.
- We have *MIXED* confidence that adding MRI to mammography improves stage-distribution at diagnosis in high-risk women.
- In modeling studies, increased risk of breast cancer substantially improves the harm-benefit ratio of screening overall, or of more sensitive but less specific strategies such as MRI alone, or MRI plus mammography compared to mammography alone.
- Within a given modeling framework, it is possible to identify thresholds of increased risk where the harm-benefit ratio for screening younger women is equivalent or better to strategies used in older women. For example, screening women aged 40-49 with a first-degree relative with a history of breast cancer, or screening even younger women who are BRCA1 and BRCA2 carriers, has a similar harm-benefit trade-off as biennial screening of 50- to 74-year-olds based on one comprehensive modeling study.


99. Dahabreh IJ, Kent DM. Can the learning health care system be educated with...


139. Gold EB, Crawford SL, Avis NE, et al. Factors related to age at natural menopause:


202. Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death 1999-2010 on CDC WONDER Online Database, released 2012. Data are from the Multiple Cause of Death Files, 1999-2010, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics.
Screening women at high risk for breast cancer


Appendix A. Exact Search Strings

PubMed® search strategy (March 6, 2014)

KQ 1 – What are the relative benefits, limitations, and harms associated with mammography screening compared to no screening in average-risk women ages 20 and older, and how do they vary by age, screening interval, and prior screening history?

<table>
<thead>
<tr>
<th>Set #</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>#4</td>
<td>#1 AND #2 AND #3</td>
</tr>
<tr>
<td>#5</td>
<td>Limits: English</td>
</tr>
</tbody>
</table>

KQ 3 – What are the benefits, limitations, and harms associated with clinical breast examination among average-risk women 20 years and older compared to no CBE, and how do they vary by age, interval, and participation rates in mammography screening?

<table>
<thead>
<tr>
<th>Set #</th>
<th>Terms</th>
</tr>
</thead>
</table>
| #3    | (randomized controlled trial[pt] OR controlled clinical trial[pt] OR randomized[tiab] OR"
KQ 4a – Among women with an increased risk of breast cancer due to factors known PRIOR to the onset of screening (e.g., family history, BRCA mutation carrier, history of chest irradiation), what are the relative benefits, limitations, and harms associated with different screening modalities compared to no screening (i.e., what ages to start and stop screening) and to each other?

<table>
<thead>
<tr>
<th>Set #</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set #</td>
<td>Terms</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>#4</td>
<td>tomosynthesis[tiab] AND #1 AND #2 AND #3</td>
</tr>
<tr>
<td>#6</td>
<td>#4 AND #5</td>
</tr>
<tr>
<td>#7</td>
<td>Limits: English</td>
</tr>
</tbody>
</table>

**CINAHL® search strategy (September 10, 2013)**

KQ 1 – What are the relative benefits, limitations, and harms associated with mammography screening compared to no screening in average-risk women ages 20 and older, and how do they vary by age, screening interval, and prior screening history?

<table>
<thead>
<tr>
<th>Set #</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>(MH &quot;Breast Neoplasms&quot;) OR ((TI breast OR AB breast) AND (TI ( neoplasms OR neoplasm OR cancer OR carcinoma OR carcinomas ) OR AB ( neoplasms OR neoplasm OR cancer OR carcinoma OR carcinomas )))</td>
</tr>
<tr>
<td>#2</td>
<td>(MH &quot;Mammography&quot;) OR TI ( mammography OR mammograms OR mammogram ) OR AB ( mammography OR mammograms OR mammogram )</td>
</tr>
<tr>
<td>#3</td>
<td>(MH &quot;Experimental Studies&quot;) OR (MH &quot;Comparative Studies&quot;) OR (MH &quot;Systematic Review&quot;) OR (MH &quot;Meta Analysis&quot;) OR (MH &quot;Decision Support Techniques&quot;) OR (MH &quot;Evaluation Research&quot;) OR (MH &quot;Case Control Studies&quot;) OR (MH &quot;Prospective Studies&quot;) OR (MH &quot;Retrospective Design&quot;) OR (MH &quot;Empirical Research&quot;) OR (MH &quot;Crossover Design&quot;) OR MW &quot;dt&quot; OR TI ( randomized OR randomised OR randomization OR randomisation OR placebo OR randomly OR trial OR groups OR &quot;clinical trial&quot; OR “clinical trials ” OR &quot;comparative study&quot; OR &quot;meta-analysis&quot; OR &quot;meta-analyses&quot; OR &quot;evaluation study&quot; OR “evaluation studies” OR &quot;intervention study&quot; OR &quot;intervention studies&quot; OR &quot;case-control&quot; OR cohort OR &quot;longitudinal&quot; OR longitudinally OR &quot;prospective” OR prospectively OR &quot;retrospective&quot; OR &quot;follow up&quot;)</td>
</tr>
</tbody>
</table>
### Set # | Terms
--- | ---
#1 | OR AB (randomized OR randomised OR randomization OR randomisation OR placebo OR randomly OR trial OR groups OR "clinical trial" OR "clinical trials" OR "comparative study" OR "meta-analysis" OR "meta-analyses" OR "evaluation study" OR "evaluation studies" OR "intervention study" OR "intervention studies" OR "case-control" OR cohort OR "longitudinal" OR longitudinally OR "prospective" OR prospectively OR "retrospective" OR "follow up")
#2 | #1 AND #2 AND #3
#3 | Limits: English

**KQ 3** – What are the benefits, limitations, and harms associated with clinical breast examination among average-risk women 20 years and older compared to no CBE, and how do they vary by age, interval, and participation rates in mammography screening?

### Set # | Terms
--- | ---
#1 | (MH "Breast Neoplasms") OR ((TI breast OR AB breast) AND (TI (neoplasms OR neoplasm OR cancer OR carcinoma OR carcinomas) OR AB (neoplasms OR neoplasm OR cancer OR carcinoma OR carcinomas)))
#2 | (MH "Breast Self-Examination") OR (MH "Physical Examination") OR (MH "Gynecologic Examination") OR (MH "Palpation") OR TI ("clinical breast examination" OR "self exam" OR "self examination") OR AB ("clinical breast examination" OR "self exam" OR "self examination")
#3 | (MH "Experimental Studies") OR (MH "Comparative Studies") OR (MH "Systematic Review") OR (MH "Meta Analysis") OR (MH "Decision Support Techniques") OR (MH "Evaluation Research") OR (MH "Case Control Studies") OR (MH "Prospective Studies") OR (MH "Retrospective Design") OR (MH "Empirical Research") OR (MH "Crossover Design") OR MW "dt" OR TI (randomized OR randomised OR randomization OR randomisation OR placebo OR randomly OR trial OR groups OR "clinical trial" OR "clinical trials" OR "comparative study" OR "meta-analysis" OR "meta-analyses" OR "evaluation study" OR "evaluation studies" OR "intervention study" OR "intervention studies" OR "case-control" OR cohort OR "longitudinal" OR longitudinally OR "prospective" OR prospectively OR "retrospective" OR "follow up")
#4 | #1 AND #2 AND #3
#5 | Limits: English

**KQ 4a** – Among women with an increased risk of breast cancer due to factors known PRIOR to the onset of screening (e.g., family history, BRCA mutation carrier, history of chest irradiation), what are the relative benefits, limitations, and harms associated with different screening modalities compared to no screening (i.e., what ages to start and stop screening) and to each other?
<table>
<thead>
<tr>
<th>Set #</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>(MH &quot;Breast Neoplasms&quot;) OR ((TI breast OR AB breast) AND (TI ( neoplasms OR neoplasm OR cancer OR carcinoma OR carcinomas ) OR AB ( neoplasms OR neoplasm OR cancer OR carcinoma OR carcinomas )))</td>
</tr>
<tr>
<td>#2</td>
<td>(MH &quot;Genetics&quot;) OR MW “fg” OR (MH &quot;Genes, BRCA&quot;) OR (MH &quot;Thorax+/RE/RA&quot;) OR (MH &quot;Carcinoma, Ductal, Breast&quot;) OR TI ( genetic OR genetics OR brca OR brca1 OR brca2 OR brca-1 OR brca-2 OR “family history” OR “chest irradiation” OR “lobular neoplasia” OR “abnormal pathology” OR “proliferative lesions” OR “proliferative lesion” OR dcis OR “ductal carcinoma in situ” OR Ashkenazi ) OR AB ( genetic OR genetics OR brca OR brca1 OR brca2 OR brca-1 OR brca-2 OR “family history” OR “chest irradiation” OR “lobular neoplasia” OR “abnormal pathology” OR “proliferative lesions” OR “proliferative lesion” OR dcis OR “ductal carcinoma in situ” OR Ashkenazi)</td>
</tr>
<tr>
<td>#3</td>
<td>(MH &quot;Mammography&quot;) OR TI ( mammography OR mammograms OR mammogram ) OR AB ( mammography OR mammograms OR mammogram ) OR (MH &quot;Breast Self-Examination&quot;) OR (MH &quot;Physical Examination&quot;) OR (MH &quot;Gynecologic Examination&quot;) OR (MH &quot;Palpation&quot;) OR TI ( &quot;clinical breast examination&quot; OR &quot;self exam&quot; OR &quot;self examination&quot; ) OR AB ( &quot;clinical breast examination&quot; OR &quot;self exam&quot; OR &quot;self examination&quot; ) OR (MH &quot;Magnetic Resonance Imaging&quot;) OR MW &quot;US&quot; OR TI ( mri OR &quot;magnetic resonance imaging&quot; OR ultrasound OR ultrasonography OR tomosynthesis ) OR AB ( mri OR &quot;magnetic resonance imaging&quot; OR ultrasound OR ultrasonography OR tomosynthesis)</td>
</tr>
<tr>
<td>#4</td>
<td>#1 AND #2 AND #3</td>
</tr>
<tr>
<td>#5</td>
<td>(MH &quot;Experimental Studies&quot;) OR (MH &quot;Comparative Studies&quot;) OR (MH &quot;Systematic Review&quot;) OR (MH &quot;Meta Analysis&quot;) OR (MH &quot;Decision Support Techniques&quot;) OR (MH &quot;Evaluation Research&quot;) OR (MH &quot;Case Control Studies&quot;) OR (MH &quot;Prospective Studies&quot;) OR (MH &quot;Retrospective Design&quot;) OR (MH &quot;Empirical Research&quot;) OR (MH &quot;Crossover Design&quot;) OR MW &quot;dt&quot; OR TI ( randomized OR randomised OR randomization OR randomisation OR placebo OR randomly OR trial OR groups OR “clinical trial” OR “clinical trials” OR &quot;comparative study&quot; OR &quot;meta-analysis&quot; OR &quot;meta-analyses&quot; OR &quot;evaluation study&quot; OR “evaluation studies” OR &quot;intervention study&quot; OR &quot;intervention studies&quot; OR &quot;case-control&quot; OR cohort OR &quot;longitudinal&quot; OR longitudinally OR &quot;prospective&quot; OR prospectively OR &quot;retrospective&quot; OR &quot;follow up&quot; ) OR AB ( randomized OR randomised OR randomization OR randomisation OR placebo OR randomly OR trial OR groups OR “clinical trial” OR “clinical trials” OR &quot;comparative study&quot; OR &quot;meta-analysis&quot; OR &quot;meta-analyses&quot; OR &quot;evaluation study&quot; OR &quot;evaluation studies&quot; OR &quot;intervention study&quot; OR &quot;intervention studies&quot; OR &quot;case-control&quot; OR cohort OR &quot;longitudinal&quot; OR longitudinally OR &quot;prospective&quot; OR prospectively OR &quot;retrospective&quot; OR &quot;follow up&quot;)</td>
</tr>
<tr>
<td>#6</td>
<td>#4 AND #5</td>
</tr>
<tr>
<td>#7</td>
<td>Limits: English</td>
</tr>
</tbody>
</table>
**PsycINFO® search strategy (September 10, 2013)**

<table>
<thead>
<tr>
<th>Set #</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>DE &quot;Breast Neoplasms&quot; OR ((TI breast OR AB breast) AND (TI neoplasms OR neoplasm OR cancer OR carcinoma OR carcinomas) OR AB neoplasms OR neoplasm OR cancer OR carcinoma OR carcinomas))</td>
</tr>
<tr>
<td>#2</td>
<td>DE &quot;Mammography&quot; OR TI mammography OR mammograms OR mammogram OR AB mammography OR mammograms OR mammogram OR DE &quot;Cancer Screening&quot; OR TI screening OR AB screening</td>
</tr>
<tr>
<td>#3</td>
<td>DE &quot;Test Anxiety&quot; AND DE &quot;Anxiety&quot; OR DE &quot;Depression (Emotion)&quot; OR DE &quot;Stress&quot; OR DE &quot;Psychological Stress&quot; OR TI (stress OR anxiety OR anxious OR depression OR depressed) AND AB (stress OR anxiety OR anxious OR depression OR depressed)</td>
</tr>
<tr>
<td>#4</td>
<td>#1 AND #2 AND #3</td>
</tr>
<tr>
<td>#5</td>
<td>Limits: English</td>
</tr>
</tbody>
</table>
Appendix B. Data Abstraction Elements

Study Characteristics

- Study Name
- Additional Articles Used in This Abstraction
- Geographic Location (Select all that apply)
  - US, Canada, UK, Nordic countries, Europe (non-Nordic Europe), S/C America, Asia, Africa, Middle East, Australia/NZ, Unclear/Not reported
- Study Dates
  - Year study intervention began
  - Year study intervention stopped
  - Year follow-up stopped
- Study Design
  - RCT
  - Prospective cohort
  - Retrospective cohort
  - Case-control
  - Cross-sectional
  - Other (specify)
- Setting (Select all that apply)
  - Organized Screening Program, Opportunistic Screening, Unclear/Not reported, Other (specify)
- Study Inclusion and Exclusion Criteria
  - Copy/paste as reported in article
- Key Question Applicability (Select all that apply)
  - KQ1, KQ2, KQ3, KQ4, KQ5
- Study Population
  - Total number of patients enrolled/included across all arms
  - Comorbidities (N and %)
  - Population Characteristics
    - Average Risk (N and %)
    - High Risk
      - Family History (N and %)
      - BrCA 1/2 Carrier (N and %)
      - Prior abnormal biopsy (N and %)
      - Unspecified (N and %)
    - Ethnicity
      - Hispanic or Latino (N and %)
      - Not Hispanic or Latino (N and %)
    - Race
      - Black/African American (N and %)
      - Native American/Alaskan (N and %)
      - Asian/Pacific Islander (N and %)
      - White (N and %)
      - Other (N and %)
• Age
  • Younger than 50 (N and %)
  • 50-74 (N and %)
  • 75 and Older (N and %)
  • Minimum Age
  • Maximum Age

• Outcomes (Check all that apply)
  o Breast cancer mortality
  o All-cause mortality
  o Quality of life
  o Overdiagnosis
  o Overtreatment
  o False positive: same day repeat examination
  o False positive: subsequent visit repeat examination (recall)
  o False positive: biopsy
  o False positive: unspecified
  o Stage distribution at diagnosis
  o Emotional impact (anxiety, depression, etc. of positive results (true and false positive))
  o Reassurance from true negatives
  o False reassurance from false negatives
  o Secondary effects of test results on health resource utilization, both breast cancer related and non-breast cancer related
  o Radiation exposure (high risk populations)
  o Recall rates
  o Sensitivity and specificity
  o Patient preferences as measured using validated quality of life measures, utilities using accepted methods such as standard gamble or time-tradeoff; stated preferences measured by conjoint analysis; revealed preference studies; etc.

• Comments

Intervention Characteristics
• Intervention Characteristics
  o Group 1, Group 2, Group3, Group 4
  • Screening Modality
    • Mammography
      o Double View/Single View/NR
      o Single Reader/Double Reader/NR
      o Digital/Film/NR
      o No CAD/CAD/NR
    • CBE
      o Family Physician
      o Nurse Practitioner
      o OBGYN
      o Other (specify)
• NR
  • Ultrasound
  • MRI
  • Tomosynthesis
  • No Screening

- Screening Interval
  • 1 year
  • 2 years
  • 3 years
  • Alternative interval (specify)

- Comments

Outcomes

- Select the critical outcome reported on this form
  • Breast cancer mortality
  • All-cause mortality
  • Quality of life
  • Overdiagnosis
  • Overtreatment
  • False positive: same day repeat examination
  • False positive: subsequent visit repeat examination ("recall")
  • False positive: biopsy
  • False positive: unspecified

- Select the non-critical outcomes reported on this form
  • Stage distribution at diagnosis
  • Emotional impact (anxiety, depression, etc. of positive results (true and false positives))
  • Reassurance from true negatives
  • False reassurance from false negatives
  • Effects of results on health resource use (BrCA related & non-BrCA related)
  • Radiation exposure (added as important outcome for high-risk population)
  • Recall rates
  • Sensitivity and specificity (only if a 2x2 table can be completed)
  • Pt. preferences using validated QOL measures and utilities w/accepted methods (see protocol)

- Outcomes definition

- Specify the timepoint for this outcome
  • Immediate (up to 12 weeks from the screening)
  • Short-term (within 12 weeks-18 months of screening)
  • Longer-term (greater than 18 months after screening)
    ▪ List all timepoints after 18 months
  • Unclear/NR

- Outcomes Data Table
  • Group 1, Group 2, Group3, Group 4
    ▪ N Analyzed
• Result
  - Mean
  - Median
  - Mean within group change
  - Mean between group change
  - Number of patients with outcome
  - % of patients with outcome
  - Events/denominator
  - Odds ratio (OR)
  - Hazard ratio (HR)
  - Relative risk (RR)
  - Other (specify)

• Variability
  - Standard Deviation (SD)
  - Standard Error (SE)
  - IQR
  - 95% CI
  - Other % CI (specify)
  - Other (specify)

• p-value between groups
• Reference group (for comparisons between groups)

• Quality
  - Study design
    - RCT
    - Cohort
    - Case-control
    - Cross-sectional
  - RCT Limitations
    - Lack of allocation concealment [Investigators potentially aware of how treatment will be allocated for a particular subject (e.g., randomization by record number, birthday, or day of the week)]
    - Lack of blinding [Subjects and/or investigators aware of treatment allocation]
    - Incomplete accounting of patients and outcome events [Don’t have CONSORT diagram or, for older studies, description of patient flow through study. High loss to follow-up (>10%)]
    - Selective outcome reporting bias [Don’t report all outcomes of interest regarding harms (e.g., would only report mortality reduction, not false positives)]
    - Stopping early for benefit
    - Use of unvalidated outcome measures
    - Carryover effects in cross-over trials
    - Recruitment bias in cluster randomized trials
  - Observation Limitations (for cohort, case-control, and cross-sectional study designs)
• Failure to develop and apply appropriate eligibility criteria [In cohort studies, major differences between screened and unscreened; in case-control studies, controls would NOT have been cases if they developed the outcome]
• Flawed measurement of both exposure and outcome [E.g.—confirmation of screening history not performed (exposure), or outcome not validated (death truly from breast cancer)]
• Failure to adequately control confounding [Didn’t use matching or multivariate analysis]
• Incomplete follow-up [Loss to follow-up >10%]
  o Indirectness
  ▪ Location
    • US
    • Non-US, opportunistic screening
    • Non-US, organized screening
  ▪ Mammography methods
    • Single/double view
    • Single/double reader
    • Film/digital
    • CAD/no CAD
    • NR/Not applicable
  o Imprecision
    • Yes/No
  o Other consideration
  o Quality Rating
    ▪ High
    ▪ Moderate
    ▪ Low
    ▪ Very Low
  • Comments

**Quality Assessment**

• Study Design:
  ▪ RCT
    ▪ High Quality
    ▪ Moderate Quality
    ▪ Low Quality
  ▪ Observational
    ▪ Observational Studies (specify study design)
      • Prospective Cohort
      • Retrospective Cohort
      • Case Control
      • Cross Sectional
      • Other
    ▪ Select the outcomes included in this study (Check all that apply)
• Breast cancer mortality
• All-cause mortality
• Quality of Life (QOL)
• Overdiagnosis
• Overtreatment

- Do you also have false positive outcomes to do a quality assessment on?
  - No
  - Yes
    - False Positive: Same day
    - False Positive: Recall
    - False Positive: Biopsy
    - False Positive: Unspecified

- Selection Bias
  - High: Historical controls; Different baseline characteristics without adjustment (Stratification, multivariate analysis)
  - Low: Concurrent controls with adjustment (Demographics, age, lead time, self-selection for screening)

- Performance Bias
  - High: Failure to adjust for secular trends in breast cancer treatment with historical controls
  - Low: Concurrent controls or specific methods to adjust for time-varying effects

- Attrition Bias
  - High: Differential length or completeness of follow-up between comparison groups Differential adherence to protocol among comparison groups (E.g., greater adherence with annual compared to biennial screening)
  - Low: Similar length, completeness, adherence between comparison groups

- Detection Bias
  - High: Different methods for assessing exposure to screening Different methods for assessing outcomes
  - Low: Similar methods for assessing exposure/outcomes Use of alternative methods and reporting both (e.g., mortality vs. underlying cause of death)

- Reporting Bias
  - High: Pre-specified outcome not reported
  - Low: All pre-specified primary outcomes reported

- Overall Quality Rating
  - High Quality
  - Moderate Quality
  - Low Quality

**High Quality:** Has the least bias, and results are considered valid. A good study has a clear description of the population, setting, interventions, and comparison groups; uses recruitment and eligibility criteria that minimizes selection bias; has a low attrition rate; and uses appropriate means to prevent bias, measure outcomes, and analyze and report results. These studies will meet the majority of items in each domain.
In general, you are confident in both the DIRECTION of the reported effect, and in the overall SIZE of the effect.

**Moderate Quality:** Is susceptible to some bias but probably not enough to invalidate the results. The study may be missing information, making it difficult to assess limitations and potential problems. As the fair-quality category is broad, studies with this rating vary in their strengths and weaknesses. The results of some fair-quality studies are possibly valid, while others are probably valid. These studies will meet the majority of items in most but not all domains. In general, you are confident in the DIRECTION of the reported effect, but not necessarily the overall SIZE of the effect.

**Low Quality:** Indicates significant bias that may invalidate the results. These studies have serious errors in design, analysis, or reporting; have large amounts of missing information; or have discrepancies in reporting. The results of a poor-quality study are at least as likely to reflect flaws in the study design as to indicate true differences between the compared interventions. In general, you have little confidence in the study’s estimate of the DIRECTION of the reported effect.
Appendix C. Modeling Methods

Estimating Absolute Effects of Breast Cancer Screening

In order to formulate recommendations about screening, the GDG will need to consider the absolute magnitude of both benefits and harms of screening for specified groups within the US. Because the majority of the available literature on screening efficacy and effectiveness, particularly regarding mortality prevention, is based on studies performed outside of the US, direct estimates of the absolute effect of screening on outcomes, particularly for mortality, for the U.S. are not available; SEER does not capture whether a particular incident case was diagnosed via screening or presentation with symptoms. Therefore, an indirect method needs to be used.

This document describes our approach to estimating the absolute effect of screening on the three critical outcomes of breast cancer mortality, overdiagnosis, and false positive rates, focusing on the estimates for each individual outcome, as well as the method used for estimating specific harm-benefit trade-offs.

I. Breast Cancer Mortality

A. Data Sources
Standard cancer-specific mortality rates published by SEER are based on death certificate data reported to the National Center for Health Statistics. Data are reported based on age at death alone. Because age-specific mortality is readily available, it is commonly used as a first approximation for estimating the impact of cancer screening or treatment changes on outcomes. For example, a recent paper estimating the absolute harms and benefits of breast cancer screening in the U.S. used this approach.

However, for cancers where late recurrence is not uncommon, such as breast cancer, using only the age at the time of death for estimating mortality (number of deaths divided by number of people alive in a given age stratum) means that some deaths occurring within a given age window (for example, 50-59), will be from cancers diagnosed prior to age 50, and that some deaths from cancers diagnosed between ages 50-59 will occur later. Using age-specific mortality rates within a given age group to estimate the potential number of deaths prevented by screening within that group is therefore subject to error—some deaths will be the result of cancers diagnosed prior to beginning screening in that group, resulting in overestimation, and some deaths occurring later will not be counted, resulting in underestimation. The net effect of over- and undercounting deaths attributable to cancer diagnosed within an age group will vary because of age-specific variation in both cancer-specific and other cause mortality.

One way to avoid this particular source of error is to use estimates of incidence-based mortality, in which only deaths among patients after a known date of diagnosis are counted. Depending on the cancer registry, incidence-based mortality can also be stratified based on age at diagnosis, year of diagnosis, method of diagnosis, cancer stage/grade, etc. Incidence-based mortality for specific cancers in the US can be estimated using SEER*stat (Surveillance Research Program, National Cancer Institute SEER*Stat software (www.seer.cancer.gov/seerstat) version 8.1.2).
Incidence-based mortality is calculated by dividing the number of deaths occurring in a given year among women who were diagnosed with cancer at some predetermined point in the past by the number of women alive in a given year.

Table 1 provides an example for a population of 100,000 women at age 50 for a hypothetical cancer with an incidence of 100 per 100,000, and a mortality rate from other causes of 500 per 100,000. This hypothetical cancer has a 50% 5 year survival, with 30% of new patients dying within the first year, 15% in year 2, 10% in year 3, and 5% in years 4 and 5.

Table 1: Example of Incidence-based Mortality Calculations

<table>
<thead>
<tr>
<th>Age</th>
<th>Number Alive</th>
<th>Number of Cancer Deaths by Age at Diagnosis</th>
<th>Other Cause Deaths</th>
<th>Incidence-based Mortality per 100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>100,000</td>
<td>30</td>
<td></td>
<td>500.0</td>
</tr>
<tr>
<td>51</td>
<td>99,470</td>
<td>15 29.8</td>
<td></td>
<td>497.4</td>
</tr>
<tr>
<td>52</td>
<td>98,928</td>
<td>10 15.0</td>
<td>29.7</td>
<td>494.8</td>
</tr>
<tr>
<td>53</td>
<td>98,378</td>
<td>5 9.9</td>
<td>15.0 29.5</td>
<td>492.1</td>
</tr>
<tr>
<td>54</td>
<td>97,827</td>
<td>5 5.0</td>
<td>9.9 15.0 29.3</td>
<td>489.4</td>
</tr>
</tbody>
</table>

- Of the 100,000 women alive at age 50, 100 will be diagnosed with the cancer, of whom 30 (30%) will die in the first year; another 500 women will die from other causes. The age-specific incidence-based mortality for 50 year olds is 30 per 100,000.
- At age 51, 99,470 women are alive (100,000 minus the 30 cancer deaths and 500 other cause deaths). Of these women, approximately 497 (500 per 100,000 multiplied by 99,470) will die of other causes, and approximately 100 will develop cancer. 30% of the newly diagnosed women will die during this first year after diagnosis, while 15% of the women diagnosed in the previous year will die. The total incidence-based mortality is then (30+20)/99,470, or 45.1 per 100,000.
- This process continues for each year, with the cumulative incidence-based mortality being the sum of the total for each year (30.0+45.1+55.3+60.4+65.7), or 256.5 per 100,000.

Tables 2-5 show the estimated incidence-based mortality for invasive breast cancer stratified by age at diagnosis and age at death, derived from 1992-2010 SEER data (Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence-Based Mortality - SEER 13 Regs Research Data, Nov 2012 Sub (1992-2010) <Katrina/Rita Population Adjustment> - Linked To County Attributes - Total U.S., 1969-2011 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2013, based on the November 2012 submission.). These particular years were chosen to optimize follow-up duration (up to 18 years), consistency of screening behavior (particularly for 1995-2010, age-specific screening rates are consistently around 65%, as described in more detail below), and relevance to current standards of treatment. Rates are per 100,000, based on the number of observed deaths within each cell and the total number of women in each single-year age category for age at death. Highlighted cells in the
tables illustrate the values used to estimate 15 year incidence-based mortality at ages 40-49, 50-59, 60-69, and 70-84 and above, as described in section I.B.

Table 2: Age-specific incidence-based mortality (per 100,000) from breast cancer by age at diagnosis, SEER 1992-2010 (age at diagnosis 40-49 years)

<table>
<thead>
<tr>
<th>Age at Death</th>
<th>Age at Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40</td>
</tr>
<tr>
<td>40</td>
<td>0.6</td>
</tr>
<tr>
<td>41</td>
<td>2.2</td>
</tr>
<tr>
<td>42</td>
<td>2.8</td>
</tr>
<tr>
<td>43</td>
<td>2.1</td>
</tr>
<tr>
<td>44</td>
<td>1.9</td>
</tr>
<tr>
<td>45</td>
<td>1.4</td>
</tr>
<tr>
<td>46</td>
<td>1.1</td>
</tr>
<tr>
<td>47</td>
<td>1</td>
</tr>
<tr>
<td>48</td>
<td>0.8</td>
</tr>
<tr>
<td>49</td>
<td>0.7</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
</tr>
<tr>
<td>51</td>
<td>0.5</td>
</tr>
<tr>
<td>52</td>
<td>0.3</td>
</tr>
<tr>
<td>53</td>
<td>0.2</td>
</tr>
<tr>
<td>54</td>
<td>0.2</td>
</tr>
<tr>
<td>55</td>
<td>0.2</td>
</tr>
<tr>
<td>56</td>
<td>0.1</td>
</tr>
<tr>
<td>57</td>
<td>0.1</td>
</tr>
<tr>
<td>58</td>
<td>0</td>
</tr>
<tr>
<td>59</td>
<td>0.1</td>
</tr>
<tr>
<td>60</td>
<td>0.3</td>
</tr>
<tr>
<td>61</td>
<td>0.3</td>
</tr>
<tr>
<td>62</td>
<td>0.1</td>
</tr>
<tr>
<td>63</td>
<td>0.1</td>
</tr>
<tr>
<td>64</td>
<td>0.1</td>
</tr>
<tr>
<td>Age at Death</td>
<td>50</td>
</tr>
<tr>
<td>--------------</td>
<td>-----</td>
</tr>
<tr>
<td>50</td>
<td>1.5</td>
</tr>
<tr>
<td>51</td>
<td>4.3</td>
</tr>
<tr>
<td>52</td>
<td>5.7</td>
</tr>
<tr>
<td>53</td>
<td>4.2</td>
</tr>
<tr>
<td>54</td>
<td>3.3</td>
</tr>
<tr>
<td>55</td>
<td>2.7</td>
</tr>
<tr>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>57</td>
<td>1.7</td>
</tr>
<tr>
<td>58</td>
<td>1.4</td>
</tr>
<tr>
<td>59</td>
<td>1.4</td>
</tr>
<tr>
<td>60</td>
<td>1</td>
</tr>
<tr>
<td>61</td>
<td>0.8</td>
</tr>
<tr>
<td>62</td>
<td>0.9</td>
</tr>
<tr>
<td>63</td>
<td>0.7</td>
</tr>
<tr>
<td>64</td>
<td>0.5</td>
</tr>
<tr>
<td>65</td>
<td>0.3</td>
</tr>
<tr>
<td>66</td>
<td>0.4</td>
</tr>
<tr>
<td>67</td>
<td>0.1</td>
</tr>
<tr>
<td>68</td>
<td>0.1</td>
</tr>
<tr>
<td>69</td>
<td>0.1</td>
</tr>
<tr>
<td>70</td>
<td>0.1</td>
</tr>
<tr>
<td>71</td>
<td>0.1</td>
</tr>
<tr>
<td>72</td>
<td>0.1</td>
</tr>
<tr>
<td>73</td>
<td>0.1</td>
</tr>
<tr>
<td>74</td>
<td>0.1</td>
</tr>
<tr>
<td>Age at Death</td>
<td>60</td>
</tr>
<tr>
<td>--------------</td>
<td>----</td>
</tr>
<tr>
<td>60</td>
<td>3.8</td>
</tr>
<tr>
<td>61</td>
<td>6</td>
</tr>
<tr>
<td>62</td>
<td>7</td>
</tr>
<tr>
<td>63</td>
<td>6.3</td>
</tr>
<tr>
<td>64</td>
<td>5.3</td>
</tr>
<tr>
<td>65</td>
<td>4</td>
</tr>
<tr>
<td>66</td>
<td>3.8</td>
</tr>
<tr>
<td>67</td>
<td>2.2</td>
</tr>
<tr>
<td>68</td>
<td>2.3</td>
</tr>
<tr>
<td>69</td>
<td>2.2</td>
</tr>
<tr>
<td>70</td>
<td>2.2</td>
</tr>
<tr>
<td>71</td>
<td>1.2</td>
</tr>
<tr>
<td>72</td>
<td>1</td>
</tr>
<tr>
<td>73</td>
<td>0.8</td>
</tr>
<tr>
<td>74</td>
<td>0.7</td>
</tr>
<tr>
<td>75</td>
<td>0.3</td>
</tr>
<tr>
<td>76</td>
<td>0.2</td>
</tr>
<tr>
<td>77</td>
<td>0.2</td>
</tr>
<tr>
<td>78</td>
<td>0.1</td>
</tr>
<tr>
<td>79</td>
<td>0.2</td>
</tr>
<tr>
<td>80</td>
<td>0.1</td>
</tr>
<tr>
<td>81</td>
<td>0.2</td>
</tr>
<tr>
<td>82</td>
<td>0.1</td>
</tr>
<tr>
<td>83</td>
<td>0.1</td>
</tr>
<tr>
<td>84</td>
<td>0.1</td>
</tr>
<tr>
<td>Age at Death</td>
<td>Age at Diagnosis</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>73</td>
<td>73</td>
</tr>
<tr>
<td>74</td>
<td>74</td>
</tr>
<tr>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>76</td>
<td>76</td>
</tr>
<tr>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td>78</td>
<td>78</td>
</tr>
<tr>
<td>79</td>
<td>79</td>
</tr>
<tr>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>81</td>
<td>81</td>
</tr>
<tr>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>83</td>
<td>83</td>
</tr>
<tr>
<td>84</td>
<td>84</td>
</tr>
<tr>
<td>85</td>
<td>85</td>
</tr>
</tbody>
</table>

As an example, the crude age-specific mortality from breast cancer during the same time period in the SEER dataset for women 55-59 years old was 51.3 per 100,000; this represents deaths from breast cancer occurring in this age range, no matter when the cancer was initially diagnosed. This is the parameter used for the overall mortality estimate by Welch and Passow.  

![Figure 1: Age-specific Breast Cancer Mortality by Single Year of Age, U.S., 1999-2010](image-url)
Estimates for mortality for single year ages are not available from SEER, but are from the National Center for Health Statistics—breast cancer mortality for 55 year olds during the period 1999-2010 was 40.9 per 100,000 (Figure 1).

The rows in Tables 2-5 represent deaths at the specified age, while the columns represent age at diagnosis. The overall mortality from breast cancer for women aged 55 will be the sum of deaths from cases diagnosed at age 55 plus deaths from cases diagnosed at age 54, plus deaths from cases diagnosed at age 53, and so on; the available SEER data is constrained to 15 years. The total mortality at age 55 calculated by summing across the row for age 55 in Table 1 (age at diagnosis 40-49) and Table 2 (age at diagnosis 50-59) is 35.7/100,000. The difference is attributable (a) the fact that SEER rates are estimates based on cases identified within the SEER registries, while the NCHS data represents all reported death certificates, (b) the contribution of cases diagnosed prior to the 15 year follow-up window in the SEER data; and (c) deaths occurring beyond the 15 year window which are not captured in the data extracted from SEER. Estimates of incidence-based mortality are consistently lower by 4-12% than estimates based on death certificate data.

Of the total incidence-based breast cancer mortality in 55 year old women (35.7 per 100,000), 27.7% (9.9 per 100,000, the sum of the 55 year old row in Table 1) is attributable to cases diagnosed prior to age 50. In an estimate of the potential effect of screening beginning at age 50 on mortality, these cases should not be counted.

Breast-cancer specific mortality moving forward from a given age is estimated by summing the mortality at each subsequent age. For women aged 40 years (Table 1), the cumulative mortality to age 45 is the sum of breast cancer deaths diagnosed at age 40 (0.6), plus the sum of deaths at age 41 that were diagnosed at ages 40 and 41 (2.2 + 0.6), plus the sum of deaths at age 42 that were diagnosed at ages 40, 41, and 42 (2.8 + 2.4 + 0.7), plus deaths at 43 diagnosed at 40-43 (2.1 + 2.7 + 2.3 + 0.8), plus deaths at 44 diagnosed at ages 40-44 (1.9 + 2.2 + 2.8 + 2.7 + 1.3), for a total of 28.1 per 100,000. Because the denominator for each age-specific mortality estimate is the number of women alive in that year, competing causes of death are captured and the sum represents cumulative cause-specific mortality in the presence of competing risks of death.

B. Generating Mortality Estimates
For the purposes of estimating the absolute breast cancer mortality reduction attributable to screening for the US, we used 15 year cumulative mortality based on starting age for screening, in 5 year increments (i.e., mortality from ages 40-54, 45-59, 50-64, etc.), for the following reasons:

- Time horizons shorter than 15 years do not capture late mortality, while time horizons longer than 15 years require either using pre-1992 data (where screening behavior and treatment options may have differed), or exacerbating potential age-period-cohort effects (for example, by assuming recent mortality rates from both breast cancer and other causes) for 70 year olds will still be applicable in 20 years to current 50 year olds.
- Changes in primary and secondary prevention strategies, treatment options, competing risks, etc. are likely over the course of the next 15 years. Although guidelines will certainly be revised based on such changes, basing current recommendations on estimates
of nearer term outcomes limits the sources of uncertainty to the available literature, without adding the unforeseeable future.

- As discussed in the main report, patient preferences for the time at which different health-related events may occur are measurable, can affect decision making, and may vary substantially between patients. A shorter time horizon somewhat mitigates the effects of

Table 6 presents estimates for cumulative breast cancer mortality over 15 years in different 10 year age groups, generated using four different methods (described below):

Note that these estimates are for the cumulative mortality for a cohort STARTING at a given age through 15 years—for example, for 40-54 year olds, the total is the cumulative mortality from age 40-54 for those who are 40 at the start of the interval, from ages 41-54 for those who are 41 at the start of the interval, for ages 42-54 for those who are 42 at the start of the interval.

**Method 1**: Cumulative 15 year mortality estimates at different starting ages, derived directly with DevCan 6.7.0, a software package developed by NCI for estimating the probability of developing or dying from specific cancers, based on SEER data from 2000-2010 and mortality data collected by NCHS from the same years; these are the source of the overall mortality estimates used by Welch and Passow.

**Method 2**: Direct estimates using life table methods, with age-specific mortality for breast cancer and other causes of death (in single year increments) taken directly from NCHS data from 1999-2010.

**Method 3**: Summing incidence-based mortality estimates based on age at diagnosis and age at death obtained from SEER*stat (from Tables 2-5) as described above.

**Method 4**: Direct estimates using life table methods (implemented as a Markov state-transition model), using

- Age-specific incidence of invasive and in situ breast cancer (SEER*stat)
- For non-cancer cases, annual age-specific probability of death from other causes (from NCHS)
- For cancer cases, conditional probability of death from cancer or other causes based on age at diagnosis and number of years since diagnosis (SEER*stat)

<table>
<thead>
<tr>
<th>Age Interval</th>
<th>Cumulative Mortality per 100,000</th>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
<th>Method 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death Certificate (Crude Age-specific Mortality)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40-49</td>
<td>322</td>
<td>322.8</td>
<td>327</td>
<td>244.3</td>
<td>248</td>
</tr>
<tr>
<td>45-54</td>
<td>454</td>
<td>462.7</td>
<td>465</td>
<td>328.5</td>
<td>322</td>
</tr>
<tr>
<td>50-59</td>
<td>630</td>
<td>648.1</td>
<td>645</td>
<td>400.7</td>
<td>413</td>
</tr>
<tr>
<td>55-64</td>
<td>806</td>
<td>791.1</td>
<td>798</td>
<td>458.9</td>
<td>482</td>
</tr>
<tr>
<td>60-69</td>
<td>969</td>
<td>950.3</td>
<td>987</td>
<td>524.1</td>
<td>568</td>
</tr>
<tr>
<td>65-74</td>
<td>1,114</td>
<td>1,091.4</td>
<td></td>
<td>596</td>
<td>713</td>
</tr>
<tr>
<td>70-79</td>
<td>1,258</td>
<td>1,216.2</td>
<td></td>
<td>553.5</td>
<td>622</td>
</tr>
</tbody>
</table>

Method 1: Estimated cumulative mortality from SEER, 2000-2010, using DevCan 6.7.0, similar to
Method 2: Derived from age-specific breast cancer and other cause mortality (in one year increments) from NCHS, 1999-2010
Welch and Passow: Estimates for specified age-interval presented in the paper.  
C-8
Method 4: SEER age-specific incidence (13 Registries, 1992-2010), SEER survival (conditioned on age at diagnosis and time since diagnosis up to 15 years, from SEER 18 Registries, Nov 2012 submission, 1992-2010), and other cause mortality (SEER and NCHS)

As expected, methods that are based on age-specific mortality alone, without consideration of age at diagnosis (Methods 1 and 2), result in higher estimates, because of the inclusion of deaths from cancer diagnosed prior to the start of “follow-up” time. Estimates based on either direct incidence-based mortality from SEER*stat or life table estimates based on combined incidence, survival, and mortality probabilities (Methods 3 and 4) are generally lower. Cumulative incidence-based mortality somewhat underestimates 15 year mortality starting at age 70, because SEER truncates estimates at age 85 (less than 5 per 10,000), with the exception of the interval 70-84 years, where incidence-based mortality results in a substantially higher number of deaths. This underestimation of mortality at advanced ages with survival-based estimates is a common finding in initial cancer models, and is due to different assumptions about when the breast cancer mortality rates are applied (because the risk of competing risks of death is so high in the older population, the size of the population at risk varies depending on the modeling method, resulting in different absolute numbers of events).

**Method 3 is the one used as the basis for overall mortality estimates with and without screening presented in the main report, while Method 4 is the one used for estimating harm-benefit trade-offs.**

C. Estimating the Effect of Mammography on Mortality

Although SEER*stat provides detailed data on a number of patient and tumor characteristics (such as age, race, insurance status, tumor stage, size, estrogen and progesterone receptor status, etc), it does NOT provide any data on how the initial tumor diagnosis was made, so that direct comparisons of incidence-based mortality between screened and unscreened women are not possible.

The difference in event rates between people exposed and unexposed to a particular risk factor can be derived as a function of the overall event rate, the prevalence of the exposure, and the relative risk associated with the exposure. Although this approach (which is commonly used in epidemiology to estimate the proportion and absolute number of cases attributable to a specific exposure) is identical to the one used by Welch and Passow, the difference is that our estimate of mortality is derived from incidence-based mortality rather than crude age-specific mortality, as described above.

Estimates of the prevalence of exposure to screening mammography are provided by the National Health Information Survey. Estimates for the proportion of women who report a mammogram within the past 2 years, by age, are provided in Table 76 for the 2008 NHIS (the last year where direct access to the data is readily available); other published data from the survey suggests that these age-specific rates have been remarkably stable between 1995 and 2010, which incorporates the majority of the available incidence-based mortality data.
Table 7: Proportion of US women reporting a mammogram within the past two years (National Health Interview Survey, 2008)

<table>
<thead>
<tr>
<th>Age</th>
<th>Percent (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40→44</td>
<td>65.3% (61.2 to 69.2%)</td>
</tr>
<tr>
<td>45→49</td>
<td>64.5% (60.5 to 68.3%)</td>
</tr>
<tr>
<td>50→54</td>
<td>65.8% (61.9 to 69.5%)</td>
</tr>
<tr>
<td>55→59</td>
<td>69.5% (65.8 to 73.0%)</td>
</tr>
<tr>
<td>60→64</td>
<td>67.7% (64.0 to 71.3%)</td>
</tr>
<tr>
<td>65→69</td>
<td>71.7% (67.7 to 75.4%)</td>
</tr>
<tr>
<td>70→74</td>
<td>65.6% (64.3 to 70.6%)</td>
</tr>
<tr>
<td>75→79</td>
<td>62.7% (57.3 to 67.8%)</td>
</tr>
<tr>
<td>80→84</td>
<td>53.7% (47.8 to 59.6%)</td>
</tr>
</tbody>
</table>

Using these age-specific values for the prevalence of screening, direct incidence-based mortality estimates from SEER*stat (Method 3), and 3 different point estimates for mortality reduction that are consistent with the range reported in the randomized trials and observational studies, we generated estimates of the absolute difference in breast cancer mortality per 100,000 women over 15 years at different starting intervals, along with the number needed to screen (NNS) to prevent one breast cancer death (1 over the absolute difference).

To illustrate, the 15 year cumulative mortality for women 40-49 for method 3 in Table 5 above is 244.3 per 100,000, for women 60-69 524.1 per 100,000. Age-specific screening prevalence is 65.3% for women 40-44, and 67.7% for women 60-64; for simplicity here, we will apply these values to the entire age range, although the results presented in the main report vary the rates by 5-year age group.

For women 40-49, the mortality in unscreened women with a relative mortality reduction from screening of 0.6 is estimated by

\[
\text{Overall Mortality} / \left( (RR_{Mortality,Screened} \times Prevalence_{Screened} ) + (1 - Prevalence_{Screened}) \right)
\]

Or

\[
0.00243 / [(0.6 \times 0.653) + (1 - 0.653)] = 0.00329
\]

The estimate for screened women is 0.6*0.00329, or 0.00197, for an absolute difference of 0.0013, or an NNS (1 divided by the absolute difference) of 760. For 60-69 year olds, the mortality per 100,000 is 524.1, with a 67.7% screening prevalence, resulting in estimates of 718.6 per 100,000 mortality in unscreened women, 413.2 per 100,000 in screened women, and absolute difference of 0.0029, and NNS of 347.9.

Estimates of NNS for 40-54 year olds are quite consistent with estimates from the UK AGE trial, with a point estimate for 10 year mortality reduction with annual screening from 39-48 years compared to usual care of 0.83, and an estimated NNS over 10 years of 2512. Applying this method to screening for ages 50-70, with follow-up to age 84, at a relative mortality reduction of 0.8 and screening prevalence of 0.65, results in a NNS of 432, again quite consistent with other estimates.
II. Overdiagnosis

A. Cumulative Overall Diagnoses

Table 8 depicts the 15 year cumulative incidence of malignant and in situ breast cancers in SEER, estimated using DevCan.

**Table 8: Cumulative Incidence of Malignant and In Situ Cancers, SEER 2000-2010**

<table>
<thead>
<tr>
<th>Ages</th>
<th>Cumulative Incidence per 100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Malignant</td>
</tr>
<tr>
<td>40→54</td>
<td>2488</td>
</tr>
<tr>
<td>45→59</td>
<td>3143</td>
</tr>
<tr>
<td>50→64</td>
<td>3848</td>
</tr>
<tr>
<td>55→69</td>
<td>4647</td>
</tr>
<tr>
<td>60→74</td>
<td>5277</td>
</tr>
<tr>
<td>65→79</td>
<td>5305</td>
</tr>
</tbody>
</table>

Using SEER*Stat, we obtained age-specific incidences for all malignant breast cancer, in-situ, and T1N0M0 cancers, and derived the estimate for all cancers 20 mm or greater and/or with lymph node involvement or distant metastases by subtracting age-specific T1N0M0 from all malignant cases (Table 9).

<table>
<thead>
<tr>
<th>Age</th>
<th>In Situ</th>
<th>T1N0M0</th>
<th>All Others</th>
<th>Age</th>
<th>In Situ</th>
<th>T1N0M0</th>
<th>All Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>31.1</td>
<td>35.5</td>
<td>66.5</td>
<td>66</td>
<td>96.6</td>
<td>206.9</td>
<td>203.1</td>
</tr>
<tr>
<td>41</td>
<td>32.8</td>
<td>38.3</td>
<td>67</td>
<td>67</td>
<td>101.1</td>
<td>207.8</td>
<td>195.3</td>
</tr>
<tr>
<td>42</td>
<td>34</td>
<td>44.4</td>
<td>73.2</td>
<td>68</td>
<td>99.1</td>
<td>212.6</td>
<td>195.6</td>
</tr>
<tr>
<td>43</td>
<td>39.2</td>
<td>48.6</td>
<td>81.2</td>
<td>69</td>
<td>95.9</td>
<td>218</td>
<td>201.3</td>
</tr>
<tr>
<td>44</td>
<td>44.7</td>
<td>56.5</td>
<td>91</td>
<td>70</td>
<td>96</td>
<td>217.5</td>
<td>203.8</td>
</tr>
<tr>
<td>45</td>
<td>49.1</td>
<td>63.7</td>
<td>99.5</td>
<td>71</td>
<td>94.6</td>
<td>217.7</td>
<td>202.9</td>
</tr>
<tr>
<td>46</td>
<td>52.7</td>
<td>69.7</td>
<td>105.4</td>
<td>72</td>
<td>95.7</td>
<td>222.9</td>
<td>199.1</td>
</tr>
<tr>
<td>47</td>
<td>59</td>
<td>76</td>
<td>112.8</td>
<td>73</td>
<td>92.8</td>
<td>223.2</td>
<td>208.2</td>
</tr>
<tr>
<td>48</td>
<td>61.7</td>
<td>81.6</td>
<td>119.6</td>
<td>74</td>
<td>93.1</td>
<td>226.9</td>
<td>206.4</td>
</tr>
<tr>
<td>49</td>
<td>64.5</td>
<td>88.5</td>
<td>122.8</td>
<td>75</td>
<td>95.3</td>
<td>227</td>
<td>222</td>
</tr>
<tr>
<td>50</td>
<td>69.8</td>
<td>92.4</td>
<td>129</td>
<td>76</td>
<td>91</td>
<td>230.1</td>
<td>217.4</td>
</tr>
<tr>
<td>51</td>
<td>69.8</td>
<td>94.5</td>
<td>125.4</td>
<td>77</td>
<td>87.3</td>
<td>229.1</td>
<td>218.1</td>
</tr>
<tr>
<td>52</td>
<td>69</td>
<td>98.3</td>
<td>130.4</td>
<td>78</td>
<td>87.6</td>
<td>230.5</td>
<td>221.9</td>
</tr>
<tr>
<td>53</td>
<td>71.1</td>
<td>101.6</td>
<td>135.5</td>
<td>79</td>
<td>81.3</td>
<td>226</td>
<td>224.3</td>
</tr>
<tr>
<td>54</td>
<td>69.4</td>
<td>109.8</td>
<td>137.9</td>
<td>80</td>
<td>72.7</td>
<td>210.3</td>
<td>226.4</td>
</tr>
<tr>
<td>55</td>
<td>72.9</td>
<td>117</td>
<td>141.9</td>
<td>81</td>
<td>70.7</td>
<td>207</td>
<td>228.2</td>
</tr>
<tr>
<td>56</td>
<td>73.7</td>
<td>121.7</td>
<td>150.4</td>
<td>82</td>
<td>72.9</td>
<td>205</td>
<td>228.7</td>
</tr>
<tr>
<td>57</td>
<td>77.5</td>
<td>131.4</td>
<td>156.9</td>
<td>83</td>
<td>62.7</td>
<td>193.8</td>
<td>235.2</td>
</tr>
<tr>
<td>58</td>
<td>79.8</td>
<td>139.7</td>
<td>161</td>
<td>84</td>
<td>59.6</td>
<td>183.3</td>
<td>230.8</td>
</tr>
<tr>
<td>59</td>
<td>81.3</td>
<td>149.4</td>
<td>167.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>82.5</td>
<td>158.4</td>
<td>175.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>87.9</td>
<td>166.6</td>
<td>175.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>88.6</td>
<td>172.6</td>
<td>180.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
To estimate the effect of screening on the distribution of in situ, T1N0M0, and all other stages, we used published estimates of relative risks with screening, and, as with mortality, disaggregated based on estimates of prevalence of screening.

Estimates for the relative risk of a small tumor come from a recent systematic review: RR of having a tumor <2 cm with no nodes is 1.5 with screening, and for having nodes 1.25 with no screening, based on a recent systematic review.\(^5\)

Estimates for the relative risk of a diagnosis of DCIS come from age-specific data from the BCSC by screening status (Table 10).\(^6\) Crude age-specific relative risks were estimated by dividing the incidence in screened women by the incidence in screened women:

Table 10. Screen-Detected and Non-Screen-Detected DCIS among Women in the BCSC* 

<table>
<thead>
<tr>
<th>Age</th>
<th>DCIS Rate per 1000 Mammograms (95% CI)</th>
<th>RR (Calculated from Mean Incidence)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Screen-Detected</td>
<td>Non-Screen-Detected</td>
</tr>
<tr>
<td>40-49</td>
<td>0.56 (95% CI, 0.41 to 0.70)</td>
<td>0.08 (95% CI, 0.02 to 0.13)</td>
</tr>
<tr>
<td>50-59</td>
<td>0.68 (95% CI, 0.52 to 0.85)</td>
<td>0.09 (95% CI, 0.03 to 0.05)</td>
</tr>
<tr>
<td>60-69</td>
<td>1.03 (95% CI, 0.83 to 1.23)</td>
<td>0.19 (95% CI, 0.11 to 0.28)</td>
</tr>
<tr>
<td>70-84</td>
<td>1.07 (95% CI, 0.87 to 1.27)</td>
<td>0.22 (95% CI, 0.13 to 0.31)</td>
</tr>
</tbody>
</table>

*Adapted from Table 4 in Ernster et al\(^6\)

As a sensitivity analysis, we also used a relative risk of 3 based on recent analyses of the Norwegian screening program.\(^7,8\) Since DCIS rates are substantially lower in Norway than in the US,\(^9\) this seems like a reasonable lower bound.

To estimate possible rates of overdiagnosis, we applied a range of proportions of non-progression to the estimates of screen-detected DCIS and T1N0 invasive cancers.

**III. False positives**

We used published estimates of false positive rates by age, screening interval, and time since previous screen derived from the Breast Cancer Screening Consortium\(^10\) and CISNET\(^11\) (Table 11).

<table>
<thead>
<tr>
<th>Age</th>
<th>First Screen</th>
<th>Subsequent Screens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Lower 95% CI</td>
</tr>
<tr>
<td>40-44</td>
<td>2.0%</td>
<td>--</td>
</tr>
<tr>
<td>45-49</td>
<td>2.8%</td>
<td>2.5%</td>
</tr>
<tr>
<td>50-54</td>
<td>3.5%</td>
<td>3.1%</td>
</tr>
<tr>
<td>55-59</td>
<td>3.0%</td>
<td>2.5%</td>
</tr>
<tr>
<td>60-64</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>
Assumptions included:

- For false positive recall, we assumed that there was no age effect for subsequent screens, since confidence intervals for the adjusted odds ratios all included 1.0 in the Hubbard paper.
- For probabilistic analyses, we used the published odds ratios and 95% CIs for age to characterize a lognormal distribution. False positive probabilities for specific ages where relative risks were significantly increased were estimated by multiplying the estimate for women aged 40-44 by the value for the OR drawn from the distribution.
- For annual vs biennial screening, we used the published odds ratio (characterized as a lognormal distribution) to reduce the per-screen probability of either type of false positive.

We did not attempt to model the effect of variability in radiologists’ false positive rates, the effects of family history and breast density, or the availability of prior films on cumulative false positive rates.

We estimated lifetime risks of false positive biopsies using both a simple model assuming independence of risks, using the approach described by the UK Age trial investigators, which includes an assumption that the probability of a false positive at any given examination is independent of previous examinations (which the BCSC data clearly show is not the case and will overestimate). The cumulative probability, and calculate the cumulative risk over \( n \) screening examinations as:

\[
(1 - \text{Probability}_{\text{False Positive First Exam}}) \times (1 - \text{Probability}_{\text{False Positive Subsequent Exams}}) \times n^{-1}
\]

For this simple estimate, we also assumed that the probability of a false positive biopsy on subsequent exam is not related to age (which will underestimate the cumulative probability), although we do vary it based on screening interval as estimated in Hubbard et al. We also used these probabilities, with an age-specific component, in the Markov model described below, used primarily for estimating harm-benefit trade-offs.
IV. Estimating Cumulative Probabilities under Different Scenarios

We developed a simple semi-Markov state-transition model to estimate the probabilities of relevant outcomes under different scenarios of screening. States, transitions, transition probabilities, and how screening modifies the probabilities are shown in Table 12.
<table>
<thead>
<tr>
<th>STATE</th>
<th>ALLOWED TRANSITION</th>
<th>TRANSITION PROBABILITY</th>
<th>MODIFIED BY SCREENING</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Cancer</td>
<td>Cancer</td>
<td>Age-specific cancer incidence</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>DCIS</td>
<td>Age-specific DCIS probability</td>
<td>Yes, age-specific RR</td>
</tr>
<tr>
<td>False-positive</td>
<td></td>
<td>Age-specific probability from BCSC, modified by screen type (first vs subsequent), screening interval</td>
<td>Yes, only possible with screening</td>
</tr>
<tr>
<td>Death from Other Cause</td>
<td></td>
<td>Age-specific other cause mortality, derived by subtracting age-specific breast cancer mortality from age-specific all-cause mortality</td>
<td>No</td>
</tr>
<tr>
<td>DCIS</td>
<td>Death from Other Cause</td>
<td>Age-specific conditional other cause mortality for years 1-15 after diagnosis, derived from SEER</td>
<td>No</td>
</tr>
<tr>
<td>Invasive Cancer</td>
<td>Death from Cancer</td>
<td>Age-specific conditional survival for years 1-15 after diagnosis, from SEER (see Table 10)</td>
<td>Yes, hazard ratio resulting in relative 15 year mortality reduction attributable to screening applied to yearly conditional probability of cancer specific death</td>
</tr>
<tr>
<td></td>
<td>Death from Other Cause</td>
<td>Age-specific conditional other cause mortality for years 1-15 after diagnosis, derived from SEER</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Long-term Survivor</td>
<td>100% after 15 years of follow-up</td>
<td>No</td>
</tr>
<tr>
<td>Long-Term Survivor</td>
<td>Death from Other Cause</td>
<td>Age-specific other cause mortality, derived by subtracting age-specific breast cancer mortality from age-specific all-cause mortality</td>
<td>No</td>
</tr>
</tbody>
</table>
Briefly, the model works as follows:

- All women start at age 40 in the No Cancer State. During the first year long cycle, they are at risk of having a false positive result (leading to either a repeat examination or biopsy, modeled using two separate sets of probabilities), a noncancer death, or having an incident case of invasive cancer or DCIS diagnosed. The probability of DCIS is conditioned on whether screening has occurred, the probability of invasive cancer is not. This likely results in an underestimate of cancer incidence among screened women early in during the screening period, and an overestimate later.

- Women who are diagnosed with invasive cancer are then subject to two possible causes of death, either breast-cancer specific or other cause. The conditional probability of dying of breast cancer or another cause during a given year post-diagnosis having survived up to that point in time is obtained directly from SEER (Table 13), and is stratified by age at diagnosis. In essence, as the simulation progresses, the effect of age-specific incidence and post-diagnosis survival conditioned on age at diagnosis result in incidence-based mortality.
Table 13: Illustrative Age-specific Post-Diagnosis Conditional Probabilities of Breast Cancer Death and Other Cause Death

<table>
<thead>
<tr>
<th>Years Post-Diagnosis</th>
<th>AGE AT DIAGNOSIS</th>
<th>40</th>
<th>50</th>
<th>55</th>
<th>60</th>
<th>65</th>
<th>70</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1.5%</td>
<td>0.1%</td>
<td>1.6%</td>
<td>0.3%</td>
<td>2.2%</td>
<td>0.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>3.0%</td>
<td>0.2%</td>
<td>2.6%</td>
<td>0.3%</td>
<td>2.7%</td>
<td>0.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>3.3%</td>
<td>0.2%</td>
<td>2.5%</td>
<td>0.4%</td>
<td>2.8%</td>
<td>0.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>2.8%</td>
<td>0.2%</td>
<td>2.3%</td>
<td>0.4%</td>
<td>2.4%</td>
<td>0.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>3.0%</td>
<td>0.2%</td>
<td>1.9%</td>
<td>0.4%</td>
<td>1.6%</td>
<td>0.7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>2.2%</td>
<td>0.2%</td>
<td>1.6%</td>
<td>0.5%</td>
<td>1.7%</td>
<td>0.7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>2.4%</td>
<td>0.2%</td>
<td>1.1%</td>
<td>0.5%</td>
<td>1.5%</td>
<td>0.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>2.1%</td>
<td>0.2%</td>
<td>1.4%</td>
<td>0.5%</td>
<td>1.2%</td>
<td>0.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>1.7%</td>
<td>0.3%</td>
<td>1.1%</td>
<td>0.6%</td>
<td>1.1%</td>
<td>0.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>1.5%</td>
<td>0.3%</td>
<td>1.1%</td>
<td>0.6%</td>
<td>1.4%</td>
<td>1.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>1.6%</td>
<td>0.3%</td>
<td>0.9%</td>
<td>0.7%</td>
<td>1.1%</td>
<td>1.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>1.2%</td>
<td>0.3%</td>
<td>1.1%</td>
<td>0.8%</td>
<td>0.8%</td>
<td>1.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>1.4%</td>
<td>0.4%</td>
<td>1.1%</td>
<td>0.8%</td>
<td>0.5%</td>
<td>1.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>1.3%</td>
<td>0.4%</td>
<td>1.2%</td>
<td>0.9%</td>
<td>0.3%</td>
<td>1.4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>0.8%</td>
<td>0.4%</td>
<td>0.5%</td>
<td>1.0%</td>
<td>1.0%</td>
<td>1.5%</td>
</tr>
</tbody>
</table>
• The post-diagnosis survival probabilities for all patients with cancer represent the weighted average of the survival probabilities across all stages. The stage shift resulting from screening results in a greater proportion of women with higher survival, which, after sufficient follow-up, results in decreased mortality. One way to model these effects is to use an underlying model of the natural history of breast cancer, with stage distribution without screening being a function of disease progression and the probability of developing symptoms and having a detected case at a given stage, and distribution with screening a function of disease progression, test sensitivity, and interval—this is the approach used by the CISNET group. Alternatively, one could model the effect of screening on stage distribution, and generate age- and stage-specific survival curves. A third approach is to use estimates of overall mortality reduction and impute a screen-attributable hazard ratio for all cancers; we elected to use this approach to make it easier to use estimates of overall mortality generated by randomized trials and observational studies to U.S based populations.

• The hazard ratios were applied to all incident cancers detected through screening for 15 years; because of the lack of data on longer follow-up, we assumed women were no longer at risk for cancer death beyond this point. This may underestimate true mortality. Because the reduction in annual mortality probability was applied throughout the entire 15 year period, this means that women with cancers detected by screening late in the screening ages retained benefits after overall screening stopped—for example, a woman with cancer detected by screening at age 70 would still benefit from a reduced risk of breast cancer death through age 84, even if screening stopped after age 74.

Key assumptions included

• False positives are only attributable to screening. In the absence of mammographic screening, women can undergo breast biopsy if they develop symptoms and have a mass detected, or if they have an asymptomatic mass detected on clinical breast examination. In the first case, a false positive breast biopsy in the presence of symptoms would, by definition, be from a benign condition, and there is no reason to think that mammographic screening would make women more or less likely to develop benign breast disease. In the second case, it is true that women not undergoing screening might undergo clinical breast examination and have a false positive biopsy, but it is unclear how this might substantially affect the incremental false positive biopsy attributable to mammography. First, it seems unlikely that there is a large pool of women undergoing regular clinical breast examination for screening who are not also getting mammography, particularly within the context of the BCSC. Second, if enough women are undergoing clinical breast examination in the absence of mammography to substantially affect false positive rates, this may affect the applicability of estimates of mortality reduction based on screening versus unscreened.

• For the bulk of the analyses, we allowed only one false positive per patient in the microsimulation. This resulted in estimates of the probability of “at least one” false positive, rather than total false positives across a population. This had the largest effect on total false positives—without restriction, population estimates were always greater than 100%.
Women diagnosed with DCIS were not at risk for breast cancer death. Although there is a small risk of breast cancer mortality among women diagnosed with DCIS, we assumed women with DCIS were not at risk for subsequent breast cancer death (or incident invasive cancer) for simplicity. Since the model results in much higher incidences of DCIS among screened women than unscreened women, this has the effect of reducing the pool of women at risk for having invasive cancer diagnosed, and of reducing breast cancer mortality. Any resulting bias is in favor of screening.

The probability that a case of DCIS was “overdiagnosed” was estimated using progression probabilities of 20%, 50%, and 80% to accommodate the wide range in the literature. This rate was applied only to screen-detected cases of DCIS—since non-screen detected cases were presumably detected through the presence of symptoms, by definition they cannot be “overdiagnosed”.

The model was run for a cohort of 40 year old women through age 100 as a Monte Carlo microsimulation. For key parameters including relative risk of mortality with screening and false positive probability, the value for each parameter was drawn from a probability distribution. For the harm/benefit trade-off analyses, we sampled each parameter 500 times, and performed 20,000 simulations.

V. Harm/Benefit Acceptability

Benefits and harms frequently do not share common metrics, an issue common to many medical and public health decision-making problems beyond mammography. We have been working on adapting methods using in health economic evaluations, specifically value-of-information (VOI) analysis, to help decision makers view the joint effects of uncertainty about the likelihood of different harms and benefits, and uncertainty about the “appropriate” balance needed to justify a particular recommendation for or against a given recommendation. The initial inspiration of our group for exploring the possibility of adapting value-of-information methods as an aid for visualizing uncertainty about harm/benefit trade-offs came from our experience during the 2012 revision of ACS’ cervical cancer guidelines, which used a GRADE framework. As part of those guidelines, the panelists had agreed to using colposcopies per CIN3+ detected as the primary measure of harm/benefit, with estimates of the impact of different strategies derived either directly from the literature or from modeling. Both during the background work of the panel on developing specific recommendations, and in the large stakeholder conference, there was considerable discussion of how to weight these two surrogate measures, with an explicit recommendation that different patients, and other key stakeholders, would place different values on each outcome (as well as the more direct outcomes for which they served as surrogates, such as preterm birth from unnecessary treatment, or prevented morbidity or mortality from cervical cancer by treatment of true preinvasive diseases). During the stakeholder conference, which was attended by representatives of the US Preventive Services Task Force (USPSTF), it became clear that some of the differences between the draft recommendations of the USPSTF and those of the ACS panel reflected different implicit weightings for harms. During this same time period, the Duke Evidence-based Practice Center was working on projects for AHRQ and the Patient-Centered Outcomes Research Institute (PCORI) on the potential use of VOI for research prioritization. One of the specific issues for PCORI was whether VOI had a role within their research agenda and methodology standards, given the statutory limitations on their use of cost-effectiveness analysis. A common theme in both the cervical cancer guidelines work and the
VOI projects was the difficulty of assessing uncertainty about harm/benefit trade-offs when common metrics do not exist, and when different stakeholders place different weights on specific harms and benefits. We have since explored the technical aspects of our proposed approach in an AHRQ-CDC funded project on the use of oral contraceptives for primary prevention of ovarian cancer and believe that VOI can explicitly help groups understand how uncertainty about evidence and the choice of, and values placed on, specific harms and benefits affect the strength of recommendations.

Rationale and approach: The underlying rationale for this approach is that, by definition, guidelines are meant to help with decisions by patients, clinicians, and other stakeholders, and that GRADE explicitly rates the overall quality of evidence in terms of confidence that the evidence reflects the true effect (Table 12). Note that the previous GRADE definition explicitly framed the level of confidence in terms of the value of future research—this definition helped inspire our group’s interest in the potential applicability of VOI, which was explicitly developed as a tool for estimating the value of future research to guideline development using GRADE.

<table>
<thead>
<tr>
<th>Quality Level</th>
<th>Current Definition</th>
<th>Previous Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>We are very confident that the true effect lies close to that of the estimate of the effect</td>
<td>Further research is very unlikely to change our confidence in the estimate of effect</td>
</tr>
<tr>
<td>Moderate</td>
<td>We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different</td>
<td>Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate</td>
</tr>
<tr>
<td>Low</td>
<td>Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect</td>
<td>Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate</td>
</tr>
<tr>
<td>Very low</td>
<td>We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect</td>
<td>Any estimate of effect is very uncertain</td>
</tr>
</tbody>
</table>

Essentially, VOI addresses two simultaneous questions: (1) “What are the chances of making a ‘wrong’ decision with the available evidence?” and (2) “Do the consequences of making a wrong decision justify collecting further evidence?” This has typically been done in the context of traditional health economic analysis, where the optimal decision is based on cost-effectiveness expressed as monetary units per quality-adjusted life year (QALY) gained. Using probabilistic (stochastic) decision modeling, where multiple simulations are performed drawing from statistical distributions, the probability that a given option will be optimal at a given willingness-to-pay (WTP) threshold is estimated across a range of WTP thresholds (typically from $0 to $100,000 per QALY). Both to facilitate calculations, and to avoid some of the issues involved with estimating incremental cost-effectiveness ratios when there are multiple potential options, net monetary benefits (NMB), defined as:

\[ NMB = Effectiveness \times WTP - Costs \]
are used as the primary measure, with the optimal choice being the one with the highest net benefit at a given WTP threshold. This is depicted graphically on an acceptability curve, where the Y-axis represents the proportion of simulations that a given option is optimal, and the X-axis the WTP threshold.

The certainty that a given option is optimal may vary based on WTP. At a WTP of 0, the option with the lowest cost will have the highest NMB, while at high levels of WTP, the option with the highest effectiveness will be favored. The proportion of simulations where a given option is optimal reflects the certainty in the evidence, particularly with respect to the precision of estimates (i.e., if confidence intervals are wide and overlapping, no single option is likely to be optimal more than 50-60% of the time, meaning that choosing that option based on the evidence carries a 40-50% chance of being the “wrong” decision).

The potential utility of this approach for guideline development using GRADE, or another formal process that links evidence quality to the balance harms and benefits, is that, conceptually, the balance between harms and benefits is equivalent to “willingness-to-pay.” The “costs” of harms can be varied, either formally (using methods for eliciting preferences) or informally, and the effect of excluding specific harms and benefits from the equation and the WTP threshold on the certainty about the optimal decision can be readily visualized using a net benefits approach:

\[
Net \text{ Benefits} = \text{Benefits} \times "WTP" - \text{Harms}
\]

Figure 1 depicts a generic example (see Figures at the end of this document). Panel members need to reach consensus on (a) what level of certainty (95%, 90%, 85%, etc.) regarding the balance of benefits and harms would lead to a strong recommendation for or against a particular intervention, (b) what level of certainty (50%, 60%, 70%) about the balance of benefits and harms would lead to a weak recommendation for or against an intervention (Figure 1a), and (c) what is an appropriate upper limit for the ratio of harm to benefit. This upper limit obviously depends on the relative weights assigned by panelists (and patients) to the different outcomes, but this discussion must take place, regardless of the method used to present the uncertainty. The probability that a given option is optimal at any given harm/benefit ratio can be displayed graphically (Figure 1b). By adding in lines indicating different upper limits of an “acceptable” ratio, the impact of choice of threshold on the strength of recommendation can be readily shown on the same graph (Figure 1c).
Figures

Figure 1: Harm/benefit acceptability and GRADE

a. Panel reaches consensus on the approximate levels of certainty required for strong and weak recommendations for or against an option, and preliminary consensus on thresholds (or a range of thresholds) for acceptable ratio of harm to benefit.

b. Using a decision model (depending on the questions, this can be a single model, or multiple models), conduct a probabilistic analysis based on the available evidence and show results on acceptability curve.

c. Illustrate how changing WTP threshold, or inclusion of different harms and benefits, might change certainty about evidence and thus strength of recommendation (Threshold X = strong recommendation for Option B, Threshold Y = weak recommendation for Option A).
Figure 1: Harm/benefit acceptability and GRADE

a. Panel reaches consensus on the approximate levels of certainty required for strong and weak recommendations for or against an option, and preliminary consensus on thresholds (or a range of thresholds) for acceptable ratio of harm to benefit

b. Using a decision model (depending on the questions, this can be a single model, or multiple models), conduct a probabilistic analysis based on the available evidence and show results on acceptability curve
c. Illustrate how changing WTP threshold, or inclusion of different harms and benefits, might change certainty about evidence and thus strength of recommendation (Threshold X = strong recommendation for Option B, Threshold Y = weak recommendation for Option A)
References to Appendix C:


Appendix D. List of Included Studies by Key Question

Key Question 1


**Key Question 2**


Key Question 3

Key Question 4


**Key Question 5**


**Background Articles**


113. Howard M, Agarwal G and Lytwyn A. Accuracy of self-reports of Pap and


Modeling Articles


39. Fett MJ. Computer modelling of the Swedish two county trial of mammographic


**Systematic Reviews/Meta-Analyses**


D-34


Appendix E. List of Excluded Studies

All studies listed below were reviewed in their full-text version and excluded for the reasons cited. Reasons for exclusion signify only the usefulness of the articles for this study and are not intended as criticisms of the articles.

Not available in English


Abstract only or not original peer-reviewed data


Dean PB. Final comment. The articles by Gøtzsche and Olsen are not official Cochrane reviews and lack scientific merit. Lakartidningen. 2000;97(25):3106. PMID: 10911710.


Hense HW. The trade-off between population and individual benefit of screening. Z Arztl Fortbild Qualitatssich. 2006;100(7):505-13. PMID: 17137063.


Kane KY, Lindbloom EJ and Stevermer JJ. Does mammography add any benefit to a thorough clinical breast examination (CBE)? J Fam Pract. 2000;49(12):1078. PMID: 11132055.


Mayor S. Row over breast cancer screening shows that scientists bring “some subjectivity into their work. BMJ. 2001;323(7319):956. PMID: 11679382.


Yaffe MJ, Barnes GT and Orton CG. Point/Counterpoint. Film mammography for breast cancer screening in younger women is no longer appropriate because of the demonstrated superiority of digital mammography for this age group. Med Phys. 2006;33(11):3979-82. PMID: 17153375.


No population of interest


E-17


E-20


E-23


No screening modality of interest


No outcomes of interest


Blanks RG, Moss SM and Wallis MG. Monitoring and evaluating the UK National Health Service Breast Screening Programme: evaluating the variation in radiological performance


**Observational study with <100 patients (High Risk); <1,000 patients (Average Risk)**


No direct or indirect comparison of outcomes


Anonymous. Mammographic surveillance in women younger than 50 years who have a family history of breast cancer: tumour characteristics and projected effect on mortality in the prospective, single-arm, FH01 study. Lancet Oncol. 2010;11(12):1127-34. PMID: 21093374.


Evans DG, Lenard F, Pointon LJ, et al. Eligibility for magnetic resonance imaging screening in the United Kingdom: effect of strict selection criteria and anonymous DNA testing on breast


Leung JW and Sickles EA. Developing asymmetry identified on mammography: correlation with imaging outcome and pathologic findings. AJR Am J Roentgenol. 2007;188(3):667-75. PMID: 17312052.


Scaranello A. Breast screening with magnetic resonance imaging. CMAJ. 2012;184(16):E877. PMID: 22802387.


## Appendix F. Key to Included Primary and Companion Articles

<table>
<thead>
<tr>
<th>Study Designation</th>
<th>Primary Abstracted Article</th>
<th>Companion Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canadian National Breast Screening Study-1 aged 40–49 (CNBSS-1)</td>
<td>Miller, 2014&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Miller, 2000&lt;sup&gt;6&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Goel, 1998&lt;sup&gt;9&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Miller, 1997&lt;sup&gt;4&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Miller, 1992&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Miller, 1992&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Baines, 1986&lt;sup&gt;7&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Miller, 2002&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Elmore, 2005&lt;sup&gt;9&lt;/sup&gt;</td>
<td>Fenton, 2005&lt;sup&gt;9&lt;/sup&gt;</td>
</tr>
<tr>
<td>Cancer Research Network</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edinburgh Trial</td>
<td>Alexander, 1999&lt;sup&gt;11&lt;/sup&gt;</td>
<td>Anonymous, 1999&lt;sup&gt;16&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alexander, 1997&lt;sup&gt;14&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Anonymous, 1992&lt;sup&gt;14&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Roberts, 1990&lt;sup&gt;15&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Anonymous, 1988&lt;sup&gt;16&lt;/sup&gt;</td>
</tr>
<tr>
<td>Gothenburg Breast Cancer Screening Trial</td>
<td>Bjurstam, 2003&lt;sup&gt;13&lt;/sup&gt;</td>
<td>Bjurstam, 1997&lt;sup&gt;19&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bjurstam, 1997&lt;sup&gt;19&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Larsson, 1997&lt;sup&gt;20&lt;/sup&gt;</td>
</tr>
<tr>
<td>Health Insurance Plan (HIP) of Greater New York Study</td>
<td>Shapiro, 1997&lt;sup&gt;17&lt;/sup&gt;</td>
<td>Chu, 1988&lt;sup&gt;22&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shapiro, 1982&lt;sup&gt;23&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shapiro, 1977&lt;sup&gt;24&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strax, 1967&lt;sup&gt;25&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Habbema, 1986&lt;sup&gt;26&lt;/sup&gt;</td>
</tr>
<tr>
<td>Malmö Mammographic Screening Trial (MMST1)</td>
<td>Andersson, 1997&lt;sup&gt;15&lt;/sup&gt;</td>
<td>Zackrisson, 2006&lt;sup&gt;20&lt;/sup&gt;</td>
</tr>
<tr>
<td>Malmö II Trial (MMST2)</td>
<td></td>
<td>Andersson, 1988&lt;sup&gt;9&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Andersson, 1987&lt;sup&gt;20&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Andersson, 1985&lt;sup&gt;14&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Andersson, 1984&lt;sup&gt;12&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Andersson, 1979&lt;sup&gt;13&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Larsson, 1997&lt;sup&gt;20&lt;/sup&gt;</td>
</tr>
<tr>
<td>Mass Screening Registry of Finland</td>
<td>Sarkeala, 2008&lt;sup&gt;27&lt;/sup&gt;</td>
<td>Sarkeala, 2005&lt;sup&gt;28&lt;/sup&gt;</td>
</tr>
<tr>
<td>Norwegian Breast Cancer Screening Program</td>
<td>Kalager, 2010&lt;sup&gt;38&lt;/sup&gt;</td>
<td>Falk, 2013&lt;sup&gt;37&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kalager, 2012&lt;sup&gt;39&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hofvind, 2007&lt;sup&gt;34&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Olsen, 2013&lt;sup&gt;40&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Roman, 2013&lt;sup&gt;41&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hofvind, 2013&lt;sup&gt;42&lt;/sup&gt;</td>
</tr>
<tr>
<td>Screening Mammography Programme of British Columbia</td>
<td>Coldman, 2008&lt;sup&gt;43&lt;/sup&gt;</td>
<td>Phillips, 2008&lt;sup&gt;44&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Coldman, 2007&lt;sup&gt;45&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wai, 2005&lt;sup&gt;46&lt;/sup&gt;</td>
</tr>
<tr>
<td>Stockholm Mammographic Screening Trial</td>
<td>Frisell, 1997&lt;sup&gt;30&lt;/sup&gt;</td>
<td>Frisell, 1997&lt;sup&gt;30&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Frisell, 1991&lt;sup&gt;49&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Frisell, 1989&lt;sup&gt;50&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Larsson, 1997&lt;sup&gt;20&lt;/sup&gt;</td>
</tr>
<tr>
<td>Study Designation</td>
<td>Primary Abstracted Article</td>
<td>Companion Articles</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>----------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>UK Age Trial</td>
<td>Johns, 2010</td>
<td>Moss, 2006&lt;sup&gt;69&lt;/sup&gt; Moss, 2005&lt;sup&gt;70&lt;/sup&gt; Moss, 2005&lt;sup&gt;71&lt;/sup&gt; Anderson, 2004&lt;sup&gt;72&lt;/sup&gt;</td>
</tr>
<tr>
<td>Women's CARE (Contraceptive and Reproductive Experiences) Study</td>
<td>Norman, 2007</td>
<td>Norman, 2006&lt;sup&gt;69&lt;/sup&gt;</td>
</tr>
<tr>
<td>None</td>
<td>Anonymous, 2006&lt;sup&gt;73&lt;/sup&gt; Anonymous, 2006&lt;sup&gt;74&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Hellquist, 2011&lt;sup&gt;75&lt;/sup&gt; Hellquist, 2012&lt;sup&gt;76&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>van Schoor, 2011&lt;sup&gt;77&lt;/sup&gt; Verbeek, 1984&lt;sup&gt;78&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Abuidris, 2013&lt;sup&gt;79&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Allgood, 2008&lt;sup&gt;80&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Barton, 2004&lt;sup&gt;81&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Blanchard, 2006&lt;sup&gt;82&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Bleyer, 2012&lt;sup&gt;83&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Braithwaite, 2013&lt;sup&gt;84&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Broeders, 2002&lt;sup&gt;85&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Chiarelli, 2009&lt;sup&gt;86&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Catto, 2013&lt;sup&gt;87&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Coldman, 2013&lt;sup&gt;88&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>De Gelder, 2011&lt;sup&gt;89&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Dittus, 2013&lt;sup&gt;90&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Domingo, 2013&lt;sup&gt;91&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Duffy, 2002&lt;sup&gt;92&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Duffy, 2010&lt;sup&gt;93&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Evans, 2014&lt;sup&gt;94&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Fielder, 2004&lt;sup&gt;95&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Gabe, 2007&lt;sup&gt;96&lt;/sup&gt; None</td>
<td></td>
</tr>
<tr>
<td>Study Designation</td>
<td>Primary Abstracted Article</td>
<td>Companion Articles</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>None</td>
<td>Haas, 2013&lt;sup&gt;1&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Hakama, 1997&lt;sup&gt;2&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Hofvind, 2012&lt;sup&gt;3&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Honjo, 2007&lt;sup&gt;4&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Hubbard, 2011&lt;sup&gt;5&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Jonsson, 2000&lt;sup&gt;6&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Jonsson, 2005&lt;sup&gt;7&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Jonsson, 2003&lt;sup&gt;8&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Jonsson, 2003&lt;sup&gt;9&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Jonsson, 2007&lt;sup&gt;10&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Jorgensen, 2009&lt;sup&gt;11&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Kerlikowske, 2013&lt;sup&gt;12&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Kikuchi, 2014&lt;sup&gt;13&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>King, 2013&lt;sup&gt;14&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Krieger, 2004&lt;sup&gt;15&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Lund, 2013&lt;sup&gt;16&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Maurice, 2006&lt;sup&gt;17&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Molins, 2009&lt;sup&gt;18&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Moody-Ayers, 2000&lt;sup&gt;19&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Morrell, 2010&lt;sup&gt;20&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Ng, 2013&lt;sup&gt;21&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Nickson, 2012&lt;sup&gt;22&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Njer, 2013&lt;sup&gt;23&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Oestreicher, 2005&lt;sup&gt;24&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Ohlinger, 2006&lt;sup&gt;25&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Olsen, 2006&lt;sup&gt;26&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Olsen, 2005&lt;sup&gt;27&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>O’Meara, 2013&lt;sup&gt;28&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Otten, 2013&lt;sup&gt;29&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Otto, 2012&lt;sup&gt;30&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Paap, 2010&lt;sup&gt;31&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Paci, 2002&lt;sup&gt;32&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Paci, 2004&lt;sup&gt;33&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Paci, 2006&lt;sup&gt;34&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Paci, 2008&lt;sup&gt;35&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Parvinen, 2006&lt;sup&gt;36&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Parvinen, 2011&lt;sup&gt;37&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Port, 2007&lt;sup&gt;38&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Puliti, 2008&lt;sup&gt;39&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Puliti, 2009&lt;sup&gt;40&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Puliti, 2012&lt;sup&gt;41&lt;/sup&gt;</td>
<td>None</td>
</tr>
<tr>
<td>Study Designation</td>
<td>Primary Abstracted Article</td>
<td>Companion Articles</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>None</td>
<td>Randall, 2009*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Roder, 2008*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Sankaranarayanan, 2011*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Schonberg, 2009*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Skaane, 2013*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Skaane, 2013*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Sung, 2011*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Tabar, 2001*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Tohno, 2013*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>van Schoor, 2010*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Vutuc, 2006*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Walker, 2013*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Warner, 2011*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Weedon-F, 2014*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Yankaskas, 2005*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Yu, 2008*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Zahl, 2004*</td>
<td>None</td>
</tr>
<tr>
<td>None</td>
<td>Zahl, 2011*</td>
<td>None</td>
</tr>
</tbody>
</table>

Reference to Appendix F:


# Appendix G. Study Characteristics Tables

## Appendix Table G1. Study Characteristics—Key Question 1

<table>
<thead>
<tr>
<th>Outcome Study</th>
<th>Study Design Country</th>
<th>Total N Age Range Risk Population</th>
<th>Intervention Interval</th>
<th>Comparator Interval</th>
<th>Risk of bias</th>
<th>Overall Quality</th>
<th>Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Criterion</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>H</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>Breast Cancer Mortality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RCT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miller, 2014* Canada</td>
<td>RCT</td>
<td>Canada</td>
<td>1980-2005</td>
<td>89,835 40-59 NR</td>
<td>Mammography + CBE</td>
<td>1 yr</td>
<td>CBE NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yen, 2012* Swedish Two-County</td>
<td>RCT</td>
<td>Sweden</td>
<td>1977-2005</td>
<td>134,867 40-74 2.4% High Risk</td>
<td>Mammography 24 mo (40-49) 33 mo (50+)</td>
<td>No screening/ Mammography (offered after 6-8 yr)</td>
<td>Selection X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Johns, 2010* UK Age</td>
<td>RCT</td>
<td>UK</td>
<td>1991-2004</td>
<td>160,921 NR NR</td>
<td>Mammography 1 yr</td>
<td>Mammography 1 yr (1 false positive; &gt;1 false positive)</td>
<td>Selection X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design</td>
<td>Total N</td>
<td>Age Range</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>---------</td>
<td>------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>--------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Bjurstam, 2003</td>
<td>RCT Goteborg</td>
<td>51,611</td>
<td>39-59 NR</td>
<td>Mammography 18 mo</td>
<td>No screening</td>
<td>High</td>
<td>RR 0.77 (0.60-1.00)</td>
</tr>
<tr>
<td></td>
<td>RCT Goteborg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>39-44 yo 0.70 (0.39 to 1.28)</td>
</tr>
<tr>
<td></td>
<td>RCT Goteborg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>45-49 yo 0.67 (0.39 to 1.23)</td>
</tr>
<tr>
<td></td>
<td>RCT Goteborg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50-54 yo 1.06 (0.66 to 1.72)</td>
</tr>
<tr>
<td></td>
<td>RCT Goteborg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>55-59 yo 0.67 (0.66 to 1.72)</td>
</tr>
</tbody>
</table>

<p>| Criterion     | H | L |          |                |                   |
|---------------|---|---|----------|-----------------|
| Selection     |   | X | Selection X |
| Detection     |   | X | Detection X |
| Performance   |   | X | Performance X |
| Attrition     |   | X | Attrition X |
| Reporting     |   | X | Reporting X |</p>
<table>
<thead>
<tr>
<th>Outcome Study</th>
<th>Study Design</th>
<th>Total N</th>
<th>Age Range</th>
<th>Intervention Interval</th>
<th>Comparator Interval</th>
<th>Risk of bias</th>
<th>Overall Quality</th>
<th>Effect (95% CI)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Relative</td>
<td>Absolute</td>
</tr>
<tr>
<td>Alexander, 1999</td>
<td>RCT</td>
<td>54,654</td>
<td>NR</td>
<td>Mammography + CBE (45-49)</td>
<td>Mammography + CBE (55-59)</td>
<td>Selection X</td>
<td>Low</td>
<td>Unadjusted RR 0.87 (95% CI 0.70-1.06)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edinburgh</td>
<td>UK</td>
<td></td>
<td>NR</td>
<td>2 yr</td>
<td>2 yr</td>
<td>Detection X</td>
<td></td>
<td>SES adjusted RR 0.79 (0.60-1.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NR</td>
<td>Mammography + CBE (50-54)</td>
<td>Mammography + CBE (60-64)</td>
<td>Performance X</td>
<td></td>
<td>Late deaths censored RR 0.71 (0.53-0.95)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2 yr</td>
<td></td>
<td>Attrition X</td>
<td></td>
<td>By age (SES adjusted) RR 45-49 yo 0.70 (0.41-1.20)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td>50-54 yo RR 0.99 (0.62-1.58)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>55-59 yo RR 0.65 (0.43-0.99)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>60-64 yo RR 0.80 (0.51-1.25)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andersson, 1997</td>
<td>RCT</td>
<td>25,770</td>
<td>44-50</td>
<td>Mammography 18-24 mo</td>
<td>No screening</td>
<td>Selection X</td>
<td>High</td>
<td>RR 0.64 (95% CI 0.45-0.89)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malmo</td>
<td>Sweden</td>
<td></td>
<td>100% Average Risk</td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----------------------------</td>
<td>-----------------------------------</td>
<td>------------------------</td>
<td>---------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>---------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frisell, 1997</td>
<td>Stockholm: RCT Sweden 1981-1995</td>
<td>60,261 40-64 100% Average Risk</td>
<td>Mammography 2 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>High</td>
<td>RR 0.74 (95% CI 0.5 to 1.10)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shapiro, 1997</td>
<td>HIP: RCT US 1963-1996</td>
<td>~62,000 40-64 NR</td>
<td>Mammography + CBE 1 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>RR 0.77 (95% CI not provided and not calculable from data in paper)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observational</td>
<td>Weedon-Fekjaer, 2014</td>
<td>NR 50-79 NR</td>
<td>Mammography 2yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>RR 0.72 (95% CI 0.64 to 0.79)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nickson, 2012</td>
<td>Case-Control Australia 1995-2006</td>
<td>4077 50-69 100% Average Risk</td>
<td>Mammography (Breast cancer deaths) 2 yr</td>
<td>Mammography (Controls) 2 yr</td>
<td>Selection X</td>
<td>Moderate</td>
<td>All 0.48 (0.38-0.59)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td>50-59 yo 0.52 (0.37-0.72)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td>60-69 yo 0.44 (0.33-0.59)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design</td>
<td>Country</td>
<td>Years</td>
<td>Total N</td>
<td>Age Range</td>
<td>Risk Population</td>
<td>Intervention</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>---------</td>
<td>-------</td>
<td>---------</td>
<td>------------</td>
<td>-----------------</td>
<td>--------------</td>
<td>----------------------</td>
<td>--------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Otto, 2012</td>
<td>Case-Control</td>
<td>Netherlands</td>
<td>1990-2003</td>
<td>4494</td>
<td>50-75</td>
<td>NR</td>
<td>Mammography NR</td>
<td>No screening</td>
<td>Selection</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Puliti, 2012</td>
<td>Retrospective cohort</td>
<td>Italy</td>
<td>1991-2008</td>
<td>51,096</td>
<td>50-69</td>
<td>100% Average Risk</td>
<td>Mammography</td>
<td>2 yr</td>
<td>No screening</td>
<td>Selection</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Hellquist, 2011</td>
<td>Retrospective cohort</td>
<td>Sweden</td>
<td>1986-2005</td>
<td>7,261,415</td>
<td>40-49</td>
<td>NR</td>
<td>Mammography</td>
<td>18-24 mo</td>
<td>No screening</td>
<td>Selection</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>van Schoor, 2011</td>
<td>Case-Control</td>
<td>Netherlands</td>
<td>1975-2008</td>
<td>1410</td>
<td>50-69</td>
<td>NR</td>
<td>Mammography</td>
<td>2 yr</td>
<td>No screening</td>
<td>Selection</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------</td>
<td>----------------------------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duffy, 2010</td>
<td>Retrospective cohort UK, Sweden 1977-2004</td>
<td>NR 50-70</td>
<td>Mammography 24-33 mo</td>
<td>Mammography 3 yr</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Sweden RR 0.62 (95% CI 0.51 to 0.75)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td>No screening</td>
<td>No screening</td>
<td>Detection X</td>
<td></td>
<td>UK RR 0.72 (0.70 to 0.74)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kalager, 2010</td>
<td>Prospective cohort Norway 1996-2005</td>
<td>462,306 50-69</td>
<td>Mammography 2 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>RR 0.9 2.4/100,000 person-years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>100% Average Risk</td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paap, 2010</td>
<td>Case-referent Netherlands 1989-2006</td>
<td>118 cases, 118 referents 50-75</td>
<td>Mammography 2 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>OR 0.24 (95% CI 0.10 to 0.58)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td>(adjusted for self-selection bias; unadjusted 0.30 (0.14 to 0.63)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>van Schoor, 2010</td>
<td>Case-Control Netherlands 1975-1990</td>
<td>1632 51% &lt;50; 35% 50-74</td>
<td>Mammography 2 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>40-49 yo OR 0.50 (95% CI 0.3 to 0.82)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NR</td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td>50-59 yo 0.54 (0.35 to 0.85)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td>60-69 yo 0.65 (0.38 to 1.13)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schonberg, 2009</td>
<td>Retrospective cohort US 1994-2006</td>
<td>2011 &gt;80 100% Average Risk</td>
<td>Mammography NR</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Screened: 0.10%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td>Unscreened: 0.20%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td>p = 0.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design</td>
<td>Country</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
<td>Relative</td>
<td>Absolute</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>---------</td>
<td>----------------------------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>Allgood, 2008**</td>
<td>Case-control</td>
<td>UK</td>
<td>852 (284 cases and 568 controls) 50-70 100% Average Risk</td>
<td>invited to attend breast screening once in every 3 yr</td>
<td>No screening</td>
<td>Selection X, Detection X, Performance X, Attrition X, Reporting X</td>
<td>Moderate; adjusted for self-selection bias but unclear if all residual confounding accounted for</td>
<td>OR 0.52 (95% CI 0.32 to 0.84)</td>
<td>Not calculable for case-control</td>
<td></td>
</tr>
<tr>
<td>Coldman, 2008**</td>
<td>Prospective cohort</td>
<td>Canada</td>
<td>658,151 40-79 NR</td>
<td>Mammography 13-14 mo</td>
<td>Mammography 18-29 mo</td>
<td>Mammography &gt;30 mo</td>
<td>Selection X, Detection X, Performance X, Attrition X, Reporting X</td>
<td>Low</td>
<td>All ages RR 0.60 (95% CI 0.55-0.65) 40-49 yo RR 0.61 (0.52-0.71) 50-59 yo RR 0.59 (0.50-0.69) 60-69 yo RR 0.60 (0.52-0.70)</td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------</td>
<td>---------------------------------</td>
<td>----------------------</td>
<td>-------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Criterion</td>
<td></td>
<td>Relative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>H</td>
<td>L</td>
<td>Absolute</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paci, 2008**</td>
<td>Retrospective cohort Italy NR-2005</td>
<td>14,262 50-69 NR</td>
<td>Mammography (Screen detected)</td>
<td>No screening (no response to invitation)</td>
<td>Selection X</td>
<td></td>
<td>Mod 0-5 yr follow up</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mammography (Clinically diagnosed)</td>
<td>No screening (not invited)</td>
<td>Detection X</td>
<td></td>
<td>50-54 yr: HR 1.04 (95% CI, 0.81 to 1.33)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td>55-69 yr: HR 1.04 (95% CI, 0.83 to 1.30)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td>60-64 yr: HR 0.87 (95% CI, 0.70 to 1.08)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td>65-69 yr: HR 0.65 (95% CI, 0.52 to 0.81)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5-10 yr follow up</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50-54 yr: HR 1.00 (95% CI, 0.65 to 1.52)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>55-69 yr: HR 0.88 (95% CI, 0.60 to 1.28)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>60-64 yr: HR 1.09 (95% CI, 0.74 to 1.60)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>65-69 yr: HR 0.92 (95% CI, 0.63 to 1.35)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------</td>
<td>----------------------------------</td>
<td>----------------------</td>
<td>-------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>---------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Criterion</td>
<td>H</td>
<td>L</td>
<td>Relative</td>
<td>Absolute</td>
<td></td>
</tr>
<tr>
<td>Puliti, 2008*</td>
<td>Case-Control Italy 1988-2002</td>
<td>8750 50-74 100% Average Risk</td>
<td>No screening</td>
<td>Mammography NR</td>
<td>Selection</td>
<td>X</td>
<td></td>
<td>Moderate</td>
<td>OR 0.50 (95% CI, 0.42 to 0.60)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roder, 2008**</td>
<td>Case-Control Australia 2002-2005</td>
<td>1964 40-80 100% Average Risk</td>
<td>No screening</td>
<td>Mammography Various</td>
<td>Selection</td>
<td>X</td>
<td></td>
<td>Moderate</td>
<td>All ages OR 0.59 (95% CI, 0.47-0.74)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design</td>
<td>Country</td>
<td>Total N</td>
<td>Age Range</td>
<td>Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>---------</td>
<td>---------</td>
<td>------------</td>
<td>-----------------</td>
<td>----------------------</td>
<td>---------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All ages RR 0.78 (95% CI 0.70-0.87)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50-54 yo RR 0.74 (0.55-0.96)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>55-59 yo RR 0.84 (0.69-1.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>60-64 yo RR 0.71 (0.58-0.87)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>65-69 yo RR 0.84 (0.66-1.04)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gabe, 2007[2]</td>
<td>Case-Control</td>
<td>Iceland</td>
<td>1128</td>
<td>43-83</td>
<td>100% Average Risk</td>
<td>Mammography 2 yr</td>
<td>No screening</td>
<td>Selection X, Detection X, Performance X, Attrition X, Reporting X</td>
<td>Moderate</td>
<td>OR 0.59 (95% CI, 0.41-0.84)</td>
</tr>
<tr>
<td>Jonsson, 2007[2]</td>
<td>Prospective cohort</td>
<td>Sweden</td>
<td>185,000</td>
<td>40-74</td>
<td>NR</td>
<td>Mammography 18-24 mo</td>
<td>No screening</td>
<td>Selection X, Detection X, Performance X, Attrition X, Reporting X</td>
<td>Moderate</td>
<td>All ages RR 0.74 (95% CI 0.62-0.88)</td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------</td>
<td>----------------------------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norman, 2007</td>
<td>Case-Control US 1994-2005</td>
<td>4569 1913N &lt;50 NR</td>
<td>Mammography (40-49) 2 yr</td>
<td>Mammography (pre-menopausal) 2 yr</td>
<td>Selection X</td>
<td>Low</td>
<td>Any screening in last 2 yr vs none OR 0.63 (95% CI, 0.50 to 0.78)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mammography (50-64) 2 yr</td>
<td>Mammography (post-menopausal) 2 yr</td>
<td>Detection X</td>
<td></td>
<td>40-49 yr: OR 0.89 (95% CI, 0.65 to 1.23)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td>50-64 yr: OR 0.47 (95% CI, 0.35 to 0.63)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td>Premenopausal: OR 0.74 (95% CI, 0.53 to 1.04)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td>Postmenopausal: OR 0.45 (95% CI, 0.33 to 0.62)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parvinen, 2006</td>
<td>Retrospective cohort Finland 1987-2001</td>
<td>1,980,026 55-69 NR</td>
<td>No screening (Helsinki)</td>
<td>Mammography (Tampere) NR</td>
<td>Selection X</td>
<td>Moderate</td>
<td>All Ages (Tampere): RR 0.77 (95% CI, 0.57 to 1.06)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mammography (Turku) NR</td>
<td>Detection X</td>
<td></td>
<td>(Turku): RR 0.58 (95% CI, 0.41 to 0.83)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swedish Organised Service Screening Evaluation Group, 2006</td>
<td>Retrospective cohort Sweden 1980-2001</td>
<td>1,108,610 &lt;70 NR</td>
<td>Mammography NR</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Overall (ORs) 0.63 (0.50-0.78)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td>40-49 yo 0.89 (0.65-1.23)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td>50-64 yo 0.47 (0.35-0.63)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td>Premenopausal:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------</td>
<td>----------------------------------</td>
<td>-----------------------</td>
<td>--------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vutuc, 2006</td>
<td>Prospective cohort Austria, Finland, Sweden 1980-2002</td>
<td>NR NR NR</td>
<td>Mammography (Austria) 1-2 yr</td>
<td>No screening (prescreening era in all countries)</td>
<td>Selection X</td>
<td>Low</td>
<td>0.74 (0.53-1.04) Postmenopausal: 0.45 (0.33-0.62)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elmore, 2005</td>
<td>Case-Control US 1983-1998</td>
<td>3852 40-65 71% Average Risk 19% High Risk</td>
<td>No screening (Cases-Average Risk)</td>
<td>No screening (Cases-High Risk)</td>
<td>Selection X</td>
<td>Low</td>
<td>40-65 yo OR 0.86 (95% CI 0.74-1.04) 40-49 yo OR 0.80 (0.62-1.01) 50-65 yo OR 1.02 (0.74-1.39)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Criterion</th>
<th>H</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Detection</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Attrition</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Reporting</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design</td>
<td>Country</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>---------</td>
</tr>
<tr>
<td>Olsen, 2005</td>
<td>Prospective cohort</td>
<td>Denmark</td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design</td>
<td>Total N</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Fielder, 2004</td>
<td>Case-Control</td>
<td>1136</td>
</tr>
<tr>
<td></td>
<td>UK</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1991-2001</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jonsson, 2003</td>
<td>Retrospective cohort</td>
<td>423</td>
</tr>
<tr>
<td></td>
<td>Sweden</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1974-1998</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range Risk Population</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>Jonsson, 2003</td>
<td>Prospective cohort Sweden 1986-1998</td>
<td>125,438 100% 50-74 NR</td>
</tr>
<tr>
<td>Broeders, 2002</td>
<td>Case-referent Netherlands 1987-1997</td>
<td>157 cases 785 referents 35-79 NR</td>
</tr>
<tr>
<td>Duffy, 2002</td>
<td>Retrospective cohort Sweden 1958-1998</td>
<td>7.5 million NR NR</td>
</tr>
<tr>
<td>Paci, 2002</td>
<td>Cohort (incidence-based mortality) Italy 1990–1999</td>
<td>~60,000 (low compliance) 50-69 100% Average Risk</td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country</td>
<td>Total N Age Range Risk Population</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>Tabar, 2001</td>
<td>Prospective cohort Sweden 1968-1996</td>
<td>1,939,348 person years 20-69 NR</td>
</tr>
<tr>
<td>Jonsson, 2000</td>
<td>Retrospective cohort Sweden 1987-1996</td>
<td>439,431 100% &lt;50 NR</td>
</tr>
<tr>
<td>Moody-Ayers, 2000</td>
<td>Retrospective cohort US 1988-1994</td>
<td>233 NR NR</td>
</tr>
<tr>
<td>Hakama, 1997</td>
<td>Retrospective cohort Finland 1987-1992</td>
<td>158,755 48-60 NR</td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range Risk Population</td>
</tr>
<tr>
<td>------------------------</td>
<td>---------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overdiagnosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miller, 2014*</td>
<td>RCT Canada 1980-2005</td>
<td>89,835 40-59 NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yen, 2012*</td>
<td>RCT Nordic 1977-2005</td>
<td>134,867 40-74 2.4% High Risk</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observational</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coldman, 2013**</td>
<td>Retrospective cohort Canada 1988-2009</td>
<td>39 million yrs at risk NR NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Outcome Study**: Study Design, Country, Years, Total N Age Range Risk Population, Intervention Interval, Comparator Interval, Risk of bias, Overall Quality, Effect (95% CI)

**Risk of bias**: Selection, Detection, Performance, Attrition, Reporting

**Overall Quality**: High, Moderate

**Effect (95% CI)**: Difference in number of breast cancer cases between mammography and control arm

- 666/44925 (mammography + CBE)
- 524/44910 (CBE)

**Overdiagnosis rate**: Participation method:
- 5.4% - invasive cancer only
- 17.3% - invasive and in situ

Population method:
- 0.7% invasive only
- 6.7% invasive and in situ
<table>
<thead>
<tr>
<th>Outcome Study</th>
<th>Study Design Country Years</th>
<th>Total N Age Range Risk Population</th>
<th>Intervention Interval</th>
<th>Comparator Interval</th>
<th>Risk of bias</th>
<th>Overall Quality</th>
<th>Effect (95% CI)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lund, 2013\textsuperscript{a}</td>
<td>Prospective cohort Norway 2005-2010</td>
<td>53,363 52-79 NR</td>
<td>Mammography 2 yr</td>
<td>Mammography (outside program) Unclear interval No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>RR of cancer in not screened in program 0.93 (95% CI 0.79 to 1.15) Never screened 0.77 (0.59 to 1.01)</td>
<td>Relative</td>
</tr>
<tr>
<td>Njor, 2013\textsuperscript{a}</td>
<td>Retrospective cohort Denmark 1991-2009</td>
<td>57,763 56-70 NR</td>
<td>Mammography (Copenhagen) 2 yr</td>
<td>Mammography (Funen) 2 yr</td>
<td>Selection X</td>
<td>Moderate</td>
<td>RRs: Copenhagen study: Invasive cancer: 1.05 (0.88-1.24) Invasive and in situ: 1.06 (0.90-1.25) Funen study: Invasive cancer: 1.01 0.92-1.10) Invasive and in situ: 1.01 (0.93-1.10) Summary overdiagnosis estimate: 2.3% (-3% to 8%)</td>
<td>Relative</td>
</tr>
<tr>
<td>Bleyer, 2012\textsuperscript{a}</td>
<td>Registry (cohort)—SEER US 1976-2008</td>
<td>US Population (extrapolated) ≥40 100% Average Risk</td>
<td>Mammography 1-2 yr (opportunistic)</td>
<td>Variable assumptions about incidence derived from pre-screening age-specific rates</td>
<td>Selection X</td>
<td>Low (no direct estimate of proportion of women screened)</td>
<td>31% of all breast cancers</td>
<td>Relative</td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------</td>
<td>----------------------------------</td>
<td>-----------------------</td>
<td>--------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>Hofvind, 2012</td>
<td>Retrospective cohort Norway 1996-2007</td>
<td>640,247 50-69 NR</td>
<td>Mammography 2 yr</td>
<td>No screening (Invited non-participants; Before invitation)</td>
<td>Selection X</td>
<td>Moderate</td>
<td>RR 3.0</td>
<td></td>
</tr>
<tr>
<td>Puliti, 2012</td>
<td>Retrospective cohort Italy 1991-2008</td>
<td>51,096 50-69 100% Average Risk</td>
<td>Mammography 2 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>RR 1.05; 95% CI 0.93-1.18</td>
<td></td>
</tr>
<tr>
<td>De Gelder, 2011</td>
<td>Microsimulation model Netherlands 1990-2006</td>
<td>NA</td>
<td>Mammography 2 yr</td>
<td>No Screening</td>
<td>Selection</td>
<td>Moderate (risk of bias not applicable)</td>
<td>2.8% (wide variation based on choice of denominators)</td>
<td></td>
</tr>
<tr>
<td>Zahl, 2011</td>
<td>Retrospective cohort Sweden 1986-2009</td>
<td>646,331 40-74 NR</td>
<td>Mammography 1 yr (40-49) 2 yr (50-74)</td>
<td>No screening</td>
<td>Selection X X</td>
<td>Moderate</td>
<td>RR Invasive breast cancer 6 yr follow up: 1.14 (95% CI 1.1-1.18) 4 yr follow up: 1.49 (1.41-1.58)</td>
<td></td>
</tr>
<tr>
<td>Duffy, 2010</td>
<td>Retrospective cohort UK, Sweden 1977-2004</td>
<td>NR 50-70 NR</td>
<td>Mammography 24-33 mo</td>
<td>No screening</td>
<td>Selection</td>
<td>Moderate</td>
<td>Overdiagnosis estimates: Sweden 4.3/1000 cases screened for 20 yr UK 2.3 cases/1000 women screened for 20 yr</td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------</td>
<td>----------------------------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Criterion</td>
<td></td>
<td>Relative</td>
<td>Absolute</td>
</tr>
<tr>
<td>Kalager, 2010&lt;sup&gt;31&lt;/sup&gt;</td>
<td>Prospective cohort Norway 1996-2005</td>
<td>39,888 50-69 (467,343 for 50-74) 100% Average Risk</td>
<td>Mammography 2 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Lead time/temporal adjustment: 25% (95% CI 19% to 31%) Alternative: 2 year lead time 15% (8% to 23%) 5 year lead time 20% (13% to 28%)</td>
<td>6-10 overdiagnoses per 2500 women invited</td>
</tr>
<tr>
<td>Morrell, 2010&lt;sup&gt;32&lt;/sup&gt;</td>
<td>Retrospective cohort Australia 1999-2001</td>
<td>NR 50-69 NR</td>
<td>Mammography 2 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Varies by methodology Using longer lead time, 30-42% of cancer cases overdiagnosed</td>
<td>Varies by methodology Using longer lead time, approximately 1,380 per 100,000 overdiagnosed (1.3% risk ages 50-69)</td>
</tr>
<tr>
<td>Jorgensen, 2009&lt;sup&gt;33&lt;/sup&gt;</td>
<td>Retrospective cohort Denmark 1991-2003</td>
<td>NR 50-69 NR</td>
<td>Mammography 2 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>RR Invasive and in situ: 1.34 (95% CI 1.29-1.40)</td>
<td></td>
</tr>
<tr>
<td>Puliti, 2009&lt;sup&gt;34&lt;/sup&gt;</td>
<td>Prospective cohort Italy 1990-2004</td>
<td>61,568 50-69 100% Average Risk</td>
<td>Mammography 2 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>Low</td>
<td>RR 1.01 (95% CI 0.95-1.07) invasive and in situ 0.99 (0.94-1.05) invasive cancer only</td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------</td>
<td>-----------------------------------</td>
<td>-----------------------</td>
<td>--------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>Olsen, 2006</td>
<td>Multistate modeling (lead time approach) Denmark 1991</td>
<td>35,123 50–69 100% Average Risk</td>
<td>Mammography 2 yr</td>
<td>No screening, based on estimates derived from observed prevalence and incidence at first and subsequent screen</td>
<td>Selection X</td>
<td>Moderate</td>
<td>First screen 7.8% (95% CI 0.3 to 26.5%) Second screen 0.5% (0.02 to 2.1%)</td>
<td></td>
</tr>
<tr>
<td>Paci, 2006</td>
<td>Before and After Italy 1986-2001</td>
<td>27,518 50-74 Average risk</td>
<td>Mammography 2 yr</td>
<td>No screening, estimated based on Poisson regression of incidence prior to introduction of screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>4.6% (95% CI 2% to 7%) invasive plus DCIS 3.2% (1% to 6%)</td>
<td></td>
</tr>
<tr>
<td>Jonsson, 2005</td>
<td>Retrospective cohort Sweden 1986-1999</td>
<td>463,000 40-74 NR</td>
<td>No screening Mammography NR</td>
<td></td>
<td>Selection X</td>
<td>Low</td>
<td>RR Invasive breast cancer Age 40-49 0.96 (0.77-1.21) Age 50-59 1.54 (1.33-1.79) Age 60-69 1.21 (1.04-1.41) Age 70-74 1.02 (0.82-1.30)</td>
<td></td>
</tr>
<tr>
<td>Paci, 2004</td>
<td>Retrospective cohort Italy 1990-1999</td>
<td>2626 50-69 NR</td>
<td>Mammography 2 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Overdiagnosis estimate: Invasive cancer 2% Invasive + in situ 5%</td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------</td>
<td>----------------------------------</td>
<td>-----------------------</td>
<td>--------------------</td>
<td>-------------</td>
<td>-----------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>Zahl, 2004&lt;sup&gt;79&lt;/sup&gt;</td>
<td>Retrospective cohort Norway, Sweden 1991-2000</td>
<td>NR &gt;30 NR</td>
<td>Mammography (Norway) NR</td>
<td>Mammography (Sweden) NR</td>
<td>Selection X</td>
<td>Low</td>
<td>RR Invasive breast cancer Age 50-69 Norway 1.54 (1.42-1.66) Sweden 1.45 (1.41-1.49)</td>
<td></td>
</tr>
<tr>
<td>Andersson, 1997&lt;sup&gt;6&lt;/sup&gt; Malmo</td>
<td>RCT Sweden 1977-1993</td>
<td>25,770 44-50 100% Average Risk</td>
<td>Mammography 18-24 mo</td>
<td>No screening</td>
<td>Selection X</td>
<td>Low</td>
<td>Treatment of clinically insignificant cancer=10 cancers/100,000 person years One clinically insignificant cancer/2 breast cancer deaths prevented</td>
<td></td>
</tr>
<tr>
<td>False positive: Recall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Johns, 2010&lt;sup&gt;7&lt;/sup&gt; UK Age</td>
<td>RCT UK 1991-2004</td>
<td>160,921 NR NR</td>
<td>Mammography 1 yr</td>
<td>Mammography 1 yr (1 false positive; &gt;1 false positive)</td>
<td>Selection X</td>
<td>High</td>
<td>85.4% (no FP) 14.6% (at least one FP) 2.1% more than one FP</td>
<td></td>
</tr>
<tr>
<td>Andersson, 1997&lt;sup&gt;6&lt;/sup&gt; Malmo</td>
<td>RCT Sweden 1977-1993</td>
<td>25,770 44-50 100% Average Risk</td>
<td>Mammography 18-24 mo</td>
<td>No screening</td>
<td>Selection X</td>
<td>High</td>
<td>1.26% (mammo) No FP for unscreened</td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------</td>
<td>----------------------------------</td>
<td>-----------------------</td>
<td>--------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>Frisell, 1997' Stockholm</td>
<td>RCT Sweden 1981-1995</td>
<td>60,261 40-64 100% Average Risk</td>
<td>Mammography 2 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>High</td>
<td>355/100,000 woman years (mammo) No FP for unscreened group</td>
<td></td>
</tr>
<tr>
<td>Observational</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kikuchi, 2014 Retrospective cohort Japan 2008-2008</td>
<td>Retrospective cohort Japan 2008-2008</td>
<td>12,823 &gt;40 NR</td>
<td>Mammography (40-49) Unclear interval</td>
<td>Mammography (≥50) Unclear interval</td>
<td>Selection X</td>
<td>Moderate</td>
<td>40-49 years 2.5% (95% CI 2.1% to 3.0%) 50 and older 1.4% (1.1% to 1.7%) Not stratified by first or subsequent</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ciato, 2013 Prospective cohort Italy 2011-2012</td>
<td>Prospective cohort Italy 2011-2012</td>
<td>7292 48-71 NR</td>
<td>Mammography + Tomosynthesis 2 yr</td>
<td>Mammography 2 yr</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Overall 1% (M+T) 2% (M) Density 3-4 1.7% (M+T) 2.7% (M) Density 1-2 0.9% (M+T) 1.8% (M) Age&lt;60 yo 1.0% (M+T) 2.2% (M) Age 60+ yo 1.0% (M+T) 1.6% (M)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design</td>
<td>Country</td>
<td>Years</td>
<td>Total N</td>
<td>Age Range</td>
<td>Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>---------</td>
<td>-------</td>
<td>---------</td>
<td>------------</td>
<td>-----------------</td>
<td>----------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Domingo, 2013*</td>
<td>Retrospective cohort</td>
<td>Denmark</td>
<td>1991-2008</td>
<td>716,875</td>
<td>50-69</td>
<td>NR</td>
<td>Mammography (Copenhagen) 2 yr</td>
<td>Mammography (Fyn) 2 yr</td>
</tr>
<tr>
<td>Haas, 2013*</td>
<td>Retrospective cohort</td>
<td>US</td>
<td>2011-2012</td>
<td>13,158</td>
<td>31% &lt;50</td>
<td>79% Average Risk 21% High Risk</td>
<td>Mammography + Tomosynthesis NR</td>
<td>Mammography</td>
</tr>
<tr>
<td>Kerlikowske, 2013*</td>
<td>Prospective cohort</td>
<td>US</td>
<td>1994-2008</td>
<td>934,098</td>
<td>40-74</td>
<td>21.7% High Risk</td>
<td>Mammography 1 yr</td>
<td>Mammography 2 yr</td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------</td>
<td>-----------------------------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Risk of bias (95% CI):**
  - Relative
  - 0.76 (0.44 to 1.33)
  - 0.99 (0.45 to 2.18)
  - 1.03 (0.76 to 1.41)
  - 0.72 (0.41 to 1.24)

- **Overall Quality: 3 yr**

- **Effect (95% CI):**
  - 1yr interval 60.0% (58.6 to 61.3)
  - 2yr interval 38.5% (37.8 to 39.3)
  - 3yr interval 27.0% (26.3 to 27.6)
  - 50-74 (No HRT) 49.8% (49.0 to 50.6)
  - 2yr interval 30.7% (30.2 to 31.2)
  - 3yr interval 21.9% (21.3 to 22.4)

- **Absolute risk of FP biopsy lower for women with fatty breast density, higher for women with heterogeneously or extremely dense breasts, or on HRT; effect of interval similar**
<table>
<thead>
<tr>
<th>Outcome Study</th>
<th>Study Design Country Years</th>
<th>Total N Age Range Risk Population</th>
<th>Intervention Interval</th>
<th>Comparator Interval</th>
<th>Risk of bias</th>
<th>Overall Quality</th>
<th>Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Otten, 2013</td>
<td>Retrospective cohort Netherlands 1975-2006</td>
<td>&gt;11,000 48-52 NR</td>
<td>Mammography (Historic cohort) 2 yr</td>
<td>Mammography (Current cohort) 2 yr</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Cumulative chance of recall 4.2% (95% CI 3.3 to 5.1%)</td>
</tr>
<tr>
<td>Skaane, 2013</td>
<td>Prospective cohort Norway 2010-2011</td>
<td>12,631 50-69 NR</td>
<td>Mammography + Tomosynthesis 2 yr</td>
<td>Mammography 2 yr</td>
<td>Selection X</td>
<td>Moderate</td>
<td>6.11% (mammo) 5.31% (mammo + tomo)</td>
</tr>
<tr>
<td>Skaane, 2013</td>
<td>Prospective cohort Norway 2010-2011</td>
<td>12,621 50-69 NR</td>
<td>Mammography (2D) 2 yr</td>
<td>Mammography (2D + 3D) 2 yr</td>
<td>Selection X</td>
<td>Moderate</td>
<td>2D 2.18% (95% CI 1.93% to 2.44%) 3D 2.73% (2.45% to 3.02%)</td>
</tr>
<tr>
<td>Tohno, 2013</td>
<td>Retrospective cohort Japan 2011-2012</td>
<td>11,753 55% &lt;50 NR</td>
<td>Ultrasound + Mammography NR</td>
<td>Mammography NR</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Mammography: 213/4528=4.7% Mammography + ultrasound: 22/974=2.2%</td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design</td>
<td>Country</td>
<td>Years</td>
<td>Total N</td>
<td>Age Range</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
</tr>
<tr>
<td>--------------------</td>
<td>---------------------</td>
<td>---------</td>
<td>----------------</td>
<td>---------</td>
<td>------------</td>
<td>------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Hubbard, 2011&lt;sup&gt;27&lt;/sup&gt;</td>
<td>Prospective cohort</td>
<td>US</td>
<td>1994-2007</td>
<td>173,948</td>
<td>40-59</td>
<td>Mammography 9-18 mo</td>
<td>Mammography 19-30 mo Mammography &gt;30 mo</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.5% High Risk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Molins, 2009&lt;sup&gt;28&lt;/sup&gt;</td>
<td>Prospective cohort</td>
<td>Spain</td>
<td>1996-2003</td>
<td>27,960</td>
<td>50-69</td>
<td>Mammography (4 rounds) 2 yr</td>
<td>Mammography (1-3 rounds) NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100% Average Risk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schonberg, 2009&lt;sup&gt;19&lt;/sup&gt;</td>
<td>Retrospective cohort</td>
<td>US</td>
<td>1994-2006</td>
<td>2011</td>
<td>&gt;80</td>
<td>Mammography NR</td>
<td>No screening</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100% Average Risk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barton, 2004&lt;sup&gt;21&lt;/sup&gt;</td>
<td>Prospective cohort</td>
<td>US</td>
<td>1999-2001</td>
<td>2390</td>
<td>&gt;39</td>
<td>Mammography (False positive)</td>
<td>Mammography (Normal)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1605N Average Risk 785N High Risk</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias Selection Detection Performance Attrition Reporting</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------</td>
<td>----------------------------------</td>
<td>-----------------------</td>
<td>-------------------</td>
<td>-----------------------------------------------------------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>False positive: Biopsy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Moderate</td>
<td>40-49 yo 0.76% (95% CI 0.51% to 1.0%)</td>
</tr>
<tr>
<td>Observational</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50-59 yo 0.42% (0.28% to 0.59%)</td>
</tr>
<tr>
<td>Kikuchi, 2014</td>
<td>Retrospective cohort Japan 2008-2008</td>
<td>12,823 &gt;40 NR</td>
<td>Mammography (40-49) Unclear interval</td>
<td>Mammography (≥50) Unclear interval</td>
<td>Selection X Detection X Performance X Attrition X Reporting X</td>
<td>Moderate</td>
<td>40-49 yo 0.76% (95% CI 0.51% to 1.0%)</td>
</tr>
</tbody>
</table>

G-28
<table>
<thead>
<tr>
<th>Outcome Study</th>
<th>Study Design Country Years</th>
<th>Total N Age Range Risk Population</th>
<th>Intervention Interval</th>
<th>Comparator Interval</th>
<th>Risk of bias</th>
<th>Overall Quality</th>
<th>Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kerlikowske, 2013&lt;sup&gt;64&lt;/sup&gt;</td>
<td>Prospective cohort US 1994-2008</td>
<td>934,098 40-74 21.7% High Risk</td>
<td>Mammography 1 yr</td>
<td>Mammography 2 yr</td>
<td>Selection</td>
<td>Moderate</td>
<td>Scatter</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection</td>
<td></td>
<td>fibroglandular</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance</td>
<td></td>
<td>densities</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition</td>
<td></td>
<td>Cum probability</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting</td>
<td></td>
<td>over 10 yr</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40-49:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1yr interval</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.3% (95% CI 8.3% to 10.4%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2yr interval</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.9% (4.6% to 5.3%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3yr interval</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.4% (3.1% to 3.7%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50-74, no HRT:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1yr interval</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.1% (7.6% to 8.6%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2yr interval</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.5% (4.3% to 4.8%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3yr interval</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.4% (3.2% to 3.7%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pattern by interval similar, but FP rate lower with fatty breast density, higher with heterogeneously or extremely dense breasts, on HRT</td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------</td>
<td>----------------------------------</td>
<td>-----------------------</td>
<td>--------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Hubbard, 2011</td>
<td>Prospective cohort US 1994-2007</td>
<td>173,948 40-59 5.5% High Risk</td>
<td>Mammography 9-18 mo</td>
<td>Mammography 19-30 mo Mammography &gt;30 mo</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Adjusted OR for FP biopsy for 19-30 months compared to 9-18 months: 1.22 (95% CI 1.05 to 1.41)</td>
</tr>
<tr>
<td>Kalager, 2010</td>
<td>Retrospective cohort Norway 1996-2005</td>
<td>231,310 50-69 100% Average Risk</td>
<td>Mammography 2 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Cumulative probability over 10 years 40-49 yo Annual 7.0% (95% CI 6.1% to 7.8%) Biennial 4.8% (4.4% to 5.2%)</td>
</tr>
<tr>
<td>Schonberg, 2009</td>
<td>Retrospective cohort US 1994-2006</td>
<td>2011 &gt;80 100% Average Risk</td>
<td>Mammography NR</td>
<td>No screening</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Screened group = 1.84% FP (biopsy) No FP for unscreened group</td>
</tr>
<tr>
<td>Ohlinger, 2006</td>
<td>Prospective cohort Germany 1994-2003</td>
<td>448 21-89 100% Average Risk</td>
<td>Ultrasound Once</td>
<td>Mammography Once Mammography + CBE Once</td>
<td>Selection X</td>
<td>Low</td>
<td>1.12% (ultrasound) 0.67% (mammo) 1.56% (ultrasound + mammo)</td>
</tr>
</tbody>
</table>
## Appendix Table G2. Study Characteristics—Key Question 2

<table>
<thead>
<tr>
<th>Outcome Study</th>
<th>Study Design Country Years</th>
<th>Total N Age Range (Y) Risk Population</th>
<th>Intervention Interval</th>
<th>Comparator Interval</th>
<th>Risk of bias Criterion</th>
<th>Overall Quality</th>
<th>Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Breast Cancer Mortality</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Observational</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parvinen, 2011</td>
<td>Prospective cohort Finland 1987-2007</td>
<td>14,765 40-49 NR</td>
<td>Mammography 3 yr</td>
<td>Mammography 1 yr</td>
<td>Selection X</td>
<td>High</td>
<td>RR 1.14 (0.59 to 1.27)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coldman, 2008</td>
<td>Prospective cohort Canada 1988-2005</td>
<td>658,151 40-79 NR</td>
<td>Mammography median 13-14 mo</td>
<td>Mammography 18-29 mo</td>
<td>Selection X</td>
<td>Low</td>
<td>RR 1.06 (0.76 to 1.46)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Overdiagnosis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Observational</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dittus, 2013</td>
<td>Prospective cohort US NR</td>
<td>1,891,039 40-74 NR</td>
<td>Mammography 9-18 mo</td>
<td>Mammography &gt;18-30 mo</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Invasive vs DCIS Normal weight women Premenopausal 0.71 (95% CI 0.48 to 1.06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td>Postmenopausal 1.43 (1.02 to 2.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td>Results qualitatively similar for overweight, obese women by menopausal status, but CIs include 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range (Y) Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias Criterion</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------</td>
<td>---------------------------------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>------------------------</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>False positive:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Observational</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dittus, 2013**</td>
<td>Prospective cohort US NR</td>
<td>1,891,039 40-74 NR 9-18 mo</td>
<td>Mammography &gt;18-30 mo</td>
<td>Mammography</td>
<td>Selection X Detection X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country</td>
<td>Years</td>
<td>Total N</td>
<td>Age Range (Y)</td>
<td>Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------</td>
<td>-------</td>
<td>---------</td>
<td>---------------</td>
<td>-----------------</td>
<td>-----------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>O'Meara, 2013</td>
<td>Prospective cohort US</td>
<td>1994-2008</td>
<td>1,276,612</td>
<td>36% &lt;50</td>
<td>64% 50-74 NR</td>
<td>Mammography 1 yr (9-18 mo)</td>
<td>Mammography</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mammography 2 yr (&gt;18-30 mo)</td>
<td>Mammography 3 yr (&gt;30-42 mo)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Study Design</td>
<td>Total N</td>
<td>Age Range (Y)</td>
<td>Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
<td>---------</td>
<td>---------------</td>
<td>-----------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>--------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>H</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design</td>
<td>Total N</td>
<td>Age Range (Y)</td>
<td>Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>---------</td>
<td>---------------</td>
<td>-----------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Kerlikowske, 2013&lt;sup&gt;64&lt;/sup&gt;</td>
<td>Prospective cohort</td>
<td>US</td>
<td>1994-2008</td>
<td>934,098</td>
<td>40-74</td>
<td>21.7% High Risk</td>
<td>Mammography 1 yr</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03 (0.76 to 1.41)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.72 (0.41 to 1.24)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40-49:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2yr interval 38.5% (37.8 to 39.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3yr interval 27.0% (26.3 to 27.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50-74:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2yr interval 30.7 (30.2 to 31.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3yr interval 21.9 (21.3 to 22.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pattern by interval similar, but FP rate lower with fatty breast density, higher with heterogeneously or extremely dense breasts</td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country</td>
<td>Total N Age Range (Y) Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias Criterion</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------</td>
<td>--------------------------------------</td>
<td>-----------------------</td>
<td>--------------------</td>
<td>------------------------</td>
<td>-------------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>Hubbard, 2011 &quot;</td>
<td>Prospective cohort US 1994-2007</td>
<td>173,948 40-59 5.5% High Risk</td>
<td>Mammography 9-18 mo</td>
<td>Mammography 19-30 mo</td>
<td>Selection X</td>
<td>Moderate</td>
<td>Adjusted OR for false positive recall for 19-30 months compared to 9-18 months 1.13 (1.08-1.19)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False positive: Biopsy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observational</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Braithwaite, 2013 &quot;</td>
<td>Prospective cohort US 1999-2006</td>
<td>292,436 66-89 NR</td>
<td>Mammography 1 yr</td>
<td>Mammography 2 yr</td>
<td>Selection X</td>
<td>Moderate</td>
<td>66-74: 9.8% (8.4 to 11.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td>4.6% (4.2 to 5.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td>Cum probability over 10 yr</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td>75-89: 9.2% (7.5 to 11.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td>4.1% (3.7 to 4.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cum probability over 10 yr</td>
</tr>
<tr>
<td>Outcome</td>
<td>Study Design Country</td>
<td>Total N Age Range (Y) Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias Criterion</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------</td>
<td>--------------------------------------</td>
<td>-----------------------</td>
<td>--------------------</td>
<td>------------------------</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Dittus, 2013</td>
<td>Prospective cohort US NR</td>
<td>1,891,039 40-74 NR</td>
<td>Mammography 9-18 mo</td>
<td>Mammography &gt;18-30 mo</td>
<td>Selection X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kerlikowske, 2013</td>
<td>Prospective cohort US 1994-2008</td>
<td>934,098 40-74 21.7% High Risk</td>
<td>Mammography 1 yr</td>
<td>Mammography 2 yr</td>
<td>Selection X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range (Y) Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias Criterion</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>-----------------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3 yr</td>
<td></td>
<td>Reporting X</td>
<td></td>
<td>Cumulative probability over 10 yr</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40-49: 1yr interval 9.3% (8.3 to 10.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2yr interval 4.9% (4.6 to 5.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3yr interval 3.4% (3.1 to 3.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50-74: 1yr interval 8.1% (7.6 to 8.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2yr interval 4.5% (4.3 to 4.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3yr interval 3.4% (3.2 to 3.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Absolute risk of FP biopsy lower for women with fatty breast density, higher for women with heterogeneously or extremely dense breasts, or on HRT; effect of interval similar</td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design</td>
<td>Country</td>
<td>Total N</td>
<td>Age Range (Y)</td>
<td>Risk Population</td>
<td>Intervention</td>
<td>Comparator</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>---------</td>
<td>---------</td>
<td>---------------</td>
<td>-----------------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>O’Meara, 2013</td>
<td>Prospective cohort</td>
<td>US</td>
<td>1,276,612</td>
<td>36% &lt;50 64% 50-74 NR</td>
<td>Mammography 1 yr</td>
<td>Mammography 2 yr</td>
<td>Mammography 3 yr</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design &amp; Country Years</td>
<td>Total N</td>
<td>Age Range (Y)</td>
<td>Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias Criterion</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------</td>
<td>---------</td>
<td>---------------</td>
<td>-----------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Blanchard, 2006</td>
<td>Retrospective cohort US 1985-2002</td>
<td>12,972</td>
<td>NR</td>
<td>NR</td>
<td>Mammography 1 yr (across 10 screens)</td>
<td>Mammography 2 yr (across 5 screens)</td>
<td>Mammography 3 yr (across 3 screens)</td>
</tr>
</tbody>
</table>

9.2% (6.2 to 12.1)

10.3% (8.2 to 12.3)

10.7% (8.9 to 12.3)

12.2% (10.3 to 14.2)

11.3% (9.4 to 13.2)

9.5% (7.8 to 11.2)

9.9% (8.1 to 11.7)

8.1% (6.4 to 9.8)

6.8% (5.4 to 8.2)

5.7% (4.6 to 6.8)
Appendix Table G3. Study Characteristics—Key Question 3

<table>
<thead>
<tr>
<th>Outcome Study</th>
<th>Study Design</th>
<th>Total N</th>
<th>Age Range Risk Population</th>
<th>Intervention</th>
<th>Comparator</th>
<th>Risk of bias</th>
<th>Overall Quality</th>
<th>Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Relative</td>
</tr>
<tr>
<td><strong>Breast Cancer Mortality</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RCT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shapiro, 1997</td>
<td>RCT</td>
<td>~62,000</td>
<td>40-64 NR</td>
<td>Mammography</td>
<td>No screening</td>
<td>Selection</td>
<td>Low (not mortality)</td>
<td>35.2 per 100 vs. 46.7 per 100 cumulative case fatality rates —note that this is survival, not mortality</td>
</tr>
<tr>
<td></td>
<td>US</td>
<td>1963-1996</td>
<td></td>
<td>+ CBE</td>
<td></td>
<td>Detection</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 yr</td>
<td></td>
<td>Performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Observational</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elmore, 2005</td>
<td>Case-control</td>
<td>3852</td>
<td>40-65</td>
<td>CBE alone</td>
<td>Mammography</td>
<td>Selection</td>
<td>Low</td>
<td>CBE alone vs. No screening: Age 40-65: OR 0.94 (0.79 to 1.12)</td>
</tr>
<tr>
<td></td>
<td>US</td>
<td>1983-1998</td>
<td>71% Average Risk</td>
<td>Unspecified</td>
<td>and/or CBE</td>
<td>Detection</td>
<td></td>
<td>Age 40-49: OR 0.91 (0.73 to 1.13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>19% High Risk</td>
<td>Unspecified</td>
<td>Unspecified</td>
<td>Performance</td>
<td></td>
<td>Age 50-65: OR 0.98 (0.74 to 1.31)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mammography</td>
<td>Unspecified</td>
<td>Attrition</td>
<td></td>
<td>Adjusted for race, comorbidity, age at first birth</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Unspecified</td>
<td></td>
<td>Reporting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False positive: Recall</td>
<td>RCT</td>
<td></td>
<td>39,338</td>
<td>CBE Once</td>
<td>No screening</td>
<td>Selection</td>
<td>Low</td>
<td>0.9%</td>
</tr>
<tr>
<td>Abuidris, 2013</td>
<td>RCT</td>
<td>39,338</td>
<td>100% Average Risk</td>
<td></td>
<td></td>
<td>Detection</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sudan</td>
<td>2010-2012</td>
<td></td>
<td></td>
<td></td>
<td>Performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

G-41
<table>
<thead>
<tr>
<th>Outcome Study</th>
<th>Study Design Country Years</th>
<th>Total N Age Range Risk Population</th>
<th>Intervention Interval</th>
<th>Comparator Interval</th>
<th>Risk of bias Criterion</th>
<th>H</th>
<th>L</th>
<th>Overall Quality</th>
<th>Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sankarana-rayanan, 2011[1]</td>
<td>RCT India 2006-2009</td>
<td>115,652 30-69 NR</td>
<td>CBE 3 yr</td>
<td>No screening</td>
<td>Selection X</td>
<td>Detection X</td>
<td>Performance X</td>
<td>Attrition X</td>
<td>Reporting X</td>
</tr>
<tr>
<td>Observational</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chiarelli, 2009[2]</td>
<td>Retrospective cohort Canada 2002-2004</td>
<td>290,230 NR 87% Average Risk 13% High Risk</td>
<td>Mammography + CBE 2 yr</td>
<td>Mammography 2 yr</td>
<td>Selection X</td>
<td>Detection X</td>
<td>Performance X</td>
<td>Attrition X</td>
<td>Reporting X</td>
</tr>
<tr>
<td>Honjo, 2007[3]</td>
<td>Prospective cohort Japan 1999-2001</td>
<td>3453 NR 100% Average Risk</td>
<td>CBE (in combination mammography or ultrasound) 2 yr</td>
<td>Mammography 2 yr Ultrasound 2 yr</td>
<td>Selection X</td>
<td>Detection X</td>
<td>Performance X</td>
<td>Attrition X</td>
<td>Reporting X</td>
</tr>
<tr>
<td>Oestreicher, 2005[4]</td>
<td>Prospective Cohort US 1996-2001</td>
<td>61,688 Min 40 NR</td>
<td>CBE alone 1-2 yr Mammography 1-2 yr Mammography + CBE 1-2 yr</td>
<td>Selection X</td>
<td>Detection X</td>
<td>Performance X</td>
<td>Attrition X</td>
<td>Reporting X</td>
<td>High</td>
</tr>
</tbody>
</table>
### Appendix Table G4. Study Characteristics—Key Question 4a

<table>
<thead>
<tr>
<th>Outcome Study</th>
<th>Study Design Country Years</th>
<th>Total N Age Range (Y) Risk Population</th>
<th>Intervention Interval</th>
<th>Comparator Interval</th>
<th>Risk of bias Criterion</th>
<th>Overall Quality</th>
<th>Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Breast Cancer Mortality</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Observational</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evans, 2014</td>
<td>Retrospective cohort UK NR</td>
<td>1656 &lt;50 100% High Risk</td>
<td>Mammography Unclear intervals</td>
<td>No screening</td>
<td>Selection X</td>
<td>Low</td>
<td>43 of 47 (91%) screened alive with no disease</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td>803 of 1,108 (72.9%) unscreened</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td>216 (19% died)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maurice, 2006</td>
<td>Prospective cohort UK 1991-2004</td>
<td>1170 &lt;50 5.3% High Risk</td>
<td>CBE+ Mammography 12-18 mo</td>
<td>No screening</td>
<td>Selection X</td>
<td>Low</td>
<td>HR 0.24 (0.09 to 0.66)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design</td>
<td>Country</td>
<td>Years</td>
<td>Total N</td>
<td>Age Range (Y)</td>
<td>Risk Population</td>
<td>Intervention</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>---------</td>
<td>-------</td>
<td>---------</td>
<td>---------------</td>
<td>----------------</td>
<td>--------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elmore, 2005$^{22}$</td>
<td>Case-control</td>
<td>US</td>
<td>1983-1998</td>
<td>3852</td>
<td>40-65</td>
<td>71% Average Risk, 19% High Risk</td>
<td>No screening (Cases-Average Risk)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evans, 2014$^{23}$</td>
<td>Retrospective cohort</td>
<td>UK NR</td>
<td>NR</td>
<td>1656</td>
<td>&lt;50</td>
<td>100% High Risk</td>
<td>Mammography</td>
</tr>
<tr>
<td>Stage distribution at diagnosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range (Y) Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Criterion</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>H</td>
<td>L</td>
<td>Relative</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Absolute</td>
</tr>
<tr>
<td>Walker, 2013</td>
<td>Prospective cohort Canada 2005-NR</td>
<td>899 NR 100% High Risk</td>
<td>Mammography Unclear intervals</td>
<td>No screening</td>
<td>Selection</td>
<td>X</td>
<td>Stage I vs Stage II-IV</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection</td>
<td>X</td>
<td>OR 7.80 (95% CI 1.18-51.5) for unscreened</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance</td>
<td>X</td>
<td>Nodal involvement</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition</td>
<td>X</td>
<td>OR 1.77 (95% 0.36 to 8.63) for unscreened</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting</td>
<td>X</td>
<td>Tumor &gt; 15 mm</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OR 9.72 (1.01 to 93.6) for unscreened</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Higher grade, mitotic score, lymphovascular invasion, ER/PR negative also associated with no screening, but wide CIs all include 1.0</td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range (Y) Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------</td>
<td>----------------------------------------</td>
<td>------------------------</td>
<td>---------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Warner, 2011&lt;sup&gt;11&lt;/sup&gt; Prospective cohort Canada 1997-2010</td>
<td>1275 25-65 100% High Risk</td>
<td>MRI+CBE+ Mammography + Ultrasound 1 yr Usual Care (CBE+ Mammography) 1 yr</td>
<td>Selection Detection Performance Attrition Reporting</td>
<td>X X X X X</td>
<td>High</td>
<td>Tumor Size MRI: 0-5mm: 29% 6-10mm: 45% 11-20mm: 23% 21+mm: 3% No MRI: 0-5mm: 8% 6-10mm: 27% 11-20mm: 36% 21+mm: 29% Node status MRI: Node negative and &lt;2 cm: 85% Node positive or ≥2cm: 15% No MRI: Node negative and &lt;2 cm: 54% Node positive or ≥2cm: 46%</td>
<td></td>
</tr>
<tr>
<td>Yu, 2008&lt;sup&gt;12&lt;/sup&gt; Retrospective cohort US 1999-2006</td>
<td>1019 21-88 100% High Risk</td>
<td>MRI+CBE+ Mammography 1 yr CBE+ Mammography 1 yr</td>
<td>Selection Detection Performance Attrition Reporting</td>
<td>X X X X X</td>
<td>Moderate</td>
<td>MRI Screening Stg 0: 4 (44%) Stg 1: 4 (44%) Stg 2: 1 (11%) Stg 3: 0 (0%) No MRI Screening Stg 0: 6 (30%) Stg 1: 8 (40%) Stg 2: 3 (15%) Stg 3: 3 (15%)</td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range (Y) Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------</td>
<td>---------------------------------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Maurice, 2006&lt;sup&gt;16&lt;/sup&gt;</td>
<td>Prospective cohort UK 1991-2004</td>
<td>1170 &lt;50 5.3% High Risk</td>
<td>CBE+ Mammography 12-18 mo</td>
<td>No screening</td>
<td>Selection X</td>
<td>Low</td>
<td>Tumor size Family history and screened &lt;2cm: 72% 2-5cm: 26% &gt;5cm: 2% No family history and unscreened: &lt;2cm: 39% 2-5cm: 51% &gt;5cm: 10% Node involvement Family history and screened: 0: 66% 1-4: 32% &gt;4: 2% No family history and unscreened: 0: 47% 1-4: 34% &gt;4: 19%</td>
</tr>
<tr>
<td>Ng, 2013&lt;sup&gt;10&lt;/sup&gt;</td>
<td>Prospective cohort US 2005-2013</td>
<td>148 22-65 100% High Risk due to chest irradiation</td>
<td>MRI 1 yr Mammography 1 yr</td>
<td>Selection X</td>
<td>Moderate</td>
<td>MRI Yr 1: 13.4% Yr 2: 9.0% Yr 3: 2.2% Mammography Yr 1: 5.9% Yr 2: 9.0% Yr 3: 7.5%</td>
<td></td>
</tr>
</tbody>
</table>

False Positive: Biopsy

Observational
<table>
<thead>
<tr>
<th>Outcome Study</th>
<th>Study Design</th>
<th>Country</th>
<th>Total N</th>
<th>Age Range (Y)</th>
<th>Risk Population</th>
<th>Intervention Interval</th>
<th>Comparator Interval</th>
<th>Risk of bias</th>
<th>Overall Quality</th>
<th>Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kriege, 2004</td>
<td>Prospective cohort</td>
<td>Netherlands</td>
<td>1952</td>
<td>19-72</td>
<td>100% High Risk</td>
<td>CBE 6 mo</td>
<td>Mammography 1 yr</td>
<td>MRI 1 yr</td>
<td>Selection X</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td>Performance X</td>
<td>Attrition X</td>
<td>Reporting X</td>
<td>Mammography: 7/25 (28.0%)</td>
</tr>
</tbody>
</table>
## Appendix Table G5. Study Characteristics—Key Question 4b

<table>
<thead>
<tr>
<th>Outcome Study</th>
<th>Study Design Country Years</th>
<th>Total N Age Range (Y) Risk Population</th>
<th>Intervention Interval</th>
<th>Comparator Interval</th>
<th>Risk of bias Criterion</th>
<th>Overall Quality Risk</th>
<th>Effect (95% CI) Relative Absolute</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stage distribution at diagnosis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Observational</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>King, 2013&lt;sup&gt;24&lt;/sup&gt;</td>
<td>Retrospective cohort US 1999-2009</td>
<td>776 NR 100% High Risk</td>
<td>MRI screening + Mammography + CBE</td>
<td>Mammography + CBE</td>
<td>Selection X</td>
<td>Low</td>
<td>No significant difference in median tumor size, node status, or receptor/HER2 status</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Port, 2007&lt;sup&gt;25&lt;/sup&gt;</td>
<td>Retrospective cohort US 1999-2005</td>
<td>378 25-90 100% High Risk</td>
<td>MRI 1 yr</td>
<td>Mammography 1 yr</td>
<td>Selection X</td>
<td>Low</td>
<td>Absolute # of Stage 2 cancers: MRI: 0/5 Mammography: 2/7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>False Positive: Biopsy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Observational</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>King, 2013&lt;sup&gt;24&lt;/sup&gt;</td>
<td>Retrospective cohort US 1999-2009</td>
<td>776 NR 100% High Risk</td>
<td>MRI screening + Mammography + CBE</td>
<td>Mammography + CBE</td>
<td>Selection X</td>
<td>Low</td>
<td>MRI 115 false positive biopsies generated by MRI findings (455 patients, #exams not clear)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Attrition X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reporting X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mammography 41 false positive biopsies from imaging (776 patients, #exams not clear)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Study</td>
<td>Study Design Country Years</td>
<td>Total N Age Range (Y) Risk Population</td>
<td>Intervention Interval</td>
<td>Comparator Interval</td>
<td>Risk of bias</td>
<td>Overall Quality</td>
<td>Effect (95% CI)</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------------</td>
<td>---------------------------------------</td>
<td>-----------------------</td>
<td>-------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Sung, 2011**</td>
<td>Retrospective cohort US 2003-2008</td>
<td>220 27-78 100% High Risk</td>
<td>MRI Unclear intervals</td>
<td>Mammography</td>
<td>Selection X</td>
<td>Low</td>
<td>MRI: 49/220 (22.2%; 95% CI 17.0% to 28.0%)  Mammography: 20/214 (9.3%; 95% CI 5.8% to 13.6%)</td>
</tr>
<tr>
<td>Port, 2007**</td>
<td>Retrospective cohort US 1999-2005</td>
<td>378 25-90 100% High Risk</td>
<td>MRI 1 yr</td>
<td>Mammography 1 yr</td>
<td>Selection X</td>
<td>Low</td>
<td>MRI: 40/182 (22.0%; 95% CI 16.3% to 28.3%)  Mammography: 14/196 (7.1%; 95% CI 4.0% to 11.1%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Criterion</th>
<th>H</th>
<th>L</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detection</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attrition</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reporting</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** The performance criterion is marked for most studies, indicating potential bias in the performance of the interventions.
### Appendix Table G6. Study Characteristics—Key Questions 5a and 5b

<table>
<thead>
<tr>
<th>Outcome Study</th>
<th>Study Design Country Years</th>
<th>Total N Age Range Risk Population</th>
<th>Intervention Interval</th>
<th>Comparator Interval</th>
<th>Risk of bias</th>
<th>Overall Quality</th>
<th>Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage distribution at diagnosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observational</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Randall, 2009</td>
<td>Retrospective cohort Australia 1998-2004</td>
<td>590 50-69 High Risk (family history)</td>
<td>Mammography 1 yr</td>
<td>Mammography 2 yr or more</td>
<td>Selection X</td>
<td>X</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References to Appendix G:


# Appendix H. GRADE Summary Tables

## Appendix Table H1. Summary Table for GRADE Assessments by Outcome—Key Question 1

<table>
<thead>
<tr>
<th>Quality Assessment</th>
<th>Summary of Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Interventions</td>
</tr>
<tr>
<td></td>
<td>Screening Modality</td>
</tr>
<tr>
<td><strong>Breast Cancer Mortality</strong></td>
<td></td>
</tr>
<tr>
<td>No. of Studies*</td>
<td>Design</td>
</tr>
<tr>
<td>8</td>
<td>RCT</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*No. of Studies refers to the number of studies included in the analysis.
<table>
<thead>
<tr>
<th>No. of Studies*</th>
<th>Design</th>
<th>Limitations</th>
<th>Inconsistency</th>
<th>Imprecision</th>
<th>Other Considerations</th>
<th>Screening Modality</th>
<th>Comparator</th>
<th>Relative (95% CI)</th>
<th>Absolute</th>
<th>Quality</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>Cohort</td>
<td>All in context of organized screening programs in non-U.S. settings</td>
<td>Consistent direction of effect</td>
<td>Moderate</td>
<td>Summary results reported only for incidence-based mortality estimates among those accepting screening</td>
<td>Mammography, ages 50 and older, variable intervals</td>
<td>No screening</td>
<td>0.62 (0.56-0.69) (pooled summary based on published systematic review which includes 7 of the included studies)</td>
<td>50-59 year olds: 15-year reduction in mortality of 202.2/100,000 (NNS 495)</td>
<td>60-69 year olds: 15-year reduction in mortality of 264.5/100,000 (NNS 378)</td>
<td>Moderate</td>
</tr>
<tr>
<td>13</td>
<td>Case-Control</td>
<td>All in context of organized screening programs in non-U.S. settings</td>
<td>Consistent direction of effect</td>
<td>Moderate</td>
<td>Most estimates based on adjustment for self-selection</td>
<td>Mammography, ages 50 and older</td>
<td>No screening</td>
<td>0.52 (0.42-0.65) (pooled summary estimate based on published systematic review which includes 7 of the included studies)</td>
<td>50- to 59-year-olds: 15-year reduction in mortality of 279.6/100,000 (NNS 358)</td>
<td>60- to 69-year-olds: 15-year reduction in mortality of 365.7/100,000 (NNS 274)</td>
<td>Moderate</td>
</tr>
<tr>
<td>No. of Studies*</td>
<td>Design</td>
<td>Limitations</td>
<td>Inconsistency</td>
<td>Imprecision</td>
<td>Other Considerations</td>
<td>Interventions</td>
<td>Effect</td>
<td>Quality</td>
<td>Comments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>--------</td>
<td>-------------</td>
<td>---------------</td>
<td>-------------</td>
<td>---------------------</td>
<td>---------------</td>
<td>--------</td>
<td>---------</td>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Modeling</td>
<td>Inherent uncertainties in modeling approaches, parameters</td>
<td>Consistent direction of effect</td>
<td>Moderate</td>
<td></td>
<td>Model-based estimate of reduction in U.S. mortality over time attributable to screening vs. improved treatment, estimate is median mortality reduction of screening across 7 models</td>
<td>Mammography as practiced in U.S.</td>
<td>Median of 7 models: 0.85</td>
<td>Not reported</td>
<td>Moderate</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>All Designs</td>
<td>Majority of direct evidence comes from non-US studies — differences in post-screening diagnosis &amp; treatment, as well as underlying distribution of cancer subtypes, could affect both relative and absolute estimates</td>
<td>Consistent for direction of effect</td>
<td>High degree of imprecision — estimates clearly vary based on study design</td>
<td></td>
<td></td>
<td></td>
<td>Moderate (High for direction of effect, Moderate for magnitude of effect)</td>
<td>Estimates of absolute effect limited not only by quality of evidence for mortality reduction, but need to make assumptions about rates of screening in US in order to generate US-specific estimates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of Studies*</td>
<td>Design</td>
<td>Limitations</td>
<td>Inconsistency</td>
<td>Imprecision</td>
<td>Other Considerations</td>
<td>Interventions</td>
<td>Effect</td>
<td>Quality</td>
<td>Comments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>--------</td>
<td>-------------</td>
<td>---------------</td>
<td>-------------</td>
<td>----------------------</td>
<td>---------------</td>
<td>--------</td>
<td>---------</td>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Life Expectancy</td>
<td>1</td>
<td>Modeling</td>
<td>Inherent uncertainties in model structures, parameters</td>
<td>Consistent qualitative direction</td>
<td>Results only reported for single “exemplar” model, no confidence intervals for that model; results potentially convey false sense of precision</td>
<td>Mammography test characteristics, post-diagnosis survival based on U.S. data (Breast Cancer Surveillance Consortium, SEER)</td>
<td>Biennial mammography ages 50-69</td>
<td>No screening</td>
<td>Biennial mammography ages 50-69</td>
<td>36.1 days gained per woman screened</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Biennial mammography ages 45-69</td>
<td>Biennial mammography ages 50-69</td>
<td>Biennial mammography ages 45-69</td>
<td>Biennial mammography ages 45-69</td>
<td>6.2 days gained per woman screened</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Annual mammography ages 50-69</td>
<td>Annual mammography ages 50-69</td>
<td>Annual mammography ages 50-69</td>
<td>Annual mammography ages 50-69</td>
<td>1.5 days gained per woman screened</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Annual mammography ages 45-69</td>
<td>Annual mammography ages 50-69</td>
<td>Annual mammography ages 50-69</td>
<td>Annual mammography ages 50-69</td>
<td>12.0 days gained per woman screened</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Annual mammography ages 40-69</td>
<td>Annual mammography ages 50-69</td>
<td>Annual mammography ages 50-69</td>
<td>Annual mammography ages 50-69</td>
<td>7.3 days gained per woman screened</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Annual mammography ages 40-69</td>
<td>Annual mammography ages 50-69</td>
<td>Annual mammography ages 50-69</td>
<td>Annual mammography ages 50-69</td>
<td>4.4 days gained per woman screened</td>
<td></td>
</tr>
<tr>
<td>Quality Assessment</td>
<td>Summary of Findings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interventions</strong></td>
<td><strong>Effect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of Studies*</td>
<td>Screening Modality</td>
<td>Comparator</td>
<td>Relative (95% CI)</td>
<td>Absolute</td>
<td>Quality</td>
<td>Comments</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Mammography, varying intervals</td>
<td>No screening</td>
<td>Proportion of all cancers diagnosed over entire follow-up period: 10.7% (9.3-12.2%)</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Mammography, varying intervals</td>
<td>No screening</td>
<td>Crude estimates: 0-54%</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Overdiagnosis**

- **No. of Studies**: 2
- **Design**: RCT
- **Limitations**: Variable
- **Inconsistency**: High
- **Imprecision**: Based on 3 studies, synthesis by UK Independent Panel
- **Other Considerations**: Mammography, varying intervals
- **Screening Modality**: Mammography
- **Comparator**: No screening
- **Relative (95% CI)**: Proportion of all cancers diagnosed over entire follow-up period: 10.7% (9.3-12.2%)
- **Absolute**: Low
- **Quality**: Low
<table>
<thead>
<tr>
<th>No. of Studies*</th>
<th>Design</th>
<th>Limitations</th>
<th>Inconsistency</th>
<th>Imprecision</th>
<th>Other Considerations</th>
<th>Interventions</th>
<th>Effect</th>
<th>Quality</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Modeling</td>
<td>Not reported</td>
<td>Not reported</td>
<td></td>
<td>Mammography, No screening</td>
<td>Breast cancer incidence “25% higher” with screening, but no confidence intervals</td>
<td>Low</td>
<td>Other CISNET models had overdiagnosis rates, but actual estimates not reported</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>All Designs</td>
<td>Highly variable</td>
<td>High</td>
<td>Lack of consensus on definitions, methods</td>
<td></td>
<td></td>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False Positives</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>RCT</td>
<td>Consistent direction</td>
<td>High</td>
<td>Non-US setting</td>
<td>Mammography, No screening</td>
<td>20.5% over 7 years of annual screening women 40-49</td>
<td>Moderate</td>
<td>Estimates from UK Age trial</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Cohort</td>
<td>Consistent direction</td>
<td>High</td>
<td>Results vary by setting (higher in U.S.) Variability between centers in U.S</td>
<td>Mammography, varying intervals, Europe</td>
<td>Initial screen: 9.3% (range 2.2-15.6%) Subsequent screens: 4.0% (range 1.2-10.5%) Initial screen: 16.3% Subsequent screens: 9.0% 10-year cumulative probability: Recall: 41.6% (40.6-42.5%)</td>
<td>Moderate</td>
<td>Probability of individual screening test being a false positive increased with longer screening interval, but not enough to compensate for cumulative probability No direct estimates of lifetime</td>
<td></td>
</tr>
<tr>
<td>Quality Assessment</td>
<td>Summary of Findings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of Studies*</td>
<td>Design</td>
<td>Limitations</td>
<td>Inconsistency</td>
<td>Imprecision</td>
<td>Other Considerations</td>
<td>Interventions</td>
<td>Effect</td>
<td>Quality</td>
<td>Comments</td>
</tr>
<tr>
<td>Screening Modality</td>
<td>Comparator</td>
<td>Relative (95% CI)</td>
<td>Absolute</td>
<td>Probability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual screening, age 40 and above</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biennial screening, age 50 and above</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual screening, age 50 and above</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biopsy: 4.8% (4.4-5.2%)</td>
<td>10-year cumulative probability: Recall: 61.3% (59.4-63.1%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biopsy: 41.6% (40.6-42.5%)</td>
<td>10-year cumulative probability: Recall: 42.0% (40.4-43.7%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biopsy: 6.4% (5.6-7.2%)</td>
<td>10-year</td>
<td>cumulative probability: Recall: 61.3% (58.0-64.7%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biopsy: 9.4% (4.7-11.5%)</td>
<td>probability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality Assessment</td>
<td>Summary of Findings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interventions</strong></td>
<td><strong>Effect</strong></td>
<td><strong>Quality</strong></td>
<td><strong>Comments</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of Studies*</td>
<td></td>
<td></td>
<td>Low</td>
<td>Estimated population risk of any false positive result greater than 100% for biennial screening starting below age 50, for annual screening starting below age 55. No published estimates of lifetime risk of “at least one”, but estimates done for report suggest similar patterns, but much lower incidence.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Design</td>
<td>Limitations</td>
<td>Inconsistency</td>
<td>Imprecision</td>
<td>Other Considerations</td>
<td>Screening Modality</td>
<td>Comparator</td>
<td>Relative (95% CI)</td>
<td>Absolute</td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>Inherent uncertainties in model structures, parameters</td>
<td>Consistent qualitative direction</td>
<td>Results only reported for single “exemplar” model, no confidence intervals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

*Note: H-8*
<table>
<thead>
<tr>
<th>Quality Assessment</th>
<th>Summary of Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Interventions</td>
</tr>
<tr>
<td>No. of Studies*</td>
<td>Design</td>
</tr>
<tr>
<td>20</td>
<td>All designs</td>
</tr>
<tr>
<td></td>
<td>Quality-adjusted Life Expectancy</td>
</tr>
</tbody>
</table>

*For mortality outcomes, No. of studies = number of studies included in relevant systematic review.
### Appendix Table H2. Summary Table for GRADE Assessments Across an Outcome—Key Question 2

<table>
<thead>
<tr>
<th>Quality Assessment</th>
<th>Summary of Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Interventions</td>
</tr>
<tr>
<td>No. of Studies</td>
<td>Design</td>
</tr>
<tr>
<td>Breast cancer mortality</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RCT</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Cohort</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of Studies</td>
<td>Design</td>
</tr>
<tr>
<td>---------------</td>
<td>--------</td>
</tr>
<tr>
<td>1</td>
<td>Modeling</td>
</tr>
<tr>
<td>9</td>
<td>All Studies</td>
</tr>
<tr>
<td>No. of</td>
<td>Design</td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td>Studies</td>
<td></td>
</tr>
<tr>
<td>Life Expectancy</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Modeling</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| Overdiagnosis | |                                                                              |               |             |                     | Annual screening | Biennial Screening | Normal weight Pre-menopausal 0.71 (0.48, 1.06) | Low | Results similar for overweight, obese women, but confidence intervals included 1.0 |
| 1          | Cohort   | Single study, only presents DCIS results, not direct                        | N/A           | Moderate    | Uncertainty about relationship between DCIS diagnosis and “overdiagnosis” | Annual screening | Biennial Screening | Normal weight Pre-menopausal 6.0% higher with biennial |          |          |</p>
<table>
<thead>
<tr>
<th>No. of Studies</th>
<th>Design</th>
<th>Limitations</th>
<th>Inconsistency</th>
<th>Imprecision</th>
<th>Other Considerations</th>
<th>Screening Modality</th>
<th>Comparator</th>
<th>Relative (95% CI)</th>
<th>Absolute</th>
<th>Quality</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Modeling</td>
<td>Inherent uncertainty in model parameters and structure, indirect evidence</td>
<td>Imprecision in estimates not presented in paper</td>
<td>Only qualitative results presented</td>
<td>Annual screening</td>
<td>Biennial screening</td>
<td></td>
<td></td>
<td></td>
<td>Very Low</td>
<td>Biennial screening strategies reduced over-diagnosis compared to annual, “…but by much less than one half”</td>
</tr>
<tr>
<td>No. of Studies</td>
<td>Design</td>
<td>Limitations</td>
<td>Inconsistency</td>
<td>Imprecision</td>
<td>Other Considerations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>--------</td>
<td>-------------</td>
<td>---------------</td>
<td>-------------</td>
<td>----------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Cohort</td>
<td>Consistent direction of effect</td>
<td>Moderate</td>
<td>Probability varies both by patient risk (age, breast density) and radiologist, type of false positive (recall vs. biopsy), availability of prior exams</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Summary of Findings**

<table>
<thead>
<tr>
<th>Interventions</th>
<th>Effect</th>
<th>Quality</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screening Modality</td>
<td>Comparator</td>
<td>Relative (95% CI)</td>
<td>Absolute</td>
</tr>
</tbody>
</table>

10-year cumulative risk in U.S. Breast Cancer Surveillance Consortium:
- 39.8% for intermediate risk patient screened biennially,
- 51% for intermediate risk patient screened annually

Moderate 10 year cumulative risks identical for each interval for starting at age 40 vs 50, but extrapolated lifetime risks likely higher with starting at earlier age.

False positive rates higher with longer screening interval, but not enough to compensate for greater number of tests.
<table>
<thead>
<tr>
<th>No. of Studies</th>
<th>Design</th>
<th>Limitations</th>
<th>Inconsistency</th>
<th>Imprecision</th>
<th>Other Considerations</th>
<th>Screening Modality</th>
<th>Comparator</th>
<th>Relative (95% CI)</th>
<th>Absolute</th>
<th>Effect</th>
<th>Quality</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Modeling</td>
<td>Inherent uncertainty in model parameters and structure, indirect evidence</td>
<td>Imprecision in estimates not presented in paper</td>
<td>Only results from “exemplar” model presented in paper</td>
<td>Annual screening ages 50-69</td>
<td>Annual screening ages 45-69</td>
<td>Biennial screening ages 50-69</td>
<td>Biennial screening ages 45-69</td>
<td>57,000 extra false positives per 100,000 women (4000 extra false positive biopsies)</td>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>All Studies</td>
<td>Consistent direction, partly based on differences in setting, technology, patient populations</td>
<td>Imprecision, partly based on differences in setting, technology, patient populations</td>
<td></td>
<td></td>
<td>Annual screening ages 40-69</td>
<td>Annual screening ages 45-69</td>
<td>75,000 extra false positives per 100,000 women (5200 extra false positive biopsies)</td>
<td>100,000 extra false positives per 100,000 women (7000 extra false positive biopsies)</td>
<td>High (moderate for age effect)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of false positives increases with screening interval; cumulative 10-year probability high in U.S. Greater lifetime increase in younger women.
<table>
<thead>
<tr>
<th>No. of Studies</th>
<th>Design</th>
<th>Limitations</th>
<th>Inconsistency</th>
<th>Imprecision</th>
<th>Other Considerations</th>
<th>Interventions</th>
<th>Comparator</th>
<th>Relative (95% CI)</th>
<th>Absolute</th>
<th>Quality</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Modeling</td>
<td>Inherent uncertainty in model parameters and structure, indirect evidence</td>
<td>Imprecision in estimates not presented in paper</td>
<td>Only qualitative results presented</td>
<td></td>
<td>Screening Modality</td>
<td>Comparator</td>
<td></td>
<td></td>
<td>Low</td>
<td>High degree of uncertainty about appropriate utilities to use</td>
</tr>
</tbody>
</table>

**Quality Assessment**

**Summary of Findings**

**Interventions**

**Effect**

**Quality**

**Comments**
## Appendix Table H3. Summary Table for GRADE Assessments Across an Outcome—Key Question 3

<table>
<thead>
<tr>
<th>Quality Assessment</th>
<th>Summary of Findings</th>
<th>Interventions</th>
<th>Effect</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Screening Modality</td>
<td>Comparator</td>
<td>Relative (95% CI)</td>
</tr>
<tr>
<td><strong>Breast cancer mortality</strong></td>
<td></td>
<td>CBE</td>
<td>Mammography</td>
<td>Not reported or estimable from data</td>
</tr>
<tr>
<td>1</td>
<td>RCT</td>
<td>NA</td>
<td>High</td>
<td>Older mammography technology</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cumulative mortality at 8 years among screen-detected cancers 31.8% for CBE only, 14.5% for mammography only</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>95% CI not presented or estimable from data</td>
</tr>
<tr>
<td>1</td>
<td>Case-Control</td>
<td>Inherent high risk of bias, but appropriate adjustments</td>
<td>NA</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Life Expectancy</strong></td>
<td></td>
<td></td>
<td></td>
<td>Only U.S.-based study</td>
</tr>
<tr>
<td>0</td>
<td>All Designs</td>
<td></td>
<td></td>
<td>No age-related effects</td>
</tr>
<tr>
<td><strong>Overdiagnosis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>All Designs</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

H-17
<table>
<thead>
<tr>
<th>No. of Studies</th>
<th>Design</th>
<th>Limitations</th>
<th>Inconsistency</th>
<th>Imprecision</th>
<th>Other Considerations</th>
<th>Interventions</th>
<th>Comparator</th>
<th>Relative (95% CI)</th>
<th>Absolute</th>
<th>Quality</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Positives</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>RCT</td>
<td></td>
<td>High</td>
<td>Both non-U.S., developing country settings; CBE performed by trained lay health workers</td>
<td>CBE</td>
<td>No screening</td>
<td>False positives of 5.7% in India, 0.9% in Sudan</td>
<td>Moderate</td>
<td>Indirect, estimates not applicable to U.S. practice</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Cohort</td>
<td></td>
<td>U.S./Canadian results consistent for addition of CBE</td>
<td>CBE alone (Japan)</td>
<td>Mammography alone in women 40 and over</td>
<td>Approximately 55 extra false positives for each additional cancer detected in both U.S and Canadian studies</td>
<td>Moderate</td>
<td>Absolute effects of trade-off in sensitivity/specificity quite similar in both studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>All Designs</td>
<td>Variability across sites, comparators</td>
<td>CBE + mammography compared to mammography alone — consistent</td>
<td>CBE + mammography compared to mammography — precise estimates of absolute effect</td>
<td>CBE + mammography</td>
<td>Mammography alone</td>
<td>Approximately 55 extra false positives for each additional cancer detected</td>
<td>Moderate</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Quality-adjusted Life Expectancy |        |             |               |             |                      |               |            |                 |          |         |          |
| 0              | All designs |             |               |             |                      |               |            |                 |          |         | Very Low |

H-18
Appendix Table H4. Summary Table for GRADE Assessments Across an Outcome—Key Question 4

<table>
<thead>
<tr>
<th>Quality Assessment</th>
<th>Summary of Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Interventions</td>
</tr>
<tr>
<td>No. of Studies</td>
<td>Design</td>
</tr>
<tr>
<td>Breast cancer mortality</td>
<td>Cohort</td>
</tr>
<tr>
<td>1</td>
<td>Case-Control</td>
</tr>
<tr>
<td>2</td>
<td>Modeling</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

H-19
<table>
<thead>
<tr>
<th>No. of Studies</th>
<th>Design</th>
<th>Limitations</th>
<th>Inconsistency</th>
<th>Imprecision</th>
<th>Other Considerations</th>
<th>Interventions</th>
<th>Screen Modality</th>
<th>Comparator</th>
<th>Relative (95% CI)</th>
<th>Absolute</th>
<th>Effect</th>
<th>Quality</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>All Designs</td>
<td>No RCT data</td>
<td>Consistent direction of results across studies, variability in magnitude of effect</td>
<td>High</td>
<td>High degree of indirectness because of location (UK) or study design (modeling)</td>
<td>of breast cancer incidence compared to average-risk women</td>
<td>Annual mammography plus MRI in BRCA1/BRCA2 mutation carriers ages 25-69</td>
<td>of breast cancer incidence compared to average-risk women</td>
<td>Annual mammography alone in BRCA1/BRCA1 carriers ages 25-69</td>
<td>biennial screening in higher risk women 40-49</td>
<td>Relative mortality reduction with addition of MRI 38% in BRCA1 (vs. 14% for mammography alone), 38% for BRCA2 (vs. 16% for mammography alone)</td>
<td>2 fold for biennial screening in 40-49 year olds</td>
<td></td>
</tr>
</tbody>
</table>

**Life Expectancy**

<p>| 2 | Modeling | Underlying uncertainty about key model parameters | Qualitatively consistent results across different models | Not quantified; results presented only for “exemplar” model, no confidence intervals | Life expectancy is not directly estimable | Biennial screening in women 40-49 at variable levels of increased risk of breast cancer incidence | Biennial screening average-risk women ages 50-74 | At relative risks of 2 or more, false positives/death prevented equivalent to biennial screening in | Life expectancy gains not presented | Low | CISNET modeling analysis identified thresholds for relative risks above average where harm/benefit |</p>
<table>
<thead>
<tr>
<th>No. of Studies</th>
<th>Design</th>
<th>Limitations</th>
<th>Inconsistency</th>
<th>Imprecision</th>
<th>Other Considerations</th>
<th>Interventions</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>around estimates</td>
<td>Screening Modality</td>
<td>Comparator</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>compared to average risk women</td>
<td>Annual screening in women 40-49 at variable levels of increased risk of breast cancer incidence compared to average-risk women</td>
<td>Biennial screening in women 40-49 at variable levels of increased risk of breast cancer incidence compared to average-risk women</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Annual mammography plus MRI in BRCA1/BRCA2 mutation carriers ages 25-69</td>
<td>Annual mammography alone in BRCA1/BRCA1 carriers ages 25-69</td>
</tr>
</tbody>
</table>

### Quality Assessment

<table>
<thead>
<tr>
<th>Stage Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
</tr>
<tr>
<td>No. of Studies</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>MRI plus mammography</td>
</tr>
<tr>
<td>MRI</td>
</tr>
<tr>
<td>Annual mammography + MRI for BRCA1/BRCA2 carriers aged 25-69</td>
</tr>
</tbody>
</table>

**False Positives**

| 5 | Cohort | Moderate to high risk of bias | Consistent direction of effect across studies |
| 1 | Modeling | Underlying uncertainty about key model parameters | |

**Quality Assessment**

- **Interventions**
- **Effect**
- **Screening Modality**
- **Comparator**
- **Relative (95% CI)**
- **Absolute**
- **Quality**
- **Comments**
<table>
<thead>
<tr>
<th>No. of Studies</th>
<th>Design</th>
<th>Limitations</th>
<th>Inconsistency</th>
<th>Imprecision</th>
<th>Other Considerations</th>
<th>Screening Modality</th>
<th>Comparator</th>
<th>Relative (95% CI)</th>
<th>Absolute</th>
<th>Quality</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>All study designs</td>
<td>Consistent direction of effect across studies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>Consistent direction of effect, but quantitative estimates widely variable</td>
</tr>
<tr>
<td>Quality-adjusted Life Expectancy</td>
<td>2</td>
<td>Modeling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>Effect of parameters on quality-adjusted life expectancy difficult to estimate directly from published results</td>
</tr>
</tbody>
</table>
## Appendix Table H5. Summary Table for GRADE Assessments Across an Outcome—Key Question 5

<table>
<thead>
<tr>
<th>Quality Assessment</th>
<th>Summary of Findings</th>
<th>Effect</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of Studies</td>
<td>Design</td>
<td>Limitations</td>
</tr>
<tr>
<td>Breast Cancer Mortality</td>
<td>0</td>
<td>All Designs</td>
<td></td>
</tr>
<tr>
<td>Stage Distribution</td>
<td>1</td>
<td>Cohort</td>
<td>Risk of bias (nonrandomized)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>All Designs</td>
<td></td>
</tr>
<tr>
<td>Quality Assessment</td>
<td>Summary of Findings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>---------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interventions</td>
<td>Effect</td>
<td>Quality</td>
</tr>
<tr>
<td>No. of Studies</td>
<td>Design</td>
<td>Limitations</td>
<td>Inconsistency</td>
</tr>
<tr>
<td>Life Expectancy</td>
<td>Very Low</td>
<td>CISNET modeling analysis identified thresholds for relative risks above average where harm/benefit was equivalent to biennial screening ages 50-74, but no direct estimates on life expectancy; threshold RR substantially higher for annual compared to biennial in women 40-49, suggesting smaller incremental gain in life expectancy</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>All Designs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overdiagnosis</td>
<td>Very Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>All Designs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False Positives</td>
<td>1</td>
<td>All Designs</td>
<td></td>
</tr>
</tbody>
</table>

| Quality-Assessed Life Expectancy | 0 | All Designs | | | | | | Very Low |