Overdiagnosis due to screening mammography for women aged 40 years and over

Jacklyn, Gemma; Mcgeechan, Kevin; Houssami, Nehmat; Bell, Katy; Glasziou, Paul P.; Barratt, Alexandra

Published in:
Cochrane Database of Systematic Reviews

DOI:
10.1002/14651858.CD013076

Cochrane Library

Cochrane Database of Systematic Reviews

Overdiagnosis due to screening mammography for women aged 40 years and over (Protocol)

Jacklyn G, McGeechan K, Houssami N, Bell K, Glasziou PP, Barratt A

Jacklyn G, McGeechan K, Houssami N, Bell K, Glasziou PP, Barratt A.
Overdiagnosis due to screening mammography for women aged 40 years and over.
Cochrane Database of Systematic Reviews 2018, Issue 7. Art. No.: CD013076.
DOI: 10.1002/14651858.CD013076.

www.cochranelibrary.com

WILEY

TABLE OF CONTENTS

[Intervention Protocol]

Overdiagnosis due to screening mammography for women aged 40 years and over

Gemma Jacklyn[1], Kevin McGeechan[1], Nehmat Houssami[2], Katy Bell[1], Paul P Glasziou[3], Alexandra Barratt[1]

[1]Wiser Healthcare, Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia. [2]Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia. [3]Centre for Research in Evidence-Based Practice (CREBP), Bond University, Gold Coast, Australia

Contact address: Gemma Jacklyn, Wiser Healthcare, Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Edward Ford Building (A27), Sydney, NSW, 2006, Australia. gemma.jacklyn@sydney.edu.au, gemmajacklyn@hotmail.com.

Editorial group: Cochrane Breast Cancer Group.
Publication status and date: New, published in Issue 7, 2018.

Citation: Jacklyn G, McGeechan K, Houssami N, Bell K, Glasziou PP, Barratt A. Overdiagnosis due to screening mammography for women aged 40 years and over. *Cochrane Database of Systematic Reviews* 2018, Issue 7. Art. No.: CD013076. DOI: 10.1002/14651858.CD013076.

ABSTRACT

This is a protocol for a Cochrane Review (Intervention). The objectives are as follows:

To assess the effect of screening mammography for breast cancer on overdiagnosis in women aged 40 years and older at average risk of breast cancer.

BACKGROUND

Description of the condition

Breast cancer is the most common cancer in women worldwide and a leading cause of cancer death (Ferlay 2015). During the late twentieth century, screening mammography was introduced in high-income countries on the evidence that it reduced breast cancer mortality, without adequate consideration or knowledge of potential harms. During this time there was an increase in breast cancer incidence in women aged between 50 to 69 years. This was partly due to changes in risk factors such as alcohol intake, reproductive factors, obesity and hormone therapy use (Bray 2004; Jemal 2010); but also, as has now become apparent, because of widespread uptake of screening mammography and resulting overdiagnosis. In the context of cancer screening, overdiagnosis is the detection of cancer by screening that would never cause symptoms or harms in the absence of screening (Baker 2014; Marcus 2015; Welch 2010). Overdiagnosis involves the interaction between the biology of preclinical cancer and competing risks for mortality. Thus it can occur through the detection of non-progressive preclinical breast cancer, or through the detection of progressive, preclinical cancer in women with limited life expectancy.

Overdiagnosis is now acknowledged as the major harm of screening mammography (Independent UK Panel on Breast Screening 2012). It should not be confused with a false positive result: when a screening test detects an abnormality but with further investigation, no cancer is found. By contrast, overdiagnosis is a cancer diagnosis which is correct according to contemporary professional standards for pathology reporting and classification. As it is currently not possible to identify individuals who will benefit or be harmed by early detection and treatment, almost all cancer patients are offered treatment. Thus, to the extent that overdiag-

nosis of cancer occurs, it leads to overtreatment (Brawley 2017; Independent UK Panel on Breast Screening 2012) - unnecessary surgery, radiotherapy and other adjuvant therapy - that will not benefit individuals but may harm them through life-long physical and psychological consequences that can impact quality of life and life expectancy (Esserman 2014). As such, there is a scientific and public health imperative to establish the frequency of overdiagnosis. This is the evidence gap this review seeks to address.

As for all healthcare interventions, the benefit of cancer screening must be weighed against the potential harm. An earlier Cochrane Review of screening mammography for breast cancer quantified the benefit (Gøtzsche 2013). The review authors estimated that breast cancer-specific mortality was reduced by approximately 19% in randomised trials where women were invited to screening. They noted, however, that there was no significant benefit when the analysis was restricted to the best quality trials. In the review, they identified overdiagnosis and overtreatment as harms of screening that should be weighed against the benefit, but only assessed randomised controlled trials (RCTs).

Although early detection of breast cancer may lead to a mortality benefit, overdiagnosis is an unintended but inevitable risk of trying to detect pre-symptomatic cancer in age groups at significant risk of death from other causes. Thus, if women wish to be screened because they value the opportunity to reduce their risk of dying from breast cancer, that inevitably entails accepting an addition risk of diagnosis and treatment, including the risk of overdiagnosis and overtreatment. If screening catches many slow-growing cancers that would not cause symptoms or death, then the harms may outweigh the health benefits, both for populations and for individual women. Therefore, establishing the frequency of overdiagnosis is critically important to determine whether the net benefit justifies the resources required for screening, and to provide the best information possible to help healthy women weigh up the potential benefit versus the potential harm of participating in breast cancer screening.

To reliably estimate the effect of screening on breast cancer incidence we look to randomised controlled trials. Of the nine large trials undertaken, only three were suitable to accurately measure overdiagnosis (Miller 2000; Miller 2002; Zackrisson 2006). An independent meta-analysis of these trials suggests that 19% of screened women who are diagnosed with breast cancer experience overdiagnosis (Independent UK Panel on Breast Screening 2012). The authors of this analysis, however, emphasised the uncertainty around this estimates due to the scarcity of data, the small number of cases and the fact that not all women were followed to the end of their lives.

Of particular concern is that these three trials were undertaken between 1977 and 1988. Contemporary screening mammography is more sensitive than film, and the incidence of ductal carcinoma in situ (DCIS) has increased because of screening (Ernster 1996; Kerlikowske 2010; Van Steenbergen 2009; Virnig 2009) without a corresponding decrease in invasive breast cancer (Jacklyn 2017a; Jørgensen 2017; Sørum 2010). There are also differences between the screening mammography trials and international programmes in target age, screening technology, intervals, number of views and readers, and follow-up time. An analysis of non-randomised studies would have the advantage of evaluating current trends in breast cancer incidence and help quantify overdiagnosis in screening programmes to better inform individuals, clinical practice and policy. More contemporary estimates of overdiagnosis from non-randomised studies range from 0% up to 54% (Biesheuvel 2007; CTFPHC 2011; Myers 2015; Nelson 2016; Puliti 2011). The variation in these results may represent discrepancies in the choice of denominator as well as methodological differences, such as allowance for lead-time and volunteer bias (Biesheuvel 2007; de Gelder 2011).

Lead time is the amount of time screening advances the diagnosis of cancer. Lead time bias inflates survival statistics when early detection of disease does not extend lifetime, it only leads to an earlier diagnosis. Allowing for lead time is essential in studies of screening because it causes a temporary increase in cancer incidence which is a prerequisite for the intervention to work. The challenge is to separate the desirable increase in incidence due to advancement of the time of diagnosis from the undesirable increase due to overdiagnosis.

Volunteer bias (a type of selection bias) refers to the observation that people who choose to participate in screening tend to be different from those who do not volunteer. Women who choose to screen are generally healthier, have better health behaviours and their outcomes tend to be better because of this (Falk 2013; Puliti 2012). They may also represent the worried well; that is, people who do not have symptoms but are at higher risk of breast cancer (Moss 2006). Good quality studies such as randomised controlled trials help to avoid these biases, but they are difficult to control for in non-randomised studies.

Description of the intervention

Screening mammography involves an x-ray of both breasts (one-view or two-view, using film or digital mammography) in asymptomatic women to detect a suspicious abnormality and classify them at high or low risk of breast cancer. If a woman has an abnormality detected she may undergo one or a combination of further investigations such as clinical examination, diagnostic mammography, ultrasound and biopsy procedures. Women who receive a diagnosis of breast cancer are treated with surgery (breast conserving surgery or mastectomy), and may receive therapies (one of or a combination of radiotherapy, hormone therapy, chemotherapy, human epidermal growth factor receptor 2 (HER2) blockade and other biological therapy).

How the intervention might work

Screening mammography works via early detection and treatment. Rather than waiting for symptoms such as a lump to appear before treating breast cancer, we look for disease before these signs develop and thus advance in time the diagnosis. By shifting the incidence to an earlier stage, breast cancer should be more curable and require less intensive treatment, thereby reducing the incidence of cancers that first present as advanced disease (Morrison 1992). However, international data from non-randomised studies demonstrate a significant increase in early-stage disease (DCIS and localised breast cancer), with minimal or no decline in advanced breast cancer (regional and distant metastases) (Autier 2011; Autier 2017; Bleyer 2012; de Glas 2014; Harding 2015; Jacklyn 2017b; Jørgensen 2017; Kalager 2012; Lousdal 2014). We are finding that cancers behave in a variable way and do not necessarily lead to metastases and death (Welch 2010; Zahl 2008). Thus an unintended consequence of screening is the detection of preclinical cancers that are either 1) non-progressive or regressive (Lewison 1976; Zahl 2008) - that is, breast cancers that were never destined to present clinically or cause harm; or 2) breast cancers which would have progressed so slowly that women die from other causes before symptoms would have appeared.

Why it is important to do this review

One of the difficulties in making recommendations about screening mammography is that the benefits have been more extensively studied than the harms. The randomised trials were primarily designed to detect a reduction in disease-specific mortality, while overdiagnosis was generally overlooked. Furthermore, the absolute benefit to harm ratio may be finely balanced or may have become less favourable as treatments for breast cancer have improved (Birnbaum 2016). At the same time, greater mammographic sensitivity could increase benefit, while at the same time also increasing the harm from overdiagnosis. As the impact of these changes is unknown, it is increasingly important to carefully quantify the trade-offs in the current context using both randomised and non-randomised studies. Further, a recent systematic review of methods suggested that well-conducted ecological and cohort studies are the most appropriate approach for quantifying and monitoring overdiagnosis in cancer screening programmes (Carter 2015). Therefore a detailed and careful assessment of overdiagnosis is needed to inform current appraisals of screening mammography and guide decisions of consumers, clinicians and policymakers when weighing up the benefits and harms. To help provide an overview of this, we will discuss the outcomes of our analysis in the context of the existing Cochrane Review of the benefit from screening (Gøtzsche 2013).

OBJECTIVES

To assess the effect of screening mammography for breast cancer on overdiagnosis in women aged 40 years and older at average risk of breast cancer.

METHODS

Criteria for considering studies for this review

Types of studies

We will identify and evaluate all primary epidemiological studies that attempt to measure overdiagnosis resulting from screening mammography. We will include:

- randomised controlled trials, including cluster randomised controlled trials;
- cohort studies;
- case-control studies; and
- ecological studies.

We will also search for re-analyses of published incidence data that assess overdiagnosis.

We will exclude the following types of studies:

- systematic reviews that only report or combine quantitative estimates of overdiagnosis from included study types (randomised trials, cohort studies, case-control studies and ecological studies) that quantified overdiagnosis;
- non-systematic reviews;
- modelling studies; and
- pathological or imaging studies.

We have provided reasons for excluding certain study types in Appendix 1.

Types of participants

Women aged 40 years and older during the active screening period and at average (background population) risk for breast cancer.

Types of interventions

Exposure

Any form of screening mammography (one-view, two-views, film, digital).

Comparator

No screening mammography.

Types of outcome measures

There are different ways to calculate overdiagnosis and different estimates address different questions. We have found the approach of the Independent UK Panel on Breast Cancer Screening helpful (Independent UK Panel on Breast Screening 2012), and thus we will present several estimates of overdiagnosis using the panel's definitions.

Primary outcomes

- The probability of overdiagnosis from the perspective of an individual woman: the percentage risk of overdiagnosis among all cancers detected (both screen-detected and interval cancers) in women invited to screening. This outcome is the preferred method of the Independent UK Panel, Method C (Independent UK Panel on Breast Screening 2012). It reflects the probability that a breast cancer diagnosed during the active screening period represents overdiagnosis. This measure of overdiagnosis is most relevant to individual women considering screening mammography and addresses the question: if I am invited to attend screening and receive a breast cancer diagnosis, how likely is it to represent overdiagnosis?

Secondary outcomes

- Relative risk of overdiagnosis: the risk of excess breast cancer detection due to overdiagnosis in screened women compared to unscreened women. This answers the question: what is the percentage increase in risk of a breast cancer diagnosis in women invited to screening?
- Absolute risk of overdiagnosis: the probability that a woman invited to screening will be overdiagnosed. This measure answers the question: for every 1,000 women invited to screening, how many will be overdiagnosed during the active screening period?
- Percentage risk of overdiagnosis of screen-detected cancers: the probability that a screen-detected cancer represents overdiagnosis (method D, Independent UK Panel on Breast Screening 2012).
- Long-term percentage risk of overdiagnosis: the probability that a cancer diagnosed during the screening period and for the remainder of a woman's lifetime in women invited to screening will be overdiagnosed (method B, Independent UK Panel on Breast Screening 2012).
- The effect of screening mammography on the incidence of early- and advanced-stage breast cancer.

Search methods for identification of studies

Electronic searches

We will use different methods to search for randomised and non-randomised studies.

When identifying completed and ongoing randomised controlled trials, we will look to the reference list of existing systematic reviews on screening mammography for breast cancer (CTFPHC 2011; Gøtzsche 2013; Independent UK Panel on Breast Screening 2012; Nelson 2016). We do not intend to search trial registries as we are aware of only one ongoing trial (conducted in the UK) that aims to complete recruitment in 2026 (NCT01081288).

When identifying non-randomised controlled studies, we will search two databases:

- MEDLINE (via OvidSP; from 1946 to present) (Appendix 2);
- Embase (via OvidSP; from 1974 to present) (Appendix 3).

Searching other resources

Bibliographic searching

We will try to find further studies from reference lists of identified relevant non-randomised studies, trials and reviews. A copy of the full article for each reference reporting a potentially eligible study will be obtained. Where this is not possible, we will attempt to contact the study authors to provide additional information.

Grey searching

We will search grey literature for reports and conference proceedings in the following databases:

- COS Conference Papers Index through ProQuest;
- Grey Literature Report and Index, The New York Academy of Medicine;
- Health Services Research Projects in Progress (HSRProj);
- Mednar;
- NIH Research Portfolio Online Reporting Tools (RePORTER);
- OAlster;
- OpenGrey Repository;
- Papers First;
- ProQuest Dissertations and Theses Global.

Data collection and analysis

We will follow the recommended approach for data collection and management as documented in the *Cochrane Handbook for Systematic Reviews* (Higgins 2011).

Selection of studies

One review author (GJ) will screen titles and abstracts of all records retrieved by the searches for relevance. We will also use RobotAnalyst, a text mining application, to screen for potentially relevant titles and abstracts (Kontonatsios 2017; O'Mara-Eves 2015). Two review authors (GJ and AB) will independently assess full-text copies of potentially eligible articles. Studies published as abstracts only will be excluded. We will resolve any discrepancies through consensus or recourse to a third review author (KB) if we cannot reach agreement. We will list all studies excluded after full-text assessment in a 'Characteristics of excluded studies' table. Multiple publications from the same study will be included only once. If one of the review authors has contributed to a study, that author will not take part in reviewing the relevant manuscript or extracting data from the study. There will be no language restrictions and where possible articles will be translated. We will use Covidence software to screen titles and abstracts identified in our search, provide reasons for exclusions and generate a flow diagram (Covidence 2016).

Data extraction and management

Two review authors (GJ and AB) will independently extract data from the included studies and enter this information into a data extraction form using Covidence, then import the data into Review Manager 5 (RevMan 5) (RevMan 2014). We will pilot test a standardised data extraction form and modify it accordingly before use. Information collected will include study design, participants, setting, type of mammography, interval between screenings; number of screening rounds; duration of screening; co-interventions; adherence to screening; number of cancers identified, follow-up (including data sources, completeness, time frames, and sub-group analysis based on follow-up), management of lead time, calculation of overdiagnosis; sources of funding and other data relevant to the 'Risk of bias' assessments. We will resolve any discrepancies by consensus or, if we cannot agree, by consulting a third review author (KB). For those studies with more than one publication, we will extract data from all publications, with the most recent version considered as the primary reference.

Assessment of risk of bias in included studies

Two review authors (GJ and AB) will independently assess and judge the risk of bias for each included study. We will resolve any disagreements by discussion. Randomised controlled trials identified from the searches will be independently assessed for risk of bias using the Cochrane 'Risk of bias' tool (Higgins 2011). We will assess six forms of bias: selection, performance, attrition, detection, reporting and other types of bias. We will judge relevant trial characteristics as at low or high risk of bias following the guidelines outlined in Higgins 2011. If there is insufficient information to permit judgement, we will classify the domain as at 'unclear risk'.

We will independently evaluate non-randomised studies and judge them according to the Risk Of Bias In Non-randomised Studies - of Interventions (ROBINS-I) tool (Sterne 2016). We will address the eight bias domains outlined in ROBINS-I, including confounding, departures from intended interventions, missing data, selection of participants into the study, classification of interventions, measurement of outcomes, selection of the reported results and other types of bias. We will use the guidelines in ROBINS-I to classify the risk of bias as low, moderate, serious, or critical. If there is insufficient information to permit judgement, we will classify the domain as 'no information'.

Potential confounding factors when estimating overdiagnosis include:

- age;
- geographical location;
- socioeconomic factors;
- difference in baseline breast cancer incidence trends between groups;
- reproductive factors;
- hormone replacement therapy (HRT);
- postmenopausal obesity;
- alcohol consumption; and
- competing mortality risk from causes of death other than breast cancer.

Other biases

Two key areas that have been identified with risk of bias specific to studies estimating overdiagnosis are: (i) approaches used to obtain an unscreened comparator group and (ii) methods used to adjust for lead-time (Biesheuvel 2007). We will assess risk of bias (i) using existing tools as outlined above, and (ii) using our classification outlined below.

Lead time bias

Screening studies that do not allow for lead time overestimate overdiagnosis (Duffy 2008; Puliti 2012). The advance in time of cancer diagnosis due to screening (lead time) causes a temporary increase in cancer incidence. Once screening stops, the rate of detection of cancers in the previously screened group should be less than the unscreened group, compensating for the earlier increase in incidence due to lead time. This phenomenon is called the "compensatory drop". Eventually, if follow-up is longer than the distribution of lead times, the cumulative incidence in the two groups will increase at the same rate. Thus, to provide an unbiased estimate of overdiagnosis, there must be an allowance for the increase in incidence observed during the active screening period. Importantly, as Baker 2014 points out, lead time only relates to cancers that were destined to become symptomatic (progressive pre-clinical cancers). This is because the definition of lead time requires cancer to present clinically due to symptoms in the absence of screening. Thus an overdiagnosed cancer does not have a

lead time (alternatively, we can think of overdiagnosis as having an infinite lead time because it will never cause symptoms) Gøtzsche 2012.

Methodology for dealing with lead time is diverse and complex, especially in non-randomised studies (Baker 2014; Duffy 2008; Etzioni 2014; Gøtzsche 2009; Puliti 2012; Ripping 2017; Zahl 2014), but three main methods exist:

- compensatory drop method: long-term follow-up of incidence after screening stops to capture and allow for the effects of lead time (compensatory drop, also called "excess incidence", which includes cumulative incidence and early- vs advanced-stage methods); and
- statistical adjustment: using estimates of average lead time (and its distributions) from the literature to adjust for lead time; and
- steady state method: comparison of incidence (total or stage-specific incidence) in invited and control populations once screening has been established longer than the expected lead time (screening is in a "steady state").

The compensatory drop method is the preferred approach and has been used in randomised controlled trials (Miller 2014; Zackrisson 2006). Ecological and cohort studies of similar groups of screened and unscreened women that include cancers diagnosed both during the active screening period and after screening has stopped, also allow for lead time as they capture the compensatory drop. Many observational studies, however, use alternative methods or simply do not allow for lead time. Statistical adjustments may contribute to biased results because the magnitude of lead time is contested (Duffy 2013; Zahl 2014), and the distributions mostly unknown (Carter 2015; Davidov 2004). Estimates of mean lead time for breast cancer range from one to 3.3 years (Duffy 2008; Feinleib 1969; Jonsson 2005; Walter 1983; Zahl 2013), though estimates based on progressive preclinical cancers only are shorter, at around one year (Zahl 2013). Only a small percentage of breast cancer cases have been estimated to have a preclinical duration longer than five years (Shen 2001; Walter 1983; Zahl 2013). While incidence rates in screened women seem to return to the expected (control) rates within five years after screening stops (Miller 2014; Zahl 2012), we need to capture the full distribution of lead time, including the occasional progressive pre-clinical breast cancer with a lead time that is long. Lastly, when estimating overdiagnosis, the management of lead time should reflect individuals time preferences - women are likely to value more immediate consequences of overdiagnosis (e.g. harms experienced within the first few years following screening) differently to consequences that may occur well into the future (harms that occur 15 years after screening) - known as temporal discounting of future health outcomes (Drummond 2015). Therefore, for both randomised and non-randomised studies, we will classify lead time bias as outlined in Table 1.

Measures of treatment effect

For dichotomous outcomes (that is, breast cancer cases detected), we will use the extracted data from the original studies for both screened and unscreened groups to estimate cumulative incidence of newly diagnosed breast cancer, or annual incidence of newly diagnosed breast cancer, or both. We will then calculate the percentage risk of overdiagnosis with 95% confidence intervals. Any method that attempts to measure the percentage of overdiagnosis attributable to screening mammography should use the excess cancers allowing for lead time in the numerator of the percentage calculation.

Primary outcome[a]

- Percentage risk of overdiagnosis of all breast cancers detected in women who are invited to or participate in screening = (cumulative incidence in the screened group − cumulative incidence in the control group)/ total number of all breast cancers detected in screened women during the active screening period).

We will use a binomial distribution when calculating the variance (Baker 2014; Independent UK Panel on Breast Screening 2012). When conducting a meta-analysis of overdiagnosis it is important to carefully consider the denominator. Including cancers diagnosed after screening ends in the denominator dilutes the estimate of overdiagnosis, makes it dependent on the length of follow-up (Carter 2015), and does not reflect the value women may place on more immediate outcomes. Furthermore, the denominator should include screen-detected, interval and clinically detected breast cancers found in women who participate in screening for two reasons. Firstly, the ratio of screen to interval cancer detection increases as the time between the screening interval decreases. Thus excluding interval breast cancers provides an estimate of overdiagnosis that is dependent on screening frequency and applicable only to one particular programme. As different studies use different screening intervals, we need to account for this. This issue was identified by the Independent UK Panel on Breast Screening 2012, and underpins their view that this expression of overdiagnosis is the best way to present information to women who are considering participation in screening. Secondly, excluding some cancer cases from the screened group in randomised controlled trials may introduce selection bias, as those women who do not attend screening but in whom a cancer is diagnosed may differ with regards to breast cancer risk and detection rates compared to those women who do attend screening and receive a breast cancer diagnosis.

Estimates of overdiagnosis should include invasive breast cancer as well as DCIS, as DCIS is primarily detected by mammography and currently treated as cancer.

The percentage risk of overdiagnosis, regardless of method, cannot be directly compared to estimates of breast cancer mortality benefit such as relative risk. Both estimates must be converted to absolute numbers in order to provide a fair and standardised comparison.

Footnote

[a]The methodology we will use to calculate overdiagnosis is based on randomised controlled trials, where the intervention is an invitation to screening. Some non-randomised studies compare attenders versus non-attenders. Both lead-time and overdiagnosis will be smaller in an invited group compared to groups based on women who actually attend screening. The attenuation of the estimate in invited groups depends on the proportion of non-attenders. We will make estimates comparable between invited and screened groups by deattenuating trial results (Jacklyn 2016).

Secondary outcomes

- Relative risk: the ratio of cumulative incidence in the screened group to the cumulative incidence in the control group.
- Absolute risk of overdiagnosis = (cumulative incidence in the screened group − cumulative incidence in the control group)/ total number of women who are invited to screening).

This will be expressed as a natural frequency for every 1000 women screened. We will use a binomial distribution when calculating the variance.

- Percentage risk of overdiagnosis of screen-detected breast cancers = (cumulative incidence in the screened group − cumulative incidence in the control group)/ screen-detected breast cancers).

When calculating the variance of the percentage risk of overdiagnosis as a proportion of screen-detected breast cancers, the numerator includes the cumulative number of clinically detected cancers in the screened group which are not part of the denominator. To allow for the variability in the cumulative number of clinically detected cancers in the screened group we will compute bootstrap confidence intervals or asymptotic variances assuming the variables follow independent Poisson distributions.

- Long-term percentage risk of overdiagnosis = (cumulative incidence in the screened group − cumulative incidence in the control group)/ cancers diagnosed over the entire follow-up period in screened women)

We will use a binomial distribution when calculating the variance (Independent UK Panel on Breast Screening 2012).

- The effect of screening mammography on the incidence of early- and advanced-stage breast cancer.
    - We will use summary staging Young 2001 to classify cancer into one of four main categories:
        - ◇ DCIS: abnormal cells are contained in the lining of the breast duct but have not spread to nearby tissue;
        - ◇ localised: cancer is limited to the tissue of origin in which it began (primary site) with no evidence of spread;
        - ◇ regional: cancer has spread beyond the primary site to nearby lymph nodes, tissues, or organs; and
        - ◇ distant metastases: cancer has spread beyond the primary site to distant parts of the body.
    - We will calculate:
        - ◇ absolute increase in incidence of DCIS and localised breast cancer per 100,000 women in a screened population; and
        - ◇ absolute decrease in regional and distant metastases per 100,000 women in a screened population.
    - We will use one or more comparator populations:
        - ◇ current unscreened populations;
        - ◇ historical unscreened populations; or
        - ◇ younger or older unscreened populations, or both.

We will also examine trends in incidence rates of early-stage (DCIS or localised disease, summary stage I or II) and advanced-stage (regional or distant metastases, summary stage III or IV) breast cancer before and after implementation of screening.

Unit of analysis issues

If studies that incorporate clustering in their design are found, we will contact the study authors to try and obtain cluster specific counts from which we will calculate variances (adjusted for clustering). If these data are not available, we will apply the intraclass correlation coefficients estimated from the studies with available data to adjust the variances. If none of the clustered studies have cluster-specific data available then we will conduct sensitivity analyses assuming a range of intraclass correlation coefficients.

Dealing with missing data

If possible we will perform intention-to-screen analyses for randomised controlled trials by including all randomised women. When necessary we will contact authors of publications to ensure the completeness of data. If data remain unavailable, we will try to estimate the missing data using the available information. Where data are missing, we will assume that participants with missing data did not receive a breast cancer diagnosis. We will report the proportion of participants with missing outcome data and consider the potential impact of the missing data in our interpretation of the results.

For non-randomised studies we will exclude participants with missing data and perform a complete-case analysis. We will report the proportion of participants with missing outcome data and consider the potential impact of the missing data in our interpretation of the results.

Assessment of heterogeneity

We will assess clinical and methodological heterogeneity before any meta-analyses are preformed and judge whether results can be pooled. We will discuss and assess inconsistency across studies by visual inspection of the forest plots and, when relevant, assess

statistical heterogeneity by calculating the I^2 statistic with 95% confidence intervals (Higgins 2003). If the I^2 statistic is greater than 30%, we will explore causes of heterogeneity in sensitivity and subgroup analyses. We will not perform meta-analyses if we encounter unexplained heterogeneity that would give misleading results. Given the clinical and methodological diversity of non-randomised studies, we will conduct any meta-analyses using a random-effects model. We will consider using meta-regression if there are more than 10 studies in the meta-analysis.

Assessment of reporting biases

We will assess reporting bias, especially publication bias and outcome reporting bias, according to the recommendations of the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins 2011). We will explore publication bias by producing funnel plots and using Egger's test (Egger 1997). We will visually inspect the funnel plot if there are more than 10 studies in the meta-analysis.

Data synthesis

We will conduct separate meta-analyses (if appropriate) for randomised controlled trials and non-randomised studies. If re-analysis of the data provided by randomised controlled trials is required, we will perform an intention-to-screen analyses by including all randomised women invited to screening. Analysis by intention-to-screen will underestimate any real effect in women who attend screening. Thus, as a secondary analysis we will adjust the primary outcome measure for adherence to screening in individual trials using a previously published method (Jacklyn 2016). Essentially, this method divides the intention-to-screen effect by the proportion attending screening. To allow for heterogeneity we will use the DerSimonian and Laird random-effects method and will present 95% confidence intervals for both the intention-to-treat and adjusted estimates (DerSimonian 1986). In case of heterogeneity in the trial results ($P < 0.10$), we will explore possible causes. We will perform the analysis using RevMan 5 (RevMan 2014), and Microsoft Excel. For the cumulative and deattenuated meta-analyses, we will develop a spreadsheet and perform statistical analyses using Microsoft Excel software.

If clinical heterogeneity is not excessive in the non-randomised studies and pooling results is appropriate, then we will perform a meta-analysis using the DerSimonian and Laird random-effects method (DerSimonian 1986), using RevMan 5 (RevMan 2014). In order to observe changes in screening mammography and trends in overdiagnosis we will also conduct a cumulative meta-analysis (Lau 1992). We will add studies one at a time in order of the date of the active screening period and will summarise the results as each new study is included. For the cumulative meta-analyses, we will develop a spreadsheet and perform statistical analyses using Microsoft Excel software.

If it is possible to combine results, we will analyse and present the pooled estimates according to study design. We will stratify studies according to risk of bias and present three (stratified) analyses of the intervention effect incorporating:

- all studies;
- those at low or moderate risk of bias; and
- those at serious, critical or high risk of bias.

We will draw forest plots to display results across studies according to risk of bias, key study design features and date of the active screening period. In the event of important heterogeneity, we will not pool data across non-randomised studies and instead will provide a narrative review and present individual study findings in a summary table according to study design.

We will use the GRADE approach to assess the quality of the evidence for overdiagnosis separately for randomised and non-randomised studies if appropriate. We will use GRADEpro GDT software, GRADEproGDT 2015, and create a 'Summary of findings' table (GRADE Working Group 2004). Given the vast difference in quality of evidence the analysis by study type is likely to have, it may prove inappropriate to provide an overall assessment of the quality of the body of evidence.

Subgroup analysis and investigation of heterogeneity

If data are sufficient we will perform subgroup analyses by:

- age at intervention (age groups 40 to 49 years, 50 to 69 years, and $\geq$ 70 years);
- study design;
- country or geographical region of study;
- date of active screening period; and
- frequency of screening (screening interval).

Sensitivity analysis

If meta-analyses are feasible, we will conduct sensitivity analyses to determine whether findings are sensitive to decisions made during the review process such as our assessment of the level of clinical heterogeneity. We will evaluate the methods used to handle missing data by excluding these studies in a sensitivity analysis, and we will discuss the extent to which the missing data are likely to influence the results of the study.

ACKNOWLEDGEMENTS

We are grateful to Professor Russell P. Harris for his contribution to the methodology for assessing the risk of lead time bias.

REFERENCES

Additional references

Autier 2011

Autier P, Boniol M, Middleton R, Doré J-F, Héry C, Zheng T, et al. Advanced breast cancer incidence following population-based mammographic screening. *Annals of Oncology* 2011;**22**(8):1726–35.

Autier 2017

Autier P, Boniol M, Koechlin A, Pizot C, Boniol M. Effectiveness of and overdiagnosis from mammography screening in the Netherlands: population based study. *BMJ* 2017;**359**:j5224.

Baker 2014

Baker SG, Prorok PC, Kramer BS. Lead time and overdiagnosis. *Journal of the National Cancer Institute* 2014; **106**(12):dju346.

Biesheuvel 2007

Biesheuvel C, Barratt A, Howard K, Houssami N, Irwig L. Effects of study methods and biases on estimates of invasive breast cancer overdetection with mammography screening: a systematic review. *Lancet Oncology* 2007;**8**(12):1129–38.

Birnbaum 2016

Birnbaum J, Gadi VK, Markowitz E, Etzioni R. The effect of treatment advances on the mortality results of breast cancer screening trials: a microsimulation model. *Annals of Internal Medicine* 2016;**164**(4):236–43. DOI: 10.7326/M15-0754

Bleyer 2012

Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. *New England Journal of Medicine* 2012;**367**(21):1998–2005.

Brawley 2017

Brawley OW. Accepting the existence of breast cancer overdiagnosis. *Annals of Internal Medicine* 2017;**166**(5): 364–5.

Bray 2004

Bray F, McCarron P, Parkin DM. The changing global patterns of female breast cancer incidence and mortality. *Breast Cancer Research* 2004;**6**(6):229–39.

Carter 2015

Carter JL, Coletti RJ, Harris RP. Quantifying and monitoring overdiagnosis in cancer screening: a systematic review of methods. *BMJ* 2015;**350**:g7773.

Covidence 2016 [Computer program]

Veritas Health Innovation. Covidence systematic review software. Melbourne, Australia: Veritas Health Innovation, Accessed 31 May 2018.

CTFPHC 2011

Canadian Task Force on Preventive Health Care, Tonelli M, Connor Gorber S, Joffres M, Dickinson J, Singh H, et al. Recommendations on screening for breast cancer in average-risk women aged 40-74 years. *Canadian Medical Association Journal* 2011;**183**(17):1991–2001.

Davidov 2004

Davidov O, Zelen M. Overdiagnosis in early detection programs. *Biostatistics* 2004;**5**(4):603–13.

de Gelder 2011

de Gelder R, Heijnsdijk EAM, van Ravesteyn NT, Fracheboud J, Draisma G, de Koning HJ. Interpreting overdiagnosis estimates in population-based mammography screening. *Epidemiologic Reviews* 2011;**33**:111–21.

de Glas 2014

de Glas NA, de Craen AJM, Bastiaannet E, Op't Land EG, Kiderlen M, van de Water W, et al. Effect of implementation of the mass breast cancer screening programme in older women in the Netherlands: population based study. *BMJ* 2014;**349**:g5410.

DerSimonian 1986

DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986;**7**(3):177–88.

Drummond 2015

Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programmes*. 4th Edition. Oxford: Oxford University Press, 2015.

Duffy 2008

Duffy SW, Lynge E, Jonsson H, Ayyaz S, Olsen AH. Complexities in the estimation of overdiagnosis in breast cancer screening. *British Journal of Cancer* 2008;**99**(7): 1176–8.

Duffy 2013

Duffy SW, Parmar D. Overdiagnosis in breast cancer screening: the importance of length of observation period and lead time. *Breast Cancer Research* 2013;**15**(3):R41.

Egger 1997

Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;**315**(7109):629–34.

Ernster 1996

Ernster VL, Barclay J, Kerlikowske K, Grady D, Henderson IC. Incidence of and treatment for ductal carcinoma in situ of the breast. *JAMA* 1996;**275**(12):913–8.

Esserman 2014

Esserman LJ, Thompson IM, Reid B, Nelson P, Ransohoff DF, Welch HG, et al. Addressing overdiagnosis and overtreatment in cancer: a prescription for change. *Lancet Oncology* 2014;**15**(6):e234–e42.

Etzioni 2014

Etzioni R, Gulati R. Oversimplifying overdiagnosis. *Journal of General Internal Medicine* 2014;**29**(9):1218.

Falk 2013

Falk RS, Hofvind S, Skaane P, Haldorsen T. Overdiagnosis among women attending a population-based mammography screening program. *International Journal of Cancer* 2013; **133**(3):705–12.

Feinleib 1969
Feinleib M, Zelen M. Some pitfalls in the evaluation of screening programs. *Archives of Environmental Health: an International Journal* 1969;**19**(3):412–5.

Ferlay 2015
Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer* 2015;**136**(5):E359-86.

GRADE Working Group 2004
GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ* 2004;**328**(7454): 1490–4.

GRADEproGDT 2015 [Computer program]
McMaster University (developed by Evidence Prime, Inc). GRADEpro GDT: GRADEpro Guideline Development Tool [Software].. Version accessed 31 May 2018. Hamilton (ON): McMaster University (developed by Evidence Prime, Inc), 2015.

Gøtzsche 2009
Gøtzsche PC, Jørgensen KJ, Maehlen J, Zahl PH. Estimation of lead time and overdiagnosis in breast cancer screening. *British Journal of Cancer* 2009;**100**(1):219.

Gøtzsche 2012
Gøtzsche PC, Jørgensen KJ, Zahl PH, Mæhlen J. Why mammography screening has not lived up to expectations from the randomised trials. *Cancer Causes and Control* 2012;**23**(1):15–21.

Gøtzsche 2013
Gøtzsche PC, Jørgensen KJ. Screening for breast cancer with mammography. *Cochrane Database of Systematic Reviews* 2013, Issue 6. DOI: 10.1002/14651858.CD001877.pub5

Harding 2015
Harding C, Pompei F, Burmistrov D, Welch HG, Abebe R, Wilson R. Breast cancer screening, incidence, and mortality across US counties. *JAMA Internal Medicine* 2015;**175**(9): 1483–9.

Higgins 2003
Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;**327** (7414):557–60.

Higgins 2011
Higgins JPT, Green S, editor(s). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available from handbook.cochrane.org.

Independent UK Panel on Breast Screening 2012
Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. *Lancet* 2012;**380**(9855):1778–86.

Jacklyn 2016
Jacklyn G, Glasziou P, Macaskill P, Barratt A. Meta-analysis of breast cancer mortality benefit and overdiagnosis adjusted for adherence: improving information on the effects of attending screening mammography. *British Journal of Cancer* 2016;**114**(11):1269–76.

Jacklyn 2017a
Jacklyn G, Morrell S, McGeechan K, Houssami N, Irwig L, Pathmanathan N, et al. Carcinoma in situ of the breast in New South Wales, Australia: current status and trends over the last 40 year. *Breast* 2017;**37**:170–8.

Jacklyn 2017b
Jacklyn G, McGeechan K, Irwig L, Houssami N, Morrell S, Bell K, et al. Trends in stage-specific breast cancer incidence in New South Wales, Australia: insights into the effects of 25 years of screening mammography. *Breast Cancer Research and Treatment* 2017;**166**(3):843–54.

Jemal 2010
Jemal A, Center MM, DeSantis C, Ward EM. Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiology Biomarkers & Prevention* 2010;**19**(8): 1893–907.

Jonsson 2005
Jonsson H, Johansson R, Lenner P. Increased incidence of invasive breast cancer after the introduction of service screening with mammography in Sweden. *International Journal of Cancer* 2005;**117**(5):842–7.

Jørgensen 2017
Jørgensen KJ, Gøtzsche PC, Kalager M, Zahl PH. Breast cancer screening in Denmark: a cohort study of tumor size and overdiagnosis. *Annals of Internal Medicine* 2017;**166** (5):313–23.

Kalager 2012
Kalager M, Adami HO, Bretthauer M, Tamimi RM. Overdiagnosis of invasive breast cancer due to mammography screening: results from the Norwegian screening program. *Annals of Internal Medicine* 2012;**156** (7):491–9.

Kerlikowske 2010
Kerlikowske K. Epidemiology of ductal carcinoma in situ. *Journal of the National Cancer Institute Monographs* 2010; **2010**(41):139–41.

Kontonatsios 2017
Kontonatsios G, Brockmeier AJ, Przybyla P, McNaught J, Mu T, Goulermas JY, et al. A semi-supervised approach using label propagation to support citation screening. *Journal of Biomedical Informatics* 2017;**72**:67–76.

Lau 1992
Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine* 1992;**327**(4):248–54.

Lewison 1976
Lewison EF. Spontaneous regression of breast cancer. *National Cancer Institute Monograph* 1976;**44**:23–26.

Lousdal 2014
Lousdal ML, Kristiansen IS, Møller B, Støvring H. Trends in breast cancer stage distribution before, during and

after introduction of a screening programme in Norway. *European Journal of Public Health* 2014;**24**(6):1017–22.

Mandelblatt 2015
Mandelblatt JS, Cronin K, de Koning H, Miglioretti DL, Schechter CS, Stout N. Technical report: collaborative modeling of U.S. breast cancer screening strategies. Available at www.uspreventiveservicestaskforce.org/Page/Document/modeling-report-collaborative-modeling-of-us-breast-cancer-1/breast-cancer-screening1 2015.

Marcus 2015
Marcus PM, Prorok PC, Miller AB, DeVoto EJ, Kramer BS. Conceptualizing overdiagnosis in cancer screening. *Journal of the National Cancer Institute* 2015;**107**(4):djv014.

Miller 2000
Miller AB, To T, Baines CJ, Wall C. Canadian National Breast Screening Study-2: 13-year results of a randomized trial in women aged 50-59 years. *Journal of the National Cancer Institute* 2000;**92**(18):1490–9.

Miller 2002
Miller AB, To T, Baines CJ, Wall C. The Canadian National Breast Screening Study-1: breast cancer mortality after 11 to 16 years of follow-up: a randomized screening trial of mammography in women age 40 to 49 years. *Annals of Internal Medicine* 2002;**137**(5_Part_1):305–12.

Miller 2014
Miller AB, Wall C, Baines CJ, Sun P, To T, Narod SA. Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ* 2014;**348**:g366.

Morrison 1992
Morrison AS. *Screening in Chronic Disease*. New York: Oxford University Press, 1992.

Moss 2006
Moss SM, Cuckle H, Evans A, Johns L, Waller M, Bobrow L. Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years' follow-up: a randomised controlled trial. *Lancet* 2006;**368**(9552):2053-60.

Myers 2015
Myers ER, Moorman P, Gierisch JM, Havrilesky LJ, Grimm LJ, Ghate S, et al. Benefits and harms of breast cancer screening: a systematic review. *JAMA* 2015;**314**(15):1615–34.

NCT01081288
NCT01081288. Extending the National Health Service (NHS) breast screening age range (AgeX). clinicaltrials.gov/show/NCT01081288 Date first received: 5 March 2010.

Nelson 2016
Nelson HD, Pappas M, Cantor A, Griffin J, Daeges M, Humphrey L. Harms of breast cancer screening: systematic review to update the 2009 U.S. Preventive Services Task Force recommendation. *Annals of Internal Medicine* 2016;**164**(4):256–67.

O'Mara-Eves 2015
O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews* 2015;**4**:5.

Puliti 2011
Puliti D, Miccinesi G, Paci E. Overdiagnosis in breast cancer: design and methods of estimation in observational studies. *Preventive Medicine* 2011;**53**(3):131–3.

Puliti 2012
Puliti D, Duffy SW, Miccinesi G, de Koning H, Lynge E, Zappa M, et al. Overdiagnosis in mammographic screening for breast cancer in Europe: a literature review. *Journal of Medical Screening* 2012;**19 Suppl 1**:42–56.

RevMan 2014 [Computer program]
Nordic Cochrane Centre, The Cochrane Collaboration. Review Manager 5 (RevMan 5). Version 5.3. Copenhagen: Nordic Cochrane Centre, The Cochrane Collaboration, 2014.

Ripping 2017
Ripping TM, Ten Haaf K, Verbeek ALM, van Ravesteyn NT, Broeders MJM. Quantifying overdiagnosis in cancer screening: a systematic review to evaluate the methodology. *Journal of the National Cancer Institute* 2017;**109**(10):djx060.

Savage 2010
Savage CJ, Lilja H, Cronin AM, Ulmert D, Vickers AJ. Empirical estimates of the lead time distribution for prostate cancer based on two independent representative cohorts of men not subject to prostate-specific antigen screening. *Cancer Epidemiology and Prevention Biomarkers* 2010;**19**(5):1201–7.

Shen 2001
Shen Y, Zelen M. Screening sensitivity and sojourn time from breast cancer early detection clinical trials: mammograms and physical examinations. *Journal of Clinical Oncology* 2001;**19**(15):3490–9.

Sterne 2016
Sterne JAC, Higgins JPT, Reeves BC, on behalf of the development group for ROBINS-I. A tool for assessing Risk Of Bias In Non-randomized Studies of Interventions, Version 7 March 2016. Available from http://www.riskofbias.info (accessed 27 May 2016).

Sørum 2010
Sørum R, Hofvind S, Skaane P, Haldorsen T. Trends in incidence of ductal carcinoma in situ: the effect of a population-based screening programme. *Breast* 2010;**19**(6):499–505.

Van Steenbergen 2009
Van Steenbergen LN, Voogd AC, Roukema JA, Louwman WJ, Duijm LEM, Coebergh JWW, et al. Screening caused rising incidence rates of ductal carcinoma in situ of the breast. *Breast Cancer Research and Treatment* 2009;**115**(1):181–3.

Virnig 2009
Virnig BA, Shamliyan T, Tuttle TM, Kane RL, Wilt TJ. Diagnosis and management of ductal carcinoma in situ (DCIS). Evidence Report/Technology Assessment Number

185 (Prepared by the Minnesota Evidence-based Practice Center under Contract No. 290-02-10064-I). Rockville, MD. Agency for Healthcare Research and Quality (AHRQ) 2009:Publication No. 09-E018.

Walter 1983
Walter SD, Day NE. Estimation of the duration of a pre-clinical disease state using screening data. *American Journal of Epidemiology* 1983;**118**(6):865–86.

Welch 2010
Welch HG, Black WC. Overdiagnosis in cancer. *Journal of the National Cancer Institute* 2010;**102**(9):605–13.

Young 2001
Young JL Jr, Roffers SD, Ries LAG, Fritz AG, Hurlbut AA, editor(s). *SEER summary staging manual-2000: codes and coding instructions*. Bethesda: National Cancer Institute, 2001.

Zackrisson 2006
Zackrisson S, Andersson I, Janzon L, Manjer J, Garne JP. Rate of over-diagnosis of breast cancer 15 years after end of Malmö mammographic screening trial: follow-up study. *BMJ* 2006;**332**(7543):689–92.

Zahl 2008
Zahl PH, Mæhlen J, Welch HG. The natural history of invasive breast cancers detected by screening mammography. *Archives of Internal Medicine* 2008;**168**(21):2311–6.

Zahl 2012
Zahl PH, Mæhlen J. Overdiagnosis of breast cancer after 14 years of mammography screening. *Tidsskrift for den Norske Laegeforening* 2012;**132**(4):414–7.

Zahl 2013
Zahl PH, Jørgensen KJ, Gøtzsche PC. Overestimated lead times in cancer screening has led to substantial underestimation of overdiagnosis. *British Journal of Cancer* 2013;**109**(7):2014–9.

Zahl 2014
Zahl PH, Jørgensen KJ, Gøtzsche PC. Lead-time models should not be used to estimate overdiagnosis in cancer screening. *Journal of General Internal Medicine* 2014;**29**(9): 1283–6.

* *Indicates the major publication for the study*

ADDITIONAL TABLES

Table 1. Judgement of the risk of lead time bias in randomised and non-randomised studies that estimate overdiagnosis

Risk of bias	Criteria	Justification
Low	Management of lead time was based on a comparison of cumulative incidence in a screened and unscreened population after an ideal follow-up time after screening stops (≥ 10 years)	Carter 2015
Moderate	Management of lead time was based on a comparison of cumulative incidence in a screened and unscreened population after a sufficient follow-up time after screening stops (5 to 9 years)	Biesheuvel 2007; Miller 2014; Puliti 2011; Shen 2001; Walter 1983; Zahl 2012
Serious	Management of lead time was based on: • a statistical correction using directly observed data and a sufficient mean lead time; or • a statistical correction from a model which explicitly allowed for progressive and non-progressive preclinical cancers, and competing mortality.	Baker 2014; Zahl 2013
Critical	Management of lead time was based on: • an insufficient follow-up time after screening stops (< 5 years); or • a statistical correction from a model that did not allow for progressive and non-progressive cancer, and competing mortality; or	Baker 2014; Zahl 2013

Table 1. Judgement of the risk of lead time bias in randomised and non-randomised studies that estimate overdiagnosis (*Continued*)

	• no consideration of lead time.	
No information	Insufficient information on which to base a judgement about the risk of lead time bias	

APPENDICES

Appendix 1. Excluded study types

Excluded study types	Reason for exclusion
Systematic reviews that only report or combine quantitative estimates of other included study types (randomised trials, cohort studies, case-control studies and ecological studies) that quantified overdiagnosis	Systematic reviews will be excluded if they simply summarise studies that had each quantified overdiagnosis (for example, by combining data from several estimates of overdiagnosis)
Non-systematic reviews	A review of the literature that does not adhere to a protocol is subject to substantial biases and limitations and therefore insufficient for clinical decision making
Modelling studies	Modelling studies are useful for addressing research questions where direct evidence is difficult to obtain, such as with cancer screening, where the data collected or duration of follow-up may be limited. Models that aim to quantify overdiagnosis may attempt to simulate disease progression and outcomes in screened and unscreened populations. They are based on data from primary or secondary studies and can use multiple other data sources, or assumptions to extrapolate beyond the observed data. It is challenging to create valid and reliable models of cancer screening that estimate overdiagnosis due to key uncertainties in the available data, particularly with respect to the proportion of preclinical cancers which are non-progressive (Mandelblatt 2015). Further, calculation of overdiagnosis requires an estimate of the mean duration of lead time and the shape of the lead time distribution. Because lead time is unobservable in practice, assumptions have to be made, without ever being able to validate them fully (Savage 2010). As the rate of overdiagnosis is implicit in the lead time distribution assumptions of the model - particularly the tail shape of the lead time distribution and non-progressive lesions which have an infinite lead time (Baker 2014) - overdiagnosis cannot be

(*Continued*)

	estimated from models. For this reason, we will exclude modelling studies from our systematic review
Pathological or imaging studies	Studies that examine overdiagnosis resulting from non-progressive disease underestimate total overdiagnosis as they cannot account for overdiagnosis due to competing mortality

Appendix 2. MEDLINE (via OvidSP)

1	Breast Neoplasms/dg [Diagnostic Imaging]
2	exp Mammography/
3	mammogra$.tw.
4	1 or 2 or 3
5	exp Mass Screening/
6	Breast Neoplasms/pc [Prevention & Control]
7	(screen$ or (routine$ adj3 (test$ or check$ or diagnos$ or detect$))).tw
8	exp "Early Detection of Cancer"/
9	(early adj3 (detect$ or diagnos$)).tw.
10	5 or 6 or 7 or 8 or 9
11	exp Medical Overuse/
12	diagnostic errors/
13	False Positive Reactions/
14	(overdiagnos$ or overdetect$ or overtest$ or over-diagnos$ or over-detect$ or over-test$).tw
15	inciden$.tw.
16	11 or 12 or 13 or 14 or 15
17	4 and 10 and 16
18	exp Breast Neoplasms/

(*Continued*)

19	(breast adj6 cancer$).tw.
20	(breast adj6 neoplasm$).tw.
21	(breast adj6 carcinoma$).tw.
22	(breast adj6 tumo?r$).tw.
23	or/18-22
24	17 and 23
25	remove duplicates from 24
26	Animals/ not Humans/
27	25 not 26

Appendix 3. Embase (via OvidSP)

1	exp breast/
2	exp breast disease/
3	(1 or 2) and exp neoplasm/
4	exp breast tumor/
5	exp breast cancer/
6	exp breast carcinoma/
7	(breast$ adj5 (neoplas$ or cancer$ or carcin$ or tumo$ or metasta$ or malig$)).ti,ab
8	or/3-7
9	exp mammography/
10	mammogra$.tw.
11	9 or 10
12	breast cancer/pc [Prevention]

(*Continued*)

13	mass screening/
14	cancer screening/
15	(screen$ or (routine$ adj3 (test$ or check$ or diagnos$ or detect$))).tw
16	early diagnosis/
17	early cancer diagnosis/
18	(early adj3 (detect$ or diagnos$)).tw.
19	or/12-18
20	exp diagnostic error/
21	(overdiagnos$ or overdetect$ or overtest$ or over-diagnos$ or over-detect$ or over-test$).tw
22	inciden$.tw.
23	20 or 21 or 22
24	8 and 11 and 19 and 23
25	remove duplicates from 24
26	limit 25 to (human and embase)

CONTRIBUTIONS OF AUTHORS

GJ was the contact person with the editorial base.

GJ co-ordinated the contributions from the co-authors and wrote the final draft of the protocol.

All of the authors contributed to the methods sections.

GJ, KM and AB responded to the methodology and statistics comments of the referees.

All of the authors contributed to writing this protocol.

AB is the guarantor of the final review.

DECLARATIONS OF INTEREST

GJ has no known conflicts of interest.

KM has no known conflicts of interest.

NH has no known conflicts of interest.

KB has no known conflicts of interest.

PG has no known conflicts of interest.

AB has no known conflicts of interest.

SOURCES OF SUPPORT

Internal sources

- No sources of support supplied

External sources

- National Health and Medical Research Council (NHMRC) Australia, Australia.

Grant Number 1074626 and 1044904, Centres of Research Excellence Grant Number 1104136, Project Grant number 1062389