PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (http://bmjopen.bmj.com/site/about/resources/checklist.pdf) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

| TITLE (PROVISIONAL) | Long-term psychosocial consequences of false-positive screening mammography: a cohort study with 12-14-year follow-up in Denmark |
|---------------------|------------------------------------------------------------------------------------------------------------------|
| AUTHORS             | Gram, Emma Grundtvig; Siersma, Volkert; Brodersen, John                                                        |

VERSION 1 – REVIEW

| REVIEWER            | Nickel, Brooke                                                   |
|---------------------|------------------------------------------------------------------|
|                     | University of Sydney, School of Public Health                    |
| REVIEW RETURNED     | 31-Jan-2023                                                     |

GENERAL COMMENTS

This cohort study aimed to compare the long-term psychosocial consequences of mammography screening among women who were diagnosed with breast cancer, had a normal mammogram result and had a false-positive mammogram result. It is the first study to look at the long-term (12-14 years) psychosocial consequences of screening results on women. The study question is important and adds to the conversation about the benefits and harms associated with mammography screening for all women and the study methods seem robust, in particular the length of follow-up. I do, however, have a few comments and suggestions that I believe will help strengthen the manuscript.

Abstract:
- The Results could be more detailed and a bit clearer.
- The conclusions should state what the study adds e.g. the importance of this study for policy and practice and not just summarise the results.

Introduction:
- It would be good to make clear how long women typically screen for when discussing the cumulative probability of receiving a false-positive. It is also not clear why the countries differ so greatly, is this because of the ages and schedule of screening – please make clear.

Method:
- I assume women being screening in 2004 and 2005 could be at any stage of their screening process i.e. initial screeners or women who have been screened 5+ times? Please make clear. Also, did you adjust for whether women had received previous or subsequent false-positive results?

Results:
- Figure 1 – please update the title. Flowchart is too generic – please make specific.
- Suggest bolding or placing an asterisk beside p-values in Tables 1 and 2

Discussion:
- It might be good here or in one of the sections above to describe dose-response or be clearer to readers in what this really means in this context.
- It would have also been interesting to know women’s overall well-being or anxiety levels to see if this made a difference in outcomes of these women. Furthermore, as noted above if women had experience of more than one false-positive this might have more greatly impacted their long-term psychosocial consequences.
- Please add a reference to support the statement in relation to informed evidence-based choice in the Implications.
- While I agree with the final sentence I think you could add a bit more detail here.

**REVIEWER**
Crew, Katherine
Columbia, Medicine and Epidemiology

**REVIEW RETURNED**
24-Feb-2023

**GENERAL COMMENTS**
The authors conducted a prospective cohort study among women undergoing screening mammography in Denmark to determine the long-term psychosocial effects of false positive results. They found that some measures of psychosocial distress are higher among women after a false positive mammogram result compared to those with negative results. Strengths of the study include the long-term follow-up, relatively large sample size, and validated measures. The authors should address the following concerns:

1) Given the differential loss to follow-up among women with breast cancer, false positive and negative mammogram results, the authors should present some basic demographic information among the responders vs. non-responders to the follow-up surveys to determine whether there was any selection bias.

2) The women were stratified based upon results of their baseline mammogram in 2004-2005 (breast cancer vs. false positive vs. negative results). Is there any long-term follow-up data from the survey participants about whether they had subsequent breast cancer screening behaviors, false positive results or a breast cancer diagnosis during the 12-14 year follow-up period? If not, then this should be discussed as a limitation of the study.

3) Although there were some statistically significant differences in psychosocial measures comparing women with false positive and negative mammogram results, it is unclear whether these absolute differences are clinically meaningful. Is there any data from the literature about whether the absolute differences detected in these measures translate into clinically meaningful outcomes?

4) This screening study was conducted in Denmark, which may be more homogeneous in terms of racial/ethnic distribution. Therefore, these results may not be generalizable to patient populations which are more racially/ethnically diverse. The authors should note this as a study limitation.

5) The authors should clarify the overall message of their findings. Is it that women should be counselled on the potential benefits (breast cancer mortality reduction) and harms (overdiagnosis, false positive results, psychological harms, costs, etc.) of screening mammography and engage in shared decision-making with their healthcare providers? Should we conduct less frequent screening mammography (every 2-3 years vs. yearly) among
women at low-average risk for breast cancer or avoid over-screening among elderly women to minimize false positive results and the potential long-term psychosocial effects?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Dr. Brooke Nickel, University of Sydney

Comments to the Author:
This cohort study aimed to compare the long-term psychosocial consequences of mammography screening among women who were diagnosed with breast cancer, had a normal mammogram result and had a false-positive mammogram result. It the first study to look at the long-term (12-14 years) psychosocial consequences of screening results on women. The study question is important and adds to the conversation about the benefits and harms associated with mammography screening for all women and the study methods seem robust, in particular the length of follow-up. I do however have a few comments and suggestions while I believe will help strengthen the manuscript.

Abstract:
- The Results could be more detailed and a bit clearer.
  Thank you we have provided more statistical information about the results and added an example.
- The conclusions should state what the study add e.g. the importance of this study for policy and practice and not just summarise the results.
  That is a great suggestion, thank you. We have amended the conclusion in the abstract to include implications of findings rather than a summarization of findings.

Introduction:
- It would be good to make clear how long women typically screen for when discussing the cumulative probability of receiving a false-positive. It is also not clear why the countries differ so greatly, is this because of the ages and schedule of screening – please make clear.
  Thank you, that is a great point. We have added examples, from the reference, on what factors might affect the risk of a false-positive, e.g., age, breast density, and screening intervals.

Method:
- I assume women being screening in 2004 and 2005 could be at any stage of their screening process ie. initial screeners or women who have been screened 5+ times? Please make clear. Also, did you adjusted for whether women had received previous or subsequent false-positive results?
  Screening was introduced in the selected areas in 1990-1993. Women were included regardless of their stage in the screening process. We added this information to the method section under the subheading “Study population and survey administration”.
  We did not adjust for the stage of the screening process or previous or subsequent screening results. However, we adjusted for age, which could be argued to be a good proxy for the previous number of screenings as 80-90% of invited women in Denmark participate in the screening program. We have stated this more clearly in the discussion under the subheading “Strengths and limitations”. Hopefully, the discussion of the limitations associated with this will stand more clear now.
  Epidemiologically, subsequent screening (results) could be considered an intermediate variable and therefore should not be adjusted for.

Results:
- Figure 1 – please update the title. Flowchart is too generic – please make specific.
  We have changed the figure title to “Flowchart of invited women and response rate at each of the six assessments.”
- Suggest bolding or placing an asterisks beside p-values in Tables 1 and 2
  Thanks, we added asterisks for every p-value <= 0.01 in Tables 1 and 2.
Discussion:
- It might be good here or in one of the sections above to describe dose-response or be clearer to readers in what this really means in this context.
Thanks, we can clearly see how a clarification of the dose response will improve understanding of results. Thank you. We added two sentences to the summarization of results at the beginning of the discussion section.

- It would have also been interesting to know women’s overall well-being or anxiety levels to see if this made a difference in outcomes of these women. Furthermore, as noted above if women had experience of more than one false-positive this might have more greatly impacted their long-term psychosocial consequences.
Unfortunately, we did not have any pre-screening assessment. We agree that this would be of interest. However, the nature of the COS-BC supposes that the individual woman judges the item in relation to her previous level of anxiety or well-being, e.g. the items in part II are worded as “same as before” or more/less than before. This indirectly adjusts for the overall level of well-being or anxiety before the screening.
In other studies on the psychosocial consequences of screening, groups of false-positive and normal results did not seem to differ pre-screening (Rasmussen, J. F., et al. (2015). "Psychosocial consequences in the Danish randomised controlled lung cancer screening trial (DLCST)." Lung Cancer 87(1): 65-72.)

- Please add a reference to support the statement in relation to informed evidence-based choice in the Implications.
Thanks, there is so much important and interesting literature on this topic. Thank you for reminding us, we have added references.

- While I agree with the final sentence I think you could add a bit more detail here.
Thank you, we have altered the sentence to be more specific.

Reviewer: 2
Dr. Katherine Crew, Columbia
Comments to the Author:
The authors conducted a prospective cohort study among women undergoing screening mammography in Denmark to determine the long-term psychosocial effects of false positive results. They found that some measures of psychosocial distress are higher among women after a false positive mammogram result compared to those with negative results. Strengths of the study include the long-term follow-up, relatively large sample size, and validated measures. The authors should address the following concerns:

1) Given the differential loss to follow-up among women with breast cancer, false positive and negative mammogram results, the authors should present some basic demographic information among the responders vs. non-responders to the follow-up surveys to determine whether there was any selection bias.
Thank you for your comment. We agree that the differential loss to follow-up potentially could lead to bias. We have accounted for that potential bias by employing the Inverse Probability Weighing (mentioned on page 5). Therefore, potential differences between groups that could be presented in the proposed table would not result in bias in our analyses.
Clarification of how IPW accounts for potential differential attrition bias: We estimated probabilities for the COS-BC response being missing either due to death, emigration, or non-response at the various time points using logistic regression. In other words, we regressed an indicator of whether a particular
observation was observed on the observed characteristics of the subject: job, living alone, socioeconomic status, screening group, previous scores, and response at previous follow-ups as explanatory variables. As we did not have previous observations for the newly added impulsivity and empathy scale, the prediction models for these were based on previous responses on the part I sum score. From the logit function, we estimated the probability of response in this survey. We weighted observations by the inverse probability of observed scores. If the probability of being observed is for example 0.2 for a specific observed score Y, this observation will be weighted with the inverse probability, i.e., the specific score, Y, will be weighted as five.

Weights were estimated for all possible scores. However, weights were not applied to the model if the score was not observed. Therefore, the importance of certain, either low or high, scores was not overestimated using this method. Further, we checked weights to make sure that none were unrealistically large before being applied to weight data. Hereby, we account for informative missing data (differential attrition).

2) The women were stratified based upon results of their baseline mammogram in 2004-2005 (breast cancer vs. false positive vs. negative results). Is there any long-term follow-up data from the survey participants about whether they had subsequent breast cancer screening behaviors, false positive results or a breast cancer diagnosis during the 12-14 year follow-up period? If not, then this should be discussed as a limitation of the study.

Thank you, this was also mentioned by the other reviewer and we realize that we might not have been clear about this. The screening was introduced in the selected areas in 1990-1993. Women were included regardless of their stage in the screening process. We added this information to the method section under the “Study population and survey administration” subheading. We did not adjust for the stage of the screening process or previous or subsequent screening results. However, we adjusted for age, which could be argued to be a good proxy for the previous number of screenings as 80-90% of invited women in Denmark participate in the screening program. We have stated this more clearly in the discussion under the subheading “Strengths and limitations”. Hopefully, the discussion of the limitations associated with this will stand more clear now.

Epidemiologically, subsequent screening (results) could be considered an intermediate variable and therefore should not be adjusted for.

3) Although there were some statistically significant differences in psychosocial measures comparing women with false positive and negative mammogram results, it is unclear whether these absolute differences are clinically meaningful. Is there any data from the literature about whether the absolute differences detected in these measures translate into clinically meaningful outcomes?

Thank you. We have tried to relate the differences to relate the findings to real life, see for example page 7: “…the existential value scale: 0.61 (95%CI(0.15 to 1.06), p=0.009) (Table 2). This could be interpreted as 61% of the women who had false positive results answered ‘less’ or ‘more’ instead of ‘same as before’ on one item on the existential value scale, all other values fixed.”.

Scores or mean scores of questionnaires can be difficult to interpret. This is especially true when items are scored on an ordinal scale, where the distance between response categories and thus scores are not definitive.

The COS-BC does not have any established thresholds for (clinical) relevant differences, for example MID or MCID. This is due to the psychometric characteristics of the conceptualized psychosocial consequences of mammography screening and the measurement; there is not a linear relation between the latent variable and the sum score (visible in plotted location curves). We have therefore not determined a threshold for meaningful differences due to the following problems with non-linearity between latent variable and sum score: 1) we would need a threshold for each level of the latent variable, for example since a meaningful difference would be different for women who have high and low levels of the latent variable. 2) We would need a threshold for each of the scales. 3) Thresholds would depend on a number of variables such as age and health status (differential item functioning) and research has shown that patients do not perceive the same items to be equally important and
tend to disagree about the importance of the change (Liang, M., et al., Measuring Clinically Important Changes With Patient-Oriented Questionnaires. Medical Care, 2002. 40(4), Fischer, D., et al., Capturing the patient's view of change as a clinical outcome measure. Jama, 1999. 282(12): p. 1157-62.).

Despite the fact that we did not establish a MID or MCID, we argue that the COS-BC does already reflect meaningful differences due to the following arguments: 1) COS-BC has previously shown to be sensitive to changes in the status of the screening groups (Brodersen, J. and H. Thorsen, Consequences of Screening in Breast Cancer (COS-BC): development of a questionnaire. Scandinavian Journal of Public Health, 2008. 26: p. 251-256, Brodersen, J., H. Thorsen, and S. Kreiner, Validation of a Condition-Specific Measure for Women Having an Abnormal Screening Mammography. Value in Health, 2007. 10.). 2) The individual's own judgment of significant change is already an intrinsic factor of the COS-BC part II, as it is incorporated into the framing of the items: respondents are asked to rate a change from 'Same as before' on an ordinal scale. It should also be pointed out that responses of part I were tested to correlate with part II. Extrapolating this could indicate that the respondent also considers this change in the responses of part I. 3) Because the domains are verified as valid scales using Rasch analyses, and therefore we assume that we can use the numerical values to compare groups and identify meaningful differences, however, the interpretation of numbers does not immediately translate into clinical meaning.

To account for this, we included women with normal results and women diagnosed with breast cancer as benchmarks and interpreted mean differences as relevant differences.

4) This screening study was conducted in Denmark, which may be more homogeneous in terms of racial/ethnic distribution. Therefore, these results may not be generalizable to patient populations which are more racially/ethnically diverse. The authors should note this as a study limitation.

Thank you, that is a great point; we have added this to the discussion on page 10.

5) The authors should clarify the overall message of their findings. Is it that women should be counselled on the potential benefits (breast cancer mortality reduction) and harms (overdiagnosis, false positive results, psychological harms, costs, etc.) of screening mammography and engage in shared decision-making with their healthcare providers? Should we conduct less frequent screening mammography (every 2-3 years vs. yearly) among women at low-average risk for breast cancer or avoid over-screening among elderly women to minimize false positive results and the potential long-term psychosocial effects?

Thank you, we acknowledge that the section on implications is especially important in writing research. Therefore, we have amended the section to hopefully make two points clear: the information should be used to inform individual decision-making and it affects the benefit-harm ratio that often informs policy-making.

**VERSION 2 – REVIEW**

| REVIEWER       | Nickel, Brooke  |
|----------------|-----------------|
| University of Sydney, School of Public Health |                 |
| REVIEW RETURNED | 27-Mar-2023     |

| GENERAL COMMENTS | The authors have adequately responded to the editor and reviewer's previous comments and concerns. I feel that the manuscript is improved and have nothing further to add. |

| REVIEWER       | Crew, Katherine |
|----------------|-----------------|
| Columbia, Medicine and Epidemiology |                 |
| REVIEW RETURNED | 08-Apr-2023     |
The authors provided thorough responses to the prior critiques which have strengthened the paper.

VERSION 2 – AUTHOR RESPONSE