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)	Media coverage of the benefits and harms of testing the healthy: a protocol for a descriptive study
AUTHORS	O'Keeffe, Mary; Barratt, Alexandra; Maher, Christopher; Zadro, Joshua; Fabbri, Alice; Jones, Mark; Moynihan, Ray

VERSION 1 - REVIEW

REVIEWER	Bjørn Hofmann NTNU and University of Oslo, Norway
REVIEW RETURNED	01-Mar-2019

GENERAL COMMENTS	I am most thankful for the opportunity to review this manuscript, which I have read with great interest. This is a well written protocol for an interesting and important study. The inclusion and exclusion criteria are well accounted for. The protocol is of interest to other researchers who would like to supplement or complement this study. However, information on data extraction, coding and analysis may be to scarce for other researchers to apply. More information on this may be provided in the subsequent presentation of the results. However, if possible, readers would benefit from somewhat more detailed information on these issues. Some minor details, which can be considered for improving the publicatiton: The reason for selecting the five specific tests instead of random sample may need elaboration, as it is not clear that the time trends will give valuable information. However, the higher volume, will give more comprehensive information, as the authors state. False alarms should be added to the harms in the inclusion criteria. See for example doi: 10.1136/bmj.j3314. The problem of early detection is not only promoted by the media as mentioned by the authors, but also to researchers and health professionals (doi: 10.1136/bmj.j2102). The authors define overdiagnosis in the following manner: “Overdiagnosis occurs when people receive a diagnosis that does more harm than good.” The reference given for this only partly warrants this definition, while another article by Carter may give a better account of this conception of overdiagnosis. However, this definition has been criticized in http://dx.doi.org/10.1136/medethics-2015-102928. Consider for example, a person tested having 65 repetitions on the Huntington gene and who is given this information. He may be more harmed

	than benefited from the diagnosis (e.g., due to anxiety and procreative concerns). However, it is not clear that the person would be overdiagnosed. A more careful formulation could easily avoid this conundrum. In sum, this is a well-designed protocol for an interesting and important study that warrants publication. The results of the project will be of great interest to a broad audience.

REVIEWER	Kim Walsh-Childers University of Florida USA
REVIEW RETURNED	02-Mar-2019

GENERAL COMMENTS	The paper describes the protocol for a study that certainly is worthy of doing, and in general, the manuscript is well written and the research methods well thought out. I would offer just a few suggestions for strengthening the paper. 1. The authors need to explain why/how the dates January 2016-January 2019 were chosen. What is significant about this three-year period? The paper says the date range "aligns with our five justifications," but it's really not clear what this means. 2. Although there has been relatively little quantitative assessment of news coverage of overtreatment/overdiagnosis, there is at least one published study of U.S. news framing of overtreatment, which might be worth the authors' examination. (Walsh-Childers & Braddock, 2014). https://www.ncbi.nlm.nih.gov/pubmed/23537348 3. It would be helpful to know how commonly the 5 chosen tests are actually used. How many people are likely to undergo these tests each year, and to what extent does the use of the tests vary from one country to another. How many people are using the Apple Watch AFib detector? 4. It also would be helpful to have some sense of the extent to which these tests are being marketed to doctors and directly to consumers (in countries where this is allowed). 5. Is there evidence that use of the liquid biopsy tests are leading to overdiagnosis/overtreatment? The standard for overdiagnosis with a cancer test is somewhat higher, I think, than for dementia because many/most cancers can actually be treated with some degree of success. 6. The researchers intend to track "online" news media. The definition of "online news" varies widely from one perspective to the next, which means that the authors are going to have to define this much more clearly. For instance, are online newsletters considered news? Online-only sources such as Vox, Huffington Post, BuzzFeed, Yahoo, etc? What about AppleNews? What about blogs, which may include comment and opinion? 7. Why are wire and syndicated stories being excluded? This would account for a very significant portion of the news in some countries where, for instance, many news organizations rely on reporting from the Associated Press, Agence France Press, etc. Thus excluding these stories seems odd if the authors are trying to get a sense of how much news content audiences are seeing related to these tests and how the tests are being covered. 8. What standard will be used to ensure acceptable intercoder reliability? It's important to note that intercoder reliability (e.g. Cohen's kappa) should be calculated for each individual variable, rather than for the overall sample. What percentage of

	stories will be double-coded so that intercoder reliability can be calculated? 9. It would seem to be of interest to know whether coverage of the various tests differs across countries as well as over time.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name: Bjørn Hofmann

Institution and Country: NTNU and University of Oslo, Norway

Please state any competing interests or state ‘None declared’: No competing interests

Please leave your comments for the authors below

I am most thankful for the opportunity to review this manuscript, which I have read with great interest.

Thank you.

This is a well written protocol for an interesting and important study. The inclusion and exclusion criteria are well accounted for. The protocol is of interest to other researchers who would like to supplement or complement this study. However, information on data extraction, coding and analysis may be to scarce for other researchers to apply. More information on this may be provided in the subsequent presentation of the results. However, if possible, readers would benefit from somewhat more detailed information on these issues.

We thank the reviewer for the comment. We have now added our data extraction (coding) tool (Table 1). Note that the current coding tool is a draft and may be modified once we pilot test it with a sample of the included media stories.

We have now extended our data analysis section (Page 11) to include specific details on our planned analyses. Since receiving the reviewer comments, we invited a biostatistician (Dr Mark Jones) on to the study protocol to provide advice on data analysis. He has now been added as a co-author, and will be involved in the study going forward.

The data analysis now reads as:

“Descriptive statistics (means, standard deviations (SD), counts and percentages) will be used to summarise the extracted data (e.g number of stories, number of countries, number reporting benefits and harms, etc). Analysis will be performed separately for each test. Categorical data analysis will be used to investigate potential associations between overall impression of the media reports and explanatory variables including conflicts of interest, time, and type of media. We plan to use multinomial logistic regression where the dependent variable is overall impression and a neutral impression is the reference category. We will report odds ratios and 95% confidence intervals for negative impressions and positive impressions associated with the independent explanatory variables (as referred to above). Analysis will be conducted separately for each test. We will use chi-square tests to compare the distribution of categories of overall impressions across the 5 tests. We will also outline at least one example of a media story for each test in the results section.”

Some minor details, which can be considered for improving the publication:

The reason for selecting the five specific tests instead of random sample may need elaboration, as it is not clear that the time trends will give valuable information. However, the higher volume, will give more comprehensive information, as the authors state.

Thank you for this thoughtful response and we, too, have given considerable time to deliberating whether to take a random sample or proceed with our current approach. There are a number of reasons for our choice of these five specific tests:

1. We are interested in tests that have high potential for contributing to overdiagnosis. There are evidence-based concerns about the potential for overdiagnosis of our conditions of interests: atrial fibrillation, cancer, and dementia. Furthermore, there are concerns about the use of mobile monitoring devices and tests for disease biomarkers among the healthy, and how these may contribute to OD in these conditions. Taking a random sample of media coverage may mean we capture tests that do not carry the same concerns.

2. As well as high potential for overdiagnosis, we want to focus on tests that have potential for broad use or uptake. We have two potential indicators for this in our inclusion criteria: FDA approval and notable media coverage. While we are unable to specify how many people currently use, or are likely, to use some of these tests, evidence of FDA approval and notable media coverage may be good indicators of potential future population use.

3. By focusing on the coverage of specific tests over time, we can assess the nature of the media coverage, and how it changes over time.

In summary, by including FDA approved tests that have notable media coverage and have high potential for overdiagnosis, we feel we are focusing on tests that may have the largest influence on readers over time, and the highest potential to affect human health.

False alarms should be added to the harms in the inclusion criteria. See for example doi: 10.1136/bmj.j3314.

Thank you, we have now added false alarms on Page 8. We have also added extra harms.

“Inclusion and exclusion criteria

We will include media stories referring to any of our five target tests for the corresponding conditions of interest. Media stories will be included if they refer to any benefits (e.g. early detection of the condition, early treatment of the condition, prevention of the condition, saves lives) or harms of the test (e.g. overdiagnosis, inappropriate diagnostic testing, misdiagnosis, false alarms, false positives, false negatives, unnecessary and/or harmful treatment, psychological distress, healthy anxiety, costs). We will exclude media stories that only focus on tests for symptomatic people or people who already have the condition of interest (e.g. mammography for monitoring the progression of breast cancer), media stories about patent approval or business issues only, press releases, conference proceedings, trade journal reports, and scholarly journal articles. We will first pilot our screening process. Depending on the results of the pilot, we may add additional criteria, or provide more detail on the current inclusion and exclusion criteria.”

The problem of early detection is not only promoted by the media as mentioned by the authors, but also to researchers and health professionals (doi: 10.1136/bmj.j2102).

We have read and cited the article mentioned above and we have added sentences to both our introduction and discussion to reflect this.

First paragraph of introduction, Page 3:

The increasing popularity of testing is indicative of recent enthusiasm for early detection,2

Third paragraph of introduction, Page 3:

“Sustained promotion to the public and patients of the importance of early detection and testing, including via the media, is considered another driver of overdiagnosis.11 Uncritical coverage of new tests, without consideration of their potential downsides, contributes to the general lack of knowledge about the potential harms of getting tested when healthy. In fact, research has shown that only a small proportion of people are knowledgeable about overdiagnosis. This includes individuals offered tests where the potential for overdiagnosis is high.22 23 As such, patients (and clinicians) overestimate the benefits of testing, while underestimating the harms.24 25”

Discussion, Page 12:

“While other drivers, including research and professional prominence given to early detection,2 are important, sustained media coverage is likely a powerful source of influence of public attitudes towards new tests.”

The authors define overdiagnosis in the following manner: “Overdiagnosis occurs when people receive a diagnosis that does more harm than good.” The reference given for this only partly warrants this definition, while another article by Carter may give a better account of this conception of overdiagnosis. However, this definition has been criticized in https://protect-au.mimecast.com/s/z9f5C2xZYvCPv3q5C1uzMO?domain=dx.doi.org. Consider for example, a person tested having 65 repetitions on the Huntington gene and who is given this information. He may be more harmed than benefited from the diagnosis (e.g., due to anxiety and procreative concerns). However, it is not clear that the person would be overdiagnosed. A more careful formulation could easily avoid this conundrum.

We have amended the text to represent the uncertainty.

We have now added “Although an exact definition of overdiagnosis remains the subject of debate, particularly in the context of non-cancer conditions, overdiagnosis can be considered to occur when persons are labeled with a technically correct diagnosis that does not improve health outcomes.”

We have added two supporting references:

Bell KJ, Doust J, Glasziou P, et al. Recognizing the Potential for Overdiagnosis: Are High-Sensitivity Cardiac Troponin Assays an Example? Annals of internal medicine 2019;170(4):259-61.

Carter SM, Degeling C, Doust J, et al. A definition and ethical evaluation of overdiagnosis. Journal of medical ethics 2016;42(11):705-14.

In sum, this is a well-designed protocol for an interesting and important study that warrants publication. The results of the project will be of great interest to a broad audience.

We thank the reviewer for very useful feedback.

Reviewer: 2

Reviewer Name: Kim Walsh-Childers

Institution and Country: University of Florida, USA

Please state any competing interests or state ‘None declared’: None

Please leave your comments for the authors below

Reviewer 2:

BMJopen-2019-029532

The paper describes the protocol for a study that certainly is worthy of doing, and in general, the manuscript is well written and the research methods well thought out.

Thank you.

I would offer just a few suggestions for strengthening the paper.

1. The authors need to explain why/how the dates January 2016-January 2019 were chosen. What is significant about this three-year period? The paper says the date range “aligns with our five justifications,” but it’s really not clear what this means.

Our date range is mainly a pragmatic decision.

1. We are dealing with five tests that are showing notable media coverage since 2016 following various FDA approvals. Our data range therefore aligns with an appropriate time period.

2. We have broad criteria in terms of not restricting by country or type of media, extending the search range any further than this, may make our study impossible to complete.

3. Further to this, previous studies on media coverage have often used a similar length of data range. For example, one study (Walsh-Childers & Braddock, 2014) examining the news coverage of overtreatment (mentioned by the reviewer in the next point) used a similar size search range (January 1, 2007, through December 31, 2010).

Depending on time to complete this study, we may update the search range to include extra 2019 months.

In summary, we think a 3 year date range is large enough to examine whether the coverage (both volume and direction of reporting on benefits and harms) changes over time.

2. Although there has been relatively little quantitative assessment of news coverage of

overtreatment/overdiagnosis, there is at least one published study of U.S. news framing of

overtreatment, which might be worth the authors' examination. (Walsh-Childers & Braddock, 2014). https://www.ncbi.nlm.nih.gov/pubmed/23537348

We have read this study and have now referred to it in our introduction on Page 4.

"Media coverage of overtreatment has been examined in one study35 which examined the framing of medical overtreatment in US newspapers from January 2007 to December 2010. The study found that the media focussed on the harms of overtreatment relating to cancer, but the overall media coverage may have implied that overtreatment was not seen as an issue across other conditions.

To date, however, no studies have examined media coverage of new tests with significant potential for overdiagnosis."

3. It would be helpful to know how commonly the 5 chosen tests are actually used. How many

people are likely to undergo these tests each year, and to what extent does the use of the tests

vary from one country to another. How many people are using the Apple Watch AFib detector?

While we welcome the reviewer comment, unfortunately providing detailed statistics about the current use, or potential use of all these tests is beyond the scope of the current protocol and study, as it would entail significant investigation to obtain accurate and reliable data.

The notable media coverage and FDA approvals for the five tests suggest widespread use is occurring and/or may be coming in the US, and probably beyond. For example the new Apple watch with ECG is now available in 19 countries. According to a data analytics firm called Canalys, Apple shipped an estimated 3.5 million Apple Watches worldwide during the second quarter of 2018. The firm stated that this was 30% increase on 2017. Likewise 3D mammography is widely used in the US and internationally where mammography screening is established.

4. It also would be helpful to have some sense of the extent to which these tests are being

marketed to doctors and directly to consumers (in countries where this is allowed).

We agree that media is not the only driver of this type of promotion. Hofmann et al in their BMJ paper documented the surge in publications in academic journals relating to the benefits of early detection, use of mobile monitoring devices, and innovative technologies to track disease biomarkers. These publications contribute to the promotion of early detection. It seems likely that direct to consumer marketing does also. We have added a comment to address this point in the Discussion. To provide further information about these other drivers however is beyond the scope of the current study.

Page 12:

"While other drivers, including research and professional prominence given to early detection,2 are important, sustained media coverage is likely a powerful source of influence of public attitudes towards new tests."

5. Is there evidence that use of the liquid biopsy tests are leading to overdiagnosis/overtreatment?

The standard for overdiagnosis with a cancer test is somewhat higher, I think, than for dementia

because many/most cancers can actually be treated with some degree of success.

We have now added extra text to demonstrate the concerns about liquid biopsy.

Page 6:

"While liquid biopsy was initially designed for monitoring patients with cancer, there seems to be increasing interest in its use for the early detection of cancer, and that the test may eventually be used for routinely screening people and detecting cancers before they cause symptoms. In fact, some groups are currently conducting studies to see whether the test can pick up tumors in seemingly

cancer-free individuals. There is a lot of uncertainty surrounding the effectiveness of liquid biopsy for both early detection, and improvement of cancer treatment.43

There are also concerns that the detection of circulating tumour DNA cells in asymptomatic populations could lead to overdiagnosis.44 The concerns are linked to findings that circulating tumour DNA cells and cancer-related mutations have been detected in healthy individuals who never go on to develop a cancer.45 46 It has also been mentioned that the cancer-related proteins used by liquid biopsy can reflect tissue damage common in inflammatory conditions like arthritis, in the absence of cancer.47"

We have also extended the blood biomarker tests for the early detection of dementia section to include a new publication expressing concerns about overdiagnosis.

Page 6-7

"In January 2019, the Cochrane Dementia and Cognitive Improvement Group published a commentary expressing their concerns about the increasing use of biomarkers tests in dementia.51 In their commentary, they referred to research demonstrating that up to 60% of healthy people over 80 years could be labelled as having dementia under new disease defintions, even though they may never develop clinical symptoms.53 They also stated that reducing dementia to positive amyloid biomarkers is "an open invitation to overdiagnosis" .51 Further to this, they refer to the data documenting the psychological, social and legal harms of overdiagnosing or overpredicting dementia.54 Finally, they expressed concerns about the lack of data validating the proposed biomarkers for dementia.55"

We have also extended the Apple Watch series 4 ECG sensor for the early detection of atrial fibrillation section to include new publications expressing concerns about overdiagnosis.

Page 7

Many concerns about been expressed about testing the healthy for AF, and it has been suggested that overdiagnosis is "only a matter of time" with the Apple Watch.65-67 There is a concern regarding the poor specificity of testing methods for AF.66 Furthermore, AF has a low prevalence68 and screening the healthy could potentially lead to harms in the form of overdiagnosis and overtreatment.67 In fact, some researchers state that the Apple watch specifically could lead to a misdiagnosis of AF in nearly 1 million people for every 10 million screened. This may lead to harms from overtesting, bleeding from unnecessary anticoagulation, and anxiety due to having a cardiac diagnosis.67 There also seems to be a lack of knowledge around the natural history of AF. For example, there is uncertainty surrounding the outcome of untreated stroke risks, so the net benefit of treating AF with anti-coagulants is unclear. Finally, while screening for AF leads to increased detection, office visits, and prescriptions for anticoagulants,69 there is still uncertainty around its effects on patient outcomes.

Finally, we added a new line to the 3D mammography section.

Page 5

In March 2019, the FDA announced new policies to change current mammography standards in the US. The proposed changes aim to increase the use of 3D mammography screening.

6. The researchers intend to track “online” news media. The definition of “online news” varies widely from one perspective to the next, which means that the authors are going to have to define this much more clearly. For instance, are online newsletters considered news? Online only sources such as Vox, Huffington Post, BuzzFeed, Yahoo, etc? What about AppleNews? What about blogs, which may include comment and opinion?

We have now updated our explanation of media coverage.

Page 8:

“Our searches will cover all of the following media coverage: newspapers, major world publications, blogs, magazines, broadcast and podcast transcripts, wire feeds/services, and webnews. These are named categories within the LexisNexis and Proquest databases.”

We will now supplement our search with a Google News search, in line with a previous media study (doi: 10.1080/10810730.2016.1266717) See added text:

Page 8

“In line with a previous study on media coverage of medicine,70 we will supplement this database search with a Google News search, reviewing the first 20 pages of each Google search result.

7. Why are wire and syndicated stories being excluded? This would account for a very significant portion of the news in some countries where, for instance, many news organizations rely on

reporting from the Associated Press, Agence France Press, etc. Thus excluding these stories seems odd if the authors are trying to get a sense of how much news content audiences are seeing related to these tests and how the tests are being covered.

Thank you for this comment. This was unclear in the initial protocol.

Wire feeds will be included, but we will be excluding press releases.

We have now updated our explanation of how we will handle syndicated stories. We will include all the stories, but syndiciated stories will be coded once.

Page 8:

"Syndicated studies will be included but will only be coded once. For example, if there are 10 media stories about the Apple Watch where the same or extremely similar story has been run across multiple media outlets, we will code this story once, but include the number 10 as the number of media reports about the Apple Watch."

8. What standard will be used to ensure acceptable intercoder reliability? It's important to note that intercoder reliability (e.g. Cohen's kappa) should be calculated for each individual variable, rather than for the overall sample. What percentage of stories will be double-coded so that intercoder reliability can be calculated?

Every story will be double screened and double coded by pairs of independent reviewers. If there are disagreements between the 2 independent reviewers, they will be resolved by consensus with a third reviewer.

For this reason, we do not feel we need to perform Cohen's Kappa.

9. It would seem to be of interest to know whether coverage of the various tests differs across countries as well as over time.

Yes we also think this would be interesting. However, we will be limited by our inclusion of English language stories only. For example, if we find that a particular test is covered widely by the British media, but little in China. That could be because the Chinese media are not covering the test, but potentially it may be because only a small percentage of Chinese media is in English, and we will not capture the true extent of the coverage.

Further to this, some countries have more media oulets compared to others (for example, the UK have more media than Australia, New Zealand, or Canada), and so direct comparisons may not be informative.

Other changes:

Methods and Analysis. Under the heading Tests and conditions of interest, we have rephrased our third criterion so that it reads better.

Page 5:

“Concern that the results of these tests will not lead to improved health outcomes for individuals; either due to the unavailabilty of effective treatment options (e.g. dementia) or treatments that may cause more harm than benefit (e.g. early mammography).”

VERSION 2 – REVIEW

REVIEWER	Bjørn Hofmann NTNU Gjøvik and University of Oslo in Norway
REVIEW RETURNED	27-Jun-2019

GENERAL COMMENTS	I am most thankful for the opportunity to review the revised version of this manuscript. The authors show that they have taken the comments into consideration, and they have improved the manuscript. In conclusion, this is a well-designed protocol for an interesting and important study that warrants publication. The results of the project will be of great interest to a broad audience.

REVIEWER	Kim Walsh-Childers University of Florida, USA
REVIEW RETURNED	31-May-2019

GENERAL COMMENTS	The authors have addressed nearly all of my concerns with the proposed study. The only issue on which I believe they could still

	strengthen the study is that of reporting intercoder reliability data. At the least, they should report the percentage of disagreements on each variable that had to be resolved through use of a third coder. This is relevant to researchers who might want to replicate the study as it reveals which variables are most problematic in terms of achieving intercoder agreement.

VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 2

Reviewer Name: Kim Walsh-Childers

Institution and Country: University of Florida, USA

Please state any competing interests or state ‘None declared’: None

Please leave your comments for the authors below

The authors have addressed nearly all of my concerns with the proposed study. The only issue on which I believe they could still strengthen the study is that of reporting intercoder reliability data. At the least, they should report the percentage of disagreements on each variable that had to be resolved through use of a third coder. This is relevant to researchers who might want to replicate the study as it reveals which variables are most problematic in terms of achieving intercoder agreement.

Thank you for this comment. We will provide descriptive statistics on the use of the third reviewer. If we planned to code stories only once, intercoder reliability data (e.g. Kappa) would make sense to appreciate the precision of our ratings that feed into our analyses. However, if we choose to calculate Kappa for the two ratings in this instance, we would get a reliability coefficient that does not apply to any of the data in our study as we would have resolved any disagreements by consensus. Therefore, we have chosen the other option you have suggested: reporting the percentage of disagreements on each variable that had to be resolved through use of a third coder.

We have added the following piece of information to page 8-9.

“Sets of two independent reviewers will extract data and code the media stories; two independent reviewers for each test. Any disagreements in extraction or coding will be resolved by discussion to reach consensus or by consultation with a third reviewer (RM). The percentage of disagreements on each coding variable requiring resolution through use of a third reviewer will be recorded. Before

formal data extraction and coding, the sets of independent reviewers will apply the data extraction tool to code 20 media stories; four for each test. Disagreements in data extraction and coding will be resolved by discussion and subsequent revisions to the data extraction tool."

Thank you for the suggestion.

Reviewer: 1

Reviewer Name: Bjørn Hofmann

Institution and Country: NTNU Gjøvik and University of Oslo in Norway

Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below

I am most thankful for the opportunity to review the revised version of this manuscript. The authors show that they have taken the comments into consideration, and they have improved the manuscript.

In conclusion, this is a well-designed protocol for an interesting and important study that warrants publication. The results of the project will be of great interest to a broad audience.

Thank you for reviewing our revised protocol.