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)	Validity of Breast, Lung and Colorectal Cancer Diagnoses in Administrative Databases: A Systematic Review Protocol
AUTHORS	Abraha, Iosief; Giovannini, Gianni; Serraino, Diego; Fusco, Mario; Montedori, Alessandro

VERSION 1 - REVIEW

REVIEWER	Jessica Chubak Group Health Research Institute, Seattle, Washington, USA
REVIEW RETURNED	06-Dec-2015

GENERAL COMMENTS	Administrative databases are frequently used for cancer research and thus it is important to understand their validity. To this end, the authors propose a protocol for a systematic review of the validity of ICD-9 codes to identify breast, lung, and colorectal cancer diagnoses. The change from ICD-9 to ICD-10 may limit the future impact of this review; however, I believe the results will still be valuable given the ongoing need to identify cancer cases in retrospective studies. I have a few recommendations to strengthen this protocol. Introduction/general: - The authors should define early in the protocol what they consider to be administrative databases. They do so in the methods, but it would helpful to orient readers if this definition were at the beginning. It would also be helpful to more consistently use terminology (e.g., healthcare database versus administrative database). The authors should also discuss how electronic health records (EHR) fit into their protocol. Approach: - I recommend the authors clarify that they are focusing on primary cancer diagnoses, rather than cancer history, cancer progression, or recurrence. They may wish to note in the introduction that such algorithms exist. But more importantly, it would be helpful to see clearly that such algorithms would excluded in their review. - A strength of this protocol is that it will use the STARD criteria. - How will the authors handle algorithms to identify multiple cancers? - Will algorithms that use CPT codes or prescription fills be included or excluded? I.e., must the algorithm consistent only of ICD-9 codes? Data extraction and quality assessment: - I recommend authors collect information on how each algorithm was developed (e.g., expert opinion, logistic regression, CART). This should include whether the study population was divided into training and testing sets. If it was, statistics for each should be reported. (This information could be included under “split sample” in the Appendix 2 checklist.)

	- I recommend the authors assess whether algorithms have been externally validated, that is: whether they have been developed in one population and tested in another. I suspect few algorithms have been externally validated, but if they have, details on both the original development and the external validation should be abstracted (including population characteristics). Search criteria - It would be helpful to see the search strategy for databases other than Medline - It is not clear why search terms related to uterine cancer and melanoma are included - The search strategy does not show how the terms related to administrative databases (1-3) will be combined with other terms - I wonder whether the term algorithm should appear somewhere in the list Analysis: - There could be additional clarity on what will be abstracted vs. calculated: o The authors state that they will calculate confidence intervals for the reported measures. If CIs are reported, will they use reported or calculated estimates? o The authors state they will calculate likelihood ratios. Is that only if LRs are not reported? o Will PPV and NPV be calculated where possible if not reported? - The authors say that where possible they will aggregate and stratify results by administrative data source. Please see my comment on aggregation below (under Meta analysis). With respect to stratification, it is not clear why data source would be the only (and presumably therefore) most important stratification factor. Meta analysis - I do not understand the value of meta-analysis in this systematic review. Meta-analysis is useful for obtaining a more precise estimate of a specific relationship reported across multiple studies. There is no single relationship being investigated in this review; rather its goals are to identify the properties of individual algorithms. A pooled estimate will not, therefore, be useful. To be consistent with the PRISMA-P guidelines, the authors should add: Introduction Objectives 7 Provide an explicit statement of the question(s) the review will address with reference to participants, interventions, comparators, and outcomes (PICO) Methods Study records: Data management 11a Describe the mechanism(s) that will be used to manage records and data throughout the review Data collection process 11c Describe planned method of extracting data from reports (such as piloting forms, done independently, in duplicate), any processes for obtaining and confirming data from investigators Data items 12 List and define all variables for which data will be sought (such as PICO items, funding sources), any pre-planned data assumptions and simplifications Outcomes and prioritization 13 List and define all outcomes for which data will be sought, including prioritization of main and additional outcomes, with rationale Risk of bias in individual studies 14 Describe anticipated methods for assessing risk of bias of individual studies, including whether this will be done at the outcome or study level, or both; state how this information will be used in data synthesis

	Confidence in cumulative evidence 17 Describe how the strength of the body of evidence will be assessed (such as GRADE) Minor comments: - Inclusion: I suggest replacing “validated case definition” with a “gold standard case definition” or “reference standard” - I suggest replacing “validity statistics” with “performance statistics.” - Appendix 2: I don’t understand what the column “Not” is? Should that read “Not stated”?

REVIEWER	Ryan Carnahan The University of Iowa College of Public Health, Department of Epidemiology, United States
REVIEW RETURNED	07-Dec-2015

GENERAL COMMENTS	This is a clearly written protocol for a systematic review of validated algorithms to identify breast, lung, and colorectal cancer diagnoses in administrative data. The manuscript is generally clearly written. My comments generally relate to suggestions for the approach that the authors might consider. However, I also suggest a few corrections and potential edits. Under “Ethics and dissemination” (page 2) you might note that these case definitions are not only important to researchers validating administrative databases but for research on these conditions that uses these databases. This is mentioned in the introduction but seems at least as relevant in a discussion of end-users. The search strategy appears to be missing a step that combines term #3 with the additional terms. I might also suggest that the authors consider expanding this component of the search somewhat. In my experience it has been difficult to find a search strategy that comprehensively identifies these types of studies, so casting a fairly wide net may be helpful. I have found it to be helpful to use Google Scholar to perform these searches because it can search the entire text of articles, but I realize that this poses problems for systematic reviews since the results include partial matches. In any case, the authors use another search strategy that they state was developed to accurately identify the relevant studies. I don’t believe the validity of that search strategy has been tested vs. other strategies, despite appearing to have many good attributes. For other ideas on terms to consider adding the authors might review the search strategies for the systematic reviews of similar types of studies published in Pharmacoepidemiology and Drug Safety volume 21, supplement 1, 2012, and Vaccine volume 31, Supplement 10, 2013. I’m not sure it’s necessary to define PPV, NPV, sensitivity and specificity, though perhaps this is helpful for some readers unfamiliar with this administrative data validation studies. Page 8 line 33: Typo: ‘ration’ instead of ‘ratio.’ However, I’ve not seen likelihood ratios used in the context of medical record validation of administrative data and am not sure what value they add. They’re generally more applicable to screening tests applied clinically. Sensitivity/specificity is also relatively rarely reported in validation studies, except when registries are the reference standard, so these may be difficult measures to produce for many

	studies. On the meta-analysis, I suspect that it will be more feasible to meta-analyze PPV and NPV than sensitivity and specificity since PPV and NPV are more commonly presented. PPV is substantially more common in my experience because a typical context for this information is confirmation of potential cases rather than review of charts of presumed non-cases. This could be different in the cancer validation literature since registries are a common source of cancer diagnosis information. As per the earlier comment, I'm not understanding the expected utility of likelihood ratios or associated odds ratios in this context. Also, the authors may find that it's difficult to meta-analyze the results since algorithms often vary. However, to the extent that studies use the same algorithms it seems reasonable to do. To the extent that studies report the validity of subsets of codes within an algorithm I would also suggest that the authors retain the information on the validity/performance of those subsets, as this can be helpful for recommending an algorithm. Have the authors considered expanding the review to include algorithms based on ICD-10? With the transition to ICD-10 in the U.S. and the already common use of ICD-10 in other countries this may increase the impact of the work.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Jessica Chubak
Institution and Country: Group Health Research Institute, Seattle, Washington, USA

Please leave your comments for the authors below

Administrative databases are frequently used for cancer research and thus it is important to understand their validity. To this end, the authors propose a protocol for a systematic review of the validity of ICD-9 codes to identify breast, lung, and colorectal cancer diagnoses. The change from ICD-9 to ICD-10 may limit the future impact of this review; however, I believe the results will still be valuable given the ongoing need to identify cancer cases in retrospective studies.
*We thank Dr. Chuback for her positive statement. Given the suggestion from both reviewers we have decided to consider also studies that validated ICD-10 codes in our systematic evaluation.

I have a few recommendations to strengthen this protocol.
Introduction/general:
- The authors should define early in the protocol what they consider to be administrative databases. They do so in the methods, but it would helpful to orient readers if this definition were at the beginning.
*We reported a brief definition of administrative databases at the end of the second paragraph in the Introduction as follows:
"By definition, administrative databases in which data is routinely and passively collected without an a priori research question, as they are usually established for billing or, in general, administrative purposes, and not for research uses. Hence the diagnostic codes used to identify, for example cancers, must be validated according to an accepted "gold standard" reference diagnosis(16").
It would also be helpful to more consistently use terminology (e.g., healthcare database versus

administrative database).
*Now we have used the terminology “Administrative healthcare databases” throughout the protocol.
The authors should also discuss how electronic health records (EHR) fit into their protocol.
*This note regarding EHR is important. We have discussed it briefly in page 7.
Studies that used electronic health records (EHR) to validate disease of our interest will not be included. EHR consists of a digital file used by healthcare providers for patient care and generally it includes clinical notes, prescription records, and radiology and laboratory data(25). Similar to most administrative databases, EHR are not established for research purposes(26). However, studies that used validated EHR as a reference standard will be considered in our evaluation.
We also highlighted this limitation in the “Strength and limitations of this study” box: “This systematic review will not address the validity of ICD codes related to breast, lung and colorectal cancers based on electronic health record”.
Approach:
- I recommend the authors clarify that they are focusing on primary cancer diagnoses, rather than cancer history, cancer progression, or recurrence. They may wish to note in the introduction that such algorithms exist. But more importantly, it would be helpful to see clearly that such algorithms would excluded in their review.
- A strength of this protocol is that it will use the STARD criteria.
*We thank Dr. Chubak for highlighting this point.
- How will the authors handle algorithms to identify multiple cancers?
*Studies that used algorithms to identify multiple cancers will not be considered.
- Will algorithms that use CPT codes or prescription fills be included or excluded? I.e., must the algorithm consistent only of ICD-9 codes?
*This points is interesting and we have provided further explanation regarding the algorithms in the review question (page 5) and also in the inclusion/exclusion criteria (page 7)

Data extraction and quality assessment:
- I recommend authors collect information on how each algorithm was developed (e.g., expert opinion, logistic regression, CART). This should include whether the study population was divided into training and testing sets. If it was, statistics for each should be reported. (This information could be included under “split sample” in the Appendix 2 checklist.)
*Thank you for this interesting suggestion.
- I recommend the authors assess whether algorithms have been externally validated, that is: whether they have been developed in one population and tested in another. I suspect few algorithms have been externally validated, but if they have, details on both the original development and the external validation should be abstracted (including population characteristics).
*We have included in the paper the issue of algorithm as well the validation cohort in the Appendix

Search criteria
- It would be helpful to see the search strategy for databases other than Medline
*The search strategies have been revised and now are included in the supplemental file.
- It is not clear why search terms related to uterine cancer and melanoma are included
*We have deleted this part.
- The search strategy does not show how the terms related to administrative databases (1-3) will be combined with other terms
*We have amended the search strategy syntax..
- I wonder whether the term algorithm should appear somewhere in the list
*Thank you for this good idea. We have included the term either as text word or as Mesh across all the search strategies.
Analysis:
- There could be additional clarity on what will be abstracted vs. calculated:
o The authors state that they will calculate confidence intervals for the reported measures. If CIs are

reported, will they use reported or calculated estimates?
o The authors state they will calculate likelihood ratios. Is that only if LRs are not reported?
o Will PPV and NPV be calculated where possible if not reported?
*We now clarified this issue: when the outcome measures are not reported we will calculate them where possible.
- The authors say that where possible they will aggregate and stratify results by administrative data source. Please see my comment on aggregation below (under Meta analysis). With respect to stratification, it is not clear why data source would be the only (and presumably therefore) most important stratification factor.
*We have added other factor for stratification: : "Where possible, validation statistics will be aggregated and stratified by administrative data source, type of ICD code (ICD-9-CM or ICD-10), stage of disease, and country of origin"

Meta analysis
- I do not understand the value of meta-analysis in this systematic review. Meta-analysis is useful for obtaining a more precise estimate of a specific relationship reported across multiple studies. There is no single relationship being investigated in this review; rather its goals are to identify the properties of individual algorithms. A pooled estimate will not, therefore, be useful.
*We agree with this comment. We have deleted the paragraph related to data pooling.
To be consistent with the PRISMA-P guidelines, the authors should add:
Introduction
Objectives 7 Provide an explicit statement of the question(s) the review will address with reference to participants, interventions, comparators, and outcomes (PICO)
*We have added a sub-title "Research question" in which we explicitly detailed the PICO topic.

Methods
Study records:
Data management 11a Describe the mechanism(s) that will be used to manage records and data throughout the review
Data collection process 11c Describe planned method of extracting data from reports (such as piloting forms, done independently, in duplicate), any processes for obtaining and confirming data from investigators
Data items 12 List and define all variables for which data will be sought (such as PICO items, funding sources), any pre-planned data assumptions and simplifications
Outcomes and prioritization 13 List and define all outcomes for which data will be sought, including prioritization of main and additional outcomes, with rationale

Risk of bias in individual studies 14 Describe anticipated methods for assessing risk of bias of individual studies, including whether this will be done at the outcome or study level, or both; state how this information will be used in data synthesis

Confidence in cumulative evidence 17 Describe how the strength of the body of evidence will be assessed (such as GRADE)
*We have amended the points above and filled the PRISMA-P form.

Minor comments:
- Inclusion: I suggest replacing "validated case definition" with a "gold standard case definition" or "reference standard"
*We replaced it with "gold standard case definition"
- I suggest replacing "validity statistics" with "performance statistics."
*We made this change.
- Appendix 2: I don't understand what the column "Not" is? Should that read "Not stated"?

*It should have read “Not applicable”. Now it is corrected

Reviewer: 2
Reviewer Name: Ryan Carnahan
Institution and Country: The University of Iowa College of Public Health, Department of Epidemiology, United States

Please leave your comments for the authors below

This is a clearly written protocol for a systematic review of validated algorithms to identify breast, lung, and colorectal cancer diagnoses in administrative data. The manuscript is generally clearly written. My comments generally relate to suggestions for the approach that the authors might consider. However, I also suggest a few corrections and potential edits.
*We thank Dr. Ryan for his positive and helpful comments.

Under “Ethics and dissemination” (page 2) you might note that these case definitions are not only important to researchers validating administrative databases but for research on these conditions that uses these databases. This is mentioned in the introduction but seems at least as relevant in a discussion of end-users.
*This point is relevant. We have changed the sentence to: “The results will serve as a guide to identify appropriate case definitions and algorithms of breast, lung and colorectal cancers for researchers involved in validating administrative healthcare databases as well as for outcome research on these conditions that used administrative healthcare databases.”
The search strategy appears to be missing a step that combines term #3 with the additional terms. I might also suggest that the authors consider expanding this component of the search somewhat. In my experience it has been difficult to find a search strategy that comprehensively identifies these types of studies, so casting a fairly wide net may be helpful. I have found it to be helpful to use Google Scholar to perform these searches because it can search the entire text of articles, but I realize that this poses problems for systematic reviews since the results include partial matches. In any case, the authors use another search strategy that they state was developed to accurately identify the relevant studies. I don’t believe the validity of that search strategy has been tested vs. other strategies, despite appearing to have many good attributes. For other ideas on terms to consider adding the authors might review the search strategies for the systematic reviews of similar types of studies published in Pharmacoepidemiology and Drug Safety volume 21, supplement 1, 2012, and Vaccine volume 31, Supplement 10, 2013.
*We agree that it is difficult to propose a comprehensive search strategy. We agree also that the Benchimol’s search strategy has not been tested with respect to other search strategies.
We thank Dr. Carnahan for giving us the occasion to revise the searched strategy based on the Mini Sentinel’s search strategy in addition to that proposed by Benchimol. The new search strategy for each electronic database is available in Appendix.

I’m not sure it’s necessary to define PPV, NPV, sensitivity and specificity, though perhaps this is helpful for some readers unfamiliar with this administrative data validation studies.
*We have preferred to retain the definitions.

Page 8 line 33: Typo: ‘ration’ instead of ‘ratio.’ However, I’ve not seen likelihood ratios used in the context of medical record validation of administrative data and am not sure what value they add. They’re generally more applicable to screening tests applied clinically. Sensitivity/specificity is also relatively rarely reported in validation studies, except when registries are the reference standard, so these may be difficult measures to produce for many studies.
*It is true that generally LRs are applied in clinical settings but we would like to calculate them to

provide complete information to the reader.

On the meta-analysis, I suspect that it will be more feasible to meta-analyze PPV and NPV than sensitivity and specificity since PPV and NPV are more commonly presented. PPV is substantially more common in my experience because a typical context for this information is confirmation of potential cases rather than review of charts of presumed non-cases. This could be different in the cancer validation literature since registries are a common source of cancer diagnosis information. As per the earlier comment, I'm not understanding the expected utility of likelihood ratios or associated odds ratios in this context.
Also, the authors may find that it's difficult to meta-analyze the results since algorithms often vary. However, to the extent that studies use the same algorithms it seems reasonable to do.
*We agree with this thought that are partly in line with Dr. Chubak. We have decided to delete the part related to meta-analyses

To the extent that studies report the validity of subsets of codes within an algorithm I would also suggest that the authors retain the information on the validity/performance of those subsets, as this can be helpful for recommending an algorithm.

Have the authors considered expanding the review to include algorithms based on ICD-10? With the transition to ICD-10 in the U.S. and the already common use of ICD-10 in other countries this may increase the impact of the work.
*We consider this suggestion important and we have decided to consider studies that validated also ICD-10 codes.

VERSION 2 – REVIEW

REVIEWER	Jessica Chubak Group Health Research Institute, USA
REVIEW RETURNED	18-Jan-2016

GENERAL COMMENTS	The authors were very responsive to prior reviews. I have a few remaining comments and questions. - The authors frame this paper as focusing on algorithms that use ICD-9 and ICD-10 codes, but they also allow CPT codes, prescriptions, etc. It might be useful to describe the algorithms they will study more generally, e.g., "administrative data algorithms" to reflect that different kinds of codes will be used. - The approach to the use of EMRs is inconsistent .The authors say they won't use it as a gold standard, but in several places refer to medical record abstraction as the gold standard. If medical record abstraction is a gold standard, then this includes EMRs (i.e., abstractors may review EMRs just as they would paper records). There are places throughout the manuscript that will need to be made consistent, including in the Strengths and Limitations box. - The last sentence before the Research Question section (new to the revision) implies chart review is the only gold standard that will be used. What about cancer registries as the gold standard? These should definitely be included. - Accordingly, I suggest the authors list what types of gold standards they expect to use – e.g., cancer registries. - In the newly added Research Question section: the term "index tests" is unclear; also, I'm not sure I understand what it means that "the secondary outcome is the accuracy of algorithms in discriminating…" How is this different from the primary outcome? - In the data extraction section, minor point: the word "disease" can be cut from "breast, lung, and colorectal cancer diseases"

	- In the data extraction section, bullet “h” could be reworded as “use of training and testing cohorts” - In the data extraction section, I think external validation should be its own bullet, not part of the modality of development (bullet “g”) - In the data extraction section, bullet “e” is subsumed under bullet “f”; these could be combined more clearly. - For quality assessment, I believe the authors should evaluate bias, even though they are not reviewing interventions. Bias can occur in any epidemiologic study. - I do not think LR will be reported or will be usefully. They may be used clinically but are generally not used for research algorithms. - The authors say they may stratify by “administrative data source” – what are the types of administrative data sources they mean? Is this outpatient vs. inpatient data, for example?

REVIEWER	Ryan Carnahan The University of Iowa College of Public Health, Department of Epidemiology
REVIEW RETURNED	23-Jan-2016

GENERAL COMMENTS	I think the authors have added some useful information to this protocol per reviewer recommendations, but have a few additional suggestions that I hope will improve clarity and consistency. Abstract: When referring to ICD-10, I’d suggest using “10th revision” instead of “10th version.” Also, while “CM” is sometimes assumed and not specified, since the authors used CM for ICD-9 I’d suggest using the same for ICD-10 since this is also the clinical modification version. Introduction/methods: I don’t want to be overly picky about this and recognize reviewer 1 suggested ‘gold standard’ as an option, but gold standard often refers to the best available reference standard for diagnosis. Many validation studies don’t require a gold standard but just a mention of the diagnosis in the chart, deferring to clinician judgment instead of requiring adjudication against gold standard criteria. I would have rather seen ‘reference standard,’ which was used instead of gold standard throughout the rest of the manuscript. Research question: I don’t quite understand this part of the research question section: “our secondary outcome is the accuracy of algorithms in discriminating cases of breast, lung and colorectal cancer diseases.” Is it expected that there will be difficulty differentiating these specific types of cancer, and that authors will report how many algorithm-identified cases were breast cancer vs. lung cancer, for example? I’d be surprised to see this kind of work as each uses different codes, and the algorithms targeting each disease will likely be distinct. I’d suggest breaking up the following sentence or revising to focus only on the primary research question. A primary question of the extent to which a code or algorithm is valid for identifying each of these conditions seems all that’s necessary. Then you can describe how each approach varies in performance for identifying each disease. This sentence almost suggests that the validity of the individual ICD-9 codes and the validity of more complex algorithms that use these codes are separate research questions, but to me

	they're all part of figuring out the best way to identify a true case. "The target populations are patients with primary diagnosis of breast, lung or colorectal cancer, the index tests for the primary question are ICD-9- CM or ICD-10 codes related to breast, lung and colorectal cancers, the index test for the secondary question will be represented by the algorithms, the reference standard will be represented by medical charts or validated electronic health records." Methods: The following new section is a little confusing to me and seems contradictory. I think it could be edited to make it clearer that you aren't trying to validate entries in the EHR but the EHR can be reviewed to classify a potential case based on the reference standard. The first sentence seems to suggest EHRs can't be used as a reference standard, but they're often the source of medical records that are reviewed to validate potential cases. Perhaps there's a distinction that needs to be made between problem list ICD codes in an EHR and the clinical notes they contain. The mention of this in the strengths and limitations box also might be revised for clarity. "Studies that used electronic health records (EHR) to validate disease of our interest will not be included. EHR consists of a digital file used by healthcare providers for patient care and generally it includes clinical notes, prescription records, and radiology and laboratory data(25). Similar to most administrative databases, EHR are not established for research purposes(26). However, studies that used validated EHR as a reference standard will be considered in our evaluation." Quality assessment: This sentence has a duplication at the end and could use an edit. "We will not appraise risk of bias of the included articles because this is not an intervention an intervention review."

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Jessica Chubak
Institution and Country: Group Health Research Institute, USA
Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below

The authors were very responsive to prior reviews. I have a few remaining comments and questions.
- The authors frame this paper as focusing on algorithms that use ICD-9 and ICD-10 codes, but they also allow CPT codes, prescriptions, etc. It might be useful to describe the algorithms they will study more generally, e.g., "administrative data algorithms" to reflect that different kinds of codes will be used.
* We have made the suggested changes.

- The approach to the use of EMRs is inconsistent .The authors say they won't use it as a gold standard, but in several places refer to medical record abstraction as the gold standard. If medical record abstraction is a gold standard, then this includes EMRs (i.e., abstractors may review EMRs just as they would paper records). There are places throughout the manuscript that will need to be made consistent, including in the Strengths and Limitations box.
*We have amended the inconsistencies and deleted the sentence in the limitation box.

- The last sentence before the Research Question section (new to the revision) implies chart review is the only gold standard that will be used. What about cancer registries as the gold standard? These should definitely be included.
- Accordingly, I suggest the authors list what types of gold standards they expect to use – e.g., cancer registries.
* We have changed the sentence in the Research question to the following: “The target populations are patients with primary diagnosis of breast, lung or colorectal cancer, the index test will be represented by administrative data algorithms related to breast, lung and colorectal cancers, the reference standard will be represented by medical charts, validated electronic health records or cancer registries.”

- In the newly added Research Question section: the term “index tests” is unclear; also, I’m not sure I understand what it means that “the secondary outcome is the accuracy of algorithms in discriminating…” How is this different from the primary outcome?
* this protocol is about validation studies that in design are diagnostic accuracy studies. The prisma-P requires that the PICO should be clarified in the Research question. Then I should stand for Index test that can be an icd-9 code or algorithm;
* regarding the secondary outcome after the suggestion provided above we have only one primary outcome. The paragraph now is as follows: “Our primary outcome is the accuracy (expressed in terms of sensitivity, specificity, positive and negative predictive values) of administrative data algorithms in discriminating cases of breast, lung and colorectal cancer diseases.”

- In the data extraction section, minor point: the word “disease” can be cut from “breast, lung, and colorectal cancer diseases”
* we made changes

- In the data extraction section, bullet “h” could be reworded as “use of training and testing cohorts”
* thank you for this suggestion. We made this change.

- In the data extraction section, I think external validation should be its own bullet, not part of the modality of development (bullet “g”)
* thank you for this suggestion. We made this change.

- In the data extraction section, bullet “e” is subsumed under bullet “f”; these could be combined more clearly.
* we made changes.

- For quality assessment, I believe the authors should evaluate bias, even though they are not reviewing interventions. Bias can occur in any epidemiologic study.
* We modified the sentence to “The presence of potential biases within the studies will be reported in a descriptive way” changes.

- I do not think LR will be reported or will be usefully. They may be used clinically but are generally not used for research algorithms.
*We have decided to delete the paragraph related to LRs.
- The authors say they may stratify by “administrative data source” – what are the types of administrative data sources they mean? Is this outpatient vs. inpatient data, for example?
*yes.

*Many thanks again for revising the protocol.

Iosief Abraha

Reviewer: 2
Reviewer Name: Ryan Carnahan
Institution and Country: The University of Iowa College of Public Health, Department of Epidemiology, USA
Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below

I think the authors have added some useful information to this protocol per reviewer recommendations, but have a few additional suggestions that I hope will improve clarity and consistency.

Abstract:
When referring to ICD-10, I'd suggest using "10th revision" instead of "10th version." Also, while "CM" is sometimes assumed and not specified, since the authors used CM for ICD-9 I'd suggest using the same for ICD-10 since this is also the clinical modification version.
* we made the suggested version

Introduction/methods:
I don't want to be overly picky about this and recognize reviewer 1 suggested 'gold standard' as an option, but gold standard often refers to the best available reference standard for diagnosis. Many validation studies don't require a gold standard but just a mention of the diagnosis in the chart, deferring to clinician judgment instead of requiring adjudication against gold standard criteria. I would have rather seen 'reference standard,' which was used instead of gold standard throughout the rest of the manuscript.
* we changed the "gold standard" to "reference standard" throughout the protocol.

Research question:
I don't quite understand this part of the research question section: "our secondary outcome is the accuracy of algorithms in discriminating cases of breast, lung and colorectal cancer diseases." Is it expected that there will be difficulty differentiating these specific types of cancer, and that authors will report how many algorithm-identified cases were breast cancer vs. lung cancer, for example? I'd be surprised to see this kind of work as each uses different codes, and the algorithms targeting each disease will likely be distinct.
I'd suggest breaking up the following sentence or revising to focus only on the primary research question. A primary question of the extent to which a code or algorithm is valid for identifying each of these conditions seems all that's necessary. Then you can describe how each approach varies in performance for identifying each disease. This sentence almost suggests that the validity of the individual ICD-9 codes and the validity of more complex algorithms that use these codes are separate research questions, but to me they're all part of figuring out the best way to identify a true case. "The target populations are patients with primary diagnosis of breast, lung or colorectal cancer, the index tests for the primary question are ICD-9- CM or ICD-10 codes related to breast, lung and colorectal cancers, the index test for the secondary question will be represented by the algorithms, the reference standard will be represented by medical charts or validated electronic health records."
* Many thanks for this suggestion. Tthis point was also raised by dr. Chubak. We changed the research question as follows: "The primary research question is "what is the accuracy of administrative data algorithms related to breast, lung and colorectal cancers in administrative databases in correctly identifying the respective diseases?". The target populations are patients with primary diagnosis of breast, lung or colorectal cancer, the index test will be represented by administrative data algorithms related to breast, lung and colorectal cancers, the reference standard

will be represented by medical charts, validated electronic health records or cancer registries. Our primary outcome is the accuracy (expressed in terms of sensitivity, specificity, positive and negative predictive values) of administrative data algorithms in discriminating cases of breast, lung and colorectal cancer diseases."

Methods:
The following new section is a little confusing to me and seems contradictory. I think it could be edited to make it clearer that you aren't trying to validate entries in the EHR but the EHR can be reviewed to classify a potential case based on the reference standard. The first sentence seems to suggest EHRs can't be used as a reference standard, but they're often the source of medical records that are reviewed to validate potential cases. Perhaps there's a distinction that needs to be made between problem list ICD codes in an EHR and the clinical notes they contain. The mention of this in the strengths and limitations box also might be revised for clarity. "Studies that used electronic health records (EHR) to validate disease of our interest will not be included. EHR consists of a digital file used by healthcare providers for patient care and generally it includes clinical notes, prescription records, and radiology and laboratory data(25). Similar to most administrative databases, EHR are not established for research purposes(26). However, studies that used validated EHR as a reference standard will be considered in our evaluation."
* We agree there was an inconsistency regarding EHR. We made the necessary changes throughout the protocol.

Quality assessment:
This sentence has a duplication at the end and could use an edit. "We will not appraise risk of bias of the included articles because this is not an intervention an intervention review."
* we have changed the sentence to "The presence of potential biases within the studies will be reported in a descriptive way"
*Many thanks again for revising the protocol.

VERSION 3 - REVIEW

REVIEWER	Ryan Carnahan The University of Iowa College of Public Health, Department of Epidemiology, Iowa City, IA, USA
REVIEW RETURNED	10-Feb-2016

GENERAL COMMENTS	I think this protocol has come together well. My only remaining comments could easily be handled in the copy editing phase. On page 4, last paragraph, I'm thinking perhaps ICD-10 should be defined prior to using the abbreviation. It wouldn't hurt to be consistent about the use of -CM or not for both ICD-9 and ICD-10. Technically clinical diagnoses are the CM version for both, though use of -CM tends to be optional in the literature. I think there's a typo on page 5, line 28/29. It refers to studies that 'evaluated the accuracy of the validated ICD-9-CM or ICD-9 codes related to the cancer diseases.' I think the second ICD was supposed to be ICD-10 (or ICD-10-CM) instead of ICD-9.