PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cohort profile: The MCC-Spain follow-up on colorectal, breast and prostate cancers. Study design and initial results
AUTHORS	Alonso, Jessica; Molina, Antonio J; Jiménez-Moleón, Jose Juan; Pérez-Gómez, Beatriz; Martin, Vicente; Moreno, Victor; Amiano, Pilar; Ardanaz, Eva; de Sanjose, Silvia; SALCEDO, INMACULADA; Fernandez-Tardon, Guillermo; Alguacil, Juan; Salas, Dolores; Marcos-Gragera, Rafael; Chirlaque, Maria Dolores; Aragonés, Nuria; Castaño-Vinyals, Gemma; Pollán, Marina; Kogevinas, Manolis; Llorca, Javier

VERSION 1 – REVIEW

REVIEWER	Christine Friedenreich Alberta Health Services, Canada
REVIEW RETURNED	17-Jul-2019

GENERAL COMMENTS	General Comments: The authors describe the cohort follow-up that was done for breast, colorectal and prostate cancer cases who were originally recruited as part of a multi-centered case-control study in Spain that was conducted between 2008 and 2013. The cohort follow-up was initiated in 2016 and ended in 2018. This paper appears to be the first one from the cohort follow-up. The authors have included a description of the 12 provinces in Spain and the number of cases recruited for these three cancer sites by center, some tumour characteristics for the included study participants and the type of first-line treatment received by the cohort members. This manuscript is restricted to this level of description and does not include any specific research question that was addressed. It appears that the main purpose of this manuscript is describe these three patient cohorts and their survival experience to date. Given the purely descriptive nature of this paper, there is limited novelty to this manuscript nor is the value of presenting this information at this time apparent. The authors could consider undertaking some survival analyses within each of these three cancer-specific cohorts and reporting on those results. Then, the description of each of the cohorts could simply be part of the introductory sections/tables for those papers. There were many issues with English language grammar and punctuation that need to be addressed as well. There were several instances in the paper in which the writing was unclear and revisions are needed to clarify the content and meaning. Some of the major issues are highlighted below. Specific Comments:

	1. Introduction - this paper would be strengthened if this cohort was placed in context of other large population-based cohorts conducted in Europe and elsewhere. There is no mention of the EPIC cohort or others that have been conducted. It would be useful to highlight the unique aspects of this cohort that make it different from previous cohort studies and what the value added of this cohort will be. 2. Cohort Description and Methods - The wording of this paragraph is awkward and needs revision. It would be clearer if there was a section that described first the case-control study and then a separate section for the cohort follow-up. As currently written, the two studies' descriptions are combined and it is often difficult to separate them. 3. Patient recruitment - it is unclear on line 52 what is meant by "locally decided according to the hospital characteristics". The description of what kind of patients were recruited from each center needs more detail including how many participants were recruited in total in the original MCC Spain study. 4. Patient recruitment, page 6, lines 3-4: This sentence is unclear "... public involvement came into consideration..." and needs revision. 5. Data collection - it appears that the investigators only have exposure (e.g. lifestyle) data from the case-control study and have not collected any post-diagnosis exposure data. They should address if there will be additional follow-up data on modifiable lifestyle factors collected. If so, these data would be very informative and would add considerable value to this cohort and make it more interesting for survival analyses. 6. Findings to date - the first paragraph is written too succinctly and is difficult to follow. More detail is needed on the work that has been done to date from the case-control study including which associations between exposures and specific cancer sites have already been published. 7. Discussion - references to the previous papers need to be added on page 13, lines 48-54. 8. Limitations - the authors need to address the lack of post-diagnosis data in their limitations section.

REVIEWER	Jianwei Zhu Karolinska Institutet, Sweden West China Hospital of Sichuan University, China
REVIEW RETURNED	03-Aug-2019

GENERAL COMMENTS	Thanks for inviting to review the paper on an interesting and meaningful topic. Authors constructed a cohort study based on a bigger case-control study only including the three specific cancers. It is a good idea to have the cancer cohort and follow up them to investigate the either risk factors and survival. I would suggest the author reorganize the paper and make it more clear to read and follow. Abstract Line 5, author might need to explain the full mean of 'MCC-Spain'

	Line 3, Abstract. Authors have specified the participants and findings to date, but how about the method? I would suggest the authors to add some words about the methods. Line 55. Lost in cohort study is neither strengths nor limitations. Authors might need to consider statistical method to adjust the lost. Article Introduction Authors might need to reform the introduction, especial paragraph two. This part can be moved to method. And I would like to read more information on background and literature review on this specific topic. Authors have mentioned that the aim of MCC-Spain is to identify factors associated with cancers, but did not specifically mention the aim of this paper. What are the aims of the three cohorts? Cohort description and methods Since authors have named this part as 'cohort description', I would suggest to more focus on the description of the three cohorts. For example, in paragraph information at recruitment, are the percentage of biological samples for all participants or the selected three kinds of cancer? How was the information and samples collected for the three kinds of cancer? I would suggest to describe the 'cohort inception' a bit earlier in the method, and put more effort on the description of the three cohorts. Findings Authors separately show the tables and text, I would suggest to describe the text corresponding to tables. And make the text more easy to follow and understand. If this study also aimed to explore the risk factor associated with cancers, I would suggest author to show some findings on potential factors, i.e. life style, environment. Discussion I am not so agree with authors that call the efficiency is strengths. Yes, I agree that converting cases from a case-control study is efficient. But for the study design aspect, saving time is not strength. I would like to read how the study represents the population and how the follow-up was completed well. Page 15 paragraph 3. Author tried to discuss the limitation of the study, i.e. misclassification and bias. What is the 'differential' and 'misclassification' in the 'differential misclassification', author did not make it clear. And I do not agree with the authors that patients' unawareness and interviewers' lacking knowledge is good way to avoid misclassification.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

General Comments:

The authors describe the cohort follow-up that was done for breast, colorectal and prostate cancer cases who were originally recruited as part of a multi-centered case-control study in Spain that was conducted between 2008 and 2013. The cohort follow-up was initiated in 2016 and ended in 2018. This paper appears to be the first one from the cohort follow-up. The authors have included a description of the 12 provinces in Spain and the number of cases recruited for these three cancer sites by center, some tumour characteristics for the included study participants and the type of first-line treatment received by the cohort members. This manuscript is restricted to this level of description and does not include any specific research question that was addressed. It appears that the main purpose of this manuscript is describe these three patient cohorts and their survival experience to date.

Given the purely descriptive nature of this paper, there is limited novelty to this manuscript nor is the value of presenting this information at this time apparent. The authors could consider undertaking some survival analyses within each of these three cancer-specific cohorts and reporting on those results. Then, the description of each of the cohorts could simply be part of the introductory sections/tables for those papers.

According to BMJ Open's instructions for authors regarding cohort profiles, we provide the rationale for MCC-Spain cohort creation, its methods, baseline data and future plans. We acknowledge that in the first version the cohort's rationale was insufficiently explained and the methods were somewhat disordered. We think both issues have been widely improved in this new version. More details are provided when answering specific comments.

There were many issues with English language grammar and punctuation that need to be addressed as well. There were several instances in the paper in which the writing was unclear and revisions are needed to clarify the content and meaning. Some of the major issues are highlighted below.

The manuscript has been revised by a native-English speaker. We apologize for previous mistakes.

Specific Comments:

1. Introduction - this paper would be strengthened if this cohort was placed in context of other large population-based cohorts conducted in Europe and elsewhere. There is no mention of the EPIC cohort or others that have been conducted. It would be useful to highlight the unique aspects of this cohort that make it different from previous cohort studies and what the value added of this cohort will be.

The introduction section has been completely rewritten. The EPIC cohort and some clinical cohorts on cancer prognosis factors are now referenced, which allows us to highlight the important amount of basal information gathered in our cohort as well as the rational for conducting our study.

2. Cohort Description and Methods - The wording of this paragraph is awkward and needs revision. It would be clearer if there was a section that described first the case-control study and then a separate section for the cohort follow-up. As currently written, the two studies' descriptions are combined and it is often difficult to separate them.

This section has been rewritten. Now, the case-control study is signalled as the origin of the current cohort, but we immediately proceed with the cohort inception and methods.

3. Patient recruitment - it is unclear on line 52 what is meant by "locally decided according to the hospital characteristics". The description of what kind of patients were recruited from each center needs more detail including how many participants were recruited in total in the original MCC Spain study.

In order to clarify patient recruitment, we have included a new Figure 1 with a flow chart; it clarifies how many patients were recruited in the case-control study and, then, how many have been followed-up. The former Table 1 (patients recruited by hospital) is now Supplementary Table 1, in order to make room for a table describing data collection (see comment 5, below).

4. Patient recruitment, page 6, lines 3-4: This sentence is unclear "... public involvement came into consideration..." and needs revision.

The sentence on public involvement has been simplified.

5. Data collection - it appears that the investigators only have exposure (e.g. lifestyle) data from the case-control study and have not collected any post-diagnosis exposure data. They should address if there will be additional follow-up data on modifiable lifestyle factors collected. If so, these data would be very informative and would add considerable value to this cohort and make it more interesting for survival analyses.

A new Table 1 has been added to better explain the information gathered in each phase and its timing. We now acknowledge in both Methods and Strengths and Limitations sections we have no collected information on modifiable lifestyle factors after treatment.

6. Findings to date - the first paragraph is written too succinctly and is difficult to follow. More detail is needed on the work that has been done to date from the case-control study including which associations between exposures and specific cancer sites have already been published.

We have rewritten this paragraph in order to be more specific on previous results. Moreover, in the supplementary material we have added a table and a complete reference list of previous MCC-Spain publications.

7. Discussion - references to the previous papers need to be added on page 13, lines 48-54.

The paragraph indicated by the referee was redundant with the information provided in

“Findings to date”; therefore, we have deleted it. Nevertheless, in the “Strengths and Limitations” section we refer again to the supplementary information on previous results from MCC-Spain.

8. Limitations - the authors need to address the lack of post-diagnosis data in their limitations section.

We specifically acknowledge this limitation.

Reviewer: 2

Thanks for inviting to review the paper on an interesting and meaningful topic.

Authors constructed a cohort study based on a bigger case-control study only including the three specific cancers. It is a good idea to have the cancer cohort and follow up them to investigate the either risk factors and survival. I would suggest the author reorganize the paper and make it more clear to read and follow.

Abstract

Line 5, author might need to explain the full mean of 'MCC-Spain'

We have spelled-out "MCC-Spain".

Line 3, Abstract. Authors have specified the participants and findings to date, but how about the method? I would suggest the authors to add some words about the methods.

We are sorry in do not provide more method information the BMJ Open Submission Guidelines limit the abstract in cohort profiles to four paragraphs: Purpose, Participants, Findings to date (including data collected) and Future plans.

Line 55. Lost in cohort study is neither strengths nor limitations. Authors might need to consider statistical method to adjust the lost.

We have completely rewritten "Strengths and limitations" box.

Article

Introduction

Authors might need to reform the introduction, especial paragraph two. This part can be moved to method. And I would like to read more information on background and literature review on this specific topic.

The "Introduction" section has been completely rewritten. Former paragraph two is now at the beginning of "Methods".

Authors have mentioned that the aim of MCC-Spain is to identify factors associated with cancers, but did not specifically mention the aim of this paper. What are the aims of the three cohorts?

The aims of the three cohorts are now included at the end of the Introduction.

Cohort description and methods

Since authors have named this part as ‘cohort description’, I would suggest to more focus on the description of the three cohorts. For example, in paragraph information at recruitment, are the percentage of biological samples for all participants or the selected three kinds of cancer? How was the information and samples collected for the three kinds of cancer?

I would suggest to describe the ‘cohort inception’ a bit earlier in the method, and put more effort on the description of the three cohorts.

We have restructured this section. Now, we let little room to the initial case-control phase and give more details on patient recruitment and data collection. In this regard, we have included a new Table 1 with the information gathered and its timing and a new Figure 1 with a flow chart on patient recruitment. Information on percentages of biological samples is now included the text.

Findings

Authors separately show the tables and text, I would suggest to describe the text corresponding to tables. And make the text more easy to follow and understand.

If this study also aimed to explore the risk factor associated with cancers, I would suggest author to show some findings on potential factors, i.e. life style, environment.

The text now properly refers to the tables. Moreover, new Supplementary Table 2 gives a complete list of references on risk factors associated with cancers in MCC-Spain study.

Discussion

I am not so agree with authors that call the efficiency is strengths. Yes, I agree that converting cases from a case-control study is efficient. But for the study design aspect, saving time is not strength. I would like to read how the study represents the population and how the follow-up was completed well.

We completely agree with this comment, so we have deleted our sentence.

Page 15 paragraph 3. Author tried to discuss the limitation of the study, i.e. misclassification and bias. What is the ‘differential’ and ‘misclassification’ in the ‘differential misclassification’, author did not make it clear. And I do not agree with the authors that patients’ unawareness and interviewers’ lacking knowledge is good way to avoid misclassification.

Misclassification is usually categorized in differential (if it affects in different way the compared groups) and non-differential (if it affects in a similar way the compared groups). Nondifferential misclassification introduces bias towards the null (i.e.: the estimated relative risk is closer to 1 than the

true relative risk) while differential misclassification could introduce both towards and away the null, which makes it unpredictable. Because of that, non-differential misclassification is usually preferred. We have modified that sentence in order to make it clearer: "*therefore, if interviewers or patients have introduced some misclassification, it could probably have been non-differential, eventually leading to bias towards the null, which would make more robust the positive findings in this cohort study*."

VERSION 2 – REVIEW

REVIEWER	Jianwei Zhu West China Hospital, Sichuan University
REVIEW RETURNED	03-Sep-2019

GENERAL COMMENTS	This study is based on the MCC-Spain, and mainly focused on the three high incident cancers, i.e. colorectal, breast and prostate cancers. The authors aim to retrospectively collected the information of potential risk factors for cancer prognosis and analyze their association. The study has include around 4000 cancer patients and will collect information for many factors. I do not have many comments for this version, only few small points. General comments: 1. the statistical analysis method was not described thoroughly. The method to estimate association between those factors and cancer prognosis was not mentioned, especially how to control the confounder and mediators. The baseline factor or cancer characteristics might related to different cancer treatment, and treatment usually strongly associates with cancer prognosis. How the author will control the treatment? 2. How to measure those factors, use questionnaires? Have all the questionnaires been valid? Use standardized method to collect information is a better way to control information bias, but not blinding patients from study hypothesis and reviewers from study design. 3. Plenty studies have been performed to investigate the risk factors for either cancer prognosis or cancer survival. Have authors calculated the power of the study design? 4. All the risk factors were measured retrospectively, what time points was aimed to, at time of cancer diagnosis, or repeat measurement for severe time points? As I understand, author want to collect information only at time of cancer diagnosis. So how to deal with the change of some factors, i.e. diet, environmenta? Page 4, paragraph 2, author only mentioned the MCC was started from 2016, but did not mention when the patients were enrolled. Paragraph 3, author showed the result for follow-up, but how long have the patients been followed? How the five-year survival was calculated? Page 18, paragraph 1, the author talk about information bias. How ever, I still cannot agree that opinion that the two points author mentioned can help to control the bias. Most cancer patients paid lots of attention on their health care, they probably believe some factors that related to cancer prognosis and more like to provide information on that. And, interviewers were not familiar with the cohort design is not a method to control information bias.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2

This study is based on the MCC-Spain, and mainly focused on the three high incident cancers, i.e. colorectal, breast and prostate cancers. The authors aim to retrospectively collect information about potential risk factors for cancer prognosis and analyze their association. The study has included around 4000 cancer patients and will collect information for many factors. I do not have any comments for this version, only a few small points.

General comments:

1. The statistical analysis method was not described thoroughly. The method to estimate the association between those factors and cancer prognosis was not mentioned, especially how to control the confounder and mediators. The baseline factor or cancer characteristics might relate to different cancer treatments, and treatment usually strongly associates with cancer prognosis. How the author will control the treatment?

We have tried not to include too many statistical details in order to accomplish BMJ Open standards on cohort profile articles. However, we acknowledge the relevance of this point, so we have included the following sentence in the "Statistical analysis" paragraph:

> *"Further analyses should deal with confounding and modifiers using multivariate regression models (e.g.: Cox or Weibull regression). Initial treatment could be related with both basal factors and survivorship, eventually leading to confounding by indication; it would be controlled using propensity scores."*

2. How to measure those factors, use questionnaires? Have all the questionnaires been valid? Use a standardized method to collect information is a better way to control information bias, but not blinding patients from the study hypothesis and reviewers from the study design.

References to the questionnaires and the website where you can find them (http://www.mccspain.org) have been included.

3. Plenty of studies have been performed to investigate the risk factors for either cancer prognosis or cancer survival. Have the authors calculated the power of the study design?

The statistical power of each tumour cohort has been added as a new paragraph in the "follow-up information" section, on page 9.

4. All the risk factors were measured retrospectively, what time points were aimed at, at the time of cancer diagnosis, or repeat measurement for severe time points? As I understand, the author wants to collect information only at the time of cancer diagnosis. So how to deal with the change of some factors, i.e. diet, environmental?

Risk factors were all measured at diagnosis, as explained in Table 1. Some of them (e.g., medical history, familial history, occupational exposures, reproductive history, use of drugs…) refer to the whole life, and we think this point is self-explicative and does not require further explanation in the text. Diet, however, was obtained referring to one year before diagnosis in order to avoid reverse causation (i.e.: diet could have changed before diagnosis because of an already existing cancer); we have clarified it in the new version.

Page 4, paragraph 2, the author only mentioned the MCC was started in 2016 but did not mention when the patients were enrolled.

Patients were enrolled between 2008 and 2013 as they were diagnosed of their cancer; at that moment basal information was obtained. In 2016, we changed our objectives towards prognosis rather than risk factors. We have changed paragraph 2 on page 4 to try to explain it better.

Paragraph 3, the author showed the result for follow-up, but how long have the patients been followed? How the five-year survival was calculated?

The follow-up time is expressed in each tumor's section.

For colorectal cancer (Paragraph 1, page 15): The first case was recruited on 18th of March 2007 and the follow-up was closed on 23rd of August 2018, accounting for 12813.8 person-years of follow-up.
For breast cancer (Paragraph 5, page 15): The maximum span for breast cancer follow-up was nine and a half years (from 13th July 2007 to 22nd March 2017). Follow-up was obtained for 1685 out of 1738 breast cancer patients (97%), adding 10931 person-years.
For prostate cancer (Paragraph 2, page 16): the first patient was included on 26th January, 2008 and the end of follow-up was on 13th July, 2018, adding 7169.6 person-year of follow-up.

On the other hand, the five-year survival probability was calculated via Kaplan-Meier (paragraph 5, page 9).

Page 18, paragraph 1, the author talks about information bias. However, I still cannot agree with that opinion that the two-points the author mentioned can help to control the bias. Most cancer patients paid lots of attention to their health care, they probably believe some factors that related to cancer prognosis and more like to provide information on that. And, interviewers were not familiar with the cohort design is not a method to control information bias.

We understand that we have not been clear enough in this point. Under no circumstances are we claiming that blinding patients and interviewers to the study hypotheses could allow us to control information bias. What we are stating -and we hope to have clarified in this version- is it could produce a non-differential misclassification, leading to bias towards the null. We have included as supporting reference the "Encyclopedia of epidemiological methods" by Mitchell Gail and Jacques Benichou.