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)	Cohort Profile: The National Health Insurance Service - National Health Screening Cohort (NHIS-HEALS) in Korea
AUTHORS	Seong, Sang Cheol; Kim, Yeon-Yong; Park, Sue K.; Khang, Y; Kim, Hyeon Chang; Park, Jong Heon; Kang, Hee-Jin; Do, Cheol-Ho; Song, Jong-Sun; Lee, Eun-Joo; Ha, Seongjun; Shin, Soon Ae; Jeong, Seung-lyeal

VERSION 1 – REVIEW

REVIEWER	Martin Gulliford King's College London UK
REVIEW RETURNED	10-Mar-2017

GENERAL COMMENTS	This paper presents information - a cohort profile - of the Korean National Health Insurance Service - National Health Screening Cohort. This is a valuable data resource with data for 514,866 individuals screened in 2002-3 with follow-up data collected to the present. This will be a useful publication. Ethics Ethical issues could be addressed more directly. What level of consent did individuals give to participation in screening and participation in the cohort study? How was data linkage addressed in terms of governance? How are data anonymised and the confidentiality of individual participants protected? The context for the study could be better explained, in terms of the health care system in Korea and the screening programmes in current use. Even after looking at Supplementary Table 1 it was not entirely clear what screening programmes were in current use, and this needs to be addressed in the main text. Sampling: was a simple random sample taken or was the sample stratified by age, gender, region etc? The report generally lacks methodological detail concerning the measurement methods used. Descriptions are vague, for example, on page 7 it says 'Variables for specific health problems and risk factors from questionnaires and bioclinical laboratory results were included in the health screening database. Some variables changed during the follow-up period.' Greater precision would be beneficial. The report could say more about arrangements for data access. For international readers, are the data coded in Korean characters?

REVIEWER	Rodrigo M. Carrillo-Larco Universidad Peruana Cayateno Heredia, Lima, Peru
REVIEW RETURNED	17-Apr-2017

GENERAL COMMENTS	The authors present information about a population-based cohort in the Republic of Korea. They give many details about available data and present how many subjects there were at baseline, and at each follow-up round. A main concern is the lack of details about some variables definitions. The manuscript is clearly written. However, there are a few comments that need to be addressed to improve the manuscript. Abstract: although this may be a premature questioning, I wonder why the authors pointed out that data can be used to study risk for non-communicable diseases. What about communicable diseases? Is there not any screening for HIV or other sexually-transmitted diseases? Strengths and limitations: in the first bullet, I would suggest including the total baseline sample size. For example: …large sample size (N=XXX at baseline), with a… Cohort description, Participant of the cohort: [first paragraph] it seems this refers to the eligibility criteria for baseline enrollment, if so, it could be relevant to clearly state what happened if at follow-up participants did not meet these criteria. Were they still assessed and thus a follow-up record created? Were they not followed-up? Cohort description, Participant of the cohort: [second paragraph, regarding supplementary table 1] it would be informative, if possible, to statistically assess the difference, at least, between this cohort and the overall dataset. Cohort description, Participant of the cohort: [third paragraph, regarding baseline characteristics of the cohort population] it would be important to provide further details about certain variables. For example, when it refers to region “non-metropolitan area”, is this the same as rural areas? Also with regards to alcohol drinking, does this variable accounts for amount of alcohol or kind of beverage? As it currently stands, it only refers to number of times they drink alcohol, though not information is given about what they drink and how much they drink each time. Moreover, please add mean (and standard deviation) or median (and interquartile range) age. Cohort description, Participant of the cohort: [key variables] further details should be provided for certain variables. For example, in table 2 it reads “cognitive impairment, depression…mental health screening”. What did this screening include? Was it conducted by a physician (i.e., psychiatrist) or was it based on questionnaires (i.e., PHQ-9)? In association with my previous commentary, there seems no information was retrieved about kind of alcoholic beverage. It is not the same to drink a cup of wine than a cup of whisky. If in fact this information is not available, perhaps this would deserve a line in

	the limitation section. At this point it would also be important a few details about data collection: what devices were used to assess blood pressure? Was there only one measure or several and the available blood pressure variable is the mean? About body mass index, further details in this line would be appreciated (weight and height was measured or self-reported?). Likewise, what about physical activity? Was it assessed with a questionnaire or with an objective method? When it talks about disabilities, what kind of disabilities it refers to? Permanent disabilities? All these details about variables definitions could go as a supplementary table, but they are indeed needed. Findings to date: the last sentence is conflict. It reads: the NHIS-HEALS will provide evidence regarding the issues that were assessed in previous studies using the NHID. If these issues have already been addressed by the NHID, what does this new cohort add? Findings to date: [second paragraph] it would be more internationally relevant to show in the text the standardized rates according to the WHO population. Rates according to the Korean census population could then be depicted only in Table 3. Provided this is an international open-access journal, figures that are comparable internationally would be more interesting. Findings to date: [third paragraph] it would be interesting to have the differences between men and women compared with p-values. Unless the authors have other reasons for not presenting this information in the manuscript, a p-value should be included. This is even more important to support statements such as that one in line 12-13 of page 9: the prevalence of diabetes and hypertension was higher in men than in women. This suggestion would also apply for incidence and mortality rates Findings to date: [page 9, line 36] it reads screening database in 2005-2013… Then it reads that patients who have a diagnosis within the first three years were excluded from the incidence analysis. Why was this decision taken? Perhaps at baseline there was not any information about pre-existing conditions. If so, this would explain this decision. In any case, this should be clearly stated. In addition, a reason for choosing this three-year threshold would be appreciated. Strengths and limitations: further discussion should be made around the quality of the variables. For example, much is not said about mortality data. Is this information reliable? Are causes of death well-defined? Is date of death accurately recorded? Is this information as accurate for urban areas as it is for rural areas? Furthermore, once further details about other variables are provided, further discussion about their quality could be included in this section. Footnote for Supplementary Figure 1: in line 4 it reads …to those age 50 or older… It seems it should read …to those aged 50 or older…

VERSION 1 – AUTHOR RESPONSE

Reviewer #1

- Ethical issues could be addressed more directly. What level of consent did individuals give to participation in screening and participation in the cohort study? How was data linkage addressed in terms of governance? How are data anonymised and the confidentiality of individual participants protected?
- >> As suggested, detailed explanations about ethical issues were inserted. (#page 5, line 25-page 6, line 3)

- The context for the study could be better explained, in terms of the health care system in Korea and the screening programmes in current use. Even after looking at Supplementary Table 1 it was not entirely clear what screening programmes were in current use, and this needs to be addressed in the main text.
- >> As suggested, detailed explanations about health system in Korea were inserted. (#page 4, lines 14-18)

- Sampling: was a simple random sample taken or was the sample stratified by age, gender, region etc?
- >> Sampling method was a simple random sampling. Correction was made. (#page 5, line 20)

- The report generally lacks methodological detail concerning the measurement methods used. Descriptions are vague, for example, on page 7 it says 'Variables for specific health problems and risk factors from questionnaires and bioclinical laboratory results were included in the health screening database. Some variables changed during the follow-up period.' Greater precision would be beneficial.
- >> As suggested, we added detailed information. (Supplementary table 2 and #page 4, lines 14-18)

- The report could say more about arrangements for data access. For international readers, are the data coded in Korean characters?
- >> We included explanations about the language of the data. (#page 11, line 11)

Reviewer: 2

- Abstract: although this may be a premature questioning, I wonder why the authors pointed out that data can be used to study risk for non-communicable diseases. What about communicable diseases? Is there not any screening for HIV or other sexually-transmitted diseases?
- >> As the National Health Screening Program focused on the detection of high risk group of chronic, non-communicable disease, NHIS-HEALS did not include the screening test for the communicable disease. (such as HIV and STD)

- Strengths and limitations: in the first bullet, I would suggest including the total baseline sample size. For example: …large sample size (N=XXX at baseline), with a…
- >> Correction was made. (#page 10, line 14)

- Cohort description, Participant of the cohort: [first paragraph] it seems this refers to the eligibility criteria for baseline enrollment, if so, it could be relevant to clearly state what happened if at follow-up participants did not meet these criteria. Were they still assessed and thus a follow-up record created? Were they not followed-up?
- >> a more detailed information was inserted. (#page 6, lines 17-20)

- Cohort description, Participant of the cohort: [second paragraph, regarding supplementary table 1] it would be informative, if possible, to statistically assess the difference, at least, between this cohort and the overall dataset.
- >> As suggested, the differences were assessed with statistical test. (Supplementary table 1)

- Cohort description, Participant of the cohort: [third paragraph, regarding baseline characteristics of the cohort population] it would be important to provide further details about certain variables. For example, when it refers to region “non-metropolitan area”, is this the same as rural areas? Also with regards to alcohol drinking, does this variable accounts for amount of alcohol or kind of beverage? As it currently stands, it only refers to number of times they drink alcohol, though not information is given about what they drink and how much they drink each time. Moreover, please add mean (and standard deviation) or median (and interquartile range) age.
- >> a more detailed information was inserted. (#page 6, line 7; #page 7, lines 15; Table 1)

- Cohort description, Participant of the cohort: [key variables] further details should be provided for certain variables. For example, in table 2 it reads “cognitive impairment, depression…mental health screening”. What did this screening include? Was it conducted by a physician (i.e., psychiatrist) or was it based on questionnaires (i.e., PHQ-9)? In association with my previous commentary, there seems no information was retrieved about kind of alcoholic beverage. It is not the same to drink a cup of wine than a cup of whisky. If in fact this information is not available, perhaps this would deserve a line in the limitation section. At this point it would also be important a few details about data collection: what devices were used to assess blood pressure? Was there only one measure or several and the available blood pressure variable is the mean? About body mass index, further details in this line would be appreciated (weight and height was measured or self-reported?). Likewise, what about physical activity? Was it assessed with a questionnaire or with an objective method? When it talks about disabilities, what kind of disabilities it refers to? Permanent disabilities? All these details about variables definitions could go as a supplementary table, but they are indeed needed.
- >> A more detailed information about measurement methods was inserted. (Supplementary table 2)

- Findings to date: the last sentence is conflict. It reads: the NHIS-HEALS will provide evidence regarding the issues that were assessed in previous studies using the NHID. If these issues have already been addressed by the NHID, what does this new cohort add?
- >> Correction was made. (#page 8, lines 9-11)

- Findings to date: [second paragraph] it would be more internationally relevant to show in the text the standardized rates according to the WHO population. Rates according to the Korean census population could then be depicted only in Table 3. Provided this is an international open-access journal, figures that are comparable internationally would be more interesting.
- >> Correction was made. (#pages 8-10)

- Findings to date: [third paragraph] it would be interesting to have the differences between men and women compared with p-values. Unless the authors have other reasons for not presenting this information in the manuscript, a p-value should be included. This is even more important to support statements such as that one in line 12-13 of page 9: the prevalence of diabetes and hypertension was higher in men than in women. This suggestion would also apply for incidence and mortality rates
- >> As suggested, the differences were assessed with statistical test. (#pages 8-9, Tables 3-5)

- Findings to date: [page 9, line 36] it reads screening database in 2005-2013… Then it reads that patients who have a diagnosis within the first three years were excluded from the incidence analysis.

Why was this decision taken? Perhaps at baseline there was not any information about pre-existing conditions. If so, this would explain this decision. In any case, this should be clearly stated. In addition, a reason for choosing this three-year threshold would be appreciated.
- >> A more detailed information was inserted. (#page 9, lines 8-11)

- Strengths and limitations: further discussion should be made around the quality of the variables. For example, much is not said about mortality data. Is this information reliable? Are causes of death well-defined? Is date of death accurately recorded? Is this information as accurate for urban areas as it is for rural areas? Furthermore, once further details about other variables are provided, further discussion about their quality could be included in this section.
- >> As suggested, a more detailed explanation was inserted. (#page 10, lines 21-23)

- Footnote for Supplementary Figure 1: in line 4 it reads …to those age 50 or older… It seems it should read …to those aged 50 or older…
- >> Correction was made. (Supplementary figure 1)