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 Geoscience and Health Cohort Consortium (GECCO) in the Netherlands
AUTHORS	Timmermans, Erik J.; Lakerveld, J; Beulens, Joline; Boomsma, Dorret; Kramer, Sophia; Oosterman, Mirjam; Willemsen, Gonneke; Stam, Mariska; Nijpels, Giel; Schuengel, Carlo; Smit, Jan; Brunekreef, Bert; Dekkers, Jasper; Deeg, Dorly; Penninx, Brenda; Huisman, Martijn

VERSION 1 – REVIEW

REVIEWER	Daniel Lewis LSHTM, UK.
REVIEW RETURNED	13-Feb-2018

GENERAL COMMENTS	This is a cohort profile for a resource that brings together a number of pre-existing cohorts in the Netherlands and links a range of spatial data to the relevant data collection periods within each cohort. The primary reason for the existence of this resource is to facilitate socio-ecological research, providing additional heterogeneity and power to statistical analyses. It is particularly relevant to the study of social and spatial determinants of health and individual risk-factor studies. In general, the cohort profile is very good, but some aspects could be clarified, particularly some factors that are unique to the Netherlands. Dutch Postal and Adminstrative Geographies: In table 2, could you provide some additional statistics in order that the reader might better understand the zones used. In particular an average count of households and/or resident population in PC4s, PC6s and Neighbourhoods, and perhaps average measures of area for these geographies if applicable. Actually, perhaps this is satisfied by Table 3 – in which case please refer to this in the text. Can you make the spatial linkage clear? I assume that for vector data we're either talking about table joins, in which data on postcodes or neighbourhoods are simply merged with participants who share a common identifier. However, for raster data (air pollution, noise) how the linkage we achieved is less obvious – is it simply the raster cell that the centroid of the participant's geography falls within, or are you using a zonal measure, such as the average of all cells touched by the participant's neighbourhood polygon? Although you have chosen to focus on 2006 as an exemplar, is it possible to comment on which exposures are likely to yield time-varying exposure data, and which are likely to be constant over

	particular time periods owing to the update rate of that data resource? You list the dates in Table 2, but is it possible to expand upon this in the text itself for clarity? Coupled to this – is there any information on how you decide to allocate an exposure to a cohort wave where the year is not a good match? Is it simply the nearest year? Do you ever interpolate? 'Livability' should probably be 'Liveability', unless this is a proper noun used in the Netherlands. Possible limitations to add include the generic nature of linkage – there is no discussion of what the true causally relevant context might be (i.e. Kwan's uncertain geographic context problem), which is understandable given the generic nature of the resource. However, if someone approached you and said I want to study a given mechanism, over a given spatial context – let's say neighbourhood polygons and their first order neighbours – could the resource as it is currently structured facilitate that request? Coupled to that, a limitation to acknowledge is that the resource only allows for analyses of the residential neighbourhood, and doesn't encode variables based on other relevant contexts such as work or school, or the activity space or geographic life environment.

REVIEWER	Daikwon Han Department of Epidemiology & Biostatistics School of Public Health Texas A&M University College Station, TX 77843 USA
REVIEW RETURNED	14-Feb-2018

GENERAL COMMENTS	This study addresses one of the important topics in spatial and environmental epidemiology – growing need for combining geospatial and epidemiology/health data for further etiologic investigation of multiple environmentally-caused chronic health outcomes. Spatial epidemiology, as a relatively new approach within the discipline, has provided great opportunities to engage in multidisciplinary works on the environment and human health, and thus provided added values in environmental epidemiology studies. However, one of the areas for further development includes data linkages between geospatial, environmental, and epidemiology/health data, to confirm the epidemiologic associations identified from descriptive spatial epidemiology studies, and this project would certainly provide data required for such investigation. I like the overall approach, and it could be usefully strengthened in a number of ways which I detail below; This paper is well-organized, and included major subsections required for this special type of manuscript; while opportunities associated with data linkages are well described, a few issues regarding important challenges need to be further discussed. First of all, I would like to see some discussions on spatial and temporal match (or mismatch) issues overall, including rates of missing data (if you have done this already), and how would you handle those unmatched and/or missing ones?

Specifically, any spatial mismatch issues you have encountered in the project? And how would you justify spatial scales you selected, like neighborhood, pc4, pc6…listed in Table 2? Also are there any other issues of data integration/linkage such as privacy/confidentiality? Similarly, what about temporal scale? Are there any temporal match/mismatch issues? If so, how would deal with this temporal mismatch issue (again if you have done this already)? One or two examples with some discussion would certainly help. This is one of the key questions in linking and combining epidemiology/health and geospatial data as there are often geospatial data available at multiple spatial and temporal scales.

It would be more impactful if the authors could discuss/provide additional pieces of information on dissemination plan and/or type of analytical epidemiology investigation (i.e. added values) that could be conducted with the completion of this project. For example, this is a special type of manuscript that requires cohort to be longer-term and/or prospective projects, while there are many types of cohort studies (including prospective cohort study, retrospective cohort study) in epidemiology; so what types of cohort study (or any other study designs) would be well suited to take advantage of the linked data available from this project, and why?

Minor comments:

Page 5 lines 15-17: Expand this sentence and give a few examples (or citation) for “objectively measured geospatial data” available there, maybe with some commonly available ones in many countries and/or unique ones in your own setting if any.

Page 12, line 5: Need to tell the readers who are not familiar with your study region what is the “neighborhood” defined by statistics Netherland?

Table 2: Add a footnote for those terms like pc4, pc6….

VERSION 1 – AUTHOR RESPONSE

REVIEWER 1
This is a cohort profile for a resource that brings together a number of pre-existing cohorts in the Netherlands and links a range of spatial data to the relevant data collection periods within each cohort. The primary reason for the existence of this resource is to facilitate socio-ecological research, providing additional heterogeneity and power to statistical analyses. It is particularly relevant to the study of social and spatial determinants of health and individual risk-factor studies. In general, the cohort profile is very good, but some aspects could be clarified, particularly some factors that are unique to the Netherlands.

Reviewer 1, Comment 1
Dutch Postal and Administrative Geographies: In table 2, could you provide some additional statistics in order that the reader might better understand the zones used. In particular an average count of households and/or resident population in PC4s, PC6s and Neighbourhoods, and perhaps average measures of area for these geographies if applicable. Actually, perhaps this is satisfied by Table 3 – in which case please refer to this in the text.

Response:
We agree and have added these details as well as the corresponding references into the revised manuscript (page 9), as follows: “In the Netherlands, 6-digit postal code areas (average area size: 0.0025km2), 4-digit postal code areas (average area size: 8.3km2) and neighbourhoods (average area size: 3.1km2) are geographically delineated areas within municipalities and include, on average, approximately 15, 1,870 and 630 households, respectively [16-21].”

Reviewer 1, Comment 2
Can you make the spatial linkage clear? I assume that for vector data we’re either talking about table joins, in which data on postcodes or neighbourhoods are simply merged with participants who share a common identifier. However, for raster data (air pollution, noise) how the linkage we achieved is less obvious – is it simply the raster cell that the centroid of the participant’s geography falls within, or are you using a zonal measure, such as the average of all cells touched by the participant’s neighbourhood polygon?

Response:
Thank you for pointing this out. We agree and have added more details about the spatial linkage into the manuscript.

As described on page 9, environmental data on the 6-digit postal code- and 4-digit postal code-level could be directly merged with individual-level cohort data using respondents’ postal codes.

As described on page 12, the Geographic Information Systems (GIS) technique of spatial joining was used to link the layer with environmental data on the neighbourhood-level to the layer with 6-digit postal codes. Subsequently, we were able to merge neighbourhood data with individual-level cohort data using respondents’ 6-digit postal codes. In the revised manuscript, we now explicitly mention ‘spatial joining’ as the used GIS technique (page 12).

For the assessment of raster data on air pollution and the linkage of these data to individual addresses, we refer to references 40-43 in the revised manuscript. In GECCO, the address-level concentrations of air pollutants were aggregated to mean values of 6-digit postal code areas. This aggregation of data facilitated the linkage of these data to individual-level data of the various cohort studies. We have added this to the manuscript (page 11).

For the assessment of raster data on traffic noise, we refer to reference 47 in the revised manuscript. We have now added to the manuscript that the noise level of a particular raster cell was linked to the point locations of all addresses that fall within that specific raster cell (page 11). The point locations of all addresses in the Netherlands were obtained from the Register of Addresses and Buildings (BAG-register, June 2015) of the Netherlands’ Cadastre, Land Registry, and Mapping Agency and the linkage was performed by using GeoDMS software (Object Vision BV, Amsterdam, the Netherlands). In GECCO, the address-level traffic noise data were aggregated to mean values of 6-digit postal code areas. This aggregation of data was needed to facilitate the linkage of these data to individual-level data of the various cohort studies (page 11).

Reviewer 1, Comment 3
Although you have chosen to focus on 2006 as an exemplar, is it possible to comment on which exposures are likely to yield time-varying exposure data, and which are likely to be constant over particular time periods owing to the update rate of that data resource? You list the dates in Table 2, but is it possible to expand upon this in the text itself for clarity? Coupled to this – is there any information on how you decide to allocate an exposure to a cohort wave where the year is not a good match? Is it simply the nearest year? Do you ever interpolate?

Response:
Thank you for pointing this out. We now expand upon this in the revised manuscript (page 16-17): "Some geo-data (e.g., traffic noise) have been suggested to vary more over time than other geo-data (e.g., air pollution) [44-46,66]. A strength of GECCO is that a variety of geo-data have been collected for different years. This makes it possible to link most geo-data to the exact assessment period of the cohort studies. For this cohort profile, geo-data were linked over periods as closely matched to the assessment period of the various cohort studies, resulting in a temporal mismatch of a maximum of 5 years for some of the participants. This particular mismatch is related to the linkage of 2009-data on air pollution to data from NESDA in 2004, and can still be considered as an accurate match [44-46]."

Reviewer 1, Comment 4
'Livability' should probably be 'Liveability', unless this is a proper noun used in the Netherlands.

Response:
As suggested by the reviewer, we have changed 'livability' into 'liveability' throughout the manuscript.

Reviewer 1, Comment 5
Possible limitations to add include the generic nature of linkage – there is no discussion of what the true causally relevant context might be (i.e. Kwan's uncertain geographic context problem), which is understandable given the generic nature of the resource. However, if someone approached you and said I want to study a given mechanism, over a given spatial context – let's say neighbourhood polygons and their first order neighbours – could the resource as it is currently structured facilitate that request? Coupled to that, a limitation to acknowledge is that the resource only allows for analyses of the residential neighbourhood, and doesn't encode variables based on other relevant contexts such as work or school, or the activity space or geographic life environment.

Response:
In GECCO, geo-data on the address-, postal code- and neighbourhood-level have been collected. The data about these areas (e.g., neighbourhood polygons) can easily be linked using postal code information from cohort respondents. Data about 'first-order neighbour areas' are available within GECCO, but it requires additional steps, including the identification of these 'first-order neighbour areas' in ArcGIS.

Although some environmental data were collected on the very detailed address-level, we now acknowledge in the manuscript's 'Strengths and limitations' section that most of the collected geo-data in GECCO are only related to the administrative areas where people are living, and that these data are not related to other specific contexts that might also impact health and wellbeing of individuals (e.g., work environment or exact geographic life environment) (page 17).

REVIEWER 2
This study addresses one of the important topics in spatial and environmental epidemiology – growing need for combining geospatial and epidemiology/health data for further etiologic investigation of multiple environmentally-caused chronic health outcomes. Spatial epidemiology, as a relatively new approach within the discipline, has provided great opportunities to engage in multidisciplinary works on the environment and human health, and thus provided added values in environmental epidemiology studies. However, one of the areas for further development includes data linkages between geospatial, environmental, and epidemiology/health data, to confirm the epidemiologic associations identified from descriptive spatial epidemiology studies, and this project would certainly provide data required for such investigation. I like the overall approach, and it could be usefully strengthened in a number of ways which I detail below. This paper is well-organized, and included

major subsections required for this special type of manuscript; while opportunities associated with data linkages are well described, a few issues regarding important challenges need to be further discussed.

Reviewer 2, Comment 1
First of all, I would like to see some discussions on spatial and temporal match (or mismatch) issues overall, including rates of missing data (if you have done this already), and how would you handle those unmatched and/or missing ones? Specifically, any spatial mismatch issues you have encountered in the project? And how would you justify spatial scales you selected, like neighborhood, pc4, pc6…listed in Table 2? Also are there any other issues of data integration/linkage such as privacy/confidentiality? Similarly, what about temporal scale? Are there any temporal match/mismatch issues? If so, how would deal with this temporal mismatch issue (again if you have done this already)? One or two examples with some discussion would certainly help. This is one of the key questions in linking and combining epidemiology/health and geospatial data as there are often geospatial data available at multiple spatial and temporal scales.

Response:
Thank you for these suggestions to improve our manuscript.

As described on page 6, missing geo-data or postal code information (e.g., as a result of living abroad) were the main reasons for unsuccessful data-linkage. In total, 93 respondents without postal code information were excluded from the present analyses. The number of respondents without postal code information per cohort can be derived from the cohort descriptions on pages 6-8. As described on page 14, we could link geo-data to 44,657 individuals. As indicated in the manuscript, the sample size may vary for some environmental variables, because of missing values. Due to item missing values the full sample varied from n=43,046 (96.4%; for number of disability insurances per 1000 residents) to n=44,567 (100.0%; for number of men and women).

As suggested by the reviewer, we have to deal with several other methodological issues in GECCO, including temporal mismatching and maintaining privacy of respondents. We now discuss these methodological issues in more detail in the manuscript's 'Strengths and limitations' section (pages 16-17):

"Some geo-data (e.g., traffic noise) have been suggested to vary more over time than other geo-data (e.g., air pollution) [44-46,66]. A strength of GECCO is that a variety of geo-data have been collected for different years. This makes it possible to link most geo-data to the exact assessment period of the cohort studies. For this cohort profile, geo-data were linked over periods as closely matched to the assessment period of the various cohort studies, resulted in a temporal mismatch of a maximum of 5 years for some of the participants. This particular mismatch was related to the linkage of 2009-data on air pollution to data from NESDA in 2004, and can still be considered as an accurate match [44-46]. For this cohort profile, the linkage of geo-data with individual-level cohort data was done locally without confidential information (e.g., residential addresses) leaving the research premises, so that the privacy of respondents is safeguarded at all times. Although some geo-data were collected on the address-level, it should be acknowledged that most collected geo-data in GECCO are related to administrative residential areas, and are not related to specific contexts (e.g., work environment or exact geographic life environment) that might also impact health and wellbeing of individuals [67]."

In the Netherlands, 6-digit postal code areas, 4-digit postal code areas and neighbourhoods are geographically delineated areas within municipalities. In the Netherlands, geo-data are mainly available for these administrative areas. Therefore, we selected these spatial scales. To make the selected areas more clear, we now have described them in more detail and added additional information (i.e., average area size and average number of households) about these areas to the revised manuscript (page 9):

“In the Netherlands, 6-digit postal code areas (average area size: 0.0025km2), 4-digit postal code areas (average area size: 8.3km2) and neighbourhoods (average area size: 3.1km2) are geographically delineated areas within municipalities and include, on average, approximately 15, 1870 and 629 households, respectively [16-21].”

Reviewer 2, Comment 2
It would be more impactful if the authors could discuss/provide additional pieces of information on dissemination plan and/or type of analytical epidemiology investigation (i.e. added values) that could be conducted with the completion of this project. For example, this is a special type of manuscript that requires cohort to be longer-term and/or prospective projects, while there are many types of cohort studies (including prospective cohort study, retrospective cohort study) in epidemiology; so what types of cohort study (or any other study designs) would be well suited to take advantage of the linked data available from this project, and why?

Response:
The main objective of GECCO was to enrich databases of prospective longitudinal cohort studies with existing geodata. However, each type of study can benefit from linkage of environmental from GECCO, including cross-sectional studies and intervention studies. To illustrate the added value of GECCO to this field of research, we now discuss recent findings and current efforts in GECCO in the “Findings to date” section (page 15).

In the ‘Strengths and limitations’ section (page 16), we further emphasize that GECCO facilitates and enables researchers from various disciplines to address research questions on the complex relationships between the environment and health outcomes. The collaboration between the cohort studies in GECCO increases the power of analyses and ensures sufficient geographical variation in environmental determinants.

Reviewer 2, Comment 3
Page 5 lines 15-17: Expand this sentence and give a few examples (or citation) for “objectively measured geospatial data” available there, maybe with some commonly available ones in many countries and/or unique ones in your own setting if any.

Response:
We agree and have now added a few examples of objectively measured environmental data (page 5). These examples include air pollution, traffic noise and area demographics.

Reviewer 2, Comment 4
Page 12, line 5: Need to tell the readers who are not familiar with your study region what is the “neighborhood” defined by statistics Netherland?

Response:
Statistics Netherlands defines a neighbourhood as a geographically delineated (administrative) area within a municipality. We have now added this definition to the revised manuscript (page 9): “In the Netherlands, 6-digit postal code areas (average area size: 0.0025km2), 4-digit postal code areas (average area size: 8.3km2) and neighbourhoods (average area size: 3.1km2) are geographically delineated areas within municipalities and include, on average, approximately 15, 1,870 and 630 households, respectively [16-21].”

Reviewer 2, Comment 5
Table 2: Add a footnote for those terms like pc4, pc6.

Response:
Thank you for pointing this out. We have now added a footnote regarding these terms in Table 2 (page 32): “a Abbreviations: PC4= 4-digit postal code area; PC6= 6-digit postal code area.”

VERSION 2 – REVIEW

REVIEWER	Daniel Lewis London School of Hygiene and Tropical Medicine, UK.
REVIEW RETURNED	27-Mar-2018

GENERAL COMMENTS	I am satisfied that the authors have adequately responded to the minor points raised. I have no further comments on this revision.

REVIEWER	Daikwon Han Texas A&M University USA
REVIEW RETURNED	02-Apr-2018

GENERAL COMMENTS	Thank you for your careful revision of the paper and your response to my comments. You have answered all the questions I had in my initial review.