PEER REVIEW HISTORY

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

(This paper received three reviews from its previous journal but only two reviewers agreed to published their review.)

ARTICLE DETAILS

TITLE (PROVISIONAL)	Association of Neighborhood Socioeconomic Status and Diabetes Burden using Electronic Health Records in Madrid (Spain): The Heart Healthy Hoods Study
AUTHORS	Bilal, Usama; Hill-Briggs, F; Sanchez-Perruca, Luis; del Cura, Isabel; Franco, M

VERSION 1 – REVIEW

REVIEWER	Nida I. Shaikh Emory University, United States
REVIEW RETURNED	17-Jan-2018

GENERAL COMMENTS	Dear Authors, Interesting article that sheds light on the neighborhood SES and diabetes burden in Spain. There are, however, several concerns, with the paper. Overall concerns/comments 1. The discussion is weak and barely compares the results with similar studies that were stated to have been done in US and other parts of Europe. The implications of the results need to be explained. There is also no clear justification to tie the results with social determinants of health and unhealthy food environments as done on page 18. What are the implications of the SES-based diabetes prevalence, incidence, and diabetes control for diabetes management, prevention and care/treatment, and policy? Should future intervention or policy work in this area be SES and/or gender driven? These are some of the things that discussion should include and would strengthen the paper. 2. The strength of the paper is in the NSES index, however, it is poorly explained and its components aren't fully justified. I recommend including some of the description from the supplementary page in the manuscript on page 7. Also, having 2 of 7 indicators on education (high and low levels) would makes the index unstable; a strong justification is needed to include both. Lastly, the index needs to be described; what is the range, why does it have negative values (Figure 1)

	Minor concerns 1. EHR is mentioned in the title and paper, but not in the abstract. 2. The use of the term tertiles as a 'categorical variable' as done on page 7 is incorrect. Please correct to 'categorical variable (NSES index tertiles)'. 3. Analysis has many different approaches. Are all required? 4. Is the logistic regression, polytomous logistic regression? If so, please add that in. 5. There are several grammatical errors. E.g.: Textile on page 10 6. After the first mention of a word along with its abbreviation, only the abbreviation should be used. Please correct for NSES and EHR. 7. On page 20, there is inconsistent use of author name (see funding and acknowledgement) and author initials (see author contributions). 8. If the sample is restricted to adults 40 years and older, why is the employment rate calculated for age 16-64? (page 7). 9. On page 2, please add 'socioeconomic status (SES)'.... 10. Page 2, line 19: please add '7 indicators from 4 domains of' education.... 11. Page 2, line 47: 'socially patterned by contextual SES' is a vague phrase. 12. Page 4, first 2 paragraphs need a strong justification for the need for this study to be conducted in Southern Europe. 13. Page 8, line 5: define 'continuous variable' as done on page 12 line 51-52. 14. Figure 1 was small and hard to read. 15. Table 1 needs 1) a footnote; each table should be self-explanatory and 2) p values - is there is statistical significance between the three groups (low, medium, and high NSES)? 16. Page 10 lines 25-30: Variable CVD, hypertension, etc have not been described nor defined in the methods.

REVIEWER	Hongjiang Wu Usher Institute of Population Health Sciences and Informatics
REVIEW RETURNED	31-Jan-2018

GENERAL COMMENTS	This paper describes the association between neighbourhood SES and prevalence, incidence and control of diabetes in Madrid (Spain). There are several concerns for the authors to consider. 1. More detailed information on how diabetes was diagnosed for both prevalence and incidence is required; and please comment whether SES would influence the ascertainment of diabetes. 2. Much stronger argument is needed why the authors restricted the analysis to people aged 40 or above. Given the large sample size, low prevalence of type 2 diabetes in young individuals would not be a problem. In addition, it is very interesting to explore whether SES gradients converge or diverge among older adults. Many previous publications have reported age modifies the association between SES and health outcome. 3. The validity of composite measure of SES is questionable. The authors restricted the study population in people aged 40 or above, and the median age is 56.2. What's the retirement age in Spain? What's the proportion of people who are already retired in the study population? Among 7 indicators, 4 of them are related to occupation. However, occupation is not a very useful SES indicator for people who are retired.

	4. Please state how many people died during follow-up. Have authors considered about the completing risk from death? Also, please state any loss to follow-up. 5. Have the authors tried to put the risk factors (hypertension, CVD, CKD, and retinopathy) into the models to examine the mechanisms linking SES and diabetes? 6. The number incident diabetes in each SES group needs to be stated. 7. Any missing data in variables? Please clarify how missing were handled. 8. Does the 'average age' refer to 'median age'? Please clarify it. 9. Please comments on whether you think Madrid is a representative of other cities in Spain and how it maybe different.

REVIEWER	Cruz Velasco University of Arkansas for Medical Sciences, USA
REVIEW RETURNED	01-Mar-2018

GENERAL COMMENTS	• Overall manuscript is well presented. • Page 7, lines 29-40: It is unclear how the index of NSES was constructed. Make it explicit in manuscript or in Online Resource. Fragment "three standardized indicators" on line 36 seems incorrect (should it be "seven standardized indicators"?). Use of a weighted mean as composite index seems simplistic (a search of "composite index of Neighborhood Socioeconomic Status" shows principal components is a common approach); if authors want to stick with weighted mean, an argument for it should be included. • Page 8, lines 28-30: Make explicit age range limits used; for example 40 to 50 and 50 to 60; to what interval patients with age=50 belong? (could use [40,50), [50,60) the "[" indicates patients with age=50 belong in second group). • Page 10, line 9: "tertile" instead of "textile"? • Description of results: usage of decrease/increase in the odds/hazard gives the impresion of subjects changing neighborhoods. In my opinion, for example on page 12 line 15 "had a 10% decrease in the odds of" could be re-expressed as "had a 10% decrease in estimated odds of" or "had 10% lower odds of". • Page 15, line 11: provide a p-value associated with term "gradient". • Page 15. Disclose whether proportional hazards assumption for Cox models was checked.

REVIEWER	C Leigh Blizzard Menzies Institute for Medical Research University of Tasmania Australia
REVIEW RETURNED	08-Mar-2018

GENERAL COMMENTS	You have made a brave claim for the novelty of this study as a precursor to the studies of "mechanistic insights" that must follow. But can you attempt to sharpen the arguments, please. The discussion of the limitations of this study is honest and commendable.

	Please explain how or at least why these particular four districts from the city of Madrid were chosen. At it presently stands, there is room for suspicion that they were a non-random selection and that casts doubt on whether or not statistical tests based on independently and identically distributed random sampling variables are appropriate. In addition, non-random sampling would skewer the claimed advantages that this study of administrative records could have over "research-driven cohort studies ... [that are] ... derived from a non-random sampling". Please provide more details about the analyses of incidence, explaining when the period of observation commences and provide reassurances about the completeness of ascertainment if that period reaches back into the past. It is not inappropriate to estimate prevalence odds ratios but, at least in my opinion, prevalence ratios are more readily interpretable. be preferable. On page 18 you refer to a hazard ratio of 0.66 for women when I think it should be 0.69 I will not complain about the use of the word "tertiles" when you mean "thirds". Nor will I complain about references about odds ratios or hazard ratios “in men” or “in women”, when an odds ratio or a hazard ratio cannot be in a man or in a women. They are lost causes. But surely we can agree that "data" is the plural form of "datum", and that the singular usage "data is" (several instances) is to be avoided.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Nida I. Shaikh
Institution and Country: Emory University, United States
Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
Dear Authors,
Interesting article that sheds light on the neighborhood SES and diabetes burden in Spain. There are, however, several concerns, with the paper.

Overall concerns/comments
1. The discussion is weak and barely compares the results with similar studies that were stated to have been done in US and other parts of Europe. The implications of the results need to be explained. There is also no clear justification to tie the results with social determinants of health and unhealthy food environments as done on page 18. What are the implications of the SES-based diabetes prevalence, incidence, and diabetes control for diabetes management, prevention and care/treatment, and policy? Should future intervention or policy work in this area be SES and/or gender driven? These are some of the things that discussion should include and would strengthen the paper.

RESPONSE: We thank the reviewer for this important issue. We have now improved the discussion section comparing with other studies, including those using EHR vs cohorts or surveys, and those studying prevalence, incidence or control of diabetes. We have also

strengthened the section on implications, tightening the connection between the results and the proposed actions. See Pages 18, 20 and 21 of the discussion.

2. The strength of the paper is in the NSES index, however, it is poorly explained and its components aren't fully justified. I recommend including some of the description from the supplementary page in the manuscript on page 7. Also, having 2 of 7 indicators on education (high and low levels) would makes the index unstable; a strong justification is needed to include both. Lastly, the index needs to be described; what is the range, why does it have negative values (Figure 1)

RESPONSE: Again, thank you for this substantial comment and improvement of the article. We have now expanded the methods section focusing on the NSES index. We have also described its main statistics in the results section, and have added a row in Table 1. To clarify, the index has negative values because it's the average Z-score of each indicator, obtained after centering each indicator by its mean and dividing by its standard deviation. This means that census sections with levels of (for example) unemployment below the mean would have a negative value in the unemployment indicator. See pages 6, 7, and 8 of the methods section.

Minor concerns
1. EHR is mentioned in the title and paper, but not in the abstract.

RESPONSE: We have fixed this accordingly.

2. The use of the term tertiles as a 'categorical variable' as done on page 7 is incorrect. Please correct to 'categorical variable (NSES index tertiles)'.

RESPONSE: We have fixed this accordingly.

3. Analysis has many different approaches. Are all required?

RESPONSE: Our analysis consists of two types of outcomes (binary and time-to-event) and therefore we require two types of regressions (logistic and Cox proportional hazards). If the editor believes we should make the manuscript leaner we can, for example, remove the part related to the analysis of NSES as a continuous variable.

4. Is the logistic regression, polytomous logistic regression? If so, please add that in.

RESPONSE: this is a binary logistic regression, where the outcome is prevalence or lack of control of diabetes.

5. There are several grammatical errors. E.g.: Textile on page 10

RESPONSE: Thanks for noticing, we have fixed several of these accordingly.

6. After the first mention of a word along with its abbreviation, only the abbreviation should be used. Please correct for NSES and EHR.

RESPONSE: We have also removed all references to neighborhood SES or electronic health records after the first mention in the methods section, and proceeded to use the abbreviation.

7. On page 20, there is inconsistent use of author name (see funding and acknowledgement) and author initials (see author contributions).

RESPONSE: We have now made this consistent.

8. If the sample is restricted to adults 40 years and older, why is the employment rate calculated for age 16-64? (page 7).

RESPONSE: NSES is trying to capture a contextual measure of socioeconomic status, and therefore is not restricted to the same age restrictions as the EHR data. This also applies to education data (measured in people aged 25 or above).

9. On page 2, please add 'socioeconomic status (SES)'....

10. Page 2, line 19: please add '7 indicators from 4 domains of' education....

11. Page 2, line 47: 'socially patterned by contextual SES' is a vague phrase.

12. Page 4, first 2 paragraphs need a strong justification for the need for this study to be conducted in Southern Europe.

13. Page 8, line 5: define 'continuous variable' as done on page 12 line 51-52.

RESPONSE (9-14): Thanks for these suggestions. We have adopted them and fixed the errors or needs for stronger justification.

14. Figure 1 was small and hard to read.

RESPONSE: We have improved the readability of the figures by splitting figure 1 into two figures (one with the prevalence and the other with the two control outcomes).

15. Table 1 needs 1) a footnote; each table should be self-explanatory and 2) p values - is there is statistical significance between the three groups (low, medium, and high NSES)?

16. Page 10 lines 25-30: Variable CVD, hypertension, etc have not been described nor defined in the methods.

RESPONSE (15-16): We have included a footnote in table 1, reordered the stable slightly and added a p-value column. We have removed other CVD risk factors from table 1.

Reviewer: 2
Reviewer Name: Hongjiang Wu
Institution and Country: Usher Institute of Population Health Sciences and Informatics
Please state any competing interests or state 'None declared': None

Please leave your comments for the authors below
This paper describes the association between neighbourhood SES and prevalence, incidence and control of diabetes in Madrid (Spain). There are several concerns for the authors to consider.

1. More detailed information on how diabetes was diagnosed for both prevalence and incidence is required; and please comment whether SES would influence the ascertainment of diabetes.

RESPONSE: All the health data in our study come from the electronic health records used by primary care physicians during their regular practice in Madrid. Diabetes was diagnosed by physicians during their usual clinical practice, and coded using the International Classification

for Primary Care version 2 (ICPC-2). We have no reason to believe that there would be differential measurement error by SES since Spain has a Universal Health Care system, and even people whose main physician is in a private practice will ultimately have their records in the public primary care system for prescription purposes. We have added more details on this important issue in the text. See page 19 of the discussion.

2. Much stronger argument is needed why the authors restricted the analysis to people aged 40 or above. Given the large sample size, low prevalence of type 2 diabetes in young individuals would not be a problem. In addition, it is very interesting to explore whether SES gradients converge or diverge among older adults. Many previous publications have reported age modifies the association between SES and health outcome.

RESPONSE: We thank the reviewer for this insight. While we agree with the reviewer that understanding SES-diabetes associations in younger people is of interest, the main reason behind our age restriction is that CVD risk factor data is only collected systematically in people aged 40 or above. We have added more details about this issue in the text. See page 6 of the methods section.

3. The validity of composite measure of SES is questionable. The authors restricted the study population in people aged 40 or above, and the median age is 56.2. What's the retirement age in Spain? What's the proportion of people who are already retired in the study population? Among 7 indicators, 4 of them are related to occupation. However, occupation is not a very useful SES indicator for people who are retired.

RESPONSE: Retirement age is 65 in Spain, but this does not factor into our Neighborhood SES index, that is a contextual index of SES. That is, the index is constructed with census or analogous data (see appendix), not with individual-level data, so retired people do not affect the calculation of the index.

4. Please state how many people died during follow-up. Have authors considered about the completing risk from death? Also, please state any loss to follow-up.

RESPONSE: Around 1.2% of the sample died during follow-up and 0.8% moved out of the study area. We have now included this important information in Table 1. These would be the only two sources of lost to follow-up, otherwise the data are complete. The analysis of incidence we conducted is equivalent to estimating cause-specific hazards (in competing risks terms), so they are interpretable as the hazard of diabetes in people that do not die or move out of the area. We have now included more details in the text. See pages 8 and 9 of the methods section. We have also tested a sub-distribution hazard approach for competing risks, leading to the same results as in our main analyses. If the editor feels it warranted, we can include the details of that sensitivity analysis in the paper as an appendix.

5. Have the authors tried to put the risk factors (hypertension, CVD, CKD, and retinopathy) into the models to examine the mechanisms linking SES and diabetes?

RESPONSE: This is an interesting suggestion, but we believe that those factors (especially CVD, CKD and Retinopathy) are consequences of diabetes, not intermediary links between SES and Diabetes. In future research, we will examine the association of SES with those factors and the mediating role of diabetes incidence, but this may not be appropriate for this analysis and manuscript. As suggested, we have now removed those factors for table 1 to clear it, given that they don't provide much information.

6. The number incident diabetes in each SES group needs to be stated.

RESPONSE: We have included this in Table 1 now.

7. Any missing data in variables? Please clarify how missing were handled.

RESPONSE: Data is complete in terms of age, sex and diabetes diagnosis. We have now included how many diabetics do not have an HbA1c % measurement during the study period (21%). Our main analysis is a complete case analysis. We have now included details on a second analysis using conditional mean imputation (using age, sex, diagnosis of other cardiovascular conditions [hypertension, CVD, retinopathy, CKD, dyslipidemia], SES index of the area, and health care center), showing that the inferences do not change when correcting for missing data. See page 15 of results and Appendix Figure 2.

8. Does the 'average age' refer to 'median age'? Please clarify it.

RESPONSE: Yes, we have now corrected this.

9. Please comments on whether you think Madrid is a representative of other cities in Spain and how it maybe different.

RESPONSE: Madrid is the largest city in Spain, and comparisons with small cities in Spain may be challenging. However, comparison with mid to large cities (e.g., Barcelona, Valencia, Bilbao, or Seville) may be feasible. We have now commented this on the text. See page 20 of the discussion.

Reviewer: 3
Reviewer Name: Cruz Velasco
Institution and Country: University of Arkansas for Medical Sciences, USA
Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below
• Overall manuscript is well presented.
• Page 7, lines 29-40: It is unclear how the index of NSES was constructed. Make it explicit in manuscript or in Online Resource. Fragment "three standardized indicators" on line 36 seems incorrect (should it be "seven standardized indicators"?). Use of a weighted mean as composite index seems simplistic (a search of "composite index of Neighborhood Socioeconomic Status" shows principal components is a common approach); if authors want to stick with weighted mean, an argument for it should be included.

RESPONSE: We have now expanded the section where we explain the details on this index. We made the choice to use a weighted mean decision based on previous studies in the same context showing that using data-driven weights (through principal component analysis) does not change inferences as compared to a weighted mean (Please see our previous publication Gullón et al, IJHG 2017). See pages 6, 7 and 8 of the methods section.

• Page 8, lines 28-30: Make explicit age range limits used; for example 40 to 50 and 50 to 60; to what interval patients with age=50 belong? (could use [40,50), [50,60) the "[" indicates patients with age=50 belong in second group).

RESPONSE: We have now made this explicit (40 to 49, 50 to 59, etc.).

• Page 10, line 9: “tertile” instead of “textile”?

RESPONSE: We have now fixed this typo and other mentions to “Textile”.

• Description of results: usage of decrease/increase in the odds/hazard gives the impresion of subjects changing neighborhoods. In my opinion, for example on page 12 line 15 “had a 10% decrease in the odds of” could be re-expressed as “had a 10% decrease in estimated odds of” or “had 10% lower odds of”.

RESPONSE: We have now fixed the wording of these sentences. We have changed our models from logistic to log-binomial based on the suggestion of Reviewer #4, and have elected to use the “had 10% lower prevalence of diabetes” formula for all of these descriptions.

• Page 15, line 11: provide a p-value associated with term “gradient”.

RESPONSE: We have now included a p-value.

• Page 15. Disclose whether proportional hazards assumption for Cox models was checked.

RESPONSE: The proportionality assumption was assessed by looking at Schoenfeld residuals. We have now included this in the text.

Reviewer: 4
Reviewer Name: C Leigh Blizzard
Institution and Country: Menzies Institute for Medical Research, University of Tasmania, Australia
Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
You have made a brave claim for the novelty of this study as a precursor to the studies of "mechanistic insights" that must follow. But can you attempt to sharpen the arguments, please.

RESPONSE: We have now included specific ideas for future research, looking at previous research we have conducted in the same environment (Madrid) that studies the potential mechanisms linking NSES and diabetes. See pages 20 and 21 of the discussion.

The discussion of the limitations of this study is honest and commendable.

RESPONSE: Thanks!

Please explain how or at least why these particular four districts from the city of Madrid were chosen. At it presently stands, there is room for suspicion that they were a non-random selection and that casts doubt on whether or not statistical tests based on independently and identically distributed random sampling variables are appropriate. In addition, non-random sampling would skewer the claimed advantages that this study of administrative records could have over "research-driven cohort studies ... [that are] ... derived from a non-random sampling".

RESPONSE: These districts were chosen because they all belonged to the same “Health Area”, an organizational division of the Madrid Health Care System. This health area was the one where electronic health records were first implemented, achieving a high degree of

standardization of data collection and also allowing for longitudinal data analyses. We have now included a figure in the appendix showing how these districts compare to the rest of Madrid, and have clarified this in the text. See page 6 of the methods section.

Please provide more details about the analyses of incidence, explaining when the period of observation commences and provide reassurances about the completeness of ascertainment if that period reaches back into the past.

RESPONSE: We have now included details on the beginning and end of risk sets. See page 9 of the methods section.

It is not inappropriate to estimate prevalence odds ratios but, at least in my opinion, prevalence ratios are more readily interpretable. be preferable.

RESPONSE: We agree with the reviewer, and have changed our models to log-binomial models and reported prevalence ratios instead.

On page 18 you refer to a hazard ratio of 0.66 for women when I think it should be 0.69

RESPONSE: We have now reviewed all numbers and fixed any inconsistencies.

I will not complain about the use of the word "tertiles" when you mean "thirds". Nor will I complain about references about odds ratios or hazard ratios “in men” or “in women”, when an odds ratio or a hazard ratio cannot be in a man or in a women. They are lost causes. But surely we can agree that "data" is the plural form of "datum", and that the singular usage "data is" (several instances) is to be avoided.

RESPONSE: We have now fixed all verbs acting on “data” to be in their plural form. We have also reworded the reference to hazard ratios.

VERSION 2 – REVIEW

REVIEWER	Cruz Velasco Ochsner Health System, USA
REVIEW RETURNED	09-May-2018

GENERAL COMMENTS	The reviewer completed the checklist but made no further comments.