PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Association of community types and features in a case-control analysis of new onset type 2 diabetes across a diverse geography in Pennsylvania
AUTHORS	Schwartz, BS; Pollak, Jonathan; Poulsen, Melissa; Bandeen-Roche, Karen; Moon, Katherine; DeWalle, Joseph; Siegel, Karen; Mercado, Carla; Imperatore, Giuseppina; Hirsch, Annemarie

VERSION 1 – REVIEW

REVIEWER	Kelly Hunt Medical University of South Carolina, United States
REVIEW RETURNED	29-Sep-2020

GENERAL COMMENTS	The manuscript entitled, “Association of community types and features in a case-control analysis of new onset type 2 diabetes across a diverse geography in Pennsylvania” is well written and of interest to readers of BMJ Open. I have very few comments. 1. Please add a section in Table 1 that indicates how many people are in the 8 categories that combine Administrative community type of residence and setting of residence. It would be helpful to have a sense of how large those 8 categories are. If you don’t have room for that with what is included, I would include only the n for the 8 combined categories. 2. In Table 2 for Model 3, please write out who is in each of the eight categories rather than using acronyms. 3. The issues of undiagnosed diabetes was discussed to some extent and a comment was made in the discussion concerning undiagnosed diabetes rates not likely to be different across different community types. As a reader, I am concerned that there is more of an issue with undiagnosed diabetes than what was discussed. While everyone in the study received medical care, are there any metrics available on routine testing for diabetes in this population. Also, a reason was not included as to why you would not expect undiagnosed diabetes to be more common in rural areas. I would think that rates would be higher in rural areas if transportation to appointments was a factor, but this was not discussed. 4. To ensure incident diabetes cases were similar across community types for the incident cases, it might be helpful to add some information comparing incident diabetes cases by community type. If you have HbA1c levels at diagnosis it would be helpful to compare. Also, it would be helpful to look at age, and other demographic characteristics by community type. I would only pick one characterization of community type to examine. But this would provide some evidence that incident cases were identified at similar time points in the progression of disease across the different community types.

	5. Please clarify how you defined 'without evidence of diabetes'. Was it only that they did not meet the definition of having incident diabetes? What about border line people? Please clarify your definition for controls. 6. In the write up of the methods it is a little confusing when the 'four definitions of community' are discussed. This becomes clear when you look at table 2, but please clarify in the methods section what the four definitions are. 7. Page 11, paragraph 2. When you refer to 'all community types' on line 239 it is confusing as to which community definition you are talking about. Please clarify which community definition you are talking about.

REVIEWER	Dr Jesse Wiki University of Canterbury, New Zealand
REVIEW RETURNED	15-Nov-2020

GENERAL COMMENTS	This is an interesting article, it has the potential to offer a novel contribution to the field of spatial epidemiology and health geography. Overall, it provides a good theoretical basis and rationale for the study and a clear methodology. What is a lacking is reference to literature on the implications of geographic scale/resolution (particularly in health research), there is plenty of work discussing this and it seems like an integral aspect of this study. This should be included in the introduction. On a similar note, you rightly point out that the spatial scales and measures you have used are more robust than county-level, but you do not discuss that these are still limited in the sense that they are administrative boundaries. For example, such areas still have limitations such as edge and boundary effects and there is no mention of the Modifiable Areal Unit Problem. Furthermore, as the areas used are not community defined these may not actually be reflective of how populations view the area in which they reside which may also have implications for the associations found. An acknowledgement of these aspects is needed within the discussion. You also state that over 40% of individuals within the study resided in rural areas, how might this have affected your results? A point about this is needed in the discussion. Methods section, line 154, add 'the' ("township/rural was the reference group"). Discussion section, line 280 ("while prior studies have evaluated county differences"), citations needed.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
The manuscript entitled, "Association of community types and features in a case-control analysis of new onset type 2 diabetes across a diverse geography in Pennsylvania" is well written and of interest to readers of BMJ Open. I have very few comments.

We thank the reviewer for this feedback.

1. Please add a section in Table 1 that indicates how many people are in the 8 categories that combine Administrative community type of residence and setting of residence. It would be helpful to have a sense of how large those 8 categories are. If you don't have room for that with what is included, I would include only the n for the 8 combined categories.
In the original submission, we provided the percentages of subjects in the nine categories (eight were in the analysis because one category was very small so was combined with another) in **Table S2** (online supplemental table). As requested, we have added the counts of persons in the analysis to this supplement table in the revised version.

2. In Table 2 for Model 3, please write out who is in each of the eight categories rather than using acronyms.

We have revised **Table 2** as requested.

3. The issues of undiagnosed diabetes was discussed to some extent and a comment was made in the discussion concerning undiagnosed diabetes rates not likely to be different across different community types. As a reader, I am concerned that there is more of an issue with undiagnosed diabetes than what was discussed. While everyone in the study received medical care, are there any metrics available on routine testing for diabetes in this population. Also, a reason was not included as to why you would not expect undiagnosed diabetes to be more common in rural areas. I would think that rates would be higher in rural areas if transportation to appointments was a factor, but this was not discussed.

In the original submission, we thought we addressed these issues by showing data that persons in all our communities, including rural areas, had regular access to health care, with similar average numbers of encounters across community types. The average number of encounters by person within each community type was high. This information was in **Table S3** (online supplemental table). This table showed that both cases and controls had high average numbers of encounters before diagnosis or the control selection date, by administrative community type and also by Medical Assistance status. For example, on average, the number of encounters before diagnosis or the control selection date per person ranged from a low of 31.6 for cases in city census tracts to a high of 36.8 for cases in townships (our most rural areas). Values for controls were similar. We thus believe that access to health care was unlikely to explain our findings. We did not think that undiagnosed diabetes was likely in rural areas because persons in rural areas had an average number of encounters per person that was slightly higher than for other areas. We believe that with this long-tern regular care, which did not differ by administrative community type, that persons who developed diabetes would be identified.

4. To ensure incident diabetes cases were similar across community types for the incident cases, it might be helpful to add some information comparing incident diabetes cases by community type. If you have HbA1c levels at diagnosis it would be helpful to compare. Also, it would be helpful to look at age, and other demographic characteristics by community type. I would only pick one characterization of community type to examine. But this would provide some evidence that incident cases were identified at similar time points in the progression of disease across the different community types.

We believe that much of the requested information was in the original submission in the Online Supplemental File. For example, in **Table S2** we provided the average age, average duration of contact with the health care system, sex, race, ethnicity, average BMI, and other features by persons residing in the different administrative community types. In **Table S4**, we provided the number of measures, the number of persons with the measure both in the one year before and the one year after diagnosis or control selection date, and the average value among those with the measure both in the one year before diagnosis or control selection date and the one year after, for each of HbA1c, LDL cholesterol, triglycerides, fasting glucose, random glucose, and BMI. We found that only 14.3% of cases and 4.7% of controls had a HbA1c measured in the year before diagnosis or control selection date, respectively, with means of 5.9 and 5.6. In the year after diagnosis, 75.5% of cases and 4.8% of controls had a HbA1c, with mean values of 7.5 and 5.6, respectively. Conclusions from the other measures in the table are similar in evaluating the validity of our EHR algorithm for diabetes cases and controls. However, we were surprised that there was not more common screening for diabetes in these patients (another example is that only 23.6% of cases had a fasting glucose in the year prior to

diagnosis, with a mean of 108.5, while 58.3% had a value in the year after diagnosis, with a mean of 147.9). The data in **Table S4** clearly show the validity of our algorithms for type 2 diabetes and controls, but we believed there were too many persons without the relevant data to show these data by administrative community type.

5. Please clarify how you defined 'without evidence of diabetes'. Was it only that they did not meet the definition of having incident diabetes? What about border line people? Please clarify your definition for controls.

In the original submission we wrote that controls were "… persons who never met any diabetes criteria." Criteria were based on diagnoses (controls could never have a diabetes diagnosis), medication orders (controls could never have a diabetes medication order), and laboratory tests (controls could never have a HbA1c, fasting glucose, or random glucose that exceeded definitional thresholds). We have revised this sentence to try to improve its clarity as requested. In our experience, it is difficult to operationalize a definition for pre-diabetes or borderline diabetes using EHR data because the laboratory data that would be used for this purpose are not obtained regularly on all patients (as shown in **Table S4**). The U.S. Preventive Services Task Force has recently changed their recommendations for diabetes screening but still do not specify the frequency of the recommended screening, so our EHR data on diabetes screening are not unexpected.

6. In the write up of the methods it is a little confusing when the 'four definitions of community' are discussed. This becomes clear when you look at table 2, but please clarify in the methods section what the four definitions are.

We have identified the four definitions of community type earlier in this paragraph, as requested.

7. Page 11, paragraph 2. When you refer to 'all community types' on line 239 it is confusing as to which community definition you are talking about. Please clarify which community definition you are talking about.

We have clarified this sentence as requested.

Reviewer: 2
This is an interesting article, it has the potential to offer a novel contribution to the field of spatial epidemiology and health geography. Overall, it provides a good theoretical basis and rationale for the study and a clear methodology.

We thank the reviewer for this feedback.

What is a lacking is reference to literature on the implications of geographic scale/resolution (particularly in health research), there is plenty of work discussing this and it seems like an integral aspect of this study. This should be included in the introduction.

We have added references on geographic scale and resolution in health research to the Introduction and Discussion as requested.

On a similar note, you rightly point out that the spatial scales and measures you have used are more robust than county-level, but you do not discuss that these are still limited in the sense that they are administrative boundaries. For example, such areas still have limitations such as edge and boundary effects and there is no mention of the Modifiable Areal Unit Problem. Furthermore, as the areas used are not community defined these may not actually be reflective of how populations view the area in which they reside which may also have implications for the associations found. An acknowledgement of these aspects is needed within the discussion.

We have added these limitations to the Discussion section. However, we note that our approach, with using multiple definitions of community and nested scales, did indirectly MAUP concerns.

You also state that over 40% of individuals within the study resided in rural areas, how might this have affected your results? A point about this is needed in the discussion.

We have addressed this concern in our response to reviewer 1.

Methods section, line 154, add 'the' ("township/rural was the reference group").

We have made this revision as requested.

Discussion section, line 280 ("while prior studies have evaluated county differences"), citations needed.

We have added citations in the requested location.