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.

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 Between Access to Social Service Resources and Cardiometabolic Risk Factors: A Machine Learning and Multi-Level Modeling Analysis
AUTHORS	Berkowitz, Seth; Basu, Sanjay; Venkataramani, Atheendar; Reznor, Gally; Fleegler, Eric W; Atlas, Steven

VERSION 1 – REVIEW

REVIEWER	Matthew B. Lanktree University of Toronto Toronto, Ontario, Canada
REVIEW RETURNED	09-Oct-2018

GENERAL COMMENTS	The authors compare body mass index (BMI), systolic blood pressure (SBP), LDL cholesterol and HbA1C obtained from medical charts of 123,355 people to the "employment resources", "food resources", "nutrition resources", among others, of the participant's home zip code. The authors use access to "big data" and a random forest machine learning technique to come to their conclusions. They find association between improved resource access and lower BMI, but no association with SBP, LDL or HbA1c. Overall, I think the analysis is careful and measured. I do worry about confounding from the inverse relationship between resource availability and socio economic status of the zip codes, which would be expected to reduce the estimate of the effect of the resources.
	Major comments: 1) I struggle with the definition and meaning of terms used throughout the paper. How are "Area resources" defined and tabulated? How does food resources compare to nutrition resources? Employment resources? I believe greater detail regarding the explanatory variables is required for the reader to understand the methods and results of this study. 2) I think it would be valuable to compare the results of the current VSURF analysis to a more traditional model building regression approach. 3) The authors should be commended for their efforts at robustness checks. The authors use "after school resources" as a negative control to test for unmeasured confounding, with the reasoning that childhood afterschool resources should not have causative effects on adult health. However, I am not sure this is an appropriate negative control, as I don't think it is implausible that

the well being of a person's child could impact one's health.
4) Table 2 is uninterpretable on its own. Are these counts present in each percentile?
5) Table 3: how are these results created? Are covariates included in the calculation?
6) Pg 14: "were less well-off". Is this referring to average annual income? Poverty rate? Would you not expect to see that resources would be inversely related to most health outcomes in this case? (ie. "well-off" adults would have less resources in their neighborhood, but better outcomes) I think ideally you would be comparing ZIP codes with similar socioeconomic status, but with different resource availability.
7) How can you explain association between resources and lower blood pressure in those without hypertension, but not in those with hypertension? Is there really evidence to support heterogeneity of the effect of resources?
Minor comments:
1) Abstract "tested the association of selected resources": This sentence does not help define objective or methods of the study. How many resources (explanatory variables) were tested? How were they defined?

REVIEWER	Rachel Gold
	KPCHR / OCHIN, USA
REVIEW RETURNED	15-Nov-2018

| GENERAL COMMENTS | Overall – Fascinating paper. Needs more emphasis on the fact that just because resources are in a community, it does not mean that they are being accessed. Needs some details added to the Methods.

Abstract – please note how 'area' was defined

Background

- Clarify / expound on how unmet needs are associated with these risk factors. This sentence sounds like you already know the answer that this analysis seeks. In other words, how do you know the needs are unmet? Are these associations with patient-reported need?
- 2nd para could be revised for clarity – especially, 'the connection between resource type and need.' Do you mean to say, 'The specific resources that best address specific social needs may not always be straightforward.'
- Overall I think the introduction could be a bit tighter.
- I do not think area-resources should be hyphenated.

Methods

- Describe the asset mapper more clearly. How was it created, is it up to date, etc. If patient data are from 2015, when are the asset data from?
- Needs more explanation of patient address relative to resources. What geographic area is involved here? Meaning, how close to the patient do the resources need to be to count?
- Related to that, I assume you are looking at resources that are near where patients live, not near the clinics themselves – is that correct? This needs to be far more clear. If my assumption is wrong and the association is with clinic location, I think that is a much weaker approach and would recommend looking at patient |

	residence. Either way, this part of the methods needs to be clarified. - Can you confirm the date of patient address – was it where they lived at any point in 2015? - I assume patient records were geocoded but that should be stated explicitly. - I'd think 'food desert' status would confound access to food resources – no? Please explain. Discussion - There are some real limitations to this analysis, which should be noted. Notably, the existence on resources does not mean that they are used. Yes, this is mentioned, but it needs to be discussed further. - Also, there are likely covariates that impact this relationship – for example, local culture – that could not be included in these analyses. - The quality and type of resources might be important and could not be measured here.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Matthew B. Lanktree
Institution and Country: University of Toronto - Toronto, Ontario, Canada Please state any competing interests or state 'None declared': None

The authors compare body mass index (BMI), systolic blood pressure (SBP), LDL cholesterol and HbA1C obtained from medical charts of 123,355 people to the "employment resources", "food resources", "nutrition resources", among others, of the participant's home zip code. The authors use access to "big data" and a random forest machine learning technique to come to their conclusions. They find association between improved resource access and lower BMI, but no association with SBP, LDL or HbA1c. Overall, I think the analysis is careful and measured.
Response: Thank you

I do worry about confounding from the inverse relationship between resource availability and socio economic status of the zip codes, which would be expected to reduce the estimate of the effect of the resources.
Response: We agree that socioeconomic status is a key potential confounder. We have attempted to address this by adjusting for an indicator of individual-level socioeconomic status (education), and several indicators of area-level socioeconomic status (median household income, unemployment rate, food 'desert' status, and number of individuals living close quarters [a measure of housing 'crowding']). We have further adjusted for residential segregation, which may still confound the relationship between number of resources and the outcome (while not a socioeconomic status indicator per se). We include this entire set of variables as we socioeconomic status is understood to be a multidimensional concept, and we were concerned that adjusting for only a single measure (e.g., area-level income or education, for example) would leave open confounding from other dimensions of socioeconomic status. Nevertheless, we recognize that residual confounding could still occur, so we have added a note to this effect to the limitations section (page 19, excerpted below):

> **"While we adjusted for several individual-level and area-level socioeconomic status indicators in order to capture the multidimensional nature of socioeconomic status and, thus, reduce confounding, it is possible that residual confounding, owing to unmeasured characteristics, exists, which would tend to reduce the observed associations between area-resources and outcomes."**

Major comments:

1)   I struggle with the definition and meaning of terms used throughout the paper. How are "Area resources" defined and tabulated? How does food resources compare to nutrition resources? Employment resources? I believe greater detail regarding the explanatory variables is required for the reader to understand the methods and results of this study. **Response: We have added more detail regarding the classification of area resources (pages 8-9, excerpted below):**

> **"HelpSteps (www.helpsteps.com) is a web and mobile screening and referral system for social needs. Originally launched in 2010, the system uses a database of social services throughout the greater Boston area to connect families to appropriate services. The database is maintained in collaboration between Boston Children's Hospital and the Mayor's Health Line at the Boston Public Health Commission. Every agency is contacted at least once per year to maintain the accuracy of the data and to grow the database. HelpSteps contains information on area-resources across 16 non-mutually exclusive domains: health, housing, food, employment, violence, safety, substance abuse, mental health, education, parenting, nutrition, afterschool, sexual health, transportation, diabetes, and care transitions. An example of organizations that would be in the food domain are food pantries. The employment domain would consist of job placement or job training services. And the nutrition domain would include organizations that provide food counseling. Agencies providing multiple resources could be included in more than one domain. Because individual-level data for this study came from 2015, we used information from HelpSteps that was current as of 2015. For this study, 'area resources' are defined as the number of organizations found in**
> **the HelpSteps database providing assistance for a given domain and within a given geographic area."**

2)   I think it would be valuable to compare the results of the current VSURF analysis to a more traditional model building regression approach.
Response: We appreciate that some of our analytic techniques may not be familiar to some readers. To clarify, our manuscript uses both VSURF and regression in our analysis. We use VSURF for variable selection, and then test the selected variables (in a separate sample) using regression. We do not believe there is a regression-based alternative for variable selection in this context, as the structure of the data used in this project violate the fundamental assumptions that underlie a parametric, regression-based, variable selection process (particularly the assumption of statistical independence of the errors). While regression-based variable selection techniques such as forwards, backwards, or stepwise selection are useful in some circumstances, they break down in the conditions seen here—namely situations where some variables are components of other variables. For example, when selecting the geographic area to focus on (e.g., ZIP code tabulation area (ZCTA) vs. census tract), our counts of ZCTA food resources are inherently aggregates of the number of food resources in the census tracts that constitute the ZCTA, so the two variables (ZCTA food resources and census tract food resources) are highly correlated with each other. Further, this high degree of correlation can also occur when selecting what resource type to investigate (e.g., food vs. housing resources) as a given social service agency can provide services for more than one issue and thus be counted in both categories. Overall, the extreme multicollinearity induced by the problem we are trying to investigate means that regression-based variable selection cannot be used. Instead, to get meaningful results, a non-parametric approach is required. We have added a discussion of the reasons regression-based variable selection cannot be used to the methods section (page 11, excerpted below):

> **"Because the nested structure of our data violate the statistical independence assumption that underlies parametric, regression-based variable selection approaches (such as forward, backwards, or stepwise selection), and to avoid multiple hypothesis testing that may lead to the identification of spurious associations, we employed a non-parametric machine learning technique called variable selecting using random forest (VSURF) to screen through variables in the derivation set.**

3)   The authors should be commended for their efforts at robustness checks. The authors use "after school resources" as a negative control to test for unmeasured confounding, with the reasoning that childhood afterschool resources should not have causative effects on adult health. However, I am not sure this is an appropriate negative control, as I don't think it is implausible that the well being of a person's child could impact one's health.

Response: We agree that, in a broad sense, the well-being of one's children could affect one's health. However, our negative control in this analysis was more limited— we specifically tested whether there was an association between number of area after-school resources and, for example, body mass index. Because we are aware of no causal mechanisms whereby after-school resources for children would reduce body mass index in adults, we believe that this is an appropriate negative control because any association seen would likely reflect residual confounding (for example, by the socioeconomic status of the neighborhood). We did not observe any association between after-school resources and body mass index, which suggests that there is not a strong association (independent of the covariates) between these factors, and that there is not strong residual confounding between the factors either. We have clarified our justification for using this negative control in the methods section (page 12, excerpted below):

> **"Our reasoning was that, since there was unlikely to be any direct effect of after school resources for children on adult body mass index, any observed association would reflect unmeasured area-characteristics not appropriately adjusted for in our model (such as high levels of civic engagement or community organization, or other beneficial resources)."**

4)   Table 2 is uninterpretable on its own. Are these counts present in each percentile?

Response: We apologize for the lack of clarity. We have re-titled the table to make clear that these are the number of resources. The new title is: "Distribution of the Number of Resources in the Selected Resource Categories".

5)   Table 3: how are these results created? Are covariates included in the calculation?

Response: We apologize for the lack of clarity. These are model-based estimates created using least-squares means. The estimates are created using the fitted models, which includes all covariates. We have added a footnote to the table clarifying this (excerpted below):

> **"Estimates created using least-squares means from fitted multi-level models. The models used fixed effects to adjust for age, gender, race/ethnicity, education, insurance, number of clinic visits, language, clinic connectedness, comorbidity, and census tract level median household income, poverty rates, 'food desert' status, unemployment, numbers living in group quarters, vehicle access, and segregation. To account for clustering within practices, we included a practice-level random effects term. To account for area-level clustering, we used a ZIP-level random effects term. These were fit as crossed effects models (i.e., we did not nest practices within ZIP codes) to allow for the fact that patients are often seen in practices outside of their ZIP code of residence."**

6)   Pg 14: "were less well-off". Is this referring to average annual income? Poverty rate?

Response: We apologize for the lack of clarity. We meant less well-off in the sense of social deprivation. We have clarified this (page 14, excerpted below).

> **"In general, individuals living in areas with more resources had lower educational attainment and higher rates of Medicaid insurance coverage (eTable 4)."**

Would you not expect to see that resources would be inversely related to most health outcomes in this case? (ie. "well-off" adults would have less resources in their neighborhood, but better outcomes) I think ideally you would be comparing ZIP codes with similar socioeconomic status, but with different resource availability.

Response: We agree that socioeconomic status is an important potential confounder. We have adjusted for a number of individual-level and area-level socioeconomic status indicators as described above. The regression adjustment in our models does exactly as the reviewer suggests—produces estimates of the associations between area-resources and the outcome among individuals with similar individual-level socioeconomic status indicators living in ZIPs with similar socioeconomic-status indicators. As above, however, we did add a discussion of the potential for residual confounding.

7)   How can you explain association between resources and lower blood pressure in those without hypertension, but not in those with hypertension? Is there really evidence to support heterogeneity of the effect of resources?

Response: Though this study is not able to determine mechanisms that can explain this difference, our hypothesis (and the reason we examined these groups separately) was that there is likely a different association between area-resources and cardiometabolic factors that are, versus are not, frequently targeted in clinical care. With regard to blood pressure, among those with hypertension, clinical management is likely much more important than area resources. For example, our estimate was that 1 additional area resource is associated with a 0.1mm hg reduction in systolic blood pressure. In comparison, we know from pharmaceutical approval studies that a blood pressure lowering medication lowers blood pressure by approximately 10 mm hg on average. We think this difference in the size of the associations is plausible as connection between area resources and blood pressure is much more indirect than with a medication (though it is likely still important on the population level). Thus once clinical management (e.g., prescribing medications) comes into play, area resources are likely much less important. However, for individuals without diagnosed hypertension, blood pressure is not under clinical management, and thus area resources may play a more important role. We discuss these issues in the methods (pages 12-13, excerpted below), and the discussion section (page 18, excerpted below):

> **"To help explore whether variations in the intensity of clinical management could explain whether community resources were associated with health outcomes, we also used the above modeling approach to test whether area resources were associated with SBP in those without a diagnosis of hypertension. The primary care network in the study has quality improvement program that emphasize the importance of SBP, LDL, and HbA1c control in appropriate clinical populations. Since BMI (in any population) and SBP control in those without a diagnosis of hypertension are not included in these programs, we reasoned that area-resources may be more important when clinicians are not intensively attempting to impact an outcome."**

> **"This is reinforced by the finding that SBP, among those without hypertension, is lower in those living in areas with more resources. Since SBP does not come under clinical management for those without hypertension, this finding supports the potential for area resources to impact population health, and is consistent with guidelines that recommend lifestyle, rather than pharmacologic, approaches to pre-hypertension treatment."**

Minor comments:
1)   Abstract "tested the association of selected resources": This sentence does not help define objective or methods of the study. How many resources (explanatory variables) were tested? How were they defined?

Response: We have added a definition of area resources to the Objectives section. We have also added the number of resource categories that were tested to the methods section (since this came from the VSURF analyses). The relevant portions of the abstract are excerpted below:

> **"We sought to determine if area-level resources, defined as organizations that assist individuals with meeting health-related social needs, are associated with lower levels of cardiometabolic risk factors."**

"We then tested the association of selected resources (three for BMI, two each for SBP and HbA1c analyses, and one for LDL analyses) with these outcomes, using multi-level models to account for individual-level, clinic-level, and other area-level factors."

Reviewer: 2
Reviewer Name: Rachel Gold
Institution and Country: KPCHR / OCHIN, USA Please state any competing interests or state 'None declared': None declared

Overall – Fascinating paper.
Response: Thank you.

Needs more emphasis on the fact that just because resources are in a community, it does not mean that they are being accessed.
Response: We have added additional discussion of this limitation (page 19, excerpted below), and have added it as a bullet point in the Strengths and Limitations of the study section:

> **"Our finding should be interpreted in light of several limitations. We did not have access to data regarding use of the resources. This means that we do not know whether individuals made use of the resources in their community. In light of this, the association between ZIP-level resources and outcomes could be viewed analogously to an 'encouragement design' intervention. This means that the association estimated in this study is likely different than the association that would be estimated if analyzing those who were known to use the resource. That association is clearly of policy interest, and should be examined in future work."**

Needs some details added to the Methods.
Response: We have added details to the methods in response to the specific comments below (see below for a more in depth discussion of the additions). Please let us know if additional details are needed.

Abstract – please note how 'area' was defined
Response: We have clarified that we were reporting results from ZIP-level analyses in the abstract, excerpted below:

> **"We then tested the association of ZIP-level selected resource types (three for BMI, two each for SBP and HbA1c analyses, and one for LDL analyses) with these outcomes, using multi-level models to account for individual-level, clinic-level, and other area-level factors."**

Background

-   Clarify / expound on how unmet needs are associated with these risk factors. This sentence sounds like you already know the answer that this analysis seeks. In other words, how do you know the needs are unmet? Are these associations with patient-reported need? **Response: We have revised to make clear that this paragraph is describing known associations between patient reported health-related social needs (e.g., food insecurity or housing instability) and poor health (e.g., higher BMI, worse diabetes control). Our goal in this study was to investigate whether the presence of resources that can be used to meet health-related social needs (e.g., food pantries) is associated with better health outcomes (page 6, excerpted below):**

> **"Cardiometabolic disease remains the most common cause of morbidity and mortality in the U.S. Though better control of cardiometabolic risk factors could substantially reduce this morbidity and mortality, individuals with low socioeconomic status (SES) are less likely to achieve recommended goals. Among the reasons for this are patient-reported health-related social needs, including such factors as food insecurity, housing instability, and lack of transportation. These health-related social needs have been associated with higher levels of important cardiometabolic risk factors including increased body mass index (BMI), systolic blood pressure (SBP), low density lipoprotein cholesterol (LDL), and hemoglobin A1c (HbA1c), even after adjusting for factors like race/ethnicity, income, and education. Proposed mechanisms linking health-related social needs to cardiometabolic risk factors including reduced dietary quality, cost-related medication underuse, reduced cognitive 'bandwidth' to attend to health, and disruptions in clinical care.**

- 2nd para could be revised for clarity – especially, 'the connection between resource type and need.' Do you mean to say, 'The specific resources that best address specific social needs may not always be straightforward.'

Response: We have revised the 2nd paragraph as suggested.

- Overall I think the introduction could be a bit tighter.

Response: We have revised the introduction to shorten it and make it more focused on the goals of the study.

- I do not think area-resources should be hyphenated.

Response: We have removed this hyphen.

Methods

- Describe the asset mapper more clearly. How was it created, is it up to date, etc. If patient data are from 2015, when are the asset data from?

Response: We had added more detail to the methods regarding the source of resource data (pages 8-9, excerpted below):

> **"HelpSteps (www.helpsteps.com) is a web and mobile screening and referral system for social needs. Originally launched in 2010, the system uses a database of social services throughout the greater Boston area to connect families to appropriate services. The database is maintained in collaboration between Boston Children's Hospital and the Mayor's Health Line at the Boston Public Health Commission. Every agency is contacted at least once per year to maintain the accuracy of the data and to grow the database. HelpSteps contains information on area-resources across 16 non-mutually exclusive domains: health, housing, food, employment, violence, safety, substance abuse, mental health, education, parenting, nutrition, afterschool, sexual health, transportation, diabetes, and care transitions. An example of organizations that would be in the food domain are food pantries. The employment domain would consist of job placement or job training services. And the nutrition domain would include organizations that provide food counseling. Agencies providing multiple resources could be included in more than one domain. Because individual-level data for this study came from 2015, we used information from HelpSteps that was current as of 2015. For this study, 'area resources' are defined as the number of organizations found in**
> **the HelpSteps database providing assistance for a given domain and within a given geographic area."**

- Needs more explanation of patient address relative to resources. What geographic area is involved here? Meaning, how close to the patient do the resources need to be to count?

Response: We have added more detail to address these questions (page 9, excerpted below):

> **"After geocoding the addresses for both individuals and the area resource organization, we created counts, for each individual, of how many resources for each domain were within the same geographic area as they were. We did this at 4 geographic levels in roughly increasing order of size: census tract (using U.S. Census 2010 boundaries), ZIP code tabulation area (which we refer to throughout this paper as 'ZIP' level, owing to common use of the term, again using U.S Census 2010 boundaries), 'neighborhood' (e.g. Allston, Roxbury, a designation based on Boston city planning that may better capture actual movement patterns), and county."**

- Related to that, I assume you are looking at resources that are near where patients live, not near the clinics themselves – is that correct? This needs to be far more clear. If my assumption is wrong and the association is with clinic location, I think that is a much weaker approach and would recommend looking at patient residence. Either way, this part of the methods needs to be clarified.

Response: We clarify that we are indeed counting how many resources are within a geographic area based on the patient's home address. For example, if there were 4 food resources within an individual's an individual's ZIP, they would get a score of 4 for that analysis. We have added a description of this to the methods (page 9, excerpted below):

> **"After geocoding the addresses for both individuals and the area resource organization, we created counts, for each individual, of how many resources for each domain were within the same geographic area as they were. We did this at 4 geographic levels in roughly increasing order of size: census tract (using U.S. Census 2010 boundaries), ZIP code tabulation area (which we refer to throughout this paper as 'ZIP' level, owing to common use of the term, again using U.S Census 2010 boundaries), 'neighborhood' (e.g. Allston, Roxbury, a designation based on Boston city planning that may better capture actual movement patterns), and county."**

- Can you confirm the date of patient address – was it where they lived at any point in 2015?

Response: For individuals with more than one address in the dataset, we used the most recent address for the individuals (page 7, excerpted below):

> **"Data were current on December 31, 2015. The most recent patient address was geocoded for the study."**

- I assume patient records were geocoded but that should be stated explicitly. **Response: We clarify explicitly that patient addresses were geocoded (page 7, excerpted below):**

> **"The most recent patient address was geocoded for the study."**

- I'd think 'food desert' status would confound access to food resources – no? Please explain.

Response: We agree that food desert status is a potential confounder (although we note that we use food desert to refer to lack of food retailer (e.g., supermarkets) and food resources to refer to organizations that provide food assistance (e.g., food pantries or assistance with enrolling in SNAP), not food retailers. Because food desert status is a potential confounder, we adjusted for 'food desert' status in our multivariable regression models testing the association between food resources and, for example, BMI.

Discussion

-     There are some real limitations to this analysis, which should be noted. Notably, the existence on resources does not mean that they are used. Yes, this is mentioned, but it needs to be discussed further.

Response: We have added additional discussion of this limitation (page 19, excerpted below):

> **"Our finding should be interpreted in light of several limitations. We did not have access to data regarding use of the resources. This means that we do not know whether individuals made use of the resources in their community. In light of this, the association between ZIP-level resources and outcomes could be viewed analogously to an 'encouragement design' intervention. This means that the association estimated in this study is likely different than the association that would be estimated if analyzing those who were known to use the resource. That association is clearly of policy interest, and should be examined in future work."**

-     Also, there are likely covariates that impact this relationship – for example, local culture – that could not be included in these analyses.

Response: We agree, we have added a discussion of this to the limitations section, (page 19, excerpted below):

> **"Additional unmeasured covariates that could affect the observed associations include local culture, and the quality of the resources available. Devising methodology to determine the quality of the services provided to help meet health-related social needs is pressing, and will be an important direction for future investigation."**

-     The quality and type of resources might be important and could not be measured here.

Response: We agree, we have added a discussion of this to the limitations section, (page 19, excerpted below):

> **"Additional unmeasured covariates that could affect the observed associations include local culture, and the quality of the resources available. Devising methodology to determine the quality of the services provided to help meet health-related social needs is pressing, and will be an important direction for future investigation."**

VERSION 2 – REVIEW

REVIEWER	Matthew B. Lanktree University of Toronto, Canada
REVIEW RETURNED	19-Dec-2018

GENERAL COMMENTS	I believe the manuscript is improved. The authors have addressed my concerns.

REVIEWER	Rachel Gold KPCHR; OCHIN Inc; both USA None declared. I have served on a conference panel with the lead author.
REVIEW RETURNED	26-Dec-2018

GENERAL COMMENTS	This is so strong now! The revisions made a huge difference. Great paper.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

I believe the manuscript is improved. The authors have addressed my concerns.

Response: Thank you.

Reviewer: 2 This is so strong now! The revisions made a huge difference. Great paper.

Response: Thank you.