

## 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

|                            |  |
|----------------------------|--|
| <b>TITLE (PROVISIONAL)</b> | Cross-sectional association between ZIP code-level gentrification and homelessness among a large community-based sample of people who inject drugs in 19 US cities   |
| <b>AUTHORS</b>             | Linton, Sabriya; Cooper, Hannah; Kelley, Mary; Karnes, Conny; Ross, Zev; Wolfe, Mary; Friedman, Samuel; Des Jarlais, Don; Semaan, Salaam; Tempalski, Barbara; Sionean, Catlainn; DiNenno, Elizabeth; Wejnert, Cyprian; Paz-Bailey, G |

### VERSION 1 - REVIEW

|                        |   |
|------------------------|---|
| <b>REVIEWER</b>        | Julia Dickson-Gomez<br>Medical College of Wisconsin<br>United States of America |
| <b>REVIEW RETURNED</b> | 30-Aug-2016   |

|                         |   |
|-------------------------|---|
| <b>GENERAL COMMENTS</b> | <p>This is a well-written paper on a neglected topic, the influence of place-based characteristics such as housing affordability, inadequate housing and gentrification on homeless among PWID. There are a few areas the manuscript could be strengthened which I mention below.</p> <p>The first bulleted comment in strengths is difficult to understand. In particular, it is not clear what you mean by "place-level housing and economic characteristics." This is described more clearly in other places in the manuscript. For example in paragraph 6 on page 4 you list, unaffordable housing, economic deprivation, inadequate housing, overcrowding and gentrification. I would recommend listing these. However, it is not clear what different point you wish to make in bullet point 1 as opposed to bullet point 2.</p> <p>The first sentence in paragraph 2, page 4 is very long and should be edited.</p> <p>You might want to consider providing a definition of the characteristics in your theoretical model and how they relate to one another. The introduction makes clear the connection between gentrification, economic deprivation, housing unaffordability, and homelessness. However, the role of incarceration and social networks is not really discussed in the introduction. It is also not clear why characteristics not examined in the study are included in the model.</p> <p>The county and zip code measures are described well in your table so is probably not needed in the text.</p> <p>There seems to be a word or phrase missing on page 8, line 1 "the potential high costs that may result..." It is not clear whether you are</p> |
|-------------------------|---|

|  |  |
|--|--|
|  | <p>referring to homelessness, injection drug use or both.</p> <p>The last line of the conclusion suggests that further research could "inform urban planning strategies and community mobilization campaigns designed to curb potential negative effects of gentrification and provide access to adequate housing among low-income and marginalized populations." Some concrete suggestions would be helpful, particularly since it seems that you do not feel like mixed income housing units have been successful. Would more supportive housing help?</p> <p>The tables could be reformatted to make them easier to read. For example, indentation in Table 1 of the types of sexual partners would increase readability.</p> |
|--|--|

|                        |   |
|------------------------|---|
| <b>REVIEWER</b>        | Danya Keene<br>Yale School of Public Health |
| <b>REVIEW RETURNED</b> | 08-Sep-2016                                 |

|                         |  |
|-------------------------|--|
| <b>GENERAL COMMENTS</b> | <p>1. This paper uses cross-sectional data to examine the association between gentrification and homelessness among PWID. As the authors note, this analysis represents an important contribution to the literature given that prior research has focused predominantly on individual versus structural/spatial determinants of homelessness. I think that this structural perspective is particularly important given that the paper focuses on a population (PWID) whose homelessness may be presumed to result from drug use. I recommend adding this point to the introduction. I also recommend drawing on literature that examines the stigmatization of homelessness to further strengthen the rationale for a structural analysis.</p> <p>2. The conceptual model describing the relationship between gentrification and homelessness could be clarified and strengthened (both in the introduction and in the graphical representation of this model).</p> <p>a. The paragraph on page 4, line 47, that describes the causes of homelessness is not clear. Some of the factors listed seem to be parallel housing outcomes that may be associated with homelessness, but are not part of the causal pathway (for example, overcrowding may occur when there is a shortage of available and affordable housing. It may be associated with homelessness at the ecological level, but both homelessness and overcrowding arise from the same lack of housing availability). I recommend simplifying this paragraph and the graphical model to describe the main hypothesized pathways between gentrification and homelessness (for example, rising housing costs). This paragraph could also discuss the demolition/revitalization of public housing that has occurred in the context of gentrification, and which may limit the number of available low income units.</p> <p>b. Related to the above point, the statement "increase demand for social services" is not clear, as it relates to gentrification and homelessness.</p> <p>c. Regarding the role of policing: it seems possible that increased policing in gentrifying areas could lead to a decrease in homelessness (rather an increase, as hypothesized by the authors) if a) local ordinances make it harder for homeless to stay outside and force them to go elsewhere b) if the homeless are incarcerated at higher rates and therefore in jail/prison, vs on the streets.</p> |
|-------------------------|--|

|  |  |
|--|--|
|  | <p>3. I have two concerns about the use of “inadequate housing” as a mediator in this paper. First, conceptually, the causal link between inadequate housing and homelessness is not clear. Inadequate housing seems like a parallel outcome---it is associated with homelessness, because it shares the root cause of not enough affordable housing. Second, it is not clear how gentrification would lead to an increase in inadequate housing. I recommend omitting this variable from the paper.</p> <p>4. The main hypothesized mediators (rent burden and availability of subsidized housing) did not pan out in the analyses. As the authors note, the rent burden measure is limited because it only captures moderate rent burden. The authors do not discuss the subsidized housing measure in their results or discussion, but this measure also seems to have some limitations given that in many cities, waiting lists are closed to new applicants, so length of time on the waitlist is unlikely to reflect actual availability of subsidies relative to need. To address these issues, the authors may want to consider using an additional zipcode/county level measure used by the Urban Institute: “available and affordable housing units for every 100 extremely low income households”. This measure (which includes subsidized units) captures the decline in available housing that is hypothesized to accompany gentrification and lead to increases in homelessness.</p> |
|--|--|

|                        |  |
|------------------------|--|
| <b>REVIEWER</b>        | Becky Genberg<br>Brown University, USA |
| <b>REVIEW RETURNED</b> | 12-Sep-2016                            |

|                         |   |
|-------------------------|---|
| <b>GENERAL COMMENTS</b> | <p>This paper reports on a study of the association between place-level factors and homelessness among people who inject drugs in the United States. The study is clearly described and well written. The investigation of area-level factors with respect to homelessness is interesting and adds to the literature on an under-studied topic and population (people who inject drugs).</p> <p>The major concern I have regarding the analysis as presented is some confusion regarding the assessment of the potential mediating relationship of inadequate housing, housing affordability and economic deprivation between gentrification and homelessness. The results state that the inclusion of these three potential mediators did not impact the results of the association between gentrification and homelessness, however, I do not think this is evidence for establishing the mediation. The authors may want to consider a more unified model, such as a structural equation model, to examine the relationships between all variables of interest without the known challenges of examining mediation using traditional logistic regression analysis. Taking this one step further, are the authors suggesting that because the main relationship between gentrification and homelessness is not impacted by the inclusion of the three mediators mentioned above, that the proposed relationship between gentrification and homelessness must then be due to other mediating factors or a direct effect? More explanation of this in the discussion may be warranted. I think the statement in the concluding paragraph: “future longitudinal studies should explore whether these associations are casual, and identify potential mediators,” is necessary, but calls into question the investigation of mediation as presented in the methods and results.</p> |
|-------------------------|---|

|  |  |
|--|--|
|  | <p>Additionally the correlations between the area-level variables could be examined and reported (there is mention of this in the methods, but nothing reported in the results on this investigation). It might be interesting to include information regarding levels of inadequate housing, housing affordability, and economic deprivation by varying levels of gentrification to continue to tease apart the relationship between these factors.</p> <p>There are some variables included in the conceptual model that are not discussed in the paper (e.g., social networks) and others that are referenced (i.e., police activity), but not included in the conceptual model (unless the authors are proposing that individual incarceration represents this area-level factor? If so this may need additional justification and clarification). Additional thought should also be given to the temporality of the area-level variables and how they are related over time.</p> <p>It may be useful to have a map or an appendix detailing the 19 MSAs that were included in the study. It is now difficult to assess the limitation as stated regarding PWID living outside of the MSAs that were included.</p> |
|--|--|

|                        |   |
|------------------------|---|
| <b>REVIEWER</b>        | Jo Røislien, Associate professor of medical statistics<br>Department of Health Studies, University of Stavanger, Norway |
| <b>REVIEW RETURNED</b> | 17-Dec-2016   |

|                         |  |
|-------------------------|--|
| <b>GENERAL COMMENTS</b> | <p><b>GENERAL</b></p> <p>The study aims to quantify the association between homelessness among people who inject drugs and various place-based factors, such as gentrification and inadequate housing. In order to achieve this, the authors use quantitative data on multiple levels, e.g. individual level data and ZIP code data.</p> <p>I have been asked to review this manuscript with a particular emphasis on the statistical methods and analyses used. This has been interesting, but not straightforward, as statistical methods are only vaguely presented, resulting in some brain gymnastics to untangle what has been done, and why. Below is a set of comments that hopefully will result in an improved paper, which attacks an important, yet analytically complex, topic.</p> <p><b>POINT-BY-POINT COMMENTS</b></p> <p><i>Page 3, line 29:</i><br/>Large numbers, e.g. 5,394 and 8992, must be written in the same way.</p> <p><i>Page 3, line 32:</i><br/>The abbreviation AOR is neither introduced nor explained. I assume the abbreviation refers to adjusted OR? Adjusted how? Please clarify.</p> |
|-------------------------|--|

*Page 3, line 37:*

I struggle to see how the conclusion is a valid conclusion based on the results presented in this study. Firstly, how does the evaluation of a planning strategy lead to less homelessness? In order for something to change you cannot just evaluate; you have to do something. Secondly, neither evaluation nor planning strategy are covariates in any of the analyses presented in this study.

*Page 5, line 9.*

I must admit I struggled at first to understand what was the aim of the study. Before I figured out why: The last sentence in the introduction ("Figure 1 presents a conceptual..." - see below) blurs the aim of the study. Instead of the introduction ending with a punctuation mark, this sentence leads to the introduction ending with a crossfade from "what" to "how". This last sentence is not introduction or background information: It is about methodology. It is valuable information, indeed, but it belongs someplace else in the manuscript. For example page 6 line 28, just after the description of the variables, and before the Stats Anal subsection.

*Page 5, line 21.*

Respondent driven sampling (RDS) is a non-probability sampling method. That is, non-randomized, and on thus on collision course with most standard assumptions for proper statistical analysis. Papers have demonstrated that even when RDS has been used correctly, bias cannot necessarily be ruled out (McCresh et al, 2012). To the best of my knowledge the jury is still out on as to how RDS results are to be interpreted; that is, how analytical results using RDS relate to the real world they try to describe. This is a serious limitation to the paper and must be addressed.

*Page 5, line 37-38.*

The authors state that they have excluded 134 individuals due to missing on key covariates, indicating that they aim for a complete case analysis. This should be OK, as 134/9702 is approx 1%, and method studies indicate that for missing < 5% complete case usually works fine.

In light of the latter I am more concerned with the 499 individuals removed because of missing Zip code information, as 499/9702 > 5%. That is, this is enough data to introduce significant bias, of unknown order and magnitude, when simply discarded. Can the authors please clarify why they have chosen to introduce this potential bias in the results, and not, say, perform multiple imputation?

*Page 6, line 34.*

While the subsection is now termed "Analysis", this is not the subsection where the authors describe analysis in general; it is the

section that describes the handling of quantitative data, e.g. statistical analysis. I would thus prefer the heading “Statistical analysis” rather than just “Analysis” as the latter is overly vague as to what the subsection actually describes.

Page 6, line 37.

The authors start the stats subsection by stating that “The distributions of all characteristics were determined using descriptive statistics.” This is not correct. There are numerous books filled with different statistical distributions. It is truly a complex task, both empirically and mathematically, to assess the true distribution of data sets. Moreover; fully determining the distribution of a covariate is not really necessary in most regression models. What we are usually interested in is some general idea about the data, some measures describing the essential features. Like a centrality measure and a variation measure. Descriptive statistics can be many things: tell the reader *\_what\_* descriptive statistics you have used. Something along the lines of:

“Symmetric continuous data are summarized using mean and standard deviation (SD) and skewed continuous data using median and 25<sup>th</sup> and 75<sup>th</sup> percentiles, while proportions are presented as percentages.”

*Page 6, line 38.*

“Logistic multilevel analysis was used...”

Firstly, on a more general note; I am not a big fan of dismissing the statistical analysis of quantitative data as merely “analysis”. It is too vague a description of what is actually being done, in particular on more advanced statistical modelling like here. Logistic regression, which I assume the authors have used (it is not explicitly stated), is not just “analysis”: it is a probability model; a mathematical model of the real world. Regression modelling is more than just analysis: It assumes several things about the world it tries to model, and this should be communicated to the reader; regression modelling assumes that the data follow certain probability distributions, that variables affect one another in a particular way and so on. So I would suggest rewriting this sentence to something along the lines of “Multilevel logistic regression models, both univariate and multivariable, were fitted to the data, and used to ...”

Secondly; *\_how\_* did the authors carry out the multilevel modelling? What type of multilevel model was this? As some data are on an individual level, and some data on group level, this is not a standard statistical model, and some explanation of the model is needed. Mixed and multilevel data modelling often refers to a way of handling clusters within the data set, the classical example being that many school kids are given a test, but as children in the same class have had the same teacher results for these kids might be correlated (good / poor teachers), results from children from the same school might be correlated (good / poor schools), and so on. That is: Data

are on an individual level, but there is clustering of the data on multiple levels. In the present study, however, the data themselves are on varying scales; individual level data, zip code data, and so on. So some information should be added to the paper as to how different types of information, from individual to group characteristics, can be properly handled, with references to the proper statistical methods (which I assume is being applied).

*Page 6, line 41.*

“Multivariable analysis assessed the relationships of place-based factors significant ( $p < 0.05$ ) in bivariate analysis”

This sentence is a bit unclear. But I assume it means that the authors first carried out a series of bivariate analyses, and then statistically significant variables were entered into multiple regression analyses? Unless having sparse amounts of data such a two-step procedure can easily be an overly ad hoc approach for building optimal statistical models. Rather one could enter all suggested variables into the multiple model, and then do optimal model selection based on some proper objective model description criteria such as AIC, BIC, FIC etc to remove non-contributive covariates from the hypothesized full multiple model.

If the authors do want to use a two-step procedure for model building it must be carried out somewhat differently than is presented here; variables with non-significant p-values from univariate analysis must also be entered into the multiple model, as p-values change from uni- to multivariable analyses, and setting the inclusion threshold as low as 0.05 will potentially result in dismissing otherwise important predictors. A p-value cutoff of  $\sim 0.20$  is usually recommended (see for example Bursac et al, 2008).

*Page 6, line 44-48.*

The two sentences below (taken from the stats section) are not statistical analysis; they describe covariate background information. They do thus not belong in the Stats section, which is for stats only. Move them further up in the Methods section.

“Based on prior research, gentrification was conceptualized as reducing economic deprivation, housing affordability, and inadequate housing. Gentrification was also hypothesized as increasing law enforcement practices that would increase the arrest of PWID and thereby increase the odds of homelessness.”

*Page 6, line 54.*

As the authors do not present a novel statistical method, but use an existing method already implemented software, information about the software and the function(s) used would be valuable. There are several analytical approaches to the same statistical idea, and different ways of implementing them, and the type of mathematical framework, and corresponding implementation used, might potentially influence the results, and thus the corresponding

conclusions.

*Page 7, line 1.*

The authors write that “Approximately 50% of participants resided in counties where there was at least 87% (25th and 75th percentiles: 80.06, 90.32) unaffordable rental units...”

The numbers presented here have no consistency to them. In the same sentence there are three different precision levels: The percentiles have four digits precision (80.06, 90.32), the median has only two digits precision (87), while the first number hasn't even two digits precision (approximately 50). Similar juggling with precision levels can be found other places in the manuscript as well. The authors need to tighten up the presentation of numbers.

Minor additional comment:

It is not mentioned anywhere what the main number, 87%, actually means. And (in my personal opinion) writing out ‘25th and 75th percentiles’ in full often tends to clutter the result section, making it more difficult to get a hold of the actual results/numbers. As a statistician I often suggest being explicit in the stats section about how the format of the presentation of numbers in the results section will be. That is, letting, say, the first sentence in the stats section be something along the lines of:

“Symmetric continuous data are summarized using mean and standard deviation (SD) and skewed continuous data using median and 25<sup>th</sup> and 75<sup>th</sup> percentiles, while proportions are presented as percentages.”

Then page 7, line 1 might read “Approximately half of participants resided in counties where there was a median of 87.xx% (80.06, 90.32) unaffordable...”

*Page 7, line 3.*

“...average number of months that applicants were on waiting lists for assisted housing was 26 months (25th and 75th percentiles: 21, 38)”

This summary of numbers is inconsistent with basic mathematical results from inferential statistics. If data are symmetric (e.g. normally) distributed, then the mean and SD are sufficient statistics, and we can summarize the data using mean and SD without loss of information. If data are skewed however, the mean is not necessarily a good measure of centrality, and SD similarly does not represent the un-symmetric variation in the data, so the median and percentiles are usually preferable summary statistics.

Here, however, the authors use average and percentiles. Why? If data are symmetric, mean and percentiles are not sufficient, and if data are skewed the mean is often a misleading centrality measure.

Please clarify.

*Page 7, line 4.*

“Across ZIP codes where participants resided, the median inadequate housing score was 1.95 (25th and 75th percentiles: 1.53, 2.43). Specifically, on average, 2.25 percent of housing units lacked plumbing and 3.81 percent of housing units lacked kitchen facilities in ZIP codes that scored above the 50th percentile.”

Here the authors stir together averages, medians and percentiles. See previous comment. Summary statistics should help summarize the data, but here it is difficult to get a hold of what the (distribution of the) data really is. The authors need to straighten out the presentation of summary measures of covariates.

*Page 7, line 27.*

I find the term “bivariate analysis” to be somewhat misleading here. The term bivariate analysis is often used primarily for traditional statistical tests, such as Pearson and Spearman correlation, T-test and so on. Here the authors fit logistic regression models, which are indeed *\_models\_*, and this nuance should preferably be communicated to the reader. I would rather suggest writing something along the lines of “In univariate logistic regression models (Table 3), ...”

*Page 7, line 27.*

I am very fond of tables like table 3, where results from both univariate multiple regression models are presented alongside one another, giving a quick indication of what variables are possibly independent predictors, which are more strongly dependent on other predictors in the model, and so on. In table 3 I would thus suggest the second column is given the header “Univariate logistic regression models” and the third column the header “Multivariable logistic regression model”. See previous comment. Consider also to add another column termed “Optimal multivariable logistic regression model” where some objective mathematical measure for model selection (e.g. AIC) has been applied to the full multivariable model.

Note also that these ORs are not adjusted ORs *\_as a group\_*, as indicated in Table 3. These are still just ORs, but from a multiple rather than a univariate model. When talking about individual covariates, however, the OR from the multiple model is adjusted for the inclusion of other covariates *\_in the fitted model\_*.

*Page 7, line 27.*

See previous comment on being explicit in the stats section about how results will be presented, in order to be able to compact the presentation of numbers. Adding a sentence like

“Results from logistic regression are presented as odds ratio (OR) and 95% confidence interval (CI).“

to the stats section, allows for page 7, line 27 to be written like

“...had a higher odds ratio of homelessness in the past year (OR 1.13 (1.07, 1.18) for gentrification and 1.13 (1.05, 1.22) for inadequate housing).”

*Page 7, line 28.*

The authors use the terms odds and odds ratio interchangeably. This is incorrect. They are indeed two fundamentally different things: The odds is a one-to-one representation of probability, while the odds ratio is a comparative measure of two probabilities. The wording must be fixed in a revised version of the manuscript.

See for example this web site for an easy explanation:

<https://www.stat.ubc.ca/~rollin/teach/643w04/lec/node50.html>

*Page 7, line 34*

The authors state «...,two multivariable models were analyzed.» This I didn't get. Two multiple models were fitted? But Table 3 presents only one? Please clarify.

*Page 7, line 40*

“In the first multivariable model...” If indeed two multiple models were fitted it would be of great help to see results from both models in Table 3, not just the one presented in column three. The presentation in the text is hard to wrap one's head around.

*Page 7, line 41*

The abbreviation AOR has neither been defined nor explained. I am guessing it means adjusted OR. Note however, that while these OR are adjusted (see previous comment) it is not overly common to give the abbreviation AOR. They are just ORs from a more complex logistic regression model, where all individual covariates are adjusted for one another.

*Page 7, line 41*

I found the sentence “The addition of economic deprivation ... to the multivariable model” a bit confusing. Does this mean that the authors performed multiple model building? With stages of multiple models, by successively increasing the number of covariates in the multiple model? The stats anal section says nothing about this methodological approach. And as far as I can tell, Table 3, which presents results from univariate and multiple regression models, does not contain any of this? Please clarify.

*Page 7, line 47*

What is «Model 1:»? Model numbers have not been defined anywhere. And I cannot seem to find a Model 2. Or a Model 3. Please clarify.

*Page 7, line 51*

|  |  |
|--|--|
|  | <p>Is the information “ZIP code-level economic deprivation and percent household crowding were not significantly associated with homelessness» contained in Table 3? In that case; refer to it, e.g. “ZIP code-level economic deprivation and percent household crowding were not significantly associated with homelessness (Table 3).»</p> <p><b>References:</b></p> <p>Bursac Z, Gauss CH, Williams DH, Hosmer DW. Purposeful selection of variables in logistic regression. Source Code Biol Med. 2008; 3:17. <a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2633005/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2633005/</a></p> <p>McCreesh, N et. al. Evaluation of respondent-driven sampling. Epidemiology. 2012 Jan;23(1):138-47. doi: 10.1097/EDE.0b013e31823ac17c.</p> |
|--|--|

|                        |   |
|------------------------|---|
| <b>REVIEWER</b>        | Dr. Brittany B. Dennis<br>Peter Boris Centre for Addiction Research, McMaster University,<br>Hamilton ON Canada |
| <b>REVIEW RETURNED</b> | 01-Jan-2017   |

|                         |   |
|-------------------------|---|
| <b>GENERAL COMMENTS</b> | <p><b>General Comments:</b></p> <p>Terminology such as “people who inject drugs” sounds colloquial. Acknowledging this term is controversial, I would still suggest using the still commonly applied medical terminology such as IV drug users (IVDU) or patients with addiction who inject substances/substances drugs intravenously.</p> <p>The authors use the acronym MSA, however it is not outlined at any point within the manuscript.</p> <p>I would caution the authors about presenting homelessness as an outcome, the authors have maintained this is a cross-sectional study and as such they are assessing the associations among numerous variables collected as part of the National HIV Behavioral Surveillance initiative. I would be inclined to suggest the authors instead introduce the study as an exploration into the factors associated with homelessness (instead of stating “homelessness is our primary outcome.”)</p> <p><b>Specific Comments:</b></p> <p>The introduction of this paper provides powerful insight into the topic of homelessness among injecting drug users across 19 U.S. cities. Within the introduction the authors provide a detailed background emphasizing that housing and economic conditions require careful consideration when evaluating factors contributing to differential health outcomes among the general population. The authors provide an excellent segue into their discussion of injecting drug users as a vulnerable population who are greatly impacted by economic stability and housing, however that this is often understudied. The topic is introduced in a concise manner with a good background demonstration of the adverse effects of homelessness within the IVDU population (e.g. infection, relapse, hospitalization). The</p> |
|-------------------------|---|

authors have identified a large source of studies to support the associations demonstrated in the current literature.

#### Methods

The methods section of this investigation has a clear lack of reporting. I would suggest the authors use the STROBE guideline to ensure they have covered all important aspects of reporting observational research. In the methods section there is no mention of a sample size calculation. Do the authors have sufficient power to detect meaningful differences between groups (homeless vs. non-homeless) across the multiple variables selected for inclusion? In addition, the authors did not mention whether or not multiple testing error was corrected for, or how they planned to handle any missing data. These are very important statistical method components to any study and should be addressed in the methods section.

Particular attention should be paid to the management of missing data. The authors outline that "participants were excluded from the analytic sample if they had invalid/incomplete surveys (n=26); invalid or missing ZIP code information (n=499); were transgender persons who comprised too small a category to be analyzed (n=51); or were missing information on key covariates (n=134). The final analytic sample included 8992 participants." I would be interested to know how the exclusion of these participants affected the results of the study. Did IVU patients who were worse off more likely to miss specific responses? It is important to remember that many of these patients may have difficulties with properly completing these surveys, whether it be due to poor literacy or other factors, it is important to examine the characteristics of patients who fail to respond, or miss questions, especially among this population.

The study may also benefit from a participant flow diagram outlining each stage of inclusion, exclusion. This gives the audience an idea how many patients completed this national survey, and at which stage they were excluded.

The authors suggest individual-level covariates were also selected based on previous research but provide no references for this. I would be interested to know how they determined what variables to adjust for. The introduction did not address this either.

#### Statistical Methods

The authors should state how the descriptive statistics will be summarized in the paper (e.g. mean and SD).

As mentioned previously the authors did not provide information on the sample size calculations. Since the authors are developing a multi-level model this is particularly important to determine whether or not they have enough data (particularly events) to properly evaluate the impact of cluster level covariates (zip codes, countries, MSAs) on homelessness. Acknowledging the random effects within this model include both the variance between individual IVU and the variance between ZIP codes, MSAs, and country, I would be interested to know how many ZIP codes were captured by the sample. Correct me otherwise but I would assume there were many zip-codes included in the study such that it may be too many to even capture the impact of cluster.

One final clarification, the authors suggest gentrification may lead to an increase in the arrest of IVU and thereby increase the odds of homelessness. The authors planned to conduct two multivariable analyses if gentrification was proven to be significantly associated with homelessness, these models would then: 1) include potential mediators, 2) not include them. Could the authors please clarify what these potential mediators are. Exactly which variables would be included in each model.

|  |   |
|--|---|
|  | <p>Results</p> <p>The results section of this investigation is very well written, with a logical flow. While confidence intervals are fine on their own I might suggest including the p-values with each OR, audiences tend to appreciate this.</p> |
|--|---|

**VERSION 1 – AUTHOR RESPONSE**

Reviewer: 1

Reviewer Name  
 Julia Dickson-Gomez

Institution and Country  
 Medical College of Wisconsin  
 United States of America

Please state any competing interests or state ‘None declared’:  
 None declared.

Please leave your comments for the authors below

This is a well-written paper on a neglected topic, the influence of place-based characteristics such as housing affordability, inadequate housing and gentrification on homeless among PWID. There are a few areas the manuscript could be strengthened which I mention below.

Thank you so much for your feedback. We greatly appreciate the suggestions that you provided as we believe that they strengthen the paper.

The first bulleted comment in strengths is difficult to understand. In particular, it is not clear what you mean by "place-level housing and economic characteristics." This is described more clearly in other places in the manuscript. For example, in paragraph 6 on page 4 you list, unaffordable housing, economic deprivation, inadequate housing, overcrowding and gentrification. I would recommend listing these. However, it is not clear what different point you wish to make in bullet point 1 as opposed to bullet point 2.

Thank you for suggesting we revise the strengths and limitations section. We have revised the first bullet to include details on ZIP code and county-level characteristics that were analyzed. We also deleted the second bullet and incorporated some of its content with the first bullet. The first bullet now reads:

This study addresses gaps in prior literature by investigating the relationships of ZIP code-level economic deprivation and gentrification, and county-level unaffordable rental housing units and demand for public housing, to homelessness.

The first sentence in paragraph 2, page 4 is very long and should be edited.

Following your suggestion, we revised the first sentence in paragraph 2, page 4, to be more concise. It now reads:

Despite declines in percentages of unsheltered homeless persons in the United States from 40% to 31% between 2007 and 2014, a recent study by the Department of Housing and Urban Development

(HUD) reported that on a single night in 2014 more than 578,000 people experienced homelessness.<sup>5</sup>

You might want to consider providing a definition of the characteristics in your theoretical model and how they relate to one another. The introduction makes clear the connection between gentrification, economic deprivation, housing unaffordability, and homelessness. However, the role of incarceration and social networks is not really discussed in the introduction. It is also not clear why characteristics not examined in the study are included in the model.

Thank you for raising this point. Incarceration and sexual partnerships are considered individual-level covariates in analysis. These characteristics were therefore not discussed in the introduction because we intended to focus on place-level characteristics to strengthen rationale for the objective of this study. Following your suggestion and that of other reviewers, however, we removed social disorder, mental health and other characteristics that are not analyzed in this study from the conceptual framework.

The county and zip code measures are described well in your table so is probably not needed in the text.

Following your suggestion, we have removed definitions of the ZIP code and county measures from the text in the data collection and measures section.

There seems to be a word or phrase missing on page 8, line 1 "the potential high costs that may result..." It is not clear whether you are referring to homelessness, injection drug use or both.

Thank you for indicating that this sentence was unclear. We have since revised this sentence to state:

A high level of homelessness (60% in last year) was reported among this large sample of people who inject drugs, which not only highlights PWID's vulnerability to poor health outcomes but also raises concerns about the potential high societal costs that may result from homelessness, including increased health care costs.

The last line of the conclusion suggests that further research could "inform urban planning strategies and community mobilization campaigns designed to curb potential negative effects of gentrification and provide access to adequate housing among low-income and marginalized populations." Some concrete suggestions would be helpful, particularly since it seems that you do not feel like mixed income housing units have been successful. Would more supportive housing help?

Following your suggestion, we have edited the last sentence to state the following:

Growth in this line of research can inform urban planning strategies and community mobilization campaigns designed to curb potential negative effects of gentrification by strengthening access to stable and permanent housing among low-income and marginalized populations.

The tables could be reformatted to make them easier to read. For example, indentation in Table 1 of the types of sexual partners would increase readability.

Following your suggestion, we have edited Tables 1 and 2 accordingly.

Reviewer: 2

Reviewer Name

Danya Keene

Institution and Country  
Yale School of Public Health

Please state any competing interests or state 'None declared':

None

Please leave your comments for the authors below

1. This paper uses cross-sectional data to examine the association between gentrification and homelessness among PWID. As the authors note, this analysis represents an important contribution to the literature given that prior research has focused predominantly on individual versus structural/spatial determinants of homelessness. I think that this structural perspective is particularly important given that the paper focuses on a population (PWID) whose homelessness may be presumed to result from drug use. I recommend adding this point to the introduction. I also recommend drawing on literature that examines the stigmatization of homelessness to further strengthen the rationale for a structural analysis.

Thank you for this suggestion. We have now added the following sentence to the last paragraph of the introduction:

Increasing empirical evidence of the potential role of place-based factors on homelessness- above and beyond individual-level factors- may suggest potential structural interventions that should be implemented and reduce social stigma.

2. The conceptual model describing the relationship between gentrification and homelessness could be clarified and strengthened (both in the introduction and in the graphical representation of this model).

a. The paragraph on page 4, line 47 that describes the causes of homelessness is not clear. Some of the factors listed seem to be parallel housing outcomes that may be associated with homelessness, but are not part of the causal pathway (for example, overcrowding may occur when there is a shortage of available and affordable housing. It may be associated with homelessness at the ecological level, but both homelessness and overcrowding arise from the same lack of housing availability). I recommend simplifying this paragraph and the graphical model to describe the main hypothesized pathways between gentrification and homelessness (for example, rising housing costs). This paragraph could also discuss the demolition/revitalization of public housing that has occurred in the context of gentrification, and which may limit the number of available low income units.

Following your suggestion, we removed overcrowding and inadequate housing from analysis, and thereby deleted discussion of these attributes from the introduction. We also revised that paragraph to include the following sentence:

Similarly, the demolition of public housing complexes that occurred under the Housing Opportunities for People Everywhere policy in several cities, may contribute to the loss of affordable housing stock.

b. Related to the above point, the statement "increase demand for social services" is not clear, as it relates to gentrification and homelessness.

Thanks for indicating that this statement was unclear. We have now revised the sentence to state:

Urban redevelopment and gentrification may also reduce affordable housing stock by increasing rent and housing market value; increasing demand for supportive housing and housing subsidies (e.g., Section-8 vouchers); and potentially causing the needs of marginalized groups to go unmet.

c. Regarding the role of policing: it seems possible that increased policing in gentrifying areas could lead to a decrease in homelessness (rather an increase, as hypothesized by the authors) if a) local ordinances make it harder for homeless to stay outside and force them to go elsewhere b) if the homeless are incarcerated at higher rates and therefore in jail/prison, vs on the streets.

Thank you for suggesting this; we address this possibility in the limitations section.

3. I have two concerns about the use of “inadequate housing” as a mediator in this paper. First, conceptually, the causal link between inadequate housing and homelessness is not clear. Inadequate housing seems like a parallel outcome---it is associated with homelessness, because it shares the root cause of not enough affordable housing. Second, it is not clear how gentrification would lead to an increase in inadequate housing. I recommend omitting this variable from the paper.

You raise very important points about the correlation between homelessness and inadequate housing. After revisiting the literature, and seriously considering your feedback, we removed inadequate housing and overcrowding from analysis, and the conceptual framework, and other sections of the paper have been edited to reflect these changes.

4. The main hypothesized mediators (rent burden and availability of subsidized housing) did not pan out in the analyses. As the authors note, the rent burden measure is limited because it only captures moderate rent burden. The authors do not discuss the subsidized housing measure in their results or discussion, but this measure also seems to have some limitations given that in many cities, waiting lists are closed to new applicants, so length of time on the waitlist is unlikely to reflect actual availability of subsidies relative to need. To address these issues, the authors may want to consider using an additional zipcode/county level measure used by the Urban Institute: “available and affordable housing units for every 100 extremely low income households”. This measure (which includes subsidized units) captures the decline in available housing that is hypothesized to accompany gentrification and lead to increases in homelessness.

Thank you for this suggestion. Because housing affordability was measured among the lowest income category by the US. Census Bureau, we believe that we are capturing affordability among extremely low-income households. However, we now describe the limitations of the subsidized housing measures.

The measure of subsidized housing units that we used in this study is also limited and may not accurately reflect demand for subsidized housing. In many cities, waiting lists for subsidized housing are closed to applicants at specific thresholds and thus will exclude the waiting times of those who could not apply.

Reviewer: 3

Reviewer Name  
Becky Genberg

Institution and Country  
Brown University, USA

Please state any competing interests or state 'None declared':

None declared.

Please leave your comments for the authors below

This paper reports on a study of the association between place-level factors and homelessness among people who inject drugs in the United States. The study is clearly described and well written. The investigation of area-level factors with respect to homelessness is interesting and adds to the literature on an under-studied topic and population (people who inject drugs).

Thanks so much for this encouraging feedback!

The major concern I have regarding the analysis as presented is some confusion regarding the assessment of the potential mediating relationship of inadequate housing, housing affordability and economic deprivation between gentrification and homelessness. The results state that the inclusion of these three potential mediators did not impact the results of the association between gentrification and homelessness, however, I do not think this is evidence for establishing the mediation. The authors may want to consider a more unified model, such as a structural equation model, to examine the relationships between all variables of interest without the known challenges of examining mediation using traditional logistic regression analysis. Taking this one step further, are the authors suggesting that because the main relationship between gentrification and homelessness is not impacted by the inclusion of the three mediators mentioned above, that the proposed relationship between gentrification and homelessness must then be due to other mediating factors or a direct effect? More explanation of this in the discussion may be warranted. I think the statement in the concluding paragraph: "future longitudinal studies should explore whether these associations are casual, and identify potential mediators," is necessary, but calls into question the investigation of mediation as presented in the methods and results.

Thank you for these important insights. We also agree that the cross-sectional design does not allow an appropriate assessment of mediation. Following the concerns that you raised we have removed the mediation analyses from the analytic plan and have edited the text and conceptual framework accordingly.

Additionally, the correlations between the area-level variables could be examined and reported (there is mention of this in the methods, but nothing reported in the results on this investigation). It might be interesting to include information regarding levels of inadequate housing, housing affordability, and economic deprivation by varying levels of gentrification to continue to tease apart the relationship between these factors.

Excellent point. Following suggestions from other reviewers we removed inadequate housing and overcrowding due to the interrelated nature of these factors with homelessness. Because of this change and the finding that gentrification was the only significant place-level "predictor" in the final model, we chose not to include correlations as this information would not greatly add to the interpretation of findings.

There are some variables included in the conceptual model that are not discussed in the paper (e.g., social networks) and others that are referenced (i.e., police activity), but not included in the conceptual model (unless the authors are proposing that individual incarceration represents this area-level factor? If so this may need additional justification and clarification). Additional thought should also be given to the temporality of the area-level variables and how they are related over time.

Following your comments and that received from other reviewers, we have edited the conceptual

model to only include characteristics that are analyzed.

It may be useful to have a map or an appendix detailing the 19 MSAs that were included in the study. It is now difficult to assess the limitation as stated regarding PWID living outside of the MSAs that were included.

Thank you for highlighting this oversight. At the bottom of Table 2, we have now added a footnote that lists the MSAs included in the study:

The Northeast region includes the MSAs of Boston, Massachusetts; Nassau-Suffolk, New York; New York, New York; Newark, New Jersey; and Philadelphia, Pennsylvania. South region includes Atlanta, Georgia; Baltimore, Maryland; Dallas, Texas; Houston, Texas; Miami, Florida; New Orleans, Louisiana; and District of Columbia. Midwest region includes Chicago, Illinois and Detroit, Michigan. West region includes Denver, Colorado; Los Angeles, California; San Diego, California; San Francisco, California; and Seattle, Washington.

Reviewer: 4

Reviewer Name

Jo Røislien, Associate professor of medical statistics

Institution and Country

Department of Health Studies, University of Stavanger, Norway

Please state any competing interests or state 'None declared':

None declared

Please leave your comments for the authors below

See attached file for point-by-point comments.

## GENERAL

The study aims to quantify the association between homelessness among people who inject drugs and various place-based factors, such as gentrification and inadequate housing. In order to achieve this, the authors use quantitative data on multiple levels, e.g. individual level data and ZIP code data.

I have been asked to review this manuscript with a particular emphasis on the statistical methods and analyses used. This has been interesting, but not straightforward, as statistical methods are only vaguely presented, resulting in some brain gymnastics to untangle what has been done, and why. Below is a set of comments that hopefully will result in an improved paper, which attacks an important, yet analytically complex, topic.

Thank you for your feedback. We hope that we have addressed your comments in a way that improves the manuscript.

## POINT-BY-POINT COMMENTS

Page 3, line 29: Large numbers, e.g. 5,394 and 8992, must be written in the same way.

Following your suggestion, we have edited the manuscript so that large numbers are consistently presented in the same format.

Page 3, line 32: The abbreviation AOR is neither introduced nor explained. I assume the abbreviation refers to adjusted OR? Adjusted how? Please clarify.

Thank you for identifying this oversight. We have now edited the manuscript so that Adjusted Odds Ratio is presented prior to the acronym.

Page 3, line 37: I struggle to see how the conclusion is a valid conclusion based on the results presented in this study. Firstly, how does the evaluation of a planning strategy lead to less homelessness? In order for something to change you cannot just evaluate; you have to do something. Secondly, neither evaluation nor planning strategy are covariates in any of the analyses presented in this study.

To avoid further confusion, we revised the conclusion to state:

Additional research is needed to determine the mechanisms through which gentrification increases homelessness among PWID to develop appropriate community-level interventions.

Page 5, line 9. I must admit I struggled at first to understand what was the aim of the study. Before I figured out why: The last sentence in the introduction ("Figure 1 presents a conceptual..." - see below) blurs the aim of the study. Instead of the introduction ending with a punctuation mark, this sentence leads to the introduction ending with a crossfade from "what" to "how". This last sentence is not introduction or background information: It is about methodology. It is valuable information, indeed, but it belongs someplace else in the manuscript. For example page 6 line 28, just after the description of the variables, and before the Stats Anal subsection.

Following your suggestion, we have now moved the description of the conceptual framework

Page 5, line 21. Respondent driven sampling (RDS) is a non-probability sampling method. That is, non-randomized, and on thus on collision course with most standard assumptions for proper statistical analysis. Papers have demonstrated that even when RDS has been used correctly, bias cannot necessarily be ruled out (McCresh et al, 2012). To the best of my knowledge the jury is still out on as to how RDS results are to be interpreted; that is, how analytical results using RDS relate to the real world they try to describe. This is a serious limitation to the paper and must be addressed.

Thank you for raising this important point and providing a citation to support this argument. We have now revised the last sentence of the first paragraph included in the limitations section to read as follows (with the McCresh citation added at the end):

Additionally, findings may not be generalizable to PWID living outside of the MSAs captured by NHBS, and the extent to which RDS generated a representative sample in this study cannot be confirmed.

Page 5, line 37-38. The authors state that they have excluded 134 individuals due to missing on key covariates, indicating that they aim for a complete case analysis. This should be OK, as 134/9702 is approx 1%, and method studies indicate that for missing < 5% complete case usually works fine.

In light of the latter I am more concerned with the 499 individuals removed because of missing Zip code information, as 499/9702 > 5%. That is, this is enough data to introduce significant bias, of unknown order and magnitude, when simply discarded. Can the authors please clarify why they have chosen to introduce this potential bias in the results, and not, say, perform multiple imputation?

Following your feedback, we have added the following information to the study sample section:

Those excluded from analysis were more likely to be white (>10% difference) and live in the West and less likely to live in the Midwest than those included in the analytic sample. Other characteristics measured in this study were not substantially different (>10%) between those included and excluded from analysis.

Because multiple imputation can impose bias particularly when place-level data is imputed, we did not to conduct imputation.

Page 6, line 34. While the subsection is now termed “Analysis”, this is not the subsection where the authors describe analysis in general; it is the section that describes the handling of quantitative data, e.g. statistical analysis. I would thus prefer the heading “Statistical analysis” rather than just “Analysis” as the latter is overly vague as to what the subsection actually describes.

We have now revised this sub-heading accordingly.

Page 6, line 37. The authors start the stats subsection by stating that “The distributions of all characteristics were determined using descriptive statistics.” This is not correct. There are numerous books filled with different statistical distributions. It is truly a complex task, both empirically and mathematically, to assess the true distribution of data sets. Moreover; fully determining the distribution of a covariate is not really necessary in most regression models. What we are usually interested in is some general idea about the data, some measures describing the essential features. Like a centrality measure and a variation measure. Descriptive statistics can be many things: tell the reader what descriptive statistics you have used. Something along the lines of:

“Symmetric continuous data are summarized using mean and standard deviation (SD) and skewed continuous data using median and 25th and 75th percentiles, while proportions are presented as percentages.”

Thank you for raising this point. We have now edited this sentence to state:

The distributions of all characteristics were determined using descriptive statistics (i.e., frequencies and percentages and means and standard deviations).

Page 6, line 38. “Logistic multilevel analysis was used...”

Firstly, on a more general note; I am not a big fan of dismissing the statistical analysis of quantitative data as merely “analysis”. It is too vague a description of what is actually being done, in particular on more advanced statistical modelling like here. Logistic regression, which I assume the authors have used (it is not explicitly stated), is not just “analysis”: it is a probability model; a mathematical model of the real world. Regression modelling is more than just analysis: It assumes several things about the world it tries to model, and this should be communicated to the reader; regression modelling assumes that the data follow certain probability distributions, that variables affect one another in a particular way and so on. So I would suggest rewriting this sentence to something along the lines of “Multilevel logistic regression models, both univariate and multivariable, were fitted to the data, and used to ...”

Following your suggestion, we revised this sentence to state:

Multilevel logistic regression models were used to assess bivariate and multivariable relationships of place-based factors to the odds of homelessness.

Secondly; how did the authors carry out the multilevel modelling? What type of multilevel model was this? As some data are on an individual level, and some data on group level, this is not a standard statistical model, and some explanation of the model is needed. Mixed and multilevel data modelling often refers to a way of handling clusters within the data set, the classical example being that many school kids are given a test, but as children in the same class have had the same teacher results for these kids might be correlated (good / poor teachers), results from children from the same school might be correlated (good / poor schools), and so on. That is: Data are on an individual level, but there is clustering of the data on multiple levels. In the present study, however, the data themselves are on varying scales; individual level data, zip code data, and so on. So some information should be added to the paper as to how different types of information, from individual to group characteristics, can be properly handled, with references to the proper statistical methods (which I assume is being applied).

We include the following statement in the statistical analysis section:

Random intercepts were included for ZIP codes, counties, and MSAs.

Page 6, line 41. "Multivariable analysis assessed the relationships of place-based factors significant ( $p < 0.05$ ) in bivariate analysis"

This sentence is a bit unclear. But I assume it means that the authors first carried out a series of bivariate analyses, and then statistically significant variables were entered into multiple regression analyses? Unless having sparse amounts of data such a two-step procedure can easily be an overly ad hoc approach for building optimal statistical models. Rather one could enter all suggested variables into the multiple model, and then do optimal model selection based on some proper objective model description criteria such as AIC, BIC, FIC etc to remove non-contributive covariates from the hypothesized full multiple model.

If the authors do want to use a two-step procedure for model building it must be carried out somewhat differently than is presented here; variables with non-significant p-values from univariate analysis must also be entered into the multiple model, as p-values change from uni- to multivariable analyses, and setting the inclusion threshold as low as 0.05 will potentially result in dismissing otherwise important predictors. A p-value cutoff of  $\sim 0.20$  is usually recommended (see for example Bursac et al, 2008).

Thank you for this suggestion, we now use a cutoff of  $p < 0.20$  and have revised the statistical analysis section accordingly to state this.

Page 6, line 44-48. The two sentences below (taken from the stats section) are not statistical analysis; they describe covariate background information. They do thus not belong in the Stats section, which is for stats only. Move them further up in the Methods section.

"Based on prior research, gentrification was conceptualized as reducing economic deprivation, housing affordability, and inadequate housing. Gentrification was also hypothesized as increasing law enforcement practices that would increase the arrest of PWID and thereby increase the odds of homelessness."

Following your comment and those of other reviewers, we have removed this statement from the analysis section.

Page 6, line 54. As the authors do not present a novel statistical method, but use an existing method

already implemented software, information about the software and the function(s) used would be valuable. There are several analytical approaches to the same statistical idea, and different ways of implementing them, and the type of mathematical framework, and corresponding implementation used, might potentially influence the results, and thus the corresponding conclusions.

We appreciate this suggestion. However, because this paper is not a “methods” paper, we have decided not to present the actual commands that were used. Furthermore, because we specified that Stata software was used, we believe that readers have sufficient information to determine what commands were used.

Page 7, line 1. The authors write that “Approximately 50% of participants resided in counties where there was at least 87% (25th and 75th percentiles: 80.06, 90.32) unaffordable rental units...”

The numbers presented here have no consistency to them. In the same sentence there are three different precision levels: The percentiles have four digits precision (80.06, 90.32), the median has only two digits precision (87), while the first number hasn't even two digits precision (approximately 50). Similar juggling with precision levels can be found other places in the manuscript as well. The authors need to tighten up the presentation of numbers.

We have edited the manuscript so that numbers are consistently presented with the same number of digits.

Minor additional comment:

It is not mentioned anywhere what the main number, 87%, actually means. And (in my personal opinion) writing out ‘25th and 75th percentiles’ in full often tends to clutter the result section, making it more difficult to get a hold of the actual results/numbers. As a statistician I often suggest being explicit in the stats section about how the format of the presentation of numbers in the results section will be. That is, letting, say, the first sentence in the stats section be something along the lines of:

“Symmetric continuous data are summarized using mean and standard deviation (SD) and skewed continuous data using median and 25th and 75th percentiles, while proportions are presented as percentages.”

Then page 7, line 1 might read “Approximately half of participants resided in counties where there was a median of 87.xx% (80.06, 90.32) unaffordable...”

Page 7, line 3. “...average number of months that applicants were on waiting lists for assisted housing was 26 months (25th and 75th percentiles: 21, 38)”

This summary of numbers is inconsistent with basic mathematical results from inferential statistics. If data are symmetric (e.g. normally) distributed, then the mean and SD are sufficient statistics, and we can summarize the data using mean and SD without loss of information. If data are skewed however, the mean is not necessarily a good measure of centrality, and SD similarly does not represent the unsymmetric variation in the data, so the median and percentiles are usually preferable summary statistics.

Here, however, the authors use average and percentiles. Why? If data are symmetric, mean and percentiles are not sufficient, and if data are skewed the mean is often a misleading centrality measure. Please clarify.

Thank you noting that this was confusing. We have now edited the results so that means are presented instead of medians.

Page 7, line 4. “Across ZIP codes where participants resided, the median inadequate housing score was 1.95 (25th and 75th percentiles: 1.53, 2.43). Specifically, on average, 2.25 percent of housing units lacked plumbing and 3.81 percent of housing units lacked kitchen facilities in ZIP codes that scored above the 50th percentile.”

Here the authors stir together averages, medians and percentiles. See previous comment. Summary statistics should help summarize the data, but here it is difficult to get a hold of what the (distribution of the) data really is. The authors need to straighten out the presentation of summary measures of covariates.

We have now edited the results so that means are presented instead of medians.

Page 7, line 27. I find the term “bivariate analysis” to be somewhat misleading here. The term bivariate analysis is often used primarily for traditional statistical tests, such as Pearson and Spearman correlation, T-test and so on. Here the authors fit logistic regression models, which are indeed models, and this nuance should preferably be communicated to the reader. I would rather suggest writing something along the lines of “In univariate logistic regression models (Table 3).”

Thank you for this suggestion. We have now replaced the words: bivariate analyses with univariate logistic regression models.

Page 7, line 27. I am very fond of tables like table 3, where results from both univariate multiple regression models are presented alongside one another, giving a quick indication of what variables are possibly independent predictors, which are more strongly dependent on other predictors in the model, and so on. In table 3 I would thus suggest the second column is given the header “Univariate logistic regression models” and the third column the header “Multivariable logistic regression model”. See previous comment. Consider also to add another column termed “Optimal multivariable logistic regression model” where some objective mathematical measure for model selection (e.g. AIC) has been applied to the full multivariable model.

Note also that these ORs are not adjusted ORs as a group, as indicated in Table 3. These are still just ORs, but from a multiple rather than a univariate model. When talking about individual covariates, however, the OR from the multiple model is adjusted for the inclusion of other covariates in the fitted model.

Following your suggestion we have now replaced the words: bivariate analyses with univariate logistic regression models.

Page 7, line 27. See previous comment on being explicit in the stats section about how results will be presented, in order to be able to compact the presentation of numbers. Adding a sentence like

“Results from logistic regression are presented as odds ratio (OR) and 95% confidence interval (CI).“ to the stats section, allows for page 7, line 27 to be written like

“...had a higher odds ratio of homelessness in the past year (OR 1.13 (1.07, 1.18) for gentrification and 1.13 (1.05, 1.22) for inadequate housing).”

Thank you for bringing this typo to our attention. Because we have removed inadequate housing from analysis based on suggestions from other reviewers, this sentence has since been deleted.

Page 7, line 28. The authors use the terms odds and odds ratio interchangeably. This is incorrect.

They are indeed two fundamentally different things: The odds is a one-to-one representation of probability, while the odds ratio is a comparative measure of two probabilities. The wording must be fixed in a revised version of the manuscript.

Odds ratios are provided in parentheses to further provide details of the how the odds are described in the text. This is presented in a conventional manner and authors of papers published in BMJ Open have used a similar format .

See for example this web site for an easy explanation:  
<https://www.stat.ubc.ca/~rollin/teach/643w04/lec/node50.html>

Page 7, line 34 The authors state «...two multivariable models were analyzed.» This I didn't get. Two multiple models were fitted? But Table 3 presents only one? Please clarify.

Following suggestions from other reviewers, mediation was no longer explored, thus the results of this "secondary multivariable model" were removed.

Page 7, line 40

"In the first multivariable model..." If indeed two multiple models were fitted it would be of great help to see results from both models in Table 3, not just the one presented in column three. The presentation in the text is hard to wrap one's head around.

Please see response to your prior comment

Page 7, line 41 The abbreviation AOR has neither been defined nor explained. I am guessing it means adjusted OR. Note however, that while these OR are adjusted (see previous comment) it is not overly common to give the abbreviation AOR. They are just ORs from a more complex logistic regression model, where all individual covariates are adjusted for one another.

Please see the response to your previous comment

Page 7, line 41 I found the sentence "The addition of economic deprivation ... to the multivariable model" a bit confusing. Does this mean that the authors performed multiple model building? With stages of multiple models, by successively increasing the number of covariates in the multiple model? The stats anal section says nothing about this methodological approach. And as far as I can tell, Table 3, which presents results from univariate and multiple regression models, does not contain any of this? Please clarify.

Following your prior suggestion related to step-wise model building, we have now included place characteristics in the multivariable model, which were significantly associated with homelessness at a p- value <0.2 in univariate models.

Page 7, line 47 What is «Model 1:»? Model numbers have not been defined anywhere. And I cannot seem to find a Model 2. Or a Model 3. Please clarify.

Following suggestions from other reviewers, mediation was no longer explored, thus the results of this "secondary multivariable model" were removed.

Page 7, line 51 Is the information "ZIP code-level economic deprivation and percent household crowding were not significantly associated with homelessness» contained in Table 3? In that case; refer to it, e.g. "ZIP code-level economic deprivation and percent household crowding were not significantly associated with homelessness (Table 3).»

Thank you for this suggestion. We have instead decided to delete this sentence and only mention those place-based variables that were significant or marginally significant in univariate models.

#### References:

Bursac Z, Gauss CH, Williams DH, Hosmer DW. Purposeful selection of variables in logistic regression. *Source Code Biol Med.* 2008; 3:17.  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2633005/>

McCreesh, N et. al. Evaluation of respondent-driven sampling. *Epidemiology.* 2012 Jan;23(1):138-47. doi: 10.1097/EDE.0b013e31823ac17c.

Reviewer: 5

#### Reviewer Name

Dr. Brittany B. Dennis

#### Institution and Country

Peter Boris Centre for Addiction Research, McMaster University, Hamilton ON Canada

Please state any competing interests or state 'None declared':

None Declared

Please leave your comments for the authors below

#### General Comments:

Terminology such as "people who inject drugs" sounds colloquial. Acknowledging this term is controversial, I would still suggest using the still commonly applied medical terminology such as IV drug users (IVDU) or patients with addiction who inject substances/substances drugs intravenously.

Thank you for this suggestion. Because "people who inject drugs" is a widely accepted term used by the United States Centers for Disease Control and Prevention and the World Health Organization, we prefer to use this terminology in our manuscript.

The authors use the acronym MSA, however it is not outlined at any point within the manuscript.

Thank you for noticing this oversight; we have since added the full title, "metropolitan statistical areas", to appear directly before the first mention of MSA in the 1st paragraph of the methods section.

I would caution the authors about presenting homelessness as an outcome, the authors have maintained this is a cross-sectional study and as such they are assessing the associations among numerous variables collected as part of the National HIV Behavioral Surveillance initiative. I would be inclined to suggest the authors instead introduce the study as an exploration into the factors associated with homelessness (instead of stating "homelessness is our primary outcome.")

Following your suggestion, we removed all references to homelessness as the outcome from the paper.

#### Specific Comments:

The introduction of this paper provides powerful insight into the topic of homelessness among injecting drug users across 19 U.S. cities. Within the introduction the authors provide a detailed background emphasizing that housing and economic conditions require careful consideration when evaluating factors contributing to differential health outcomes among the general population. The authors provide an excellent segue into their discussion of injecting drug users as a vulnerable population who are greatly impacted by economic stability and housing, however that this is often understudied. The topic is introduced in a concise manner with a good background demonstration of the adverse effects of homelessness within the IVDU population (e.g. infection, relapse, hospitalization). The authors have identified a large source of studies to support the associations demonstrated in the current literature.

Thank you so much for this feedback!

#### Methods

The methods section of this investigation has a clear lack of reporting. I would suggest the authors use the STROBE guideline to ensure they have covered all important aspects of reporting observational research. In the methods section there is no mention of a sample size calculation. Do the authors have sufficient power to detect meaningful differences between groups (homeless vs. non-homeless) across the multiple variables selected for inclusion? In addition, the authors did not mention whether or not multiple testing error was corrected for, or how they planned to handle any missing data. These are very important statistical method components to any study and should be addressed in the methods section.

Particular attention should be paid to the management of missing data. The authors outline that “participants were excluded from the analytic sample if they had invalid/incomplete surveys (n=26); invalid or missing ZIP code information (n=499); were transgender persons who comprised too small a category to be analyzed (n=51); or were missing information on key covariates (n=134). The final analytic sample included 8992 participants.” I would be interested to know how the exclusion of these participants affected the results of the study. Did IVDU patients who were worse off more likely to miss specific responses? It is important to remember that many of these patients may have difficulties with properly completing these surveys, whether it be due to poor literacy or other factors, it is important to examine the characteristics of patients who fail to respond, or miss questions, especially among this population.

The study may also benefit from a participant flow diagram outlining each stage of inclusion, exclusion. This gives the audience an idea how many patients completed this national survey, and at which stage they were excluded.

Thank for raising this important point. We have added the following sentence to the end of the NHBS study sample section:

Those excluded from analysis were more likely to be white (>10% difference) and live in the West and less likely to live in the Midwest than those included in the analytic sample. Other characteristics measured in this study were not substantially different (>10%) between those included and excluded from analysis.

The authors suggest individual-level covariates were also selected based on previous research but provide no references for this. I would be interested to know how they determined what variables to adjust for. The introduction did not address this either.

Following your suggestion, we have added citations to follow this statement in the analysis section. Because place-based exposures were the key variables of interest, we provided detail about them in the introduction and not individual-level confounders. However, there are references that are cited

after the description of the conceptual framework, which serve as rationale for including the confounders that were analyzed.

#### Statistical Methods

The authors should state how the descriptive statistics will be summarized in the paper (e.g. mean and SD).

Following your suggestion, we have added the following information to the first sentence of the first paragraph of the analysis section:

The distributions of all characteristics were determined using descriptive statistics (i.e., frequencies and percentages and means and standard deviations).

As mentioned previously the authors did not provide information on the sample size calculations. Since the authors are developing a multi-level model this is particularly important to determine whether or not they have enough data (particularly events) to properly evaluate the impact of cluster level covariates (zip codes, countries, MSAs) on homelessness.

Formal sample size calculations were not conducted for this analysis.

Acknowledging the random effects within this model include both the variance between individual IVDU and the variance between ZIP codes, MSAs, and country, I would be interested to know how many ZIP codes were captured by the sample. Correct me otherwise but I would assume there were many zip-codes included in the study such that it may be too many to even capture the impact of cluster.

The number of ZIP codes, counties and MSAs within which participants reported residing are listed in Table 2.

One final clarification, the authors suggest gentrification may lead to an increase in the arrest of IVDU and thereby increase the odds of homelessness. The authors planned to conduct two multivariable analyses if gentrification was proven to be significantly associated with homelessness, these models would then: 1) include potential mediators, 2) not include them. Could the authors please clarify what these potential mediators are? Exactly which variables would be included in each model?

Following suggestions of other reviewers, mediation analysis has been removed from the paper, thus only one final multivariable model was analyzed.

#### Results

The results section of this investigation is very well written, with a logical flow. While confidence intervals are fine on their own I might suggest including the p-values with each OR, audiences tend to appreciate this.

Thank you for this suggestion. Because, the interpretation of the results does not change with the addition of p-values, we have only included p-values for marginally significant associations between place characteristics and homelessness.

## VERSION 2 – REVIEW

|                        |   |
|------------------------|---|
| <b>REVIEWER</b>        | Julia Dickson-Gomez<br>Medical College of Wisconsin |
| <b>REVIEW RETURNED</b> | 29-Mar-2017   |

|                         |   |
|-------------------------|---|
| <b>GENERAL COMMENTS</b> | I believe that previous concerns have been adequately addressed. One small suggestion would be to expand the explanation in the introduction of how Housing Opportunities for People Everywhere may have contributed to the loss of affordable housing. |
|-------------------------|---|

|                        |  |
|------------------------|--|
| <b>REVIEWER</b>        | Becky Genberg<br>Brown University, USA |
| <b>REVIEW RETURNED</b> | 05-Apr-2017                            |

|                         |                     |
|-------------------------|---------------------|
| <b>GENERAL COMMENTS</b> | No further comments |
|-------------------------|---------------------|

|                        |  |
|------------------------|--|
| <b>REVIEWER</b>        | Jo Røislien<br>Department of Health Studies, University of Stavanger, Stavanger, Norway. |
| <b>REVIEW RETURNED</b> | 05-Apr-2017  |

|                         |  |
|-------------------------|--|
| <b>GENERAL COMMENTS</b> | All my comments have been answered satisfactory. |
|-------------------------|--|