Reviewer's report

Title: Environmental justice and drinking water quality: Are there socioeconomic disparities in nitrate levels in U.S. drinking water?

Version: 0 Date: 30 Jul 2018

Reviewer: Kirsten Almberg

Reviewer's report:

This paper was aimed at understanding if there are socioeconomic disparities in nitrate levels in U.S. drinking water. The authors attempted to address this through regression modeling using county-level variables of racial/ethnic, land use, and economic indicators as predictors of nitrate levels. This paper suffers from several large methodological issues including a lack of model building and therefore incomplete assessment of confounding and co-linearity; lack of seasonal differences in results; lack of discussion of ecologic fallacy; and a lack of investigation into whether the CWSs in this study met regulatory requirements and how that might have affected the interpretation of the results.

Background

Line 82: The term environmental justice is used here, but is not explained. It appears to be a non-specific term that encompasses racial, ethnic, and socio-economic geographic disparities in environmental contaminant exposure. Please either define this, or use alternate phrasing.

Line 91-93: Consider rephrasing to "Environmental justice associations with drinking water have not been consistently observed, and may depend on the spatial scope and individual contaminants studied."

Lines 130-132: Citation needed for the EPA MCL - this should be the title 40 of the federal code of regulations.

Lines 130-149: This paragraph can be significantly shortened. As health endpoints were not assessed in this analysis, the details provided here are less relevant to the readers. Please condense.

Lines 158-162: The two sentences that start with "To conduct our analyses…" are methods, not background. Remove from background.
Lines 164-167: "the sentence beginning with "Socioeconomic disparities in nitrate exposures…" is better left for the discussion section.

Methods

Line 172: Remove "application program interface."

Line 173: "was" should be edited to "were CWSs…"

The section "Identifying populations served by CWSs" should be shortened.

Lines 197-199: Omit sentence referring to the San Joaquin Valley.

Lines 199-202: Consider rephrasing to "Few states provide public access to electronic records documenting the service areas of their public water systems, so we relied on the information included in SDWIS."

Lines 210-211: Omit reference to zip code as it was not used in the analysis. This may be worth mentioning in the discussion regarding limitations.

Lines 212-232: This paragraph should be shortened to 2-3 sentences only.

Lines 233-244: Similarly, this paragraph should be shortened and relay only the relevant information to your analysis which is that wholesaler contaminant levels were assigned to all downstream systems that purchased water from them.

Line 247: Consider rephrasing to "Water system characteristics and demographic data were linked by matching the names in the SDWIS city served and county served fields with the geographic names in the Census Bureau demographic data."

Lines 257-259: Consider rephrasing to "Typographical errors and differences in abbreviation conventions were corrected to improve string matching."
Lines 260-265: "For city served names that did not match or matched multiple Census Bureau records after these corrections, we made additional amendments based on information gathered from Internet searches. When we could not definitively link city served names from SDWIS with location names in Census Bureau data, names were considered non-matching and were not included in subsequent analyses." "Additional amendments based on internet searches" is non-specific and hard for an audience to understand what has happened with these records. The second sentence in this quote seems to contradict the first about non-matching records. Please clarify and condense.

Lines 281: Please clarify the statistic of "189,065 years of system activity." Consider omitting the second half of this sentence starting with "…matched to county demographics."

Lines 286-288: please provide a reference for using this regression analysis for estimating nitrate concentrations from nitrate-nitrite values.

Lines 299-300: Omit this sentence starting with "For non-detects lacking DL data…"

Lines 307-320: Why were covariates chosen a priori and no model building attempted? Was each covariate individually assessed for confounding before modeling? If so, this should be stated and results presented. If not, this is highly recommended. Based on results presented later, it appears that % livestock is a poor predictor of nitrate levels and could have been dropped from regression models.

Results

Lines 379-380: The sentence "Nearly 90% of our sample's CWSs have groundwater sources, while 65% of the Americans in our sample are served by a CWS with a surface water source." Is confusing. Please consider revising to "Nearly 90% of the CWSs in our sample have groundwater sources, however, the majority of the population in our sample area (65%) are served by a CWS with a surface water source."

Lines 388-389 and 424-426: The authors report seemingly contradictory results: "Very small systems (≤500 people) had higher nitrate concentrations than other system sizes." And "Larger water systems generally had higher nitrate concentrations than very small systems (increased nitrate levels of 426 9.0% for small systems up to 51% for very large systems)." Please clarify.
Discussion

Lines 498-500: Do not reference tables in the discussion, especially if they have not been referenced yet in the results section (Table S2).

Lines 551-556: Insufficient discussion of why the authors observed higher nitrate concentrations in smaller water systems compared to larger systems, only to have this association reversed in regression modeling. The implication here is that the size of the water system is not what is driving the association, but that there is a confounder that is masking this association in bivariate comparison. This finding is troublesome and could be address through model building, with the addition of one variable at a time. In crude regression models, was CWS size associated with nitrate levels with small systems having the highest likelihood of high nitrate? If so, this would match your descriptive analysis. Then, with the addition of one covariate at a time, the authors could see which covariate causes this association to flip. Without out this process, it is hard to interpret these results and draw meaningful conclusions.

Lines 567-568: This sentence is misleading and sounds like smaller systems have more frequent monitoring, which contradicts what was said in the previous paragraph. Please rephrase to make clear these are findings from another study, which found that smaller systems have worse reporting practices.

Lines 571-572: It seems like a Pearson correlation would not be the best way to test if smaller systems had high proportion of missing values than larger. One problem with this discussion of missing nitrate values may be addressed by discussion of the regulatory requirements for monitoring. If a CWS has demonstrated low levels of nitrate over a certain number of years, their monitoring frequency is reduced. There are also difference monitoring frequencies for CWS with surface vs groundwater sources. Discussion of this is warranted.

Level of interest
Please indicate how interesting you found the manuscript:

An article whose findings are important to those with closely related research interests

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable
**Declaration of competing interests**

Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal.