PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Does spatial proximity drive norovirus transmission during outbreaks in hospitals?
AUTHORS	Harris, John; Lopman, Benjamin; Cooper, Ben; O'Brien, Sarah

VERSION 1 - REVIEW

REVIEWER	Janneke Heijne, PhD Institute of Social and Preventive Medicine (ISPM) University of Bern Switzerland I have no competing interests to declare
REVIEW RETURNED	14-May-2013

GENERAL COMMENTS	The authors reconstruct transmission trees of who is infected by whom using data from multiple outbreaks of norovirus in hospitals. Using this method, the authors explored whether proximity (defined as sharing the same bay with a symptomatic norovirus case) is an important risk factor of acquiring symptomatic norovirus infection. Major comments 1. The authors based their method of estimating transmission trees on a previously published method, and also refer to this work. However, no description of the methods is provided here, which makes it difficult to see whether the methods are truly the same. The authors provide the R-code (which I appreciate), but the reader has to interpret the code to be able to understand the method. As far I am able to read the provided R-code, I have the feeling that the method used here, is more similar to the method described by Ypma et al (2012, Proc Biol Sci. 279(1728):444-50). I suggest that the authors provide more details about the used method either in the methods section of the paper or in an appendix. 2. As said before, I very much appreciate that the authors provide the R-code. However, the authors should make an effort in explaining the code more clearly, and provide it without any errors, so that others are able to use it. For example: - There are some "}" missing - Some variables are given clear names (such as "whoinfectswhom"), while others are not (such as pkl), please amend. - Explain the meaning of "w" (page 23 line 16 – page 24 line 20). I assume that this is the serial interval distribution? If so, why are there zero's at w[5], w[7] and w[9]. As I understand the code, the "w" is later overwritten by another function. Please explain. - Would it be possible to provide the data (or an example dataset),

	so that people can really re-use the code? 3. I found the main outcome of the study (Pk) difficult to interpret. I also have the feeling that the equation provided on page 6 is not an adequate description of Pk. But again, this is based on my understanding of Pk from the R-code. Why not write the meaning of Pk in simple words, for example “Pk is the number of times that a transmission occurred between two individuals sharing the same bay (per outbreak)”. Furthermore, I have the feeling that Pk depends on outbreak size, and has a different maximal value for every outbreak and should therefore not be summed over all outbreaks. By looking at Figure 2, values are between 80 and 160, but this is difficult to put in a context as I don’t know the outbreak sizes. I somehow expected a value between 0 and 1 (i.e. the percentage of transmissions that occurred between people sharing the same bay) for each outbreak. This would give a more meaningful number and allows for comparison between different outbreaks. 4. As the authors acknowledge in the introduction, it is already known that proximity to a (vomiting) norovirus case increases the infection risk. Therefore, I’m not entirely sure what this study adds to these findings. Since the authors simulate the transmission trees, would it be able to go a step further and try to quantify within bay and between bay transmission? The authors could also look at, for example, the bay reproduction numbers. Furthermore, if the ward types are known for these outbreaks (e.g. paediatric ward) reproduction numbers could be analysed stratified by ward-type. 5. There are some issues related to the way the authors construct the serial interval distribution. By looking at the difference in symptom onset between the first two cases of an outbreak has the tendency to over represent short serial intervals. Especially in the case of point source outbreaks, where cases developing symptoms 1 day after the “index case” are very likely related to the same point source and not to the symptoms of the index case. This limitation should be discussed. Minor comments 1. The authors used random mutations of bays to test if proximity is associated with an increased risk of acquiring norovirus. By doing this, did the authors keep the distribution of bay-sizes per ward constant (i.e. was it possible that, by chance, 20 people occupied one bay)? 2. Could a short description be given about the type of wards where the norovirus outbreaks where observed and if patients were able to walk around and have contact with people in other bays or wards? 3. Introduction, page 3, lines 50 – 52: be specific here that the aim is to assess the risk of acquiring SYMPTOMATIC norovirus gastroenteritis. 4. Page 4, lines 39 – 43: The sentence “In three … of 2007/2008” is unclear to me. Be more specific about the difference in data collection methods and questionnaires between the hospitals. 5. Page 5: to give the readers an idea about the size of the wards, please provide an average number of bays per ward. Furthermore, what is the average occupancy rate in the bays? 6. Please change “Heinje et al” to “Heijne et al” throughout the paper and the appendix table and figure

	7. Page 9, line 25: remove “very” 8. Page 9, line 38: please explain the 47% (should this perhaps be 44%, i.e. 65 / 149?) 9. Page 10, line 8: I assume that with “Figure 3” the authors mean “Figure 1_supplement”. Please amend. 10. Page 10, line 30: add to the sentence “in a hospital setting” 11. Reference 21: change “Tuenis” to “Teunis” 12. The functionality of the second column in Table 1 is unclear to me, please change. 13. The first key message is unclear to me.

REVIEWER	David Partridge, Consultant Microbiologist, Sheffield Teaching Hospitals, UK. No competing interests.
REVIEW RETURNED	15-May-2013

THE STUDY	Method of diagnosis of norovirus at the included hospitals is not recorded and nor is testing strategy e.g. test just index case or test all cases during outbreak.

VERSION 1 – AUTHOR RESPONSE

Major comments

1. The authors based their method of estimating transmission trees on a previously published method, and also refer to this work. However, no description of the methods is provided here, which makes it difficult to see whether the methods are truly the same. The authors provide the R-code (which I appreciate), but the reader has to interpret the code to be able to understand the method.
As far I am able to read the provided R-code, I have the feeling that the method used here, is more similar to the method described by Ypma et al (2012, Proc Biol Sci. 279(1728):444-50). I suggest that the authors provide more details about the used method either in the methods section of the paper or in an appendix.

As stated in the paper, we used the method described by Wallinga & Teunis (2004). We can confirm that what we wrote in the paper was correct, and we did not use the approach described by Ypma et al. We have supplied the code so reviewers (and readers) can check that our claim is correct. The key section of the code is

```
pkl<-matrix(rep(0,N^2),nrow=N) #relative likelihiood  that k was infected by l


for(l in 1:N){
 # print(l)
  for(k in 1:N){
   numerator<-wij[k,l]
   denom<-denoms[k]
   pkl[k,l]<-numerator/denom
```

```
  }
}
```

This corresponds to the first formula on page 511 of Wallinga & Teunis. The variable names are chosen to be consistent with those in Wallinga & Teunis to make it easier for the readers.

To be able to understand the method readers should refer to Wallinga & Teunis (2004) which we cite (and since this is an open access publication, all readers with internet access should be able to get hold of this paper). The aim of this paper is not to explain this method (which was explained very well by Wallinga & Teunis), but to apply it. We think a detailed technical explanation here would be both unnecessary and inappropriate. We did howerver include a brief non-technical description of the method in natural language, and we have now extended this, explaining the idea behind this approach in a little more detail. The full non-technical description now reads:

"The analysis is based on a probabilistic reconstruction of chains of transmission (trees) based on the dates of illness onset for patients affected in outbreaks. It makes use of methods developed for SARS transmission and later applied to norovirus [12-15]. If we knew with certainty who acquired infection from whom it would be straightforward to quantify the role of proximity in norovirus outbreaks, for example, by using regression analysis. However, in practice, transmission events are unobserved, so instead we consider all possible infection trees consistent with the data. We used a previously described approach to calculate the probability, π_{ij}, that patient *i* was infected by patient *j* for each pair of infected patients in each outbreak based on onset times and the serial interval distribution (the serial interval is the time from onset of symptoms in case *i* to case *j*), without using proximity data. The serial interval distribution tells us the probability of durations of 0,1, 2,…days between onset in a case and onset in secondary cases infected by this case. Given multiple possible sources for a case, we can use knowledge of this distribution to tell us how likely each is to be the true source. Full technical details are described in Wallinga & Teunis (2004) [12]."

2. *As said before, I very much appreciate that the authors provide the R-code. However, the authors should make an effort in explaining the code more clearly, and provide it without any errors, so that others are able to use it. For example:*
- There are some "}" missing
- Some variables are given clear names (such as "whoinfectswhom"), while others are not (such as pkl), please amend.
- Explain the meaning of "w" (page 23 line 16 – page 24 line 20). I assume that this is the serial interval distribution? If so, why are there zero's at w[5], w[7] and w[9]. As I understand the code, the "w" is later overwritten by another function. Please explain.
- Would it be possible to provide the data (or an example dataset), so that people can really re-use the code?

We agree that presentation of the code wasn't as clear as it could have been and have now revised this. We have also corrected some of the annotation to the code which was incorrect

Variable names were chosen to be consistent with those in Wallinga & Teunis (2004) and we think it would improve clarity to keep these names. We have however, improved and corrected the annotation of this code to address the reviewer's concerns about readability.

There are zeros in the serial interval distribution (w[i]) because the primary analysis made use of the emprical serial interval distribution derived from onset dates for the first and second cases in each outbreak. To make this clear we have added the text ", and our primary analysis made use of this empirical serial interval distribution" to the methods. As the comment in the code says , these values get overwritten to perform a sensitivity analysis using a gamma distribution instead. We have revised the code so that by default the code for this sensitivity analysis is commented out.

3. I found the main outcome of the study (Pk) difficult to interpret. I also have the feeling that the equation provided on page 6 is not an adequate description of Pk. But again, this is based on my understanding of Pk from the R-code. Why not write the meaning of Pk in simple words, for example "Pk is the number of times that a transmission occurred between two individuals sharing the same bay (per outbreak)".
Furthermore, I have the feeling that Pk depends on outbreak size, and has a different maximal value for every outbreak and should therefore not be summed over all outbreaks. By looking at Figure 2, values are between 80 and 160, but this is difficult to put in a context as I don't know the outbreak sizes. I somehow expected a value between 0 and 1 (i.e. the percentage of transmissions that occurred between people sharing the same bay) for each outbreak. This would give a more meaningful number and allows for comparison between different outbreaks.

We agree that this wasn't clear and have made the following changes. First, we have improved annotation of the code to make it clear that what is referred to as *Pk* in the main text corresponds to test.statistic in the code where tests.statistic is defined

test.statistic<-sum(realcumulativeDist) # for all outbreaks

We have also added comments to make it clear that this is calculated using the function calcOutbreakDist.

We have also clarified the interpretation of Pk by adding the text:

The value of *P* (and of P_k for individual outbreaks) should be interpretted as a measure of how much transmission occurs between patients in the same bay.

If people in the same bay pose a greater risk of infecting each other this will tend to lead to larger values of the proximity metrics, *P,* and . P_k.

While the reviewer is correct that *Pk* does depend on outbreak size, we are not interested in the absolute values of *P,* only in how the value of *P* calculated with real proximity data compares with the value calculated with randomly generated proximity data (based on a permutation of the bay identities) which will be affected in the same way by outbreak sizes. We have included this in the discussion.

4. As the authors acknowledge in the introduction, it is already known that proximity to a (vomiting) norovirus case increases the infection risk. Therefore, I'm not entirely sure what this study adds to these findings. Since the authors simulate the transmission trees, would it be able to go a step further and try to quantify within bay and between bay transmission? The authors could also look at, for example, the bay reproduction numbers. Furthermore, if the ward types are known for these outbreaks (e.g. paediatric ward) reproduction numbers could be analysed stratified by ward-type.

Although we state in the introduction that proximity is associated with infection, most of the studies are based on single events with prolonged environmental exposure, where an incident has occurred and subsequent people are infected as the come into contact with areas that have been previously contaminated. A previous study shows that patients are more likely to transmit infections compared to staff. Here we add to the weight of this by showing that patients in proximity or sharing bays are those most at risk of contracting the virus where other patients have become symptomatic. Also the health care setting to which we have applied this study is a more complex environment than the environments of the other studies. We feel that this adds to the body of knowledge in order to inform infection control strategy. In the UK there is currently debate about the need to close or not to close bays/wards and whether outbreaks can be managed by simply putting people in bays or moving people from an infected bay to another.

5. There are some issues related to the way the authors construct the serial interval distribution. By looking at the difference in symptom onset between the first two cases of an outbreak has the tendency to over represent short serial intervals. Especially in the case of point source outbreaks, where cases developing symptoms 1 day after the "index case" are very likely related to the same point source and not to the symptoms of the index case. This limitation should be discussed.

We have used more than one method of estimating the serial interval and this is raised in the discussion section. However, we feel that there is a substantial number of outbreaks contributing to the estimation of the serial interval, and given the previously estimated infectiousness of norovirus this serial interval is not felt to be unrealistic.

Minor comments

1. The authors used random mutations of bays to test if proximity is associated with an increased risk of acquiring norovirus. By doing this, did the authors keep the distribution of bay-sizes per ward constant (i.e. was it possible that, by chance, 20 people occupied one bay)?

In the random permutations we kept the number of bays constant within earch ward, and just permuted the ids of the beds of the symptomatic cases. So, yes, the number of bays per ward was constant, and it was not possible that 20 people occupied one bay.

2. Could a short description be given about the type of wards where the norovirus outbreaks where observed and if patients were able to walk around and have contact with people in other bays or wards?

We have added a breakdown of the ward types in the introduction. As information was not collected on the individual patient's illness characteristics (why they were admitted) it isn't possible to comment on how mobile or otherwise patients were

3. Introduction, page 3, lines 50 – 52: be specific here that the aim is to assess the risk of acquiring SYMPTOMATIC norovirus gastroenteritis.

We have added the term symptomatic

4. Page 4, lines 39 – 43: The sentence "In three … of 2007/2008" is unclear to me. Be more specific about the difference in data collection methods and questionnaires between the hospitals.

We have altered the wording to make this clearer. " In three other hospitals the data were downloaded from a database on which infection control specialists had recorded these data items during outbreaks of norovirus occurring in the season of 2007/2008."

5. Page 5: to give the readers an idea about the size of the wards, please provide an average number of bays per ward. Furthermore, what is the average occupancy rate in the bays?

The data were not collected on occupancy rate, so it is not possible to provide this. However, in UK hospitals the occupancy rate is usually over 90% and in some wards greater than 100%. (Ben I'll have to look into the ward sizes and put this in the introduction).

6. Please change "Heinje et al" to "Heijne et al" throughout the paper and the appendix table and figure

We have altered the spelling, apologies for this error.

7. Page 9, line 25: remove "very"

we have changed very to highly

8. Page 9, line 38: please explain the 47% (should this perhaps be 44%, i.e. 65 / 149?)

This is now corrected to 44%

9. Page 10, line 8: I assume that with "Figure 3" the authors mean "Figure 1_supplement". Please amend.

Authors: this is now corrected to figure 1 supplement as suggested.

10. Page 10, line 30: add to the sentence "in a hospital setting"

We have added this in the sentence.

11. Reference 21: change "Tuenis" to "Teunis"

We have altered the spelling, apologies for this error.

12. The functionality of the second column in Table 1 is unclear to me, please change.

We have changed the table to assist with clarity.

13. The first key message is unclear to me.

We have altered the key messaged to assist with clarity.

Reviewer: David Partridge, Consultant Microbiologist, Sheffield Teaching Hospitals, UK.
No competing interests.

Reviewer: Method of diagnosis of norovirus at the included hospitals is not recorded and nor is testing strategy e.g. test just index case or test all cases during outbreak

There is not a single policy within the UK on testing strategy, and hospitals will use their own local policies, often this is dependent upon the circumstances of each outbreak. However, in this study all of the hospitals test a number of patients (not just the index cases) and all use PCR to test for norovirus. We have added a line in the methods to state:
"The hospitals in this study all used Polymerase Chain Reaction (PCR) for detection of norovirus in stool samples."

Correction

Harris JP, Lopman BA, Cooper BS, *et al*. Does spatial proximity drive norovirus transmission during outbreaks in hospitals? *BMJ Open* 2013;3:e003060. One of the authors' affiliations is incorrect. Sarah J O'Brien's affiliation should be number 2 (University of Liverpool), not 4.

BMJ Open 2013;**3**:e003060corr1. doi:10.1136/bmjopen-2013-003060corr1