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

TITLE (PROVISIONAL)	Evaluation of person-level heterogeneity of treatment effects in published multi-person N-of-1 studies: systematic review and re-analysis
AUTHORS	Raman, G; Balk, EM; Lai, Lana; Shi, Jennifer; Chan, Jeffrey; Lutz, Jennifer; Dubois, Robert; Kravitz, Richard; Kent, David

VERSION 1 – REVIEW

REVIEWER	Michael Sobel Columbia University
REVIEW RETURNED	01-Jun-2017

GENERAL COMMENTS	Review of manuscript BMJ Open 2017-017641. This manuscript examines “the extent of evidence for person level HTE’s ” (heterogeneous treatment effects) using multi-subject crossover studies (N of 1 studies). The authors pull together many such studies across a broad range of different treatments and outcomes. They conclude that person level hetero- geneity in treatment effects is very common, also that the studies they examined rarely address this issue. It is well known that in many instances there is treatment heterogeneity, as treatment effects often vary with observed characteristics of subjects. That there is additional variability that will not generally be fully accounted for using observed characteristics is not surprising. Contrary to the introduction, I don’t believe there is much need for a demonstration that treatment effects are often heterogeneous. In addition, as the authors present no equations in the text, I cannot be sure what models they estimated. Even had they presented material to clarify this, they do not spell out the conditions under which the treatment effects estimated for the N of 1 models permit the interpretation as such. In addition, while the text is short (16 pages), the paper with tables, etc. is 91 pages, vey long. But I suppose that is ok for an online journal.

REVIEWER	Ravi Varadhan Johns Hopkins University
REVIEW RETURNED	18-Jun-2017

GENERAL COMMENTS	Systematic Review of N-of-1 Studies for Person-level HTE Major comments: This is paper makes a useful contribution to the literature on HTE. The authors seemed to have conducted a rigorous review and re-analysis of 56 crossover studies with the goal of characterizing the

person-level variation in treatment effect. I don't have any major concerns regarding the quality of the study, but I do have a statistical concern. The authors do not formally describe how they estimated the person-level treatment effects and their standard errors when they re-analyzed the original studies. For example, I randomly selected a couple of studies reported in the Appendix (Figure 1 – Emmanuel 2011; Figure 3 – Haas 2004) and tried to verify the treatment effect estimates and confidence intervals, but could not do so. Therefore, I recommend that the authors provide sufficient details of how they re-analyzed the data from papers including, but not limited to, the regression model that was used to analyze the multi-period crossover data and how time-period was accounted in the models. One or two examples of how they did it would be extremely useful.

I also have another important suggestion: the authors should make it clear that the crossover design is quite limited in the sense that it is only applicable to conditions where the disease process is relatively stable over time, treatment effects are transient, and outcomes vary and are observable over time. In more common settings, e.g., where the interest is on treatments whose effects is cumulative and the outcomes are only observed only once, this design is not applicable and hence person level HTE is not identifiable. Even though they mention this, it should be highlighted more prominently in the Box on Strengths and Limitations of this Study.

Specific comments:

- I suggest that the authors also use the phrase "crossover studies" in the title (it could be added parenthetically), which has been around longer than the more recent "N-of-1" studies.
- There are some discrepancies in the percentages calculated in the Abstract. For example, 9/56 = 16% (not 11%); 2/56 = 3.6% (not 2%); 23/56 = 41% (not 19%). Also, what is "24 data points?"
- On page 8 (lines 32-46), where the authors discuss person-level HTE estimation using fixed-effects inverse variance approach, they should provide a citation for this approach. What do they mean by method of moments estimator for variance? Do they mean simply computing the variance of the estimates of treatment effect?
- In Figure 2 depicting the study flow diagram, the number of articles not meeting the criteria is 417, but the breakup according to different reasons does not add up to 417. Were there other reasons for excluding articles?
- In Appendix Figure 9, what happened to patient 10? Why was patient 14 excluded?
- In Tables 4 and 5, I am bit puzzled by the inference based on I^2 statistic for heterogeneity. In many cases, it is estimated to be 0, and in some of these cases, it is 0 even when the P-value according the traditional LRT for interaction is small (less than 0.05). Why does this happen and what are the implications of this for using I^2 for inferring person-level HTE?
- On page 13 (lines 15-36), the authors discuss the issue of generalizability of patient-level estimates and suggest that Bayesian approaches might be used to combine individual estimates to obtain a population estimate. A more important idea is to use the information from other patients to obtain more "reliable" treatment effect estimates for an individual patient. In other words, a shrinkage estimator which shrinks the individual level estimates towards the overall population mean would be "optimal" in the sense of minimizing mean-squared error (see Henderson et al., "Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research," Heath Services Outcomes Research, 2016).

REVIEWER	Karsten Bruins Slot Oslo University Hospital, Oslo, Norway.
REVIEW RETURNED	06-Aug-2017

GENERAL COMMENTS	The authors present a paper that aimed to summarize the reporting of person-level heterogeneity of treatment in multi-person N-of-1 studies. They also examined evidence for person-level heterogeneity of treatment through re-analyses of available trial data. In all 56 multi-person N-of-1 studies with at least two subjects were identified in a systematic review. A re-analyses of the data from available from 31 of these studies suggested the presence of substantial variation in treatment effects across individuals. Based on their findings the authors conclude that person-level heterogeneity of treatment appears to be fairly common for several conditions/outcomes and large enough to be clinically meaningful. Further, they suggest that improved assessment and reporting of person-level treatment effects in multi-person N-of-1 studies are needed. The paper is generally well-written and comprehensive and provides relevant information. The Discussion section is somewhat lengthy, though, and could be more focused. Further, I missed a more criticial discussion on the generalisability of the findings, also considering that included studies were mainly performed in neurology, rheumatology and psychiatry.

VERSION 1 – AUTHOR RESPONSE

Comments from the Associate Editor:
- The search is old (March 21, 2014) and needs updating.
- The Strengths and Limitations section needs revising to include the limitations of this review.
- Please provide details of how study quality was assessed?

Response: We have updated our search and revised the manuscript.
We have updated the search and revised the manuscript accordingly. We updated our searches from March 2014 through August 17, 2017. We have revised the strengths and limitations sections. While there is a very recent guideline statement for reporting quality of N-of-1, to the best of our knowledge there are no specific guidelines to assess methodological quality on how to conduct N-of-1 trials. Therefore, we conducted an assessment of methodological quality similar to assessing randomized crossover trials. These assessments are also added to the appendix.
Editorial Requirements:
- Please update the search to include more recent literature.
- Please ensure the references section is fully up to date with the relevant literature.

Response: We have updated the search and revised the manuscript references accordingly. We updated our searches from March 2014 through August 17, 2017. Our updated search yielded an additional 2438 citations. Screening of citations identified an additional 34 potentially relevant articles that were retrieved for full-text screening from the updated search. Upon full-text screening, additional six articles met eligibility criteria and two of those provided re-analyzable data.

Reviewer(s)' Comments to Author:
Reviewer: 1
This manuscript examines “the extent of evidence for person level HTE” (heterogeneous treatment effects) using multi-subject crossover studies (N of 1 studies). The authors pull together many such studies across a broad range of different treatments and outcomes. They conclude that person level

heterogeneity in treatment effects is very common, also that the studies they examined rarely address this issue. It is well known that in many instances there is treatment heterogeneity, as treatment effects often vary with observed characteristics of subjects. That there is additional variability that will not generally be fully accounted for using observed characteristics is not surprising. Contrary to the introduction, I don't believe there is much need for a demonstration that treatment effects are often heterogeneous.

In addition, as the authors present no equations in the text, I cannot be sure what models they estimated. Even had they presented material to clarify this, they do not spell out the conditions under which the treatment effects estimated for the N of 1 models permit the interpretation as such. In addition, while the text is short (16 pages), the paper with tables, etc. is 91 pages, very long. But I suppose that is ok for an online journal.

Response: We agree with the reviewer that the presence of individual-level HTE is intuitive and many readers would not need convincing. Nevertheless, we note that at least one distinguished statistician has suggested (in the pages of BMJ) that these effects could be a myth and (rightly) pointed out the lack of empirical evidence (Senn S. BMJ. 2004:966-968). We note too that treatment effects often appear to vary at the group-level (i.e. based on observable characteristics). But even statistically significant HTE is very frequently (probably more often than not) spurious (e.g. statistically significant treatment-by-sex interactions appear in the literature roughly as frequently as one would expect by chance alone [Wallach JD et al. BMJ 2016: 5826]). We have added a sentence and reference to underscore the value of being skeptical of discovered/reported HTE.

We used mixed effect models as described in Zucker et al paper (Zucker DR et al. J Clin Epidemiol. 2010:1312-23) and we have added this citation in the methods section. This article provides detailed equations of linear mixed effect model as well as hierarchical modeling that we conducted in our paper.

In combining individual measurements within a series of N-of-1 trials, we used a linear mixed model. We determined model parameters, characterized use of fixed versus random effect model on the outcome, and included within-patient variance. In the fixed effect model, the regression parameter remained the same for each patient and regression parameters varied across patients in random-effect model. We also derived a similar model with treatment-by-participant interactions. This model allows each patient to have a different treatment effect. The statistical significance of person-level HTE was assessed by a likelihood ratio test comparing the two models. Additionally, we structured covariance matrices within patients. We investigated the model assuming an uncorrelated common variance structure for each patient with different variances across patients. We also fit a hierarchical linear or generalized linear mixed model with a random intercept and a random slope (for the treatment effect) to estimate the average treatment effect across all patients (assuming person-level HTE). The statistical significance of person-level HTE was assessed by a likelihood ratio test comparing the two models.

We have clarified this in the methods (adding the last two sentences from the paragraph above) and also included the STATA code for the models in the appendix.

Reviewer 2:

Systematic Review of N-of-1 Studies for Person-level HTE

Major comments:

This is paper makes a useful contribution to the literature on HTE. The authors seemed to have conducted a rigorous review and re-analysis of 56 crossover studies with the goal of characterizing the person-level variation in treatment effect. I don't have any major concerns regarding the quality of the study, but I do have a statistical concern.

Response: Thank you.

The authors do not formally describe how they estimated the person-level treatment effects and their standard errors when they re-analyzed the original studies. For example, I randomly selected a couple of studies reported in the Appendix (Figure 1 –Emmanuel 2011; Figure 3 – Haas 2004) and tried to verify the treatment effect estimates and confidence intervals, but could not do so. Therefore, I

recommend that the authors provide sufficient details of how they re-analyzed the data from papers including, but not limited to, the regression model that was used to analyze the multiperiod crossover data and how time-period was accounted in the models. One or two examples of how they did it would be extremely useful.
Response: Studies reported data heterogeneously and therefore, we have presented methods that were used to derive the standard errors using the most common examples. For example, Emmanuel 2011, the standard error was estimated using the following equation: SD of intervention (or control) score/square root of intervention days (or control days). Then SE of difference was estimated. For Hass 2004, we used the SE available in Table 4 of the original article. The appendix has been edited to include details on how the SE was derived. Additionally, we have provided statistical codes in the appendix.
We used mixed effect models as described in Zucker et al paper (Zucker DR et al. J Clin Epidemiol. 2010:1312-23) and we have added this citation in the methods section. In addition, we have provided statistical code for couple of examples. As already discussed in our discussion section: only fairly small numbers of patients were observed over a small number of treatment periods, thus the data do not allow the fitting of more complex (and realistic) models, for example models that account for period effects or other effects of time on the outcome.
I also have another important suggestion: the authors should make it clear that the crossover design is quite limited in the sense that it is only applicable to conditions where the disease process is relatively stable over time, treatment effects are transient, and outcomes vary and are observable over time. In more common settings, e.g., where the interest is on treatments whose effects is cumulative and the outcomes are only observed only once, this design is not applicable and hence person level HTE is not identifiable. Even though they mention this, it should be highlighted more prominently in the Box on Strengths and Limitations of this Study.
Response: Thank you; we have added to the Box on Strengths and Limitations of this Study
Specific comments:
I suggest that the authors also use the phrase “crossover studies” in the title (it could be added parenthetically), which has been around longer than the more recent “N-of-1” studies.
Response: Thank you for the suggestion. We have retained the original title because so-called cross over studies are only included in the very rare cases when they presented individual patient level data.
There are some discrepancies in the percentages calculated in the Abstract. For example, 9/56 = 16% (not 11%); 2/56 = 3.6% (not 2%); 23/56 = 41% (not 19%). Also, what is “24 data points?”
Response: We have edited the proportions. One study of total 23 studies reported both re-analyzable person-level outcomes and treatment effects data and was mentioned as 24 data points in the abstract. We have deleted “24 data points” to avoid confusion.
On page 8 (lines 32-46), where the authors discuss person-level HTE estimation using fixed-effects inverse variance approach, they should provide a citation for this approach. What do they mean by method of moments estimator for variance? Do they mean simply computing the variance of the estimates of treatment effect?
Response: We have added suggested references to these in the methods section. We have changed method of moments estimator to DerSimonian and Laird method of moments estimator. We have also added relevant citations to the described method.
In Figure 2 depicting the study flow diagram, the number of articles not meeting the criteria is 417, but the breakup according to different reasons does not add up to 417. Were there other reasons for excluding articles?
Response: We have corrected this error and revised Figure 2.
In Appendix Figure 9, what happened to patient 10? Why was patient 14 excluded?
Response: The patient 10 and 16 in Figure 9 dropped out subjects and had only one observation. Patient 14 data was excluded by output because there were issues in the data in terms of effect size and standard error estimated from the p-value was irreconcilable.

In Tables 4 and 5, I am bit puzzled by the inference based on I^2 statistic for heterogeneity. In many cases, it is estimated to be 0, and in some of these cases, it is 0 even when the P-value according the traditional LRT for interaction is small (less than 0.05). Why does this happen and what are the implications of this for using I^2 for inferring person-level HTE?
Response: We note that significant p-values for the likelihood ratio test always correspond to high I^2 for studies presenting results as treatment effects, and usually also for studies presenting results as outcomes. The few studies in which there is discordance for the LRT and the I^2 are in studies with very poor precision. These studies had small number of subjects that were tested with interventions and had sparse outcomes. We reiterate that only fairly small numbers of patients were observed over a small number of treatment periods. It is well-documented that I^2 statistic for heterogeneity need to be interpreted with caution when there are limited number of events or trials (von Hippel PT. BMC Med Res Methodol. 2015:35). To aid in interpretation, we provided 95% CI for I^2, which show that the I^2 is imprecisely estimated in these examples. We have added this as one of our limitation.
On page 13 (lines 15-36), the authors discuss the issue of generalizability of patient-level estimates and suggest that Bayesian approaches might be used to combine individual estimates to obtain a population estimate. A more important idea is to use the information from other patients to obtain more “reliable” treatment effect estimates for an individual patient. In other words, a shrinkage estimator which shrinks the individual level estimates towards the overall population mean would be “optimal” in the sense of minimizing mean-squared error (Henderson NC et al. Health Serv Outcomes Res Methodol. 2016:213-233).
Response: Thank you. The suggested reference pertains to methods using Bayesian methods to estimate group-level HTE; Similar methods have been used by Zucker et al in their approach analyzing individual effects using multiperson n-of-1 trials. We have cited both these references in the manuscript.

Reviewer: 3

The authors present a paper that aimed to summarize the reporting of person-level heterogeneity of treatment in multi-person N-of-1 studies. They also examined evidence for person-level heterogeneity of treatment through re-analyses of available trial data.
Response: Thank you.

In all 56 multi-person N-of-1 studies with at least two subjects were identified in a systematic review. A re-analyses of the data from available from 31 of these studies suggested the presence of substantial variation in treatment effects across individuals. Based on their findings the authors conclude that person-level heterogeneity of treatment appears to be fairly common for several conditions/outcomes and large enough to be clinically meaningful. Further, they suggest that improved assessment and reporting of person-level treatment effects in multi-person N-of-1 studies are needed. The paper is generally well-written and comprehensive and provides relevant information. The Discussion section is somewhat lengthy, though, and could be more focused. Further, I missed a more critical discussion on the generalisability of the findings, also considering that included studies were mainly performed in neurology, rheumatology and psychiatry.

Response: Thank you; we have added further information regarding applicability of N-of-1 studies. We have added the following in the discussion section “Many conditions are not amenable to the N-of-1 design (e.g. because treatment effects are cumulative or because outcomes are observed only once). Further, even for conditions and treatment that are potentially amenable to this design, many important disease categories lacked published N-of-1 studies.”.”

VERSION 2 – REVIEW

REVIEWER	Ravi Varadhan

	Johns Hopkins University
REVIEW RETURNED	10-Jan-2018

GENERAL COMMENTS	The authors have addressed most of my comments satisfactorily. However, I still have a few remaining issues. 1. Appendix Figure 12: the average treatment effect is reported as -18.823. This cannot be correct, since all the individual effects are much less than -18.8. 2. Some of the Appendix figures do not report the average treatment effect (e.g., Fig 8, Fig 15). This needs to be consistent for all figures. 3. I am still not clear on how the treatment effects were estimated when each individual had multiple treatments over time. On page 6, the following statement is given: "Person-level treatment effect was defined as contrasts of outcomes in individuals on one treatment versus the comparator" How was person-level effect computed for multiple time periods? How was correlation taken into account when computing the standard error for person-level effects? 4. p9: The following sentence does not make sense to me. Can you please clarify? For studies that reported person-level outcomes, we developed a linear model (for continuous outcomes) or generalized linear model (for binary or count outcomes) using the outcome of interest as the response, the intervention(s) as a covariate; indicator variables for different study participants were derived.

VERSION 2 – AUTHOR RESPONSE

1. Appendix Figure 12: the average treatment effect is reported as -18.823. This cannot be correct, since all the individual effects are much less than -18.8.
Figures 12 and 13 were swapped and we have corrected this error.

2. Some of the Appendix figures do not report the average treatment effect (e.g., Fig 8, Fig 15). This needs to be consistent for all figures.
All figures 1-16 in the Appendix are now presented with average treatment effect.

3. I am still not clear on how the treatment effects were estimated when each individual had multiple treatments over time. On page 6, the following statement is given:
"Person-level treatment effect was defined as contrasts of outcomes in individuals on one treatment versus the comparator"
How was person-level effect computed for multiple time periods? How was correlation taken into account when computing the standard error for person-level effects?

As described in our prior response, we applied methods described in Zucker et al paper.1 This study was one of the largest in our sample, with 58 patients and 6 crossover periods. Trial period did not improve model fit, nor did adding an autoregressive term for correlation among time periods or additional variance component for separate variances by treatment group. We clarified this in the methods section:

“For modeling within-patient variance, we used a common variance with an uncorrelated covariance structure, as was used in a prior n-of-1 study.1 Person-level treatment effect was assumed to be

equal across time-periods. For the treatment effect, we used more than one random slope when >2 treatments were compared."
We also note the following sentence in our limitations section: "data limitations precluded the use of more complex models, for example models that account for period effects or other effects of time on the outcome."

4. p9: The following sentence does not make sense to me. Can you please clarify?
For studies that reported person-level outcomes, we developed a linear model (for continuous outcomes) or generalized linear model (for binary or count outcomes) using the outcome of interest as the response, the intervention(s) as a covariate; indicator variables for different study participants were derived.

We have edited this sentence as follows: "For studies that reported person-level outcomes, we developed a linear model (for continuous outcomes) or generalized linear model (for binary or count outcomes) using the outcome of interest as the response, the intervention(s) as a covariate; and indicator variables for different study participants."

Reference List

(1) Zucker DR, Ruthazer R, Schmid CH. Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. J Clin Epidemiol. 2010;63(12):1312-23. PubMed PMID: PM:20863658. PubMed Central PMCID: PMC2963698

VERSION 3 – REVIEW

REVIEWER	Ravi Varadhan Johns Hopkins University, USA
REVIEW RETURNED	21-Mar-2018

GENERAL COMMENTS	I have a minor comment: there are a couple of references which seems to have been incorrectly cited. Please check them. #13 (WG Cochran's paper does not discuss I^2 statistic; you should cite Higgins and Thompson 2002) #4 (this is not the correct reference for describing the linear model)