This is a point-by-point response to [reviewer one of v1](https://f1000research.com/articles/6-1995/v1). This is *work in progress* and is drafted solely by me (still needs approval of other authors)

The paper by Hartgerink et al. sought to determine whether or not there is evidence for data manipulation in the DECREASE-I and Decrease-IV trials. First, the authors reproduced the results of a previous meta-analysis from Bouri et al., using raw event data to calculate the ln(relative risk) and its standard error. Next, they modeled the probability that the results from the DECREASE trials could have occurred from the effect distribution of the non-DECREASE trials. Finally, by creating simulated datasets that were subsequently manipulated they determined the probability that the likelihood of the manipulated data (DECREASE meta-analytic effect distribution) is larger than the likelihood of the genuine data (non-DECREASE meta-analytic effect distribution).

Even though the motivating case for the paper is the questioned reliability of the data within the DECREASE trials that tries to answer a very concrete clinical question (should we use perioperatively use beta blockers to reduce mortality?), the paper ultimately tries to answer a question that is much broader: can we use forensics statistics even when individual data records are not available. In our opinion, the authors succeeded in part to answer that underlying question.

Thank you :)

General Comments:

Some circular reasoning in step 2, In step 2 the authors provide clear evidence that the results of the DECREASE trials are quite unlikely, however, they employ some circular reasoning in their argumentation. Because we know the DECREASE values are extreme, when a p-value is calculated for the DECREASE effect is calculated based off of the non-DECREASE effect distribution it will of course be very small. The results of this approach are methodically sound but suffer from circular reasoning and do not seem to add much to the goal of determining whether data manipulation occurred. Although this step might be useful to help ease the reader into the following steps, the limited added value of these analyses should be recognized.

We recognize the circularity and adapt the text to indicate that is a quantification of the suspicion that the data points are extreme. This quantification allows us to better assess *how* extreme these data points are.

Robustness of approach In Step 3, the combination of adding simulated datasets with a likelihood calculation is an interesting approach to try to determine data manipulation when only summary statistics are available. In the second paragraph of the methods, when describing the inversion method, the authors say “We assumed that if data are manipulated, each data point is manipulated in the same way and to the same extent.” We have a feeling that this is a very strong assumption that is likely not to hold true and might drive the results. An exploration of the robustness, in text or analyses, could help strengthen the argument for the approach the authors apply. Additionally, since the DECREASE-IV trials have approximately 10x the sample size as DECREASE-I we expected a reduction in the confidence interval, yet the confidence interval still spans almost the entire range. Further deliberations on this phenomenon would help the reader understand the robustness of the applied technique.

We agree this is a strong assumption and recognize it as a limitation in the second paragraph of our discussion section. Because we are building these models solely on summary statistics, we have to compensate the lack of information with assumptions. For that same lack of information, it is difficult to assess the robustness simply because of the wide range of potential ways of fabricating data. We inspected the integrity reports for information on how data were fabricated but found insufficient indications to help in our modeling.

With respect to the confidence intervals, this could very well be a floor effect where the information difference is so minimal the difference also does not really show in the confidence interval range. We updated a sentence in the discussion to clarify this: "Nonetheless, much residual uncertainty remains and simply less information is available in summary results when compared to raw data, as indicated by the large range of the confidence intervals in Step 3"

Interpretation of results. In general, the results show that the data simulation does not give a lot of precision when it comes to pinpointing the number of manipulated datapoints. In some sense, and as also described by the authors, this is also not needed as the lower boundary of the confidence interval of manipulated datapoints should always include 0. In fact, the authors claim that “there is a high degree of confidence that data manipulation did occur based on the difference of the trial results alone”. We do follow that reasoning, but do not fully understand why the authors adhere so much to the difference between the results of the two trials. There does not seem to be a material difference in the 95% confidence intervals between DECREASE-I and DECREASE-IV since in both cases the confidence interval spans essentially the entire range of possible outcomes. Nonetheless, the authors seem to put a lot more emphasis on the analyses of DECREASE IV with its higher sample size, in part because the results show that the lower boundary of the confidence interval does not span the 0 but ranges from 3-1066 or 10-1066, depending on the method used. Given the unknown robustness of this approach, we wonder whether the implied relevance of the small difference between the results of the DECREASE I and DECREASE IV is justified.

We understand the reviewer's hesitation and hope our adjustment to the previous comment helps contextualize the findings a bit better. We note that the dichotomous question of some form of manipulation or no manipulation is more generic than estimating the exact number. Hence, there is more confidence in making that dichotomous statement than making more specific statements.

Conclusion. In the discussion section, we take issue with the sentence beginning “Our results indicate that the results from the DECREASE trials ... at least some of the data were manipulated”. The word manipulated implies an intention to manipulate data, yet there are numerous ways for data to become incorrectly classified without overt and intended manipulation. Although the authors show it is likely that something altered the data, we believe that strong claims about the provenance of that change based on the analyses provided in this paper should be avoided. However, we agree with the authors that DECREASE I and DECREASE IV should be regarded as "highly problematic and untrustworthy when assessing the effect of beta blockade in perioperative mortality", but that is only when we place the results from this paper in the context of the reports on scientific integrity of the DECREASE trials.

We explicitly refrained from using the word fabricated here because it implies intent in a legal way. Upon reading the reviewer's comment, we considered alternatives such as "altered", "misprocessed", but given the conclusions of the integrity reports that preceded this study, we also find it a misrepresentation to underplay the results of those reports.

Next to these general comments, we also have some specific suggestions or questions, mainly to help increase the readability in a potential second version of the manuscript :

It would be helpful to have a forest plot of Table 1, to better visualize the results and compare DECREASE vs. non-DECREASE.

We did generate this forest plot at some point, and saw exactly the same one as in Bouri et al. (2014). We explicitly did not include a forest plot for that reason.

In Table 2, the authors could report the relative risk (RR) instead of the log(RR) to improve the table’s readability and help the reader understand the magnitude of the different effects depicted. We would also be interested whether some more general information on this topic, especially regarding the non-DECREASE trials to put the DECREASE trials and their data into perspective. (e.g. how many non decrease trials were registered but not published; or whether there is any other evidence or suggestion for non-publication of relevant data that could alter some of the conclusions in this paper.)

We could report those numbers, but we decided to stick to the log(RR) because that is closer to the output and maintains the symmetry of the confidence intervals (asymmetric CIs might confuse the reader too). With respect to investigating how many trials had been registered, we agree it is an interesting research question on its own and consider a superficial investigation in the current paper to contextualize the findings not proportional to the complexity of the issue (e.g., a review of trial registration would be necessary given the first DECREASE trial took place in the late 90s). Hence, we do not take up this suggestion here.

A minor point, re-label the subscripts of the Likelihood function as DECREASE and non-DECREASE to avoid the loaded terminology of manipulated and genuine which seem to presuppose the conclusion.

Considering this investigation was based on the integrity reports that preceded it, we consider it reasonable to use these terms without presupposing the conclusion. We also explicitly state that these labels operate under *the assumption* that the DECREASE trials were fabricated (in step 3; in step 2 that assumption is explicitly reversed).

We suggest rewording the second paragraph under the results section, and specifically address the sentence that begins “The relatively minor difference…”. This sentence first references “estimates”, but it is unclear what these estimates are referring to.

Adjusted for clarity to "The relatively minor difference between the point- and distribution estimates indicates "

In Figure 2 please address why, in the bottom lower panel, the confidence interval for the point estimate is larger than the distribution estimate. Because the distribution estimate has more error in it, we were expecting that the confidence interval of the distribution estimate would be larger than that of the point estimate.

The reviewers are correct in finding this odd. We also found this odd and checked the code multiple times to find any mistakes. We invite anyone to take a look and correct us if we're wrong, but we are unable to find the mistake if there is one. Notably, it is the same code that gives the results for DECREASE-I that do follow the notion that the point estimates are more precise. We added a sentence to make this explicit to the reader: "We note that, for reasons unknown to us, the confidence interval of the point estimates are more uncertain than those of the distribution estimates, whereas similar code for DECREASE-I provides the reverse (the code is available for review at https://osf.io/4zs29/)."