PEER REVIEW HISTORY

BMJ Medicine publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Vibration of effects in more than sixteen thousand pooled analyses of individual participant data from twelve randomized controlled trials comparing canagliflozin and placebo for type-2 diabetes mellitus
AUTHORS	Gouraud, Henri; Wallach, Joshua D.; Boussageon, Rémy; Ross, Joseph S.; Naudet, Florian

VERSION 1 - REVIEW

REVIEWER 1	**Hoffmann, Sabine** Ludwig Maximilian University Munich Institute of Medical Information Processing Biometrics and Epidemiology. Competing Interest: None
REVIEW RETURNED	02-Mar-2022

GENERAL COMMENTS	In this manuscript, the authors apply the vibration of effects framework to individual patient data (IPD) from twelve randomized controlled trials comparing caagliflozin versus placebo for type-2 diabetes mellitus. The paper is well written and I want to commend the authors for addressing an important issue and for providing a clearly written pre-registered protocol, a statistical analysis plan, a data management report and a (for the most part) well-documented code for their analyses. In my view, this is a worthy contribution, but I think that there are some issues that should to be further specified in the manuscript. - In my view, the protocol and the statistical analysis plan provide some helpful background information that might escape a reader who only focuses on the paper. In particular, both documents mention that “There is still a tension in this field about the clinical value of the drugs that reduce chronic hyperglycemia. While there is no doubt about efficacy of these drugs on the surrogate marker of HBA1C levels, there is still a heated debate about their impact on clinical outcomes including cardiovascular one.” Why did the authors choose to not include this information in the manuscript? From my perspective, it makes the motivation and the relevance of the work more clear. In particular, it would make it more clear why the hypothesis was that VoE would not be observed for HbA1c while it would be for both MACEs and SAEs. - Throughout the manuscript and the statistical analysis plan, the authors refer to “relative end date of follow-up”, “relative first day of treatment”, “relative day of collection” and “relative day of start of the SAE”, but I was missing a more detailed definition of these quantities. If these are relative quantities, it seems as if the authors put them in relation to something, but it is not clear to what. Could the authors provide more information concerning the definition of these quantities? - I was missing a more detailed motivation for choosing the timepoints at 12, 18, 26 and 52 weeks as the most relevant

	methodological choice needing an investigation through the vibration of effects framework. In Palpacuer et al. (2019), the authors considered different inclusion/exclusion criteria, for instance based on medical condition, somatic comorbidity, psychological support and treatment duration etc. As mentioned by the authors in the discussion on page 17, line 8, the vibration of effects might be influenced by many other methodological choices. In my view, it would for instance have made sense to explore the vibration of effects concerning the imputation of missing values: For instance, the authors chose to impute missing values for HBA1c using the ``last observation carried forward'' method, but other imputation strategies could have been sensible, right? Similarly, it is not entirely clear to me how the authors determined time to occurrence of the first MACE and time to occurrence of the first SAE. In the statistical analysis plan on page 10, the authors describe that, depending on the study, they either "consider as day of death the maximal relative day between 1/ start day and 2/ end day of the event and 3/ end day of follow-up" or they "consider as relative day of death the start day of the event and control the matching with the end of follow-up relative day". For me it is not entirely clear why the authors chose these two strategies, but there could have been other strategies to determine the day of death, right? My main point is that conducting a large number of analysis (here defined by every possible combination of trials * four different time points) gives the impression that the results are robust to alternative analysis strategies, but if only one methodological choice out of many possible choices is considered, the results may substantially underestimate the vibration of effects that might have resulted from all researcher degrees of freedom in the analysis. I am not suggesting that the authors should deviate from their pre-specified protocol, I am only trying to understand why the authors chose the two specific methodological choices among the large number of potential methodological choices. - Somewhat related to the last point, I did not fully understand why the authors considered MACEs and SAE at different timepoints. In the description of the study outcomes on page 10, line 20, it seems as if the authors only considered the four different timepoints for HBA1c, but not for MACE and SAE (it says:"We explored VoE for 3 different outcomes: 1) HBA1c difference from baseline (data was extracted at baseline and at weeks 12, 18, 26, and 52), 2) time to occurrence of the first MACE, and 3) time to occurrence of the first SAE"), but on page 11 line 5 and page 12 line 35, it becomes more clear that the computations for all three outcomes were performed for the four different time points. Does it make sense to restrict the analysis of time-to-event outcomes to different timepoints? If the proportional hazards assumption holds (i.e. if there are no time-varying treatment effects), it seems to me that using different timepoints will only change the percentage of censoring and thereby make the estimates more precise, but they should not systematically change the hazard ratios whereas it is more reasonable to assume that one has to compare HBA1c difference at the same timepoint across different studies. - Why did the authors choose to decide on a fixed or a random effects model based on a two-stage meta-analysis instead of based on a one-stage IPD meta-analysis? It seems a little counterintuitive to me to make this choice based on a two-stage meta-analysis rather than in the IPD meta-analysis because the two-stage analysis might imply an unnecessary loss of information and there should be some way of deciding based on the variance of the random effect in a one-stage IPD meta-analysis, right?

	Minor points: - It should read SRMA instead of SMRA on line 20, 23 and 25 on page 6.

REVIEWER 2	**Patel, Shirag** Harvard University. Competing Interest: None
REVIEW RETURNED	22-Mar-2022

GENERAL COMMENTS	This study by Gourard and colleagues explores use of massive sensitivity analysis, dubbed vibration of effects (VoE), a type of "multiverse analysis", in pooled meta-analysis of randomized trials. If this reviewer was asked his/her prior bias before observing these results, it would be that pooled RCTs of high quality would have few opportunities to exhibit VoE, but these authors in a very important result show that the, in fact, do (albeit rarely). My comments and critique are below: 1.) A notable strength is the risk of bias estimate 2.) The authors test only inclusion and time of followup as critical parameters of VoE, but many others can be tested as the authors point out, such as subgroup analysis. How prevalent are the other modifications in study designs? Why were the other sources not examined? 3.) Besides pre-specification, how would the authors recommend use of VoE in practice when evaluating pooled RCTs? Should it be used? 4.) Can the authors report the median and the IQR for the estimates? It seems as though the associations that contribute to the "Janus Effect" may be rare. If rare, is this really a threat to conclusions that are made from pooled analysis? Are there characteristics of individual studies that are pooled that could contribute to VoE? 5.) It might be useful for the readers to describe how estimating VoE over study inclusion is different (or similar to) than assessing heterogeneity or publication bias tests, such as Egger's test: suppose we just had summary statistics - could we have attained a similar profile of "heterogeneity" between studies. A pleasure to review your work

REVIEWER 3	**Riley, Richard** Keele University, School of Medicine. Competing Interest: None
REVIEW RETURNED	02-Apr-2022

GENERAL COMMENTS	Thank you for the opportunity to assess this paper for potential publication in BMJ Medicine. It raises a lot of interesting points and will generate much debate and intrigue. Many IPD meta-analyses are done on a subset of the IPD that can be obtained, and it is hard to predict which trials will provide their IPD, and indeed researchers may deliberately only select or ask for IPD from a subset of all existing trials. As the paper shows, this can have an impact on the results. However, I have a number of points to be considered going forward, which I hope the authors can take on board. 1) A major comment is that the paper needs to better distinguish between the issue of bias and the issue of sampling variability. The

	paper shows well that the results are variable depending on the choose of trials and endpoints, and that sometimes statistical significance and even direction of pooled estimates change … but this is to be expected based on sampling variability. Fewer trials or different sets of trials will lead to variability in the meta-analysis estimates. Is this a concern? No, it is to be expected, and indeed sampling variability is an underlying premise of the design and analysis of any research. Crucially, it does not mean that anything is wrong, per se. Yes, the choice of trials will impact the findings in terms of estimates and confidence intervals (and significance in terms of p-values), but – as long as the included trials are a random sample of all trials (or representative of the populations of interest for all trials, especially in terms of any effect modifiers) – they should be unbiased and appropriate. It is a bit like the issue of sequential analysis of trials entering into a meta-analysis, and how there is variability, but this decreases as more evidence is added. A bigger concern is when there is selection or availability bias in the IPD used in the meta-analysis (See for example Ahmed et al. https://www.bmj.com/content/344/bmj.d7762 and Chapter 9 of Riley RD, et al. IPD Meta-Analysis: A Handbook for Healthcare Research. Wiley, Chichester; 2021) – as then the vibration of effects is not just due to sampling variability but also bias. So actually the framework for the paper should be about making sure the IPD entered into pooled analyses are of high quality and representative of the target population – and not a biased selection of the available evidence. 2) Given this, I think the recommendation that “Our findings suggest that results from pooled analyses should be used cautiously, and only when based on pre-specified protocols” is not helpful or misleading … too broad a statement. Rather, the focus should be on obtaining IPD from studies that are representative of the target population of interest and high quality, in order to best estimate the estimands of interest. I agree with the message that pooled analyses of IPD cannot be simply assumed the gold-standard, and need to be critically appraised still – this is in line with previous papers like Ahmed. But to give a blanket statement that all pooled analyses should be viewed cautiously is not the message – rather, it is that they should be criticially appraised, as in any research study. 3) Another point that is not discussed is that availability of IPD allows the quality of studies to be better assessed, and for analyses to be standardised. Some studies may be ignored specifically because they are identified to be of low quality, or not record key (adjustment) factors needed for appropriately estimating the estimand(s) of interest. Therefore, when done well, there may be less variability in the selected studies for the analysis and less potential for vibration of effects. 4) I do not agree with the approach of modelling heterogeneity or estimating it. Firstly, I-squared does not measure heterogeneity directly, and so using thresholds of 25%, 50% etc to define magnitude of heterogeneity is misleading and wrong. I-squared may be very high when heterogeneity is low and vice-versa. See this paper by Rucker et al. https://pubmed.ncbi.nlm.nih.gov/19036172/. Secondly, using I-squared to determine the analysis approach is not appropriate, and the authors would be better to pre-define whether heterogeneity is expected and choose a priori their modelling approach. Given that heterogeneity is a concern in their examples, and that variability across analyses will also be impacted upon heterogeneity if heterogeneity exists and it is ignored, I strongly

	recommend using a random effects model by default. 5) Vibration of effects will also be influenced by other methodology choices, including the approach to estimation of the random effects model. REML is recommended for a two-stage analysis, and for one-stage models with continuous outcomes. Also, confidence intervals need to account for uncertainty in the estimation of variances, for example using the Hartung-Knapp approach (in the second stage) or the Kenward Roger approach (in 1-stage). Variability may be less when CIs and p-values are derived appropriately. For example, CIs will be artificially narrower when there are fewer included trials and the CIs do not account for this. See for example: https://onlinelibrary.wiley.com/doi/full/10.1002/jrsm.1316 and https://onlinelibrary.wiley.com/doi/full/10.1002/sim.8555 6) One-stage models need great care when fitting, in order to make sure clustering is accounted for and to ensure that random effect variances are estimated without bias. For the former, a random intercept or stratified intercept by study is needed – and for the latter, centering of covariates and treatment groups can reduce downward bias in between-study variances. E.g. see https://onlinelibrary.wiley.com/doi/full/10.1002/sim.8555 - but there are no details in the paper of what has been done in this regard for the one-stage model specification (in addition to estimation and CI derivation). 7) It's not clear why two-stage analyses were done for some and one-stage for other investigations. 8) Isn't a Janus effect expected? That is, across all permutations of included trials, and as number of trials and sample size reduces, I would expect variability to increase and eventually lead to wide differences in estimates across the empirical distribution – such that there will often be a Janus effect, with results in opposite directions. I hope these comments are helpful to the authors. I recognise I have raised a number of critical issues that require (considerable) re-analyses and re-interpretation. Nonetheless, I look forward to reading the revision and reading the authors' response.

VERSION 1 – AUTHOR RESPONSE

We would like to thank all the 3 reviewers for their insightful comments and hope that they will be satisfied with the changes that were made.

Reviewer: 1

1. In this manuscript, the authors apply the vibration of effects framework to individual patient data (IPD) from twelve randomized controlled trials comparing caagliflozin versus placebo for type-2 diabetes mellitus. The paper is well written and I want to commend the authors for addressing

an important issue and for providing a clearly written pre-registered protocol, a statistical analysis plan, a data management report and a (for the most part) well-documented code for their analyses. In my view, this is a worthy contribution, but I think that there are some issues that should to be further specified in the manuscript.

Thank you for this kind comment and are happy that you enjoyed the manuscript. We have tried to clarify all the points that were made.

2. In my view, the protocol and the statistical analysis plan provide some helpful background information that might escape a reader who only focuses on the paper. In particular, both documents mention that “There is still a tension in this field about the clinical value of the drugs that reduce chronic hyperglycemia. While there is no doubt about efficacy of these drugs on the surrogate marker of HBA1C levels, there is still a heated debate about their impact on clinical outcomes including cardiovascular one.” Why did the authors choose to not include this information in the manuscript? From my perspective, it makes the motivation and the relevance of the work more clear. In particular, it would make it more clear why the hypothesis was that VoE would not be observed for HbA1c while it would be for both MACEs and SAEs.

Thank you, we have added this information in our manuscript. We hope that it gives now a better taste of clinical relevance for the readers.

“To gain fuller understanding of the clinical implications of the different combinations of trials and repeated endpoint measures when conducting pooled analyses of IPD from RCTs, we explored the VoE in pooled analyses in the field of type-2 diabetes mellitus. mellitus. To date, the clinical value of drugs that reduce chronic hyperglycemia, as measured by serum HbA1c, remains uncertain because of less clear effects on clinical outcomes, such as cardiovascular events (1). Canagliflozin is a drug used for glycaemic control among patients with type-2 diabetes and it has been consistently found to reduce haemoglobin A1c (HBA1c), a surrogate measure of diabetes control (11–13), and for which there is evidence for cardiovascular event reduction among patients at high cardiovascular risk (14).”

3. Throughout the manuscript and the statistical analysis plan, the authors refer to “relative end date of follow-up”, “relative first day of treatment”, “relative day of collection” and “relative day of start of the SAE”, but I was missing a more detailed definition of these quantities. If these are relative quantities, it seems as if the authors put them in relation to something, but it is not clear to what. Could the authors provide more information concerning the definition of these quantities?

Thank you, because the data set is partially anonymised, we don't know the exact date for each visit but we only have dates relative to the date of inclusion. We have added this precision in the text and in the appendix.

In the text :

“1) subject identification number, 2) treatment and dose received, 3) study in which the patient was included, 4) relative end date of follow-up (all original dates relatives to individuals subjects have been removed from the dataset, only relatives days to the inclusion are provided), 5) any deviation from trial protocol and, 6) study outcomes listed below.”

In the appendix :

“In most studies, (except studies 2, 7 and 10) two rows describe the exposition for some or all subjects. The selected row will be the row that contain the relative (all originales dates relatives to individuals subjects have been removed from the dataset, only relatives days to the inclusion are provide) first day of treatment (EXSTDY=1).”

4. I was missing a more detailed motivation for choosing the timepoints at 12, 18, 26 and 52 weeks as the most relevant methodological choice needing an investigation through the vibration of effects framework. In Palpacuer et al. (2019), the authors considered different inclusion/exclusion criteria, for instance based on medical condition, somatic comorbidity, psychological support and treatment duration etc. As mentioned by the authors in the discussion on page 17, line 8, the vibration of effects might be influenced by many other methodological choices. In my view, it would for instance have made sense to explore the vibration of effects concerning the imputation of missing values: For instance, the authors chose

to impute missing values for HBA1c using the ``last observation carried forward'' method, but other imputation strategies could have been sensible, right? Similarly, it is not entirely clear to me how the authors determined time to occurrence of the first MACE and time to occurrence of the first SAE. In the statistical analysis plan on page 10, the authors describe that, depending on the study, they either “consider as day of death the maximal relative day between 1/ start day and 2/ end day of the event and 3/ end day of follow-up” or they “consider as relative day of death the start day of the event and control the matching with the end of follow-up relative day”. For me it is not entirely clear why the authors chose these two strategies, but there could have been other strategies to determine the day of death, right? My main point is that conducting a large number of analysis (here defined by every possible combination of trials * four different time points) gives the impression that the results are robust to alternative analysis strategies, but if only one methodological choice out of many possible choices is considered, the results may substantially underestimate the vibration of effects that might have resulted from all researcher degrees of freedom in the analysis. I am not suggesting that the authors should deviate from their pre-specified protocol, I am only trying to understand why the authors chose the two specific methodological choices among the large number of potential methodological choices.

Thank you for this important clarification. We fully agree with the point raised. We only explored a small part of VoE with our design. We agree that 1/ imputation of missing values and 2/ adjudication of events are possible sources of vibration of effect. We already pointed this out in the limitation section (in bold in the following paragraph). And indeed there are many other sources (e.g. as you will see in our response to reviewer 3, there was also the possibility of VoE because of model specifications, ii.e. we made the required change and the results were slightly different). We have clarified this in the discussion section.

We also clarified the choices we made. From our point of view, it is interesting in those exercises of VoE to make choices that are plausible and relevant concerning the specific situation explored. For instance, concerning Palpacuer et al. (2) [about indirect comparisons], we performed VoE regarding study inclusion criteria because we had many overlapping meta-analyses in the litterature with slightly different inclusion criteria and contradicting results. We are conducting a somewhat similar study [about head to head meta-analyses on aggregated data] in the treatment of smoking cessation by acupuncture

(3). In the case of such a complex intervention, there is a large heterogeneity in terms of treatment modalities across studies and even in terms of control groups, etc.. Therefore it made sense to insist on exploring VoE due to different types of PICOS. Again, in the literature there are also many overlapping meta-analyses on the same topic with some variations in the choice of PICOS.

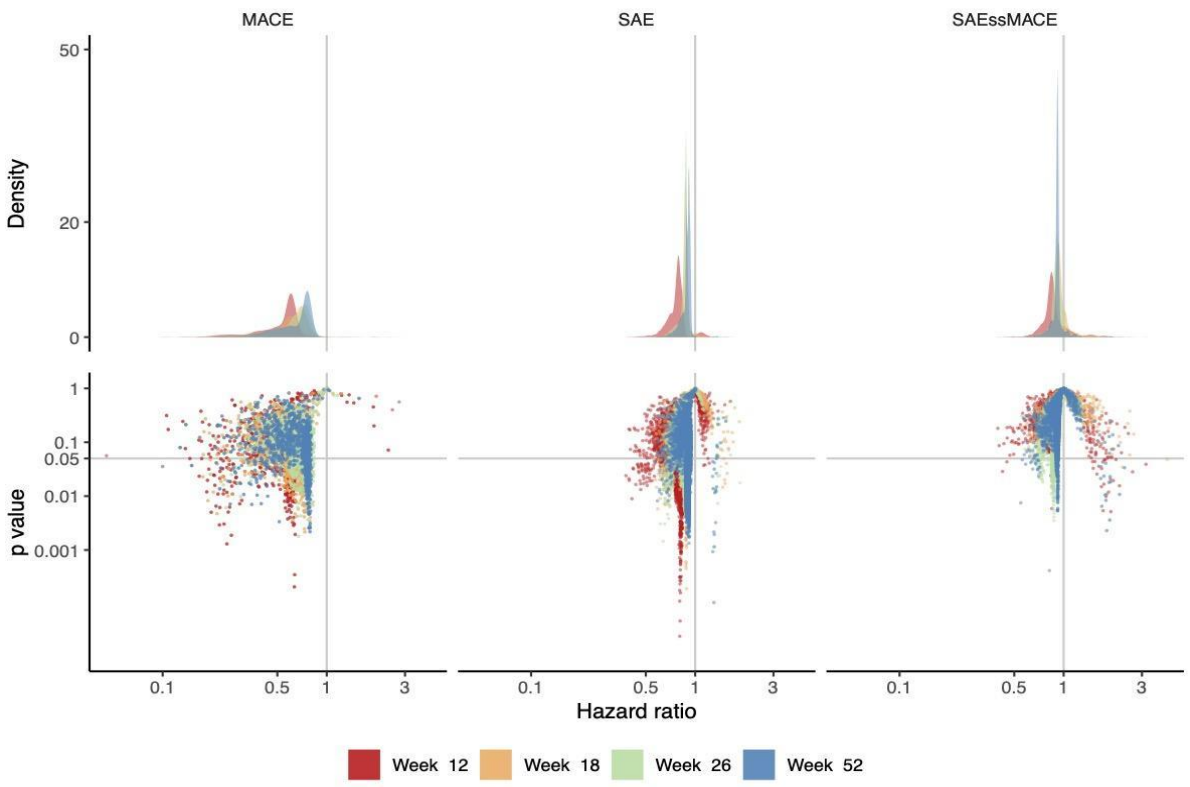
Similarly, in this new case study [this time about IPD meta-analyses and pooled analyses], it was interesting to explore a question with a certain relevance. As stated in our introduction, the literature points toward many pooled analyses run by the pharmaceutical industry (e.g. for duloxetine) and with sometimes a very selected sample of studies. We wanted to explore this very specific point because we were afraid that those pooled analyses offered a possibility for being hijacked. Our methodological choice to explore all possible combinations of studies was therefore driven by this very specific state of the art. Regarding the second point -dates- one must note that the included studies had quite different duration. Therefore it was reasonable to take this constraint into account and to rely both on combinations and on dates. We have added these clarifications in the text.

“The findings from our case study may not be generalizable across all fields. Firstly, we selected an example involving a large number of trials and therefore a large number of possible combinations of RCTs. Identifying VoE in fields with only a few trials could be more challenging. Secondly, we only considered two methodological choices, i.e. study inclusion and timing of endpoints. VoE can be influenced by many other characteristics, including subgroup analyses, different definitions of outcomes (e.g. a different construction of MACE), different groupings of dosages, and different analytical strategies (e.g. choice of one-stage vs two-stage IPD, model specification, or different handling of missing data). Therefore, our results may in fact underestimate the VoE that might have resulted from many other researcher degrees of freedom in the analysis. However, our evaluation concerned 16332 analyses and provides an idea of the impact of different trial combinations and repeated endpoint measures. Our choice to focus such combinations of studies was driven by the fact that many pooled analyses are run with the risk of manipulating the results, by selecting favorable combinations of studies (5). Still subgroup analyses may be very frequently conducted in pooled analyses (e.g. for duloxetine (5)), a consideration that deserves attention in future research. Thirdly, we relied only on studies available on the YODA platform at the time of our request. All these studies were sponsored by Janssen and we considered that this specific subset of studies was quite representative of the sample a given

sponsor would use when conducting a series of secondary analyses. Relying on IPD from such a homogeneous subset of studies allowed the quality of studies to be better assessed, and for analyses to be standardised. Therefore, there may be less variability in the selected studies for the analysis and less potential for VoE. We did not conduct a systematic search for other studies, for example, those conducted in an academic context. Whether the authors of academic studies would have shared the trial IPD necessary to conduct our VoE analyses is uncertain and including studies of this type could have added heterogeneity and VoE."

5. Somewhat related to the last point, I did not fully understand why the authors considered MACEs and SAE at different timepoints. In the description of the study outcomes on page 10, line 20, it seems as if the authors only considered the four different timepoints for HBA1c, but not for MACE and SAE (it says:"We explored VoE for 3 different outcomes: 1) HBA1c difference from baseline (data was extracted at baseline and at weeks 12, 18, 26, and 52), 2) time to occurrence of the first MACE, and 3) time to occurrence of the first SAE"), but on page 11 line 5 and page 12 line 35, it becomes more clear that the computations for all three outcomes were performed for the four different time points. Does it make sense to restrict the analysis of time-to-event outcomes to different timepoints? If the proportional hazards assumption holds (i.e. if there are no time-varying treatment effects), it seems to me that using different timepoints will only change the percentage of censoring and thereby make the estimates more precise, but they should not systematically change the hazard ratios whereas it is more reasonable to assume that one has to compare HBA1c difference at the same timepoint across different studies.

Thank you. We have clarified the text as indeed MACE and SAEs were considered at different timepoints. Again, this choice was made in order to take into account the fact that pooled analyses may include studies of different duration. We have plotted an additional figure to explore this for the purpose of peer review, but did not change the text in order to stick to our protocol. As you see in the figure, the choice of timepoints add some VoE. This is not surprising as part of the VoE could be due to random sampling. See our comment to reviewer 2 and 3.



6. Why did the authors choose to decide on a fixed or a random effects model based on a two-stage meta-analysis instead of based on a one-stage IPD meta-analysis? It seems a little counterintuitive to me to make this choice based on a two-stage meta-analysis rather than in the IPD meta-analysis because the two-stage analysis might imply an unnecessary loss of information and there should be some way of deciding based on the variance of the random effect in a one-stage IPD meta-analysis, right?

We agree that our approach was suboptimal. We have considerably edited this part in line with the comment made by the statistical reviewer (please our response to reviewer 3). Although we have performed new analyses, the results are almost the same (indeed there was some Voe due to model specification). We hope that it is now clarified.

7. Minor points: It should read SRMA instead of SMRA on line 20, 23 and 25 on page 6.

Thank you, we made this change.

“However, in recent years, concerns have been raised about the numbers of overlapping (1) and sometimes conflicting SRMAs/pooled analyses (2). In particular, in certain fields, the exponential

increase in the numbers of SRMAs/pooled analyses has resulted in nearly 1 new SRMA/pooled analysis for every new RCT (3)."

Reviewer: 2

This study by Gouraud and colleagues explores use of massive sensitivity analysis, dubbed vibration of effects (VoE), a type of "multiverse analysis", in pooled meta-analysis of randomized trials. If this reviewer was asked his/her prior bias before observing these results, it would be that pooled RCTs of high quality would have few opportunities to exhibit VoE, but these authors in a very important result show that the, in fact, do (albeit rarely).

Thank you for this kind comment, especially as your seminal paper on VoE was very insightful before deciding to investigate the theme of VoE. We have tried to clarify all the points that are mentioned below.

My comments and critique are below:

1.) A notable strength is the risk of bias estimate

Thank you. No specific edits were necessary.

2.) The authors test only inclusion and time of followup as critical parameters of VoE, but many others can be tested as the authors point out, such as subgroup analysis. How prevalent are the other modifications in study designs? Why were the other sources not examined?

We are not aware of systematic assessment of subgroups in IPDs. For duloxetine (4), we can count at least 8/43 pooled analyses that explore subgroups. We have added a few words about this. We have extensively answered to this point in our response to the first reviewer (Cf.). We have edited the text in accordance.

"The findings from our case study may not be generalizable across all fields. Firstly, we selected an example involving a large number of trials and therefore a large number of possible combinations of RCTs. Identifying VoE in fields with only a few trials could be more challenging. Secondly, we only considered two methodological choices, i.e. study inclusion and timing of endpoints. VoE can be influenced by many other characteristics, including subgroup analyses, different definitions of outcomes

(e.g. a different construction of MACE), different groupings of dosages, and different analytical strategies (e.g. choice of one-stage vs two-stage IPD, model specification, or different handling of missing data). Therefore, our results may in fact underestimate the VoE that might have resulted from many other researcher degrees of freedom in the analysis. However, our evaluation concerned 16332 analyses and provides an idea of the impact of different trial combinations and repeated endpoint measures. Our choice to focus such combinations of studies was driven by the fact that many pooled analyses are run with the risk of manipulating the results, by selecting favorable combinations of studies (5). Still subgroup analyses may be very frequently conducted in pooled analyses (e.g. for duloxetine (5)), a consideration that deserves attention in future research. Thirdly, we relied only on studies available on the YODA platform at the time of our request. All these studies were sponsored by Janssen and we considered that this specific subset of studies was quite representative of the sample a given sponsor would use when conducting a series of secondary analyses. Relying on IPD from such a homogeneous subset of studies allowed the quality of studies to be better assessed, and for analyses to be standardised. Therefore, there may be less variability in the selected studies for the analysis and less potential for VoE. We did not conduct a systematic search for other studies, for example, those conducted in an academic context. Whether the authors of academic studies would have shared the trial IPD necessary to conduct our VoE analyses is uncertain and including studies of this type could have added heterogeneity and VoE."

3.) Besides pre-specification, how would the authors recommend use of VoE in practice when evaluating pooled RCTs? Should it be used?

It is a very relevant remark. At this point we make no specific recommendation for using VoE in routine because we have only explored this in only one case study. We really think that it would be premature to recommend implementing the method in all IPD meta-analysis/pooled analyses yet. We therefore refrain from making such recommendations for this specific reason. We rather think that systematically exploring VoE in a large set of meta-analyses will provide a better sense of its relevance. There is room for new research in this area and we have clarified the text accordingly.

However, we really think that this approach has a great potential, especially when one wants to explore issues related with reproducibility, in the context of overlapping meta-analyses with divergent

conclusions. Meta-analyses (head to head, indirect, network, and IPD MA) are being more and more used with an epidemic of meta-analyses that are sometimes somewhat in contradiction. We really think that VoE could be very useful in such situations.

"These steps will continue to be important as data-sharing increases in medicine and secondary uses of this type become more popular (10,42).

We think that the VoE approach shows promise in exploring issues related with reproducibility, especially because overlapping meta-analyses with divergent conclusions are not rare in the literature (5). However, it would be premature to recommend implementing the method in all IPD meta-analysis/pooled analyses. Therefore we recommend that future research systematically explores VoE in a large set of meta-analyses in order to give a better taste of its relevance. Such a study will also help to investigate associations between VoE and "Janus Effect" with many parameters such as heterogeneity, effect size, study quality, and indeed, random sampling."

4.) Can the authors report the median and the IQR for the estimates? It seems as though the associations that contribute to the "Janus Effect" may be rare. If rare, is this really a threat to conclusions that are made from pooled analysis? Are there characteristics of individual studies that are pooled that could contribute to VoE?

We agree with this point and have tried to answer this in our response to reviewer 3 who raised a quite similar concern. We have added more nuance in our discussion insisting on the need for critical appraisal of pooled analysis finding. Concerning the characteristics, we did not explore this in this study as this would be too exploratory at this point, but, in line with reviewer 3 comments, we have edited the text.

We have now provided IQR and median in the text and in **Table 2**.

"Concerning the difference in HBA1c, …The distribution of the mean differences estimated ranged from -0.97% to -0.37% (range of 0.60%), with a median of -0.60% (interquartile range, IQR, from -0.64% to -0.57%) ..."

"Concerning MACEs, …The distribution of the HR estimated ranged from 0.05 to 2.76 (range of 2.71), with a median of 0.62 (IQR, from 0.50 to 0.73) …"

"Concerning SAEs …The distribution of the HR estimated ranged from 0.38 to 1.88 (range of 1.5), with a median of 0.87 (IQR, from 0.80 to 0.89) …"

"In the post-hoc sensitivity analysis excluding MACEs from the definition of SAEs … The distribution of the HR estimated ranged from 0.40 to 4.28 (range of 3.88), with a median of 0.91 (IQR, from 0.87 to 0.94) ..."

5.) It might be useful for the readers to describe how estimating VoE over study inclusion is different (or similar to) than assessing heterogeneity or publication bias tests, such as Egger's test: suppose we just had summary statistics - could we have attained a similar profile of "heterogeneity" between studies.

Thank you for this stimulating question. It is quite difficult to edit the text with very specific answers because it was not the initial purpose of our study. Still, we have a few ideas about this point. First, about the publication bias test. It could be quite difficult to elaborate on this as we don't have much studies in our dataset, i.e. only 12. In addition, we had, with these studies, quite a complete view of the development programme of canagliflozin and, perhaps publication bias is less of a problem than in a classic IPD meta-analysis. In other words it is a sample that is in our opinion representative of the sample used by industrialists who decide to perform pooled analyses. We have added a few words about this.

"Thirdly, we relied only on studies available on the YODA platform at the time of our request. All these studies were sponsored by Janssen and we considered that this specific subset of studies was quite representative of the sample a given sponsor would use when conducting a series of secondary analyses. Relying on IPD from such a homogeneous subset of studies allowed the quality of studies to be better assessed, and for analyses to be standardised. Therefore, there may be less variability in the selected studies for the analysis and less potential for VoE. We did not conduct a systematic search for other studies, for example, those conducted in an academic context. Whether the authors of academic studies would have shared the trial IPD necessary to conduct our VoE analyses is uncertain and including studies of this type could have added heterogeneity and VoE."

For heterogeneity, we think that this is key. In our first paper about indirect comparisons :

- Nalmefene vs placebo exhibit almost no VoE and indeed almost no heterogeneity as we used data from pivotal trial, from a very coherent development program ;
- Naltrexone vs placebo exhibit much more VoE and indeed there was more heterogeneity as the studies were conducted both before and after approval in a various settings ;

In our second paper (in revision) about acupuncture, there was, as one can imagine, a large amount of heterogeneity in those studies. And indeed, we identified a large amount of VoE.

In a sense, VoE may be seen as a visual representation of heterogeneity, as is the GOSH method (6). In addition, if VoE is expected to be associated with heterogeneity, it may also be due to other factors including sampling variability. We have tried to address this point in our response to reviewer 3. We have edited the text accordingly but don't have any specific metrics to comment on these points. We think that more research is needed to explore that. For instance, in the previous question, we answered that systematically exploring VoE in a large set of meta-analyses is necessary. This is surely the obvious next step. In such a study it will be possible to explore the relation between heterogeneity and VoE and other factors as described in our response to reviewer 3.

"These steps will continue to be important as data-sharing increases in medicine and secondary uses of this type become more popular (10,42).

We think that the VoE approach shows promise in exploring issues related with reproducibility, especially because overlapping meta-analyses with divergent conclusions are not rare in the literature (5). However, it would be premature to recommend implementing the method in all IPD meta-analysis/pooled analyses. Therefore we recommend that future research systematically explores VoE in a large set of meta-analyses in order to give a better taste of its relevance. Such a study will also help to investigate associations between VoE and "Janus Effect" with many parameters such as heterogeneity, effect size, study quality, and indeed, random sampling."

A pleasure to review your work,

Thank you so much for this kind comment.

Reviewer: 3

Thank you for the opportunity to assess this paper for potential publication in BMJ Medicine. It raises a lot of interesting points and will generate much debate and intrigue. Many IPD meta-analyses are done on a subset of the IPD that can be obtained, and it is hard to predict which trials will provide their IPD, and indeed researchers may deliberately only select or ask for IPD from a subset of all existing trials. As the paper shows, this can have an impact on the results. However, I have a number of points to be considered going forward, which I hope the authors can take on board.

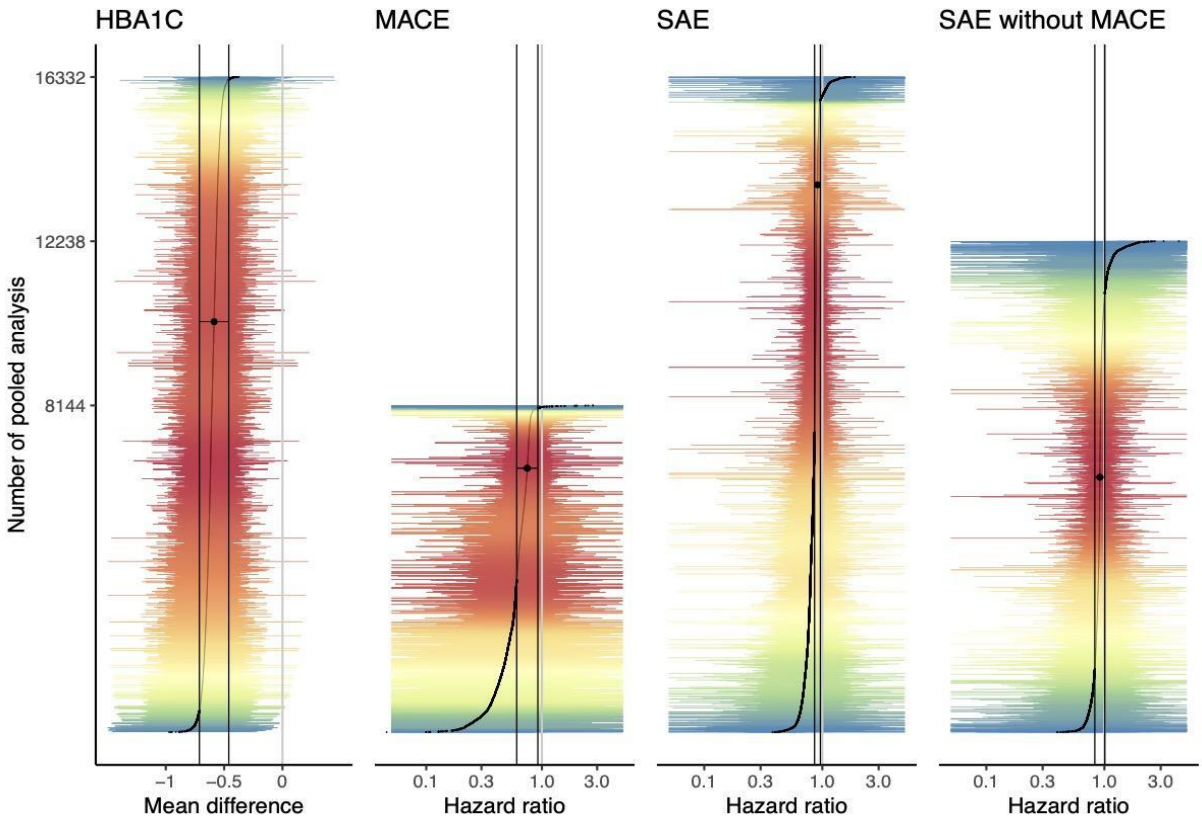
Thank you for this kind comment and for your interest. It is a chance to receive such an in depth peer review, even if it took time to answer because we had to find the best time slot to extensively re-analyse our data following your comments. We hope that you will be satisfied with the new analysis. All in all, our results were robust with some slight numerical differences. Indeed, the change in methodological choices is also a source of VoE.

1) A major comment is that the paper needs to better distinguish between the issue of bias and the issue of sampling variability. The paper shows well that the results are variable depending on the choose of trials and endpoints, and that sometimes statistical significance and even direction of pooled estimates change … but this is to be expected based on sampling variability. Fewer trials or different sets of trials will lead to variability in the meta-analysis estimates. Is this a concern? No, it is to be expected, and indeed sampling variability is an underlying premise of the design and analysis of any research.

We fully agree that this can be due to bias but also to sampling variability. We would be more nuanced about the following statement : “Is it a concern? No”. Indeed, if sampling variability is not random, it offers the opportunity to manipulate the results of those pooled analyses.

Crucially, it does not mean that anything is wrong, per se. Yes, the choice of trials will impact the findings in terms of estimates and confidence intervals (and significance in terms of p-values), but – as long as the included trials are a random sample of all trials (or representative of the populations of interest for all trials, especially in terms of any effect modifiers) – they should be unbiased and appropriate. It is a bit like the issue of sequential analysis of trials entering into a meta-analysis, and how there is variability, but this decreases as more evidence is added.

Exactly, we fully agree. We have added a figure to explore (in a very first step) this question.



"Figure 3:

Overlap in terms of points estimates and 95% confidence intervals, for all pooled analyses and for the full meta-analysis (in black). The colours represent the densities (red=high; blue=low). For clarity, x limits were set at -1.5;0.5 for continuous outcome and 0.05;5 for survival outcomes. The full figures, including extreme values are presented in Web Appendix 6."

"A total of 97.13% (15864/16332) of the meta-analyses had a HR in favour of canagliflozin (36.27% (5754/15864) were statistically significant) and 2.81% (468/16632) of the meta-analyses had a HR in favour of placebo (8.90% 39/468 were statistically significant). Figure 3 details overlap in terms of confidence interval for all pooled analyses."

A bigger concern is when there is selection or availability bias in the IPD used in the meta-analysis (See for example Ahmed et al. Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey | The BMJ and Chapter 9 of Riley RD, et al. IPD Meta-Analysis: A Handbook for Healthcare Research. Wiley, Chichester; 2021) – as then the vibration of effects is not just due to sampling variability but also bias. So actually the framework for the paper should be about making sure the IPD entered into pooled analyses are of high quality and representative of the target population – and not a biased selection of the available evidence.

Again, we fully agree and think that it is important to elaborate more on this point. See below, we have used some of your suggestions in our text.

2) Given this, I think the recommendation that "Our findings suggest that results from pooled analyses should be used cautiously, and only when based on pre-specified protocols" is not helpful or misleading … too broad a statement. Rather, the focus should be on obtaining IPD from studies that are representative of the target population of interest and high quality, in order to best estimate the estimands of interest. I agree with the message that pooled analyses of IPD cannot be simply assumed the gold-standard, and need to be critically appraised still – this is in line with previous papers like Ahmed. But to give a blanket statement that all pooled analyses should be viewed cautiously is not the message – rather, it is that they should be criticially appraised, as in any research study.

You are right, we should not throw the baby with the bathwater. We agree with the wording. We have edited the text accordingly and we tried to incorporate all ideas from the 3 previous comments.

In the first part of the discussion :

"VoE has been suggested as a standardized method that can be used to systematically evaluate the breadth and divergence of any study results (7,9,36), depending on the various methodological choices. In the context of meta-analyses, this approach is quite similar to the GOSH method (Graphical display Of Study Heterogeneity), which was proposed for meta-analyses on aggregated data (37). We believe that a method of this type, exploring all possible subsets, makes even more sense in the context of pooled analyses, as these studies do not, by nature, exhaustively cover all existing studies. In addition, the use of IPD enabled us to explore and extract outcomes (MACE and SAE) that would have been

difficult to extract from aggregated data, because they would not have been measured or reported in the initial publications, thus increasing the relevance of our study beyond the classic GOSH approach.

Lastly, we used a definition of the “Janus Effect” that is only contingent on point estimates, and not on statistical significance, as in previous work (7,9,36). It can be noted that when looking for statistical significance, our case study very rarely identified contradictory results.

“Observing changes in the direction of effect estimates and occasionally statistical significance is to be expected because of sampling variability only. Heterogeneity, bias in some of the initial studies, and the magnitude of the effect may also impact the existence of VoE and “Janus Effect”. However, we believe that the bigger concern for pooled analyses is when there is selection or availability bias in the IPD used in the meta-analysis (7).“

In the last part of the discussion:

“Our findings have several implications. Firstly, especially when performing post-hoc evaluations of published trials, pooled analyses focusing on a subset of all available studies cannot be simply assumed to be the gold-standard. In particular, our findings suggest that results from pooled analyses should be critically appraised. Health authorities, for instance, should not rely exclusively on findings from pooled analyses when approving therapeutics. Evidence suggests that findings from pooled analyses have been used to guide approvals by the European Medicine Agency (39), including that for nalmefene for alcohol use disorders (40). To enhance the quality of pooled analyses and the evidence generated by them, we suggest that pooled analyses should be planned a priori, with detailed, pre-registered study protocols, as with prospective meta-analyses (41). This would minimize any methodological changes during the analyses that could introduce VoE. When pre-registration is not possible (e.g. when the researchers conducting pooled analyses are not involved in the design or conduct of the original RCT), analytical plans should be registered prior to data analysis, and there should be full transparency regarding any decision made during the conduct of the study, such as the selection of studies to be pooled in the analysis. It is paramount that pooled analyses rely on IPD from studies that are representative of the target population of interest and high quality, in order to best estimate the estimands of interest. These steps will continue to be important as data-sharing increases in medicine and secondary uses of this type become more popular (10,42).”

3) Another point that is not discussed is that availability of IPD allows the quality of studies to be better assessed, and for analyses to be standardised. Some studies may be ignored specifically because they are identified to be of low quality, or not record key (adjustment) factors needed for appropriately estimating the estimand(s) of interest. Therefore, when done well, there may be less variability in the selected studies for the analysis and less potential for vibration of effects.

Agree we have incorporated this idea in the text that is line line with comment 4 by reviewer 1.

"Thirdly, we relied only on studies available on the YODA platform at the time of our request. All these studies were sponsored by Janssen and we considered that this specific subset of studies was quite representative of the sample a given sponsor would use when conducting a series of secondary analyses. Relying on IPD from such a homogeneous subset of studies allowed the quality of studies to be better assessed, and for analyses to be standardised. Therefore, there may be less variability in the selected studies for the analysis and less potential for VoE. We did not conduct a systematic search for other studies, for example, those conducted in an academic context. Whether the authors of academic studies would have shared the trial IPD necessary to conduct our VoE analyses is uncertain and including studies of this type could have added heterogeneity and VoE."

4) I do not agree with the approach of modelling heterogeneity or estimating it. Firstly, I-squared does not measure heterogeneity directly, and so using thresholds of 25%, 50% etc to define magnitude of heterogeneity is misleading and wrong. I-squared may be very high when heterogeneity is low and vice-versa. See this paper by Rucker et al. Undue reliance on I(2) in assessing heterogeneity may mislead. Secondly, using I-squared to determine the analysis approach is not appropriate, and the authors would be better to pre-define whether heterogeneity is expected and choose a priori their modelling approach. Given that heterogeneity is a concern in their examples, and that variability across analyses will also be impacted upon heterogeneity if heterogeneity exists and it is ignored, I strongly recommend using a random effects model by default.

We agree with you and our approach was wrong. We now report tau2. In addition, we corrected this point and indeed we performed only random effect meta-analyses now. See below for more details about the new approach and its results.

5) Vibration of effects will also be influenced by other methodology choices, including the approach to estimation of the random effects model. REML is recommended for a two-stage analysis, and for one-stage models with continuous outcomes. Also, confidence intervals need to account for uncertainty in the estimation of variances, for example using the Hartung-Knapp approach (in the second stage) or the Kenward Roger approach (in 1-stage). Variability may be less when CIs and p-values are derived appropriately. For example, CIs will be artificially narrower when there are fewer included trials and the CIs do not account for this. See for example: https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1316 and https://onlinelibrary.wiley.com/doi/full/10.1002/sim.8555.

We have now implemented the Hartung-Knapp approach (in a two stage meta-analysis, for all outcomes). See below for more details about the new approach and its results.

6) One-stage models need great care when fitting, in order to make sure clustering is accounted for and to ensure that random effect variances are estimated without bias. For the former, a random intercept or stratified intercept by study is needed – and for the latter, centering of covariates and treatment groups can reduce downward bias in between-study variances. E.g. see https://onlinelibrary.wiley.com/doi/full/10.1002/sim.8555 - but there are no details in the paper of what has been done in this regard for the one-stage model specification (in addition to estimation and CI derivation).

For sake of simplicity, we relied on two-stage analyses for all analyses now. See below for more details about the new approach and its results.

7) It's not clear why two-stage analyses were done for some and one-stage for other investigations.

We relied on two-stage analyses for all analyses now.

“Assessment of vibration of effect

All possible combinations of all combinations of the RCTs included were computed. Formula 1 was used to compute the number of possible combinations of RCTs. HbA1c measures, MACE, and SAE were analysed separately. All computations were performed for 4 different time points: 12 [or closest date], 18 [or closest date], 26 [or closest date] and 52 [or closest date] weeks. For HBA1c data, if an observation was missing at a time point (+/- 3 weeks, except for baseline where only measures at – 3 weeks were considered), it was replaced using the “last observation carried forward” (LOCF) method (20).

For each pooled analysis (defined by a given combination of individual studies and a given time point), data was pooled using a two-stage IPD meta-analysis approach (8). We used a random effects model (REML) and estimated the variance between studies using the Hartung Knapp approach. Heterogeneity was estimated using Tau2.

Effect estimates were expressed in terms of mean differences for changes in HbA1c levels and HR for MACEs and SAEs. In case of sparse events, we used the adaptation of Firth’s correction (9) to compute HR. We computed the distribution of these effect estimates and their corresponding P-values in all analytical scenarios. Pooled analyses were considered to be “nominally statistically significant” if the effect estimate had a P-value < 0.05. The presence of a “Janus Effect” was investigated by calculating the 1st and 99th percentiles of the distribution of the effect estimates (7). A “Janus Effect” is when the 1st and the 99th percentiles of the effect estimates of pooled analyses are in the opposite direction, illustrating the presence of substantial VoE (7).

All analyses were performed using R (Version 3.6.3). Two-stage analyses were undertaken using the ‘meta’ package (23) in R (24) (Version 4.15-1). Adaptation of Firth’s correction was implemented using the ‘coxphf’ package (10). All the codes necessary to reproduce the analyses are available on OSF (https://osf.io/z9cfb/).

Additional analyses and changes from the initial protocol

We performed a post-hoc sensitivity analysis for SAEs by excluding all MACEs from the definition of SAEs. This was done to focus on the SAEs that were not related to the MACE outcome, which had already been explored in our analysis and which could have indeed reflected a potential benefit of canagliflozin.

We made several minor changes to our initial protocol after receiving the data: 1) we included patients receiving a lower dose (50 mg) than initially planned (100 to 300 mg) because all doses were found to be efficacious in lowering HbA1c and were in use (26,27), 2) we excluded one study that did not match our selection criteria, 3) we only analysed intention-to-treat RCTs (no per-protocol analysis was reported, either in the study reports, or in the publications), 4) to simplify the analysis, we decided during the peer-review process to rely on a two-stage approach for all meta-analyses."

8) Isn't a Janus effect expected? That is, across all permutations of included trials, and as number of trials and sample size reduces, I would expect variability to increase and eventually lead to wide differences in estimates across the empirical distribution – such that there will often be a Janus effect, with results in opposite directions.

In line with the comments from reviewer 2, the idea that a Janus effect is necessarily expected is not fully intuitive. At least one would expect that in pooled analyses it is less important than what one would expect in VoE due to model specification in observational research. We see this comment as very related with the previous questions about random sampling and also related with the comment of reviewer 2 regarding heterogeneity. Indeed, a Janus effect is surely the result of many characteristics of the dataset. Those characteristics include heterogeneity, but also magnitude of the effect with more VoE when the effects are small/inconsistent across studies, but also as you stressed before random sampling… Our study, as planned a priori, does not explore the impact of each of those determinants on the magnitude of VoE and on a Janus effect. We think that more research is needed to explore that. For instance, in the response to reviewer 3, we answered that systematically exploring VoE in a large set of meta-analyses is necessary. This is surely the obvious next step. In such a study it will be possible to explore the relation between heterogeneity and VoE and other factors as described in our response to reviewer 2.

We have however :

- incorporated these ideas in our discussion / in line with the previous comments ;
- added a figure with all 95 % CI that may help to explore a bit further this issue (see the comment about random sampling).

I hope these comments are helpful to the authors. I recognise I have raised a number of critical issues that require (considerable) re-analyses and re-interpretation.

Thank you, yes those comments were really helpful. We really think that the paper is better now. The major changes in the analyses obviously changed slightly the numerical estimations, but not the main results and the interpretation of the results. Please find here all the changes that were made concerning the numerical results.

Nonetheless, I look forward to reading the revision and reading the authors' response.

We would like to thank you for this opportunity and hope that you will be satisfied with the changes that we made. Again, we would like to thank you for the insightful comments.

VERSION 2 – REVIEW

REVIEWER 3	**Riley, Richard** Keele University, School of Medicine. Competing Interest: None
REVIEW RETURNED	05-Aug-2022

GENERAL COMMENTS	The revision is excellent. I gave a lot of recommendations previously, and it is very clear that the authors have spent considerable time to address these. I thank them for this, and I think the paper is much stronger now. I only have minor comments remaining: 1) Suggest to change "individual patient data" to "individual participant data' throughout, as more inclusive 2) "We used a random effects model (REML) and estimated the variance between studies using the Hartung Knapp approach" – this is confusing as REML is an abbreviation for restricted maximum likelihood estimation, and the Hartung Knapp approach is for CI derivation and not for the estimation of between-study variance. So, may I suggest: "We used a random effects model estimated via restricted maximum likelihood estimation (REML), and derived

	confidence intervals using the Hartung Knapp approach". If this is correct of course. 3) Abstract conclusion, suggest make a bit clearer by: "Results from pooled analyses can be subject to vibration of effects and should be critically appraised, especially regarding the risk for selection and availability bias in IPD retrieved." 4) Key messages points should also mention specifically the issue of selection and availability bias in the IPD retrieved, I think. 5) "These findings suggest that when conducting pooled analyses of IPD from RCTs, trial selection and analytical decisions have considerable consequences on treatment effect estimation" – I think MAY have is more appropriate. Also, has the paper really shown the issue of analytical decisions? I think better to focus on the issue of analysing subsets of all trials and their selection/ availability of IPD. I look forward to seeing this published in BMJ Medicine.

VERSION 2 – AUTHOR RESPONSE

1) Suggest to change "individual patient data" to "individual participant data' throughout, as more inclusive

Thank you for your insightful remark, we will use this term instead of patient.

2) "We used a random effects model (REML) and estimated the variance between studies using the Hartung Knapp approach" – this is confusing as REML is an abbreviation for restricted maximum likelihood estimation, and the Hartung Knapp approach is for CI derivation and not for the estimation of between-study variance. So, may I suggest: "We used a random effects model estimated via restricted maximum likelihood estimation (REML), and derived confidence intervals using the Hartung Knapp approach". If this is correct of course.

Thank you for this remark, we have adopted your wording.

3) Abstract conclusion, suggest make a bit clearer by: "Results from pooled analyses can be subject to vibration of effects and should be critically appraised, especially regarding the risk for selection and availability bias in IPD retrieved."

Thank you for this remark, we have adopted your wording

4) Key messages points should also mention specifically the issue of selection and availability bias in the IPD retrieved, I think.

Thank you, we made this change.

"Our findings suggest that pooled analyses focusing on a subset of all available studies cannot be simply assumed the gold-standard. Results from pooled analyses should be critically appraised. Selection or availability bias in the IPD retrieved may impact the existence of VoE."

5) "These findings suggest that when conducting pooled analyses of IPD from RCTs, trial selection and analytical decisions have considerable consequences on treatment effect estimation" – I think MAY have is more appropriate. Also, has the paper really shown the issue of analytical decisions? I think better to focus on the issue of analysing subsets of all trials and their selection/ availability of IPD.

Thank you, we made this change.

"These findings suggest that when conducting pooled analyses of IPD from RCTs, trial selection, analysing subsets of all trials and their selection/ availability of IPD may have considerable consequences on treatment effect estimation."