Author’s response to reviews

Title: Methodological limitations in experimental studies on symptom development in individuals with idiopathic environmental intolerance attributed to electromagnetic fields (IEI-EMF) – A systematic review

Authors:
Kristina Schmiedchen (schmiedchen@femu.rwth-aachen.de)
Sarah Driessen (driessen@femu.rwth-aachen.de)
Gunnhild Oftedal (gunnhild.oftedal@ntnu.no)

Version: 1 Date: 16 Jul 2019

Author’s response to reviews:

We are very thankful to the three reviewers for their interest and the time they spent on carefully reading our manuscript. Their feedback was valuable and very constructive. We deemed almost all comments and suggestions made by the reviewers very helpful to improve the manuscript. Comments concerning the statistical analysis, including power estimates have resulted in the most extensive revisions, particularly in the methods, results (including new Table 3, additions made to Figure 3) and the discussion. Over and above these suggestions, we have changed some terminology throughout the manuscript: "key issue" has been changed to "key question" since we have reformulated the key issues into questions, "findings" has been changed to "results", "associated with imprecision" now reads "judged to have concern precision" and "psychological disorder" has been changed to "mental disorder".

Reviewer #1:

In this systematic review the authors investigated whether differences in methodologies, in particular risk of bias and impression, could account for differing statistical results. The introduction provided a good overview of the previous research and was balanced in presenting methodological limitation for studies reporting positive and negative effects. The methods used to select studies and extract data were robust and clearly described. The description of the statistical analyses needs further details and the conclusions could be expanded.

Response: We were very pleased by your positive overall assessment of our manuscript. Please find below our detailed responses to your specific comments.
P. 9 section 2.6: Please clarify the nature of the data and explain why Mann-Whitney U tests were used to analyze the data. For example, did the data consist of number of key issues per study by type of result or number of studies with a given key issue?

Response: We are thankful to the reviewer for this suggestion. In the revised version, we explain that we used non-parametrical tests because our data were not normally distributed. Please note that we re-analysed part of the data using Fisher’s exact test because the outcomes for several dependent variables were binary (see Section 2.6., Statistical analysis for details). Mann-Whitney-U tests were used for dependent variables with more than two outcomes.

Further, the nature of the data as submitted to the statistical tests is now explained in more detail: We specified that they were sorted by type of study outcome (positive or negative result) which is the dependent variable and that they consisted of the number of key questions per study judged to be at high risk of bias (for dependent variables with more than two outcomes) or the number of studies judged to be at high risk of bias/judged to have concern regarding precision (for dependent variables with binary outcomes).

P. 9 section 2.6: Please provide the estimated power for the statistical analyses. If the power is less than .80 please describe in the discussion section how this affects the interpretation of the obtained results. Specifically, if you have low power then this will limit your ability to detect difference between the groups if in reality differences exist.

Response: This is also a good point since the limited number of eligible studies for this review (especially with positive results) would require large differences in median values etc. between study outcomes to reach statistical significance. Because we are not aware of similar studies that could provide us with information needed for the power estimates, and post-hoc estimations are not advised (Dziak et al. 2018), we have used parameters from the analyses of this review, but have defined a difference in means between the two groups of studies regarding the number of key questions judged to be at high risk of bias. The method is described in Section 2.6 and the results are provided in Section 3.3.2. We have now discussed the implications of such low statistical power and the potential influence on the effect sizes in Section 4.2.

P. 13 lines 7 - 12: statistical notation is missing for the following sentence, "Also, high imprecision in data analysis was not more common among studies indicating no effect of exposure than among studies suggesting an effect of exposure (Figure 3)."

Response: This conclusion was previously based solely on visual inspection and therefore lacked statistical notion. To substantiate this conclusion, we also submitted the results for the two key questions under imprecision to Fisher’s exact test (one analysis for concern regarding
statistical power and one for concern regarding the missing adjustment for multiple comparisons) and confirmed that there are no statistically significant differences between study outcomes (see Table 3 and Section 3.3.2.). We have also added graphically the distribution for the key questions related to imprecision to Figure 4.

Section 4.1.3 is missing citations

Response: Thank you for this advice. We have added the missing citations to the studies that suggested a nocebo effect (Andersson et al. 1996; Barth 2000; Hillert et al. 2008; Kim et al., 2012; Oftedal et al., 2007; Rubin et al. 2006; Szemersky et al. 2015; Verrender et al. 2018; Wallace et al. 2010)

P. 21 lines 47 - 54 states "Nonetheless, in consideration of the fact that we identified a limited number of methodically sound studies, and even though their findings mainly indicate no effect of exposure, attempts could be made to add studies designed to detect potential weak reactions or to identify a few hypersensitive individuals." In this recommendation it would be important to address that if a study is designed to detect a weak reaction 1) what is the clinical utility of this and 2) such a study would need a large sample size to detect a weak effect. Also, if a study only includes a few hypersensitive individuals, such a study may suffer from 1) low power and 2) low external validity. In this paragraph it would be good to include information regarding how difficult it is to recruit IEI-EMF participants.

Response: The reviewer is completely right to point out that this discussion needs to be addressed in more detail. First, we have removed “designed to detect a weak reaction” from our recommendation. Although reactions to exposure may be weak and could be overshadowed by e.g. the nocebo effect, they do not have explicit clinical significance. Also, in accordance with the suggestions made by Reviewer #2, we have restructured the first paragraph of Section 4.3., Research needs. In our previous version we had already included a statement about the difficulties in recruiting study participants (formerly: “Because we lack diagnostic criteria and precise standards and because potential effects may be weak, it is particularly challenging to recruit and select study participants in order to determine if any of the individuals with IEI-EMF might actually suffer from health complaints due to a physical effect of EMF exposure and even concentrated efforts to do so may not necessarily prove successful”). In the revised version, we elaborate more on this and explain that these difficulties are mainly based on the lack of an objective case definition for IEI-EMF and that this definition is seemingly circular: a precise definition of what is considered IEI-EMF is required to conduct rigorous experimental studies, but rigorous experimental studies are required to verify whether such a precise definition actually exists. Thereafter, we suggest performing experiments at the individual level and make the reader aware of the pitfalls of such an approach (large sample size and many repetitions required
to ensure sufficient statistical power). We also added the following sentence: “Still, the potential for new insight provided by any further study, and the clinical significance of its outcome should be weighed against time investment and required resources.”

P. 22 lines 38 - 50: It would be helpful to note in this paragraph that studies using homogeneous samples of IEI-EMF participants have been unable to find any relationship between EMF exposure and well-being (e.g., Furubayashi et al., 2009; Kwon et al., 2012; Nam et al., 2009; Oftedal et al., 2007; Wilén et al., 2006).

Response: We agree that it is an interesting aspect to assess the homogeneity in the study samples and the outcome of the studies. However, the studies you listed only made an attempt to test samples of IEI-EMF that had been screened for other diseases or disorders and that reported to react to the same critical exposure sources, while we consider it most important to include all key questions under Selection bias in order to reach a judgement about homogeneity of the sample. Then one study with a positive result (Oftedal et al. 1995) and four studies with a negative result (Andersson et al. 1996, Flodin et al. 2000, Oftedal et al. 2007, Verrender et al. 2018) are relevant to mention. In Section 4.1.1., we had already referenced these 5 studies that were judged to be free from high risk of selection bias, and now have added the point that four of them were unable to find a relationship between EMF exposure and symptom development.

Reviewer #2:
This manuscript reviews experimental studies of IEI-EMF from a methodological perspective. This approach is quite necessary, as such studies have reached somewhat of a dead-end, while they remain the most obvious way to assess the alleged electromagnetic causes of IEI-EMF. One might hope that a better knowledge of these studies’ methodological drawbacks and limitations would lead to better experimental designs and more definitive answers regarding the origins of IEI-EMF. Therefore, this manuscript offers a relevant contribution to the IEI-EMF literature. It rests on a solid and clearly exposed reasoning overall. I have no problem recommending it for publication with one significant revision.

Response: Thank you very much for this positive statement and supporting acceptance of the manuscript. We have addressed all your comments below.

Indeed, I am not convinced by the authors' decision to exclude studies investigating cognitive and physiological parameters from their review, which is acknowledged but not properly
justified, and for which I can see no valid ground. As IEI-EMF symptoms are mostly subjective, it is unavoidable to rely on self-reports and subjective assessments. But for that to change, it is essential to investigate objective parameters. The failure to do so could even appear as a methodological limitation in its own right. Self-reports are not very accurate, even about well-established organic diseases or processes, and could overshadow genuine physiological reactions to EMF exposure. This does not mean, obviously, that any parameter is worth measuring. What should be discussed is rather whether the measured parameters can be connected in some physiologically plausible way with the symptoms experienced by EHS persons. In some cases at least, this connection seems straightforward, e.g., between cardiac parameters and heart palpitations, or cognitive parameters and the inability to stay focus. Studies investigating such parameters fall logically within the scope of this review and should be included. Moreover, this would yield more data to compare the methodological profiles of studies with positive and negative results, and would possibly allow to reach more definitive conclusions. Otherwise, the decision to exclude the studies investigating cognitive and physiological parameters should be justified better, and the abstract and title be modified accordingly. There is currently a mismatch between them and the manuscript's content: it is actually a review, not of experimental studies on IEI-EMF, but of experimental studies of symptom reporting and well-being assessment by EHS persons.

Response: We are very thankful to the reviewer for commenting on this and we understand the reviewer’s rationale: recordings of physiological and cognitive parameters may reveal reactions that could otherwise go unnoticed. However, there are several reasons why we did not include studies with objective outcomes in our analysis: (I) Systematic analyses of studies on alterations of physiological/cognitive parameters in IEI-EMF individuals have been provided by several authors (Hug & Röösli, 2011; Röösli et al., 2010; Rubin et al. 2011), but none of them could find a causal relation between exposure to EMF and physiological/cognitive reactions. Importantly, the review prepared by Rubin et al. (2011) included a comprehensive risk-of-bias assessment of individual studies. In our previous version, we had already referred to the paper by Rubin et al. (2011) in Section 4.2., but moved this citation to the Introduction. Also, several systematic reviews of experimental studies on symptoms and well-being in IEI-EMF individuals were published, but these reviews either considered exposure sources within a limited frequency range and/or lacked a risk-of-bias assessment in individual studies. A major aim of our review was to fill this gap.

(II) Symptoms and reduced well-being are more relevant for individuals with IEI-EMF based on their complaints in connection with EMF exposure than objective outcomes such as blood pressure, heart rate, electrical activity of the brain, or visual attention. We had stated this in Section 4.1.3, but we now elaborate on this in the Introduction. Although alterations in physiological and cognitive functioning may be linked to the symptoms, such alterations do not necessarily result in the development of symptoms. Thus, different experimental approaches are
required for studies examining symptom development and those investigating physiological or cognitive parameters.

(III) Therefore, also the systematic evaluation of studies with subjective and objective outcomes requires different approaches. Our tool for rating risk of bias and imprecision is customized to experimental studies examining symptoms and well-being and is only partly suited to evaluate studies examining physiological and cognitive parameters. Nevertheless, our analysis and our rating tool may serve as a basis for preparing an updated systematic analysis on studies investigating objective outcomes, given that relevant amendments are made to the rating tool.

We acknowledge that we did not sufficiently outline why we did not consider studies with objective outcomes and have therefore included a more elaborated explanation in the last paragraph of the introduction. Here, we also explain that a major aim of our systematic analysis was to provide a comprehensive assessment of the methodological limitations in individual studies on symptom development that would allow to judge the quality of evidence. Since we have extended the considerations about objective outcomes in the introduction, we have removed similar in the discussion where we previously considered the limitation by not including objective outcomes.

We thus agree with the reviewer that also the title of our review should be modified in order to better reflect its content and it now reads: “Methodological limitations in experimental studies on symptom development in individuals with idiopathic environmental intolerance attributed to electromagnetic fields (IEI-EMF) – A systematic review”. We have also specified in the Abstract that our analysis is about experimental studies on symptom development in IEI-EMF individuals.

There are some others points I want to raise, but none as fundamental as this one. Addressing them might help to improve the manuscript, and I look forward to reading the authors' answer.

(1) Is the existence of "other reasons for their EMF attributed health-complaints" (p.8, l.62) a relevant exclusion criterion for EHS participants? It seems logical, but the probability that EHS persons suffer from undiagnosed, well-established organic diseases is arguably low given their history of multiple medical examinations (de Graaff & Broër, 2012; Dieudonné, 2016, 2019) and the revision rate of the functional somatic symptoms diagnosis (between 0,5 and 8,8% according to a meta-analysis by Eikelboom et al., 2016). In other cases, what would constitute a "reason" for IEI-EMF symptoms is unclear. What about functional somatic syndromes of unknown etiology and similar symptoms as IEI-EMF, e.g., fibromyalgia or chronic fatigue syndrome? What about depression or anxiety, especially when EHS persons claim EMF exposure durably alters their mood? What about persons suffering from a somatic disease they consider entirely distinct from their EHS? To untangle this would require in-depth
interdisciplinary assessments, e.g., as in the study of environmental sensitivities by Herr et al. (2004), that would quite exceed the scope of provocation studies. Furthermore, excluding participants with any of these diagnoses might introduce a bias, as there is no indication that IEI-EMF should only occur alone. As the authors acknowledge later in the manuscript (p.17, l.62), the most relevant criterion is ultimately that participants report significant differences in their symptoms between situations with and without exposure. Why not draw the full conclusion from this?

Response: Thank you for raising this important issue. We acknowledge that the exclusion of individuals with somatic diseases and/or physiological disorders that may equally explain the symptoms but are unrelated to EMF is a very challenging task. Also, it is possible that hypersensitivity to EMF is comorbid in some individuals. From our point of view, however, the exclusion of participants with EMF unrelated diseases/disorders is a crucial aspect of the selection procedure, precisely because a large number of the reviewed studies actually did not verify the contrast in symptom development with and without exposure. Please note that we specify that only somatic or mental conditions that may explain the EMF-attributed symptoms should be reasons for exclusions. In the key questions defined under selection bias, we ask for a rigorous anamnesis of potential study participants. From our clinical routine, we are aware that some individuals consider themselves hypersensitive to electromagnetic fields, but it turned out that they actually suffer from a psychiatric/mental disorder or similar and needed different treatment. Furthermore, several studies indicate that IEI-EMF individuals suffer more often from mental or somatic conditions than healthy controls and these conditions may actually be the reasons for various reported symptoms (e.g. Hillert et al. 1998, Eltiti et al. 2007b; Dahmen et al. 2009). Such individuals should be excluded from the study, because as cited in 4.1.1 Risk of bias and imprecision, Baliaatsas et al. (2012) “noted that the inclusion of individuals whose symptoms are not related to EMF exposure may dilute the findings and reduce the chance of identifying individuals who suffer from health complaints due to a physical effect of EMF exposure”. That is, the inclusion of individuals with EMF-unrelated diseases and disorders would favour a non-statistically significant result and such a selection bias could lead to underestimation of an effect of exposure - if existent. We have addressed the challenges regarding this key question in more detail in Section 4.1.1 following the Baliaatsas et al. (2012) citation. Furthermore, we have re-named the key-question to be clearer about its purpose, it now reads: “Were individuals excluded whose symptoms may be caused by diseases/disorders that are unrelated to EMF exposure?”

(2) Considering that "the nocebo effect could either overshadow physical effects or may otherwise be the only reason why symptoms are experienced in everyday life" (p.19, l.26) - shouldn't the failure to prevent such effects from occurring during the experiments be regarded as a methodological limitation? For example, have precautions been taken to reduce the participants' anxiety, resulting understandably from their exposure to EMF they consider harmful? Have their perceptions of the experimental set-up been analysed? Etc. There are other
problems with using the results of provocation studies to attribute IEI-EMF to nocebo responses. First, these results only demonstrate that EHS persons are susceptible to the nocebo effect, which is not specific to them: arguably, any person knowingly exposed to a substance perceived as strongly harmful would react in the same way. Second, the nocebo hypothesis is not ecologically valid, for reasons I exposed in my latest article. In particular, it doesn’t convincingly explain how EHS persons came to fear EMF, neither why they were sufficiently motivated to go through the complex learning process that allows them to identify EMF in their environment.

Also, p.19, l.31: "Alternatively, in previous studies, only a few of the individuals believing that their symptoms are caused by EMF exposure were right, and in most cases, causes other than EMF exposure explain the symptoms." What causes?

Response: We agree that how to handle the nocebo effect in provocation studies is challenging. As long as the individuals know that they will or may be exposed, a nocebo effect cannot be excluded. It would take a completely different approach to handle that, i.e. the symptoms must be recorded in situations when the participants have no clue that they may be exposed. Probably, registering symptoms in every-day life in combination with dosimetry registration of the exposure (in a blind way) could lower the nocebo effect, but also in such an approach the nocebo effect cannot be controlled in situations when participants are aware of being exposed, e.g. when being near a base station, and WiFi antenna, or using their own mobile phone (e.g. see Bogers et al. 2018).

You also mention anxiety. The anxiety concerning the effect of the exposure on the development of symptoms is a key feature of individuals classified as IEI-EMF, which should be handled with treatment/therapy. However, we do not regard it feasible to control anxiety concerning the effect of the exposure in provocation studies. Another issue is the general anxiety and stress level in the exposure situation. Arrangements may be made to reduce this by e.g. including habituation sessions, which we have considered in the risk-of-bias-assessment tool, and by other approaches such as at-home testing. Generally high anxiety or stress levels may potentially be reasons for the symptoms to appear and they might mask any potential physical effect of the exposure. We have now included this aspect in the discussion in Section 4.3.

We agree that previous interpretations of the results in favour of a nocebo effect cannot be used to exclude a potential physical effect of the exposure, and we have now made this even clearer in the discussion (Section 4.1.3).

It is also true that the nocebo effect is not unique to IEI-EMF individuals, thus there must be other differences between those with and without IEI-EMF. We have now included examples of some possible other factors that may be part of the explanation for developing such a condition, even if EMF is not the physical cause. We only provide a few examples in the discussion since such other factors had also been specified in the Introduction, and we do not regard it relevant to elaborate on this in this review. Our conclusion is that the physical effect (if it exists) seems
weaker than the nocebo effect or that a physical relation exists only for a few individuals among those attributing their symptoms to EMF, and this is important to be aware of for potential new studies. We believe this is sufficiently clear from Section 4.1.3.

(3) I wish the content of the reviewed studies was discussed at greater length. I understand this manuscript has a different purpose, but having more insight on these studies’ protocols could shed light on some limitations of IEI-EMF literature as a whole, and help designing better experiments in the future. Consider two examples. Regarding exposure parameters, Table 1 indicates that out of 15 studies involving RF-EMF exposure, 13 used mobile telephony signals and 2 used TETRA signals. Therefore, some sources of exposure to which EHS persons attribute their symptoms have still not been studied, e.g., cordless phones and Wi-Fi routers. This is a blind spot of the literature that may be worth mentioning. Regarding symptoms reporting and well-being assessment, no precision is given on the research tools actually used by the reviewed studies, even in Table 1. Some authors consider these tools should be standardized and have developed suitable questionnaires for IEI-EMF, environmental hypersensitivity, or non-specific physical symptoms (e.g., Eltiti et al., 2007; Nordin et al., 2013; Yzermans et al., 2016). How widely used are these?

Response: We are very thankful to the reviewer for the advice to provide more insight into the study protocols. To former Table 1 (now Table 2) we also added, as you suggested, the outcome measures, i.e. which research tools were used during the experimental sessions to record the symptoms and rate their severity. In Section 4.1.1., we have further noted, as advised, that many different tools (both customized and validated tools) were used but that the use of standardized questionnaires developed for studies with IEI-EMF individuals is rare. Regarding the sources of EMF-exposure that have not been examined in the reviewed studies, we decided against mentioning cordless phones and WiFi-routers. On the one hand, cordless phones hardly exist any more as mobile phones have become the technology of choice. On the other hand, it is most important that devices were tested that individuals associate with their health complaints and this has been done in the majority of the reviewed studies.

(4) A fundamental problem remains the lack of an objective IEI-EMF case definition. This problem is somewhat circular: rigorous experimental studies should allow to characterize IEI-EMF objectively, but an objective definition is necessary to conduct rigorous experimental studies. Otherwise, it is impossible to ensure the homogeneity of test and control groups (i.e., the absence of persons believing wrongly that they are EHS in the test group, and of persons believing wrongly that they are not EHS in the control group). The surest way to make this problem tractable seems to perform and analyze experiments at the individual level. But that requires a high number of exposure sessions per subject, which might prove unrealistic. Overall,
I feel little progress has been made in the definition of IEI-EMF or in the distinction of subtypes. I have tried to address this problem in my latest paper, where I propose a behavioural definition of IEI-EMF: I don't expect to see it mention here, but it could add some food for thought.

Response: We completely agree with the reviewer’s rationale. IEI-EMF is at present circumscribed by self-report of symptoms and critical exposure sources from individuals who believe being hypersensitive to EMF. However, due to the lack of an objective case definition, it still remains unclear whether this condition actually exists, i.e., is physically related to EMF. We particularly agree with the reviewer that approaches at the individual level seem most promising, but that time and costs for realising such studies may be disproportionate. However, if it were possible to identify individuals who show a clear contrast between situations with and without exposure in the double-blind condition, an in-depth characterisation of these individuals could serve as a basis to provide evidence for a physical relation between EMF exposure and health complaints.

We have elaborated on this aspect in more detail in Section 4.3, and also highlight the fact that little progress has been made with the objective case definition for IEI-EMF. In this place, we refer the reader to your previously published paper.

Reviewer #3:

Dear Authors,

My summary of your ms is below along with my comments. I hope they are helpful.

The purpose of this systematic review was to assess the methodological quality of previous publications regarding reported symptoms during/after exposure conditions. In particular, whether methodological limitations may have led to false positive or false negative results reported for IEI-EMF participants in single- and double-blind studies that investigated acute and semi-acute effects of EMF exposure, using a customised risk-of-bias assessment tool. This customised tool identified 16 key limitations in methodology, which may have led to bias and/or imprecision in results. These 16 key limitations were categorised in terms of 7 domains, 6 for risk of bias and 1 for imprecision. The tool rated studies as high (or not) risk of bias/imprecision. The 16 key limitations were also categorised as favouring positive or negative results. All included studies were therefore rated as having a high (or not) risk of bias in favour of a positive, negative or non-directional outcome within these domains and compared in terms of their actual statistical outcomes using non-parametric statistics to assess impact.
I think the paper would be useful to researchers new to the field, those already researching in this area who want to conduct a new study and to those researching environmental illnesses more generally; it provides a useful summary of findings and pitfalls to be aware of. In my view, the review was systematic and thorough, with clear reporting of how the search was conducted and studies chosen (2.1 to 2.4 and Figure 1). You have stated that you created a search protocol a priori, which is fine. If you published your systematic review protocol in PROSPERO or similar, please provide details.

Response: We are very thankful for this nice overall assessment of our paper. Please find below our responses to your specific comments.

Although we prepared our systematic review protocol before the search for relevant articles and data extraction we did not publish this protocol in PROSPERO.

I think the idea of a customised tool is helpful, though if you have based it loosely on any existing tool, it would be a good idea to say so and provide a reference. Your 16 items for the tool are pertinent and well-considered.

Response: The missing references (assessment tools published by COCHRANE and GRADE) on which we had based the development of our rating tool are now provided in Section 2.5.

I think the paper could be improved with a clearer methods and results section. My recommendations for the manuscript are given below. This is your manuscript, so you don't have to amend verbatim as I've suggested as you might come up with better ideas. The aim with my recommendations is to draw attention to areas that I think could be clearer/more precise.

Recommendations:

General:

Be precise/clear with your chosen terms.

Regarding your customised rating tool. If I understand correctly, you are rating studies in terms of risk of bias AND whether or not they have a high risk of bias in favour of the null or a high risk of bias in favour of a statistically significant result. You need to make this explicit in your introduction when you state the aim of your systematic review. Your caption for Table A1 reads: "Description for rating 16 key issues. The key issues are grouped into six domains for risk of bias and one domain for imprecision. Alternatives marked in red are associated with high risk of bias, alternatives marked in yellow are associated with high imprecision." You don't state that
you are assigning bias in a particular direction. This table is where you have the opportunity to formally define which key limitation favours a \("+\)\), a \("-\)\) or a \("\pm\)\).

Response: We are thankful to the reviewer for this critical comment and acknowledge that the definition of the bias direction had not been sufficiently outlined in the previous version of the manuscript. It is correct that we rated studies in terms of risk of bias plus assigned to each key question a direction of bias it would have on study outcome, i.e. in favour of an effect of exposure, in favour of a null result or uncertain direction on the study outcome.

In the revised version, we elaborate more on this aspect both in the introduction where we describe the aim of our systematic review and in the methods where we describe our rating tool (Section 2.5.). To formally define the particular direction of bias, we have added a table (now Table 1) to the manuscript such that the reader can find the relevant information for each key question at a glance. This seemed more convenient than defining them in Table A1 in the supplementary material. Nevertheless, we have also added bias directions as signs to Table A1 as you suggested.

Be precise/clear with your statistical analyses.

I think, in general, the analyses are perhaps under-developed. Regression analyses might provide more depth/clarity. You indicate this in your discussion "several key issues are not independent of each other and interfere to some degree…." (Line 50-51, page 17). I've made some additional points below.

Response: We agree with the reviewer that the statistical analyses need to be developed and explained them in more detail. Please see below for our responses to more specific comments.

We evaluated whether studies with positive results differed from those with negative results with respect to the profiles of the methodological limitations. Therefore, the two groups of study outcomes were considered as independent variables, while the outcomes from the risk-of-bias and imprecision assessment were the dependent variables. With two independent variables, we do not regard regression analysis as relevant.

We have looked more carefully at the sentence cited from p. 17, lines 50-51 and acknowledge that its meaning was not clear. It now reads: “Furthermore, when revisiting the assessment tool and the results of the assessment, we became aware that in some cases the application of a method as specified for a particular key question or additional results/ information provided by the studies might have an influence on the rating of other key questions .”
More specific points:

Line 40, page 8: I would just say "i.e. none of the other alternatives were applied or reported in the paper, see Figure 2".

Response: Agreed, we shortened this explanation.

Line 54-55, page 8: I recommend you move the sentence "One domain was defined….two key issues" to line 26 on the following page before "Under imprecision…". Also, I would start a new paragraph here (starting with your 2 key issues around imprecision).

Response: We moved this sentence further down as suggested by the reviewer and also started a new paragraph here.

Line 9, page 9: In the methods and results sections try to be precise where possible, so rather than saying "most", give a percentage with (n =) in brackets.

Response: Thank you for this advice. Instead of saying “many/most” we now give the exact numbers of studies and also the percentages in brackets, e.g. “21 (75%) studies” where relevant.

Line 9, page 9: What does "adaptation with time" mean? Are you referring to habituation or something else?

Response: You are right, we meant to say “habituation” and changed wording here.

Line 9, page 9: Confounding bias: Are you saying that most of the covariates are unrelated to the exposure condition and may influence the outcome therefore they are not actually confounding variables?

Response: We have rephrased this explanation and gave three examples for possible co-variates as we are aware that this could have been misunderstood by the reader. It now reads: “The specified co-variates relevant for the key question “Appropriate control for other co-variates” under confounding bias are not related to the exposure condition but may influence the outcomes (e.g., use of an adaptation period, inclusion of pre-trial symptom levels in the analysis, or control of temperature, humidity and light). Therefore, in randomized trials that did not control any of these co-variates, we did not consider this to cause a high risk of bias” We have also specified: “If the sequence of exposure was not randomized, this was considered a source of high risk for bias and addressed under performance bias.”
Line 14, page 9: "missing control" could be replaced with "uncontrolled".

Response: Please see response to your previous comment.

Line 28-29, page 9: If you can, provide a short example of "low precision in the findings" to add a bit of colour to your point.

Response: Thank you for this advice. We acknowledge that the formulation was very imprecise, and have rephrased this paragraph in Section 2.5. and also explain in more detail the terms we use.

Line 54, page 9: 2.6 Statistical analysis: This section should be very precise and concise. Provide the statistical tests along with the IV and DV information only. What is your alpha level? If you are conducting multiple comparisons, will you be applying a correction? Any results, including your chosen ratings, should be kept to the results section.

Response: Thank you for these helpful suggestions. To be more precise, we explained the nature of the data and specified the independent/dependent variables and the significance level in the description of the statistical analysis (Section 2.6.). Further, because we did multiple comparisons, we have explained the adjustment of the significance level using the Bonferroni method. Please note that we re-analysed part of the data using Fisher’s exact test because the outcomes for several dependent variables were binary (two outcomes). Mann-Whitney-U tests were used for dependent variables with more than two outcomes.

The descriptions of the direction of bias that a key question would have on the study outcome, including the number of key questions assigned to the various directions, were actually not results, but a description of our assessment tool. This information has been moved to Section 2.5 with more explicit explanations and reference to the newly included Table 1.

Line 57-58, page 9: This is an interesting outcome. It looks like you are referring to Figure 3 here. It would be helpful to say this so that the reader can refer to the figure, whilst reading this outcome. Note also that in the context of discussing your analysis, avoid using phrases such as "associated with" when it would be more accurate to say the number of key issues rated (or judged) as having high risk of bias…
Response: You are right that the outcomes refer to Figure 3, and we have referred to the figure in the first sentence of this paragraph.

We now say "number of key questions judged to be at high risk of bias". We have changed this phrasing throughout the manuscript, including the figures.

Line 2, page 10: what do you mean by "the direction of influence was uncertain"? Please provide an example.

Response: This term has been rephrased throughout the manuscript, including the figures and now reads “uncertain direction [of bias] on the study outcome.

Results section

Line 17-18, page 10: It would be useful to provide a sentence with total number of participants, broken down into sensitives vs controls across studies. You could even break it down into studies with positive vs negative findings etc. Please provide % (n=).

Response: We appreciate this idea and have provided the relevant information on the total number of participants in Section 3.1. However, as we split the studies into outcomes later in the manuscript, we break down the numbers for IEI-EMF participants into studies with positive and negative results in Section 3.2. As specified in Section 3.2, we only focus on the results for IEI-EMF participants and therefore have not provided the number for controls according to study outcome.

E.g. Line 40-41, page 10: "Three studies were not associated with…" you could say "Three studies were not rated has having high risk of bias…"

Response: Agreed. As mentioned in response to one of your previous comment, we have changed wording throughout the manuscript, including the figures.

Line 50, page 10: start a new paragraph with "Effects of exposure…"

Response: Ok, a new paragraph has been started here.
Line 50: for clarity I would change the second sentence. Instead of "Four studies…" it might read better as "Of these seven studies, four reported…., three found… and five tested…".

Response: Agreed. We started this sentence with “Of these seven studies…” However, “…and five tested” is not part of this sentence because it refers to a different aspect and is considered in the next sentence.

3.3 Methodological quality:

Line 34, page 11: It would be helpful to tell the reader to refer to Figure 3 here to highlight that the most common methodological limitation was selection of participants - it's quite striking when you look at the figure. What is the overall proportion of studies affected?

Response: Thank you for this advice. We have referred the reader to the corresponding Figure 3 and gave the overall proportions of studies for the most common limitations: selection bias (n=23, 82%), performance bias (n=14, 50%) and imprecision (n=23, 82%).

Line 44, page 11: Please provide the percentage of studies with number in brackets e.g. 64% (n=18).

Response: As stated in response to one of your previous comments, instead of saying “many” or “most”, we now provide the number of studies and the percentages in the description of the results where relevant.

3.3.2 Distribution of key issues …

Lines 38-45/46, page 12: I would move this bit to the previous section where you talk about the overall rankings of bias….or at least put it in a separate paragraph 3.3.2 seems to be mainly examining positive vs negative bias ratings.

Response: Agreed, we have moved this paragraph to Section 3.3.1. In addition to summarizing the total number of key questions judged to be at high risk of bias across individual studies, we now also provide a short overview for key questions related to imprecision in Section 3.3.1.
Please provide your full Mann Whitney U finding with degrees of freedom rather than just p-value. Also provide the median and inter-quartile values in brackets within text (or as a supplementary table) to make it clear what the comparison is.

Response: Thank you for this recommendation. We agree that the missing information should be provided and therefore prepared an additional table (Table 3) in which we assembled the results regarding the statistical analyses, including the test statistics (U, p), medians and interquartile ranges for the Mann-Whitney-U test. Note that we ran Fisher’s exact test on three dependent variables because their outcomes were binary. Accordingly, we provided the numbers and percentages of studies judged to be at high risk of bias/judged to have concern regarding precision and the p-value in the line “test statistics”. Since the table now provides the relevant details, we have revised the text in Section 3.3.2. accordingly.

Line 48: "Also…." Did you conduct an analysis here or was this based on visual inspection? If you performed analyses, what were the results? How many comparisons were performed?

Response: We agree that this was not clearly stated. The results of the statistical comparisons including those for imprecision are presented in Table 3, and as said above the description of the results has been rephrased accordingly.

In places the wording is a bit confusing….you refer to favouring negative findings in studies with negative findings…consider saying studies with negative statistical outcomes or results.

Response: We understand that the wording used to explain the relation between study outcomes and bias direction may have been confusing and have therefore changed the terms throughout the manuscript, including the figures, as follows:

“bias favouring positive findings” → now “bias in favour of an effect of exposure”

“bias favouring negative findings” → now “bias in favour of a null result”

Because we also rephrased Section 3.3.2, doubling of the use of “positive/negative” within a term is now avoided.

Line 7-8, page 13: Double negatives can be confusing. It might be better to say that high imprecision in data analysis was comparable rather than "…not more common…"
Response: Thank you for this advice. Because we restructured the description of the statistical comparisons, this sentence has been removed from the manuscript. However, when describing results that were “similar”, we used the terms “was/were comparable” as suggested.

Discussion

I found it hard to follow when you were talking about actual positive vs negative findings vs bias towards positive/negative findings.

Response: Please see response to your comment above. We have changed wording throughout the manuscript.

Line 4-5, page 18: It appears that you did two sets of ratings to re-classify some papers' bias score. Is that right? "…indicated as such in the initial ratings in Figures 2-4"…. "Thus, the subsequent analysis of interferences between key risk-of-bias issues yielded that six studies, three with positive [45,48,49] and three with negative findings [56,57,59], have a lower number of key issues associated with high risk of selection bias compared to the initial rating…"

Response: This is correct. We did two sets of rating, one initial, systematic rating that was based on the outlined rating tool and a second revisited rating, that was not included in the protocol and was done by one of the authors after we finished the initial rating, i.e. the revisited rating was less systematic than the first. It considered possible interferences between key questions, i.e., the application of methods as specified for a particular key question or additional results/information provided by the studies might have an influence on the rating of risk of bias of other key questions. This reduced the number of key questions judged to be at high risk of bias for some papers. In the revised version of the manuscript, we describe that the revisited rating, including the statistical analysis of the results yielded comparable results to the initial rating.

Table A1, page 14: "not considered" has not been highlighted.

Response: This has not been highlighted because we did not judge a study to be at high risk of bias when none of the co-variates were considered (please see also Methods, Section 2.5.: “…, in randomized trials that did not control any of these co-variates, we did not consider this to cause a high risk of bias. If the sequence of exposure was not randomized, this was considered a source of high risk for bias and addressed under performance bias.”)
Table A1 categories: "not sufficiently considered" please provide some examples of what this means. It might be best to just say "not reported", unless you can qualify what you would consider "not sufficient".

Response: The reviewer is right to point out that this classification is not quite meaningful. We removed all “not considered” and “not sufficiently considered” categories (except for one key question) and modified the relevant explanations to fit with this. Only the methodological alternative judged to be at high risk of bias for the key question “Were individuals excluded whose symptoms may be caused by diseases/disorders that are unrelated to EMF exposure?” is labelled “not sufficiently considered/not reported”. We have provided an example for a case that we consider “not sufficient”.

Figure 4: It would be helpful to provide the names of the key issues rather than just number of key issues, given that there are not that many. For example, the "Key Issues associated with high risk of bias favouring positive findings" has one key issue (blinding) that seems comparable across studies. If you put that label in, then the reader doesn't have to look that up. Likewise for the other graphs.

Response: We found that it would take quite some space to list up to seven key questions because some of the definitions are wordy and not easy to name with just a few words. As stated above, we included an additional table (now Table 1) in which we specified the direction of bias on study outcome for each key question (in favour an effect of exposure, in favour a null result or have an uncertain direction on study outcome). This table may not only facilitate reading of Figure 4, but also reading of the methods and the results section. To each subplot in the figure, we have also added information about the maximum possible number of key questions judged to be at high risk of bias.