PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Outcome measures for assessing the effectiveness of non-pharmacological interventions in frequent episodic or chronic migraine – a Delphi study
AUTHORS	luedtke, kerstin; Basener, Annika; Bedei, Stephanie; Castien, Rene; Chaibi, Aleksander; Falla, Deborah; Fernández de las Peñas, Cesar; Gustafsson, Mirja; Hall, Toby; Jull, Gwen; Kropp, Peter; Madsen, BK; Schaefer, Benjamin; Seng, Elizabeth; Steen, Claudia; Tuchin, Peter; von Piekartz, Harry; Wollesen, Bettina

VERSION 1 – REVIEW

REVIEWER	Deena Kuruvilla Yale school of Medicine, USA
REVIEW RETURNED	12-Apr-2019

GENERAL COMMENTS	My biggest concern and limitation of this study is the delphi method itself. Since the method doesn't quantify the number of experts needed and doesn't define what an expert really is, I am not sure one can make a conclusion based on 10 people's popular opinion. This paper does not have any headache specialist or Integrative medicine specialist. The majority of experts on the paper are physiatrists so there is likely a bias here. I would highlight the general qualifications of the expert panel. I would also consider making a table with each non pharmacological approach you considered in this paper.

REVIEWER	James Odell Bournemouth University UK
REVIEW RETURNED	26-Apr-2019

GENERAL COMMENTS	I have listed it for major revision although I suspect it falls between minor and major. If I were to summarise I would say that the issues that spring out are lack of consistency, particularly in terminology and more concerning the use of statements that do not seem to be justified with the information provided. I hope my comments in the attached file are helpful – please contact pulisher for this file.

REVIEWER	Ana-Carolina Goncalves University of Southampton
REVIEW RETURNED	30-May-2019

GENERAL COMMENTS	Congratulations on this interesting research. This is a very interesting study, looking at reaching consensus on the most useful

	measurement tools to be used in clinical trials of non-pharmacological interventions to manage chronic migraines. Overall this manuscript presents interesting findings but with some important methodological limitations: a) a consensus on tools before a clear consensus on outcomes; b) absence of patients and clinicians in the Delphi panel; and c) a lack of a clear definition of consensus. Please consider the specific comments below in order to improve the clarity of the manuscript: Abstract: 1) The abstract would benefit from a short sentence explaining the rational for the study. Through the abstract the reader is left questioning why is this study important and why was it conducted. Introduction: 2) Lines 44-51. The authors argue that, to determine the effectiveness of non-pharmacological interventions it is important to agree on the use of measurement tools. What about an agreement on the most important outcomes? What is it that non-pharmacological interventions want to achieve? E.g. reductions in the use of medication? Reductions of work absence? In pain intensity? In frequency of episodes of migraines? Or a reduction of fear of symptoms? Improvements in quality of life?. Potentially all these outcomes can be measured by a different tool. Before a consensus is reached on the tools, we may need to reach a consensus on the outcomes. If this work has already been done, please make reference to it in the background. Otherwise, please justify the consensus on tools, before a consensus on outcomes in achieved. 3) Please be aware of some inconsistency in the use of the expressions “outcome” and “outcome measure”. These are two very different things (e.g. Outcome = quality of life; outcome measure or measurement tool = SF-36). The aim of the study in the abstract says “outcome measures”; the aim in the background says “outcomes”. This means that by the end of the background, the reader is still unsure if the consensus process is about outcomes, or about measurement tools. Methods 1) Page 6, Line 17 – selection of experts. Please provide a rationale for not inviting patients and clinicians to be part of the panel, as they may be key stakeholders in this consensus process. 2) Page 6, Lines 32 to 36. I would suggest presenting the number of experts who agreed to take part in the results section rather than the methods. Further, please be aware that the information in the abstract is not consistent with the information in the methods section. The abstract suggested that 12 participants completed the 3 rounds of the Delphi. In this section, it is clear that only 10 completed all 3 rounds. Please correct this information in the abstract to avoid misleading the reader. 3) Step II, page 7. The order in which the items in a Delphi survey is presented is known to influence participants’ responses – please indicate if the outcome measures were presented in the same order to all survey participants, or if it was randomised.

	4) Page 8, patient and public involvement. Please consider re-phrasing the first sentence of this paragraph. At the moment it is very long and not very clear. 5) Page 8, patient and public involvement. Please clarify if the interviews with patients were conducted as “patient and public involvement” or formal data collection. If this was a moment of formal data collection (as it appears to be based on the results section), please explain the method used for data analysis present this information is different section of the methods (not under “patient and public involvement”). 6) A clear definition of consensus is missing in the methods section. An a-priori definition of consensus is essential in Delphi studies. 7) It would add value to the manuscript to clarify if the survey included information on the psychometric properties of the tools or copies of the actual tools – knowledge of this would have influenced the choices made by the expert panel. Results: 1) First paragraph. Please note once more the interchangeable use of the expressions “outcomes” and “outcome measure” and refer to my previous comment. 2) Page 10, Line 26 “no outcome was discharged after this initial round” – what would be the criteria to exclude outcomes from round to round? This is part of the definition of consensus, which is missing in the methods section. 3) Results from patient interviews: this section really highlight the problematic of selecting tools before a consensus is reached on the outcomes. Patients expressed that none of the tools measured outcomes that matter to them (e.g. fear of a migraine attack). This imposes the question: Why are we using any of these tools if they do not capture what is important to patients? Discussion: 1) Please consider revising the discussion. It is very descriptive at the moment. I would also advise some caution making recommendations about any of these tools as the findings appear to indicate that perhaps none of the tools is ideal.

VERSION 1 – AUTHOR RESPONSE

REVIEWER: 1
Reviewer Name: Deena Kuruvilla
Institution and Country: Yale school of Medicine, USA

My biggest concern and limitation of this study is the delphi method itself. Since the method doesn't quantify the number of experts needed and doesn't define what an expert really is, I am not sure one can make a conclusion based on 10 people's popular opinion. This paper does not have any headache specialist or Integrative medicine specialist. The majority of experts on the paper are physiatrists so there is likely a bias here.

Author response: The focus of this recommendation is on non-pharmacological interventions such as physiotherapy and aerobic exercise. It was important to find experts for this field rather than for pharmacological interventions. It is made transparent in the report who was on the panel and I don't see any method which eliminates bias from expert recommendations.

I would highlight the general qualifications of the expert panel.

Author response: This was already stated in the manuscript:" These instruments were evaluated by a panel of 12 experts (8 physiotherapists, 2 chiropractors, 2 psychologists)". We now added the academic qualification: "All experts had an academic degree equivalent to PhD, except 1 expert with an MSc."

I would also consider making a table with each non-pharmacological approach you considered in this paper.

Author response: we did not consider any specific intervention. Experts were free to imagine anything non-pharmacological. For the choice of experts, we have presented a search strategy in the manuscript for publications on such interventions and contacted the first and last authors of each publication.

REVIEWER: 2

Reviewer Name: James Odell
Institution and Country: Bournemouth University, UK

I have listed it for major revision although I suspect it falls between minor and major. If I were to summarise I would say that the issues that spring out are lack of consistency, particularly in terminology and more concerning the use of statements that do not seem to be justified with the information provided. I hope my comments in the attached file are helpful.

Author response:

Point	Pg/27	Line	Text	Comment	*Author response:*
1	3	16	12 agreed to participate and completed all 3 rounds of the survey.	Only 10 completed. Not the 12 as stated here	*Thank you this was corrected to: "12 agreed to participate and 10 completed all 3 rounds of the survey."*
2	5	39-	other nonpharmacological	Lines 23-28 discuss patients	*This has been clarified by adding:*

		44	treatment strategies are effective to complement ...	who cannot take or do not want meds but this section talks about NPI being a complement – does this mean as an adjunctive or a standalone? It is unclear	*"…are effective to complement or replace the pharmacological management"*
3	6	29-35	Non-pharmacological interventions have a low risk of adverse events (AEs) and are less expensive. Intuitively, these approaches should therefore not be measured on the same scale as preventive medication since medication has a high risk	Why not? If you are examining the outcomes it may be that you include AE as an outcome measure – it seems odd to say because there are lower AE we should give NPI studies a less rigorous outcome measure unless patients are willing to trade it off. Part of the issue with NPI studies is the lack of rigour and this would seem to just support that view. I think a better explanation or justification is needed if it is part of an argument	*Methodological rigour does not require cut-off values which are out of reach for the planned investigation.* *A benefit from a medication has to be high in order to tolerate side effects. No patient would take a medication which gives a minor improvement of symptoms and still tolerate side effects. We are currently conducting a physiotherapy intervention study and are asking patients how many days reduction they feel would make the effort worthwhile and most of them respond that any single day reduction is a great relief.* *An additional point to make is, that the non-pharmacological interventions may be investigated as an adjunct to e.g. a preventive medication. In many cases of chronic or frequent episodic migraine it would be unethical to take patients off their medication for research purposes. If the preventive medication works there is often only little range for improvement. But this small range is still a factor to improve these sufferer's quality of life!* *To make it clearer, the option to use the NPI as an adjunct or stand-alone has been added throughout the text, e.g. here: "…or any other non-pharmacological intervention as an adjunct or as a stand-alone intervention should also use headache frequency as a primary outcome measure"*
4	6	45-47	for assessing the effectiveness of non-pharmacological interventions for chronic or frequent	The title just states chronic migraine and not frequent episodic (which may not be an	*Frequent episodic was added to the title*

			episodic migraine	officially recognised diagnosis). I think consistency is needed.	
5	7	31-35	Twelve experts agreed to participate and completed the first round, and ten experts completed all three rounds of the survey.	Consistency see point 1 above	*this was corrected in the abstract*
6	7-8	54-6	((((headache [Title/Abstract] OR migraine [Title/Abstract])) AND ((aerobic training or sports or exercise or acupuncture or chiropractic spinal manipulative therapy or progressive muscle relaxation or behavioural migraine management	The search seems a bit restricted. For example no osteopathy, physiotherapy , physical therapy or manual therapy?	*This is true and might give the impression of a restriction. However, you have probably realised looking at the list of authors, that the most important researchers in the field are included. Manipulative therapy and manual therapy are used within the same MeSH tree by pubmed and exercise covers most of the physiotherapy (used interchangeable by pubmed with physical therapy) field. For a Cochrane-style systematic review I agree, this is not the so-called super-sensitive search strategy, but this was never the purpose of this search.*
7	8	26-29	chronic migraine.	Just chronic migraine or frequent episodic too? See point 4 above. Just about consistency	*Changed for consistency throughout the manuscript*

2

8	9	49-54	They were further asked whether they regarded a ≥50% change in the number of headache days as an acceptable outcome level of improvement to justify a non-pharmacological treatment	Did the question ask if this was in addition to on-going medication or as a standalone NPI? It is not clear throughout the paper if you are referring to NPI as an adjunctive or standalone intervention	*This was made more clear throughout the paper by adding the option stand-alone or add-on. During the data collection process this was not used as a restricting factor.*
9	11	12-	Henry Ford-Hospital Headache	The fact that the search did not pick	*Thank you for pointing this out. I was intrigued and just did a simple search on*

		19	Disability Inventory (HDI), Headache Impact Test (Hit-6), Migraine Disability Assessment (MIDAS), Numerical Rating Scale (NRS), Pain Disability Index (PDI), Short Form-36 (SF-36), and Short Form McGill Pain Questionnaire (SF MPQ).	up MSQ 2.1 seems to support point 6 above. I know it was added later by some of the panel but did it not raise questions as to the thoroughness of the original search? Perhaps an explanation as to why this particular search was used would help readers.	*pubmed for MSQ 2.1. What does come up is a list of validation studies for the questionnaire and some recent studies on erenumab. There is not a single trial on a non-pharmacological intervention using this questionnaire.*
10	14	Table 2	Table 2	How were these views formed? Was there a suggested approach? Hit 6 has been more thoroughly evaluated than MIDAs and is the only instrument validated for CM. However there is an issue with bunching as most CM patients are in the highest category and suggestions have been made to alter it for this reason. This looks like personal guesses without a rationale. This may come across as just adding less rigour to NPI studies. Perhaps a bit more explanation or detail is possible.	*Hm, this is a comment that I find difficult to reply to. The reason why we felt we had to ask experts for their personal opinions, and this is all this is, is because there was not sufficient evidence to answer our research question. While I agree, that expert opinion is not at the top of our evidence pyramid, it was at least a combination of many opinions combined in a manner, which is recommended by the literature and conducted according to guidelines.*
11	15	32-35	A 50% reduction of headache days did not seem to be realistic for either	Given that only 4 patients with long standing migraine were interviewed I	*As much as I understand qualitative research it is not about the numbers of participants but about the quality of the data. We had to find patients, who know*

			patient for a non-pharmacological treatment	am not sure this is a valuable/valid finding. However again the question is does this relate to NPI as a standalone or adjunctive treatment? If standalone why should it not be subject to the same standards as Pharma unless patient groups come up with a trade-off they are willing to accept?	*all these outcome measures and who have been asked these questions frequently enough to form an opinion.* *The patient involvement is important because in most instances, outcome measures are applied that practitioners find useful, let's say range of motion, but that patients might not feel are particularly important because they are more interested in aspects of participation and quality of life.*
12	17	18-21	Hence, experts on non-pharmacological interventions potentially prioritise aspects of suffering and disability over the simple counting of headache days	I am not sure this statement has been made or vindicated in the results provided. They have all agreed the outcome measures best suited are the same as those used by pharma studies – Unfortunately it looks like a statement to support a pre formed view and not one that comes from the findings?	*It is stated in the discussion section that the same outcome measures as for pharma-research were suggested.*

3

13	17	28-37	Combining headache frequency and diary in the ranking task would put the combined instruments before the MIDAS and the Hit-6. The headache diary, however, while also including the number	This seems an odd approach. The point of the Delphi was to gain a consensus but this approach seems to be at odds with the methodology. Unfortunately again it could be seen as a justification for a	*The recommendation of the IHS (stated in the introduction is "* th*e use of headache frequency as the primary outcome measure, usually assessed as headache days per month documented in a headache diary." The rationale behind combining a diary and frequency is because they measure effectively the same thing: headache days per month. If you want to measure frequency there are two options, let patients have a diary for*

			of headache days, measures many additional items; therefore, it was maintained as a separate outcome measure in the current stu	pre formed view - "we wanted the diary to be in the final list and this is our justification". There are many issues with a diary which are not discussed and would add to the paper.	*a defined period of time, or let them recount retrospectively how many headache days they had during the past month (not a good idea due to recall bias).*
14	17	55-60	The highest rated outcome measures also reflected those best evaluated in the literature. MIDAS has been validated and translated in many languages [30–35] and has shown good psychometric properties [40, 43–45]. Similar research is available on HIT-6 [33,39	This appears to say the MIDAS has been better evaluated than HIT 6 which is not the case and the HIT6 is the only validated instrument for CM.	*https://www.ncbi.nlm.nih.gov/pubmed? linkname=pubmed_pubmed& from_uid=30153314 MIDAS has been developed for migraine populations while HIT 6 for general headache. And I don't agree that MIDAS has not been validated for CM, what about Bigal et al. 2003, Bagley et al. 2012. I don't particularly like the MIDAS, because it is difficult to fill out (recall and also the fact that in one question you must not add the days already counted in the previous question, ..) but it appears that HIT6 isn't much better either,….*
15	18	22-27	It is unclear why the tool was not popular with the experts in this study, since expert statements rated it as an important questionnaire.	It seems strange that you didn't go back and ask why not if you feel it is important to raise the question. Could it be because few had used it and the fact that your initial search didn't bring it up? Since this was a Delphi surely the aim/objectives was to get their views and not second guess without revisiting their views. Only 2 of the final 10 rated it as important which wouldn't	*The half-sentence that it was rated as important was a qualitative rather than a quantitative statement. Since you misunderstood it, and potentially other readers might too we deleted it. Experts are authors of this paper. Speculations were therefore changed to statements: "At this stage, there does neither seem to be neither an ideal outcome measure nor does it seem to be clear which aspects of migraine (such as it's intensity, it's impact on a person's life, it's uncertainty,…) are the most important aspects to be measured."*

				seem to qualify as 'expert statements rated as important'	
16	18	54-60	50% reduction is a target that is difficult to reach with non pharmacological interventions and lowering the target should be considered, especially since AEs are few, mild and transient [53],	I am not sure why it should be considered for lowering - only 3 of the respondents actually gave an answer and only 1 said less than 50% - assuming headache frequency is the factor in question here(it isn't clear). This again seems like a pre formed idea and not borne of the study results and it is unclear if this is as an outcome for a standalone NP intervention or an adjunctive one.	*Yes, maybe this was over interpretation when targeting frequency. It was more seen as an overall statement (for all outcome measures) but toned down to say that in general 50% is extremely high, especially when NPI is considered as an add-on to medication.*
17	19		Therefore, we promote the use of the MIDAS, the HIT-6, and headache frequency, as well as an	It is interesting that these are the same as in most pharma studies but the point is not made or felt worthy of discussion.	*Maybe you have missed it, but there is a sentence in the discussion which reads: "These are identical to the outcome measures used in pharmacological trials; "*

			outcome measure for quality of life (e.g., SF-36), which was preferred by patients and identified as a useful indicator for change in a	However I note that the option of a PRO such as the Patient Global Rating of Change (GRC) was not included or discussed either by experts or patients nor put forward by the research.	N*o and this is indeed odd and we totally agree that this should at least be mentioned in the discussion. For this purpose, the following sentence was added: This might also be the reason, why no global rating tool, such as Patient Global Rating of Change was suggested or discussed by experts.*
18	20	6-20	Conclusion section	The conclusion in the abstract does not seem to	*Thank you for pointing this out. The conclusion in the abstract has been*

				represent the actual conclusion	*changed to:* *Recommendations are for the use of the MIDAS, the HIT-6, and headache frequency, in combination with an outcome of quality of life. Associated symptoms and fear of attacks should also be considered as secondary outcome measures, if relevant for the individual target population. The cut-off level for effectiveness might be lower than for non-pharmacological trials.*
19	20	15-20	The cut-off level for effectiveness might be lower than for non- pharmacological trials since a threshold of ≥25% improvement was suggested by the expert panel.	I cannot see this as a finding in the study results. What is being discussed here? Headache frequency? It doesn't say. If that is the case then only one person said 25% so I am struggling to see how this is a conclusion. Unfortunately it seems like another pre formed view that has been inserted without justification. If there is more data to support it then I feel it should be made clearer.	*I am sorry, this is indeed not fully justifiable. We have changed this to:* *Based on the expert panel of this survey, there was no clear agreement on a specific cut-off level for any of the tools, but it seems that more realistic targets are needed to show the true effect of a non-pharmacological intervention.*

REVIEWER: 3
Reviewer Name: Ana-Carolina Goncalves
Institution and Country: University of Southampton

Congratulations on this interesting research. This is a very interesting study, looking at reaching consensus on the most useful measurement tools to be used in clinical trials of non-pharmacological interventions to manage chronic migraines.

Overall this manuscript presents interesting findings but with some important methodological limitations: a) a consensus on tools before a clear consensus on outcomes; b) absence of patients and clinicians in the Delphi panel; and c) a lack of a clear definition of consensus.

Please consider the specific comments below in order to improve the clarity of the manuscript:

Abstract:

1) The abstract would benefit from a short sentence explaining the rational for the study. Through the abstract the reader is left questioning why is this study important and why was it conducted.

Author response: Thank you. The journal guidelines did not include this, but I agree that the abstract will benefit from an explanation. We therefore added:

"This is important, since guidelines for pharmacological trials recommend measuring the frequency of headaches with 50% reduction considered a clinically meaningful effect. It is unclear whether the same recommendations apply to non-pharmacological approaches, whether the same cut-off levels need to be considered for effectiveness and whether this is meaningful to patients".

Introduction:

2) Lines 44-51. The authors argue that, to determine the effectiveness of non-pharmacological interventions it is important to agree on the use of measurement tools. What about an agreement on the most important outcomes? What is it that non-pharmacological interventions want to achieve? E.g. reductions in the use of medication? Reductions of work absence? In pain intensity? In frequency of episodes of migraines? Or a reduction of fear of symptoms? Improvements in quality of life?. Potentially all these outcomes can be measured by a different tool. Before a consensus is reached on the tools, we may need to reach a consensus on the outcomes. If this work has already been done, please make reference to it in the background. Otherwise, please justify the consensus on tools, before a consensus on outcomes in achieved.

It is difficult to respond to this comment and the methods cannot be adjusted retrospectively. We have tried to follow your recommendation by adding to the discussion that outcomes might have to be clarified. The new section reads as follows: "At this stage, there does neither seem to be neither an ideal outcome measure nor does it seem to be clear which aspects of migraine (such as it's intensity, it's impact on a person's life, it's uncertainty,…) are the most important aspects to be measured."

3) Please be aware of some inconsistency in the use of the expressions "outcome" and "outcome measure". These are two very different things (e.g. Outcome = quality of life; outcome measure or measurement tool = SF-36). The aim of the study in the abstract says "outcome measures"; the aim in the background says "outcomes". This means that by the end of the

background, the reader is still unsure if the consensus process is about outcomes, or about measurement tools.

Thank you, this has been made consistent by using outcome measure throughout the manuscript.

Methods

1) Page 6, Line 17 – selection of experts. Please provide a rationale for not inviting patients and clinicians to be part of the panel, as they may be key stakeholders in this consensus process.
Patients were involved in the process and given particular attention by not only collecting their numerical data but by listening to their opinions without pre-formed scales. They rated the same questionnaires as the researchers.

And many of the researchers were also clinicians. Since the focus of the research question was on outcome measures for non-pharma research, it was important to approach persons who are aware of outcome measures and aware of the problems arising from unrealistic effect sizes.

2) Page 6, Lines 32 to 36. I would suggest presenting the number of experts who agreed to take part in the results section rather than the methods. Further, please be aware that the information in the abstract is not consistent with the information in the methods section. The abstract suggested that 12 participants completed the 3 rounds of the Delphi. In this section, it is clear that only 10 completed all 3 rounds. Please correct this information in the abstract to avoid misleading the reader.

Thank you, this information was moved to the results section. The abstract was corrected to identify the correct numbers completing the survey.

3) Step II, page 7. The order in which the items in a Delphi survey is presented is known to influence participants' responses – please indicate if the outcome measures were presented in the same order to all survey participants, or if it was randomised.

The order was not randomised to not confuse participants. I see, that this might be interpreted as a limitation and we have added it as a limitation to the final section of the discussion. Outcome measures were presented to experts in a non-randomised order. While this might influence responses, it also helped to keep experts oriented by using a standardised order of outcome measures and response options.

4) Page 8, patient and public involvement. Please consider re-phrasing the first sentence of this paragraph. At the moment it is very long and not very clear.

Thank you. The sentence was divided and now reads: "Personal contacts with patients at the university headache clinic revealed frustration, in that access to non-pharmacological interventions is limited. Patients also repeatedly stated that there was so much more to migraine patients' suffering than the number of headache days in their diaries."

5) Page 8, patient and public involvement. Please clarify if the interviews with patients were conducted as "patient and public involvement" or formal data collection. If this was a moment of formal data collection (as it appears to be based on the results section), please explain the method used for data analysis present this information is different section of the methods (not under "patient and public involvement").

This is part of the formal data collection but shows that patients were involved in the research, I am not sure, if I understand this point correctly. The methods are described as follows: The interviews were recorded on a digital voice recorder. After the interviews, two researchers transcribed the

interview independently into personal computers to avoid potential risks of mishearing and misinterpretation [29]. Transcripts were compared and discussed before they were coded and analysed. For feasibility purpose only, the factual contents of the interviews were considered [29].

The transcript was analysed by categorisation of questions, themes and quotations. To focus on the research question, only those quotations were chosen in which the participants explained their views about the outcome measures.

6) A clear definition of consensus is missing in the methods section. An a-priori definition of consensus is essential in Delphi studies.

This was made more explicit by stating:

"Definition of consensus: no further rounds are conducted if no new outcome measures are suggested. Median values across experts will be used to identify the most useful tools. Consensus is reached if at least 75% of experts agree on the tools identified using this procedure. "

7) It would add value to the manuscript to clarify if the survey included information on the psychometric properties of the tools or copies of the actual tools – knowledge of this would have influenced the choices made by the expert panel.
Y*es, absolutely right and was forgotten to mention. We have added the following sentence: "To allow for an informed decision on test properties and previous application in research, references were provided for each outcome measure."*

Results:
1) First paragraph. Please note once more the interchangeable use of the expressions "outcomes" and "outcome measure" and refer to my previous comment.

Yes, thank you. This was made consistent by using "outcome measures" throughout the manuscript.

2) Page 10, Line 26 "no outcome was discharged after this initial round" – what would be the criteria to exclude outcomes from round to round? This is part of the definition of consensus, which is missing in the methods section.

I am sorry, you must have missed this sentence. It is stated in the methods section and reads: "All outcome measures, rated as "definitely not useful" or "probably not useful" were excluded from subsequent Delphi rounds (based on the median value across experts)."

3) Results from patient interviews: this section really highlight the problematic of selecting tools before a consensus is reached on the outcomes. Patients expressed that none of the tools measured outcomes that matter to them (e.g. fear of a migraine attack). This imposes the question: Why are we using any of these tools if they do not capture what is important to patients?

T*his cannot be adjusted at this stage. There is a tool that measures fear of attacks and I was suggested in the discussion.*

Discussion:
1) Please consider revising the discussion. It is very descriptive at the moment. I would also advise some caution making recommendations about any of these tools as the findings appear to indicate that perhaps none of the tools is ideal.

To highlight your point, we have added this section: "At this stage, there does seems to be neither an ideal outcome measure nor is it clear which aspects of migraine (such as its intensity, its impact on a person's life, its uncertainty,…) are the most important aspects to be measured."

VERSION 2 – REVIEW

REVIEWER	James Odell Bournemouth University UK
REVIEW RETURNED	13-Aug-2019

GENERAL COMMENTS	My only question is on the clarity or focus of the research question as highlighted in the abstract. At times the authors talk about Non pharmacological intervention (NPI) as an adjunctive therapy to pharmacological and at others it focuses on NPI as a stand alone. As they point out, to expect a 50% decrease over and above the pharmacological is probably unfair but if comparing a NPI against pharmacological then why wouldn't the outcome measure be the same? It is not clear throughout which situation is being considered by the experts. The conclusion doesn't really seem to address this issue despite highlighting it explicitly in the abstract "the same recommendations apply to complementary (or adjunct) non-pharmacological" - the conclusion simply says "measures might be lower for NP intervention" I am raising this as I think the authors have highlighted an important issue which is key for the future of NPI research in migraine as more often than not NPIs are used as an adjunct and as such appropriate outcome measures to establish the adjunctive benefit are needed.

REVIEWER	Ana-Carolina Goncalves University of Southampton, UK
REVIEW RETURNED	12-Sep-2019

GENERAL COMMENTS	I would like to congratulate the authors on their corrections. I believe they have addressed the reviewers' comments thoroughly and the manuscript has indeed improved. Please find below a few additional minor comments for your consideration. ABSTRACT AND INTRODUCTION: • Both sections are much clearer now and I have no further comments. METHODS: • Page 8: definition of consensus: thank you for adding this. It really improved the manuscript. May I ask to clarify the sentence: "Consensus was reached if at least 75% of experts agreed on the tools identified using this procedure?" Do these 75% correspond to 75% of the participants selecting a particular tool to their "top 3"? Is that right? If so, please specify in the definition. Further, I would also suggest you add here the definition of "consensus out". In page 9, you explain that outcome measures rated as "definitely not useful" or

	"probably not useful" were excluded. Was it the case even if just 1 participant rated the tool that way? I would suggest this is clarified and added to the definition of consensus in page 8. I understand you may not be able to change the definition of consensus retrospectively. But, if it is possible according to the data you have collected, I would suggest a definition such as: "Consensus on the relevance of an outcome measure was assumed if 80% of the participants selected the tool to their "top 3" with less than 20% of participants classifying the tool as "definitely not useful" or "probably not useful". If multiple tools meet this definition of consensus, a hierarchical ranking was used to determine the tool considered most useful from the perspective of the experts in the panel." I hope this suggestion is helpful. • Page 9: Patient and public involvement. I understand there was some confusion with my previous comment. Apologies for not making my point clearer. The expression "Patient and public involvement" means something different than including patients in interviews. "Patient and public involvement" means patients were included not as participants, but as co-researchers and they helped designing and implementing the research. Here a couple of links with more information about "patient and public involvement": https://www.invo.org.uk/find-out-more/what-is-public-involvement-in-research-2/ or http://www.healthtalk.org/peoples-experiences/improving-health-care/patient-and-public-involvement-research/what-patient-and-public-involvement-and-why-it-important I understand that what the authors have done here were patient interviews. Not patient and public involvement. To avoid confusion with readers familiar with the terminology "patient and public involvement – or PPI", I would suggest changing the subtitle from "patient and public involvement" to "Patient interviews" or "Patients' views". • Page 10, line 25: this paragraph provides a very generic description of the analysis of qualitative data. I would recommend you specify what method was used and provide an appropriate reference (e.g. thematic analysis, content analysis, framework analysis etc.) RESULTS • Page 12, line 31 "predominantly rated as don't know". Please provide specific data. What do the authors mean by "predominantly"? 7/10 participants? 5/10? 8/10? Please be specific. Lastly (and perhaps not as important) please use "do not" rather than "don't". • Please consider presenting results on the way participants ranked their top 3 tools in hierarchical order. This is described in the methods. For consistency it is important to present results of all the methods described. If the authors prefer not to present these results, please consider deleting it from the methods, whilst making sure the definition of consensus is still coherent with the methods, and results in the manuscript. • Table 1: I would leave a suggestion of adding some quantitative information about the levels of consensus against each tool presented in table 1. Like this, the reader can see at one glace how a particular tool did in the consensus exercise, and what comments were made by the expert panel.

	DISCUSSION • Page 19, lines 41-42: please note that there are still come inconsistency between the terms “outcome” and “outcome measure”. “Associated symptoms” and “fear of attacks” are not outcome measures. These are outcomes. Please revise the conclusion as well for the same inconsistency. If these concepts are still confusing, a possible suggestion would be to use the term “measurement tool” instead of “outcome measure”.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2
Reviewer Name: James Odell
Institution and Country: Bournemouth University, UK

My only question is on the clarity or focus of the research question as highlighted in the abstract. At times the authors talk about non-pharmacological intervention (NPI) as an adjunctive therapy to pharmacological and at others it focuses on NPI as a stand alone.

As they point out, to expect a 50% decrease over and above the pharmacological is probably unfair but if comparing a NPI against pharmacological then why wouldn't the outcome measure be the same? It is not clear throughout which situation is being considered by the experts.

The conclusion doesn't really seem to address this issue despite highlighting it explicitly in the abstract "the same recommendations apply to complementary (or adjunct) non-pharmacological" - the conclusion simply says "measures might be lower for NP intervention"
I am raising this as I think the authors have highlighted an important issue which is key for the future of NPI research in migraine as more often than not NPIs are used as an adjunct and as such appropriate outcome measures to establish the adjunctive benefit are needed.

Thank you very much for pointing out this interesting discussion point. When a patient decides to take e.g. topiramate or CGRP antibodies or Botox injections as a prophylactic intervention for migraine, he/she will have to consider whether the amount of the effect (e.g. reduction in headache days per month) exceeds the suffering from the side effects including constipation, cognitive deficits, cardiovascular events, nerve damage,…. Non-pharmacological interventions commonly have no or less of such unwanted effects. There is therefore not as much to trade off. This obviously doesn’t mean that NPIs should be less effective but it implies that the effect can be smaller and the patient will still be happy to receive this intervention. Currently, patients have little access to NPI because it is not seen as “effective”. This is true, if compared to CGRP antibodies which are significantly more effective, but it is untrue if the research community starts to accept that NPIs do not necessarily have to be compared to medication but can be measured on a different scale.

In this survey, we did not distinguish between these two options and after reading your comments, I am aware that this is a limitation of our study. Since this cannot be changed retrospectively, I have made this clear in the discussion:

A limitation of the survey design was that cut-off levels for effectiveness were not distinguished for studies using non-pharmacological interventions as an adjunct treatment to e.g. prophylactic medication and studies using non-pharmacological interventions as a stand-alone treatment.

However, in both situations a lower target should be used, since non-pharamcological interventions do not (or to a much lower extent) have to consider the trade-off between effect and side-effects.

Reviewer: 3
Reviewer Name: Ana-Carolina Goncalves
Institution and Country: University of Southampton, UK

I would like to congratulate the authors on their corrections. I believe they have addressed the reviewers' comments thoroughly and the manuscript has indeed improved. Please find below a few additional minor comments for your consideration.

Thank you very much for your time and effort reviewing our manuscript. We have responded to each of your comments below.

METHODS:

- Page 8: definition of consensus: thank you for adding this. It really improved the manuscript. May I ask to clarify the sentence: "Consensus was reached if at least 75% of experts agreed on the tools identified using this procedure?" Do these 75% correspond to 75% of the participants selecting a particular tool to their "top 3"? Is that right? If so, please specify in the definition. Further, I would also suggest you add here the definition of "consensus out". In page 9, you explain that outcome measures rated as "definitely not useful" or "probably not useful" were excluded. Was it the case even if just 1 participant rated the tool that way? I would suggest this is clarified and added to the definition of consensus in page 8.

I understand you may not be able to change the definition of consensus retrospectively. But, if it is possible according to the data you have collected, I would suggest a definition such as: "Consensus on the relevance of an outcome measure was assumed if 80% of the participants selected the tool to their "top 3" with less than 20% of participants classifying the tool as "definitely not useful" or "probably not useful". If multiple tools meet this definition of consensus, a hierarchical ranking was used to determine the tool considered most useful from the perspective of the experts in the panel." I hope this suggestion is helpful.

I am glad you are pointing out that this section needs to be clarified. No, 75% agreement on a tool means that at least 75% of the experts rated a suggested tool as either "useful" or "extremely useful". This was clarified by adapting the sentence you suggested to:

Consensus on the relevance of an outcome measure was assumed if 75% of the participants rated the tool as "useful" or "extremely useful". All outcome measures, rated as "definitely not useful" or

"probably not useful" were excluded from subsequent Delphi rounds (based on the median value across experts). If multiple tools meet this definition of consensus, a hierarchical ranking was used to determine the tool considered most useful from the perspective of the experts in the panel."

- Page 9: Patient and public involvement. I understand there was some confusion with my previous comment. Apologies for not making my point clearer. The expression "Patient and public involvement" means something different than including patients in interviews. "Patient and public involvement" means patients were included not as participants, but as co-researchers and they helped designing and implementing the research. Here a couple of links with more information about "patient and public involvement": https://www.invo.org.uk/find-out-more/what-is-public-involvement-in-research-2/ or http://www.healthtalk.org/peoples-experiences/improving-health-care/patient-and-public-involvement-research/what-patient-and-public-involvement-and-why-it-important

I understand that what the authors have done here were patient interviews. Not patient and public involvement. To avoid confusion with readers familiar with the terminology "patient and public involvement – or PPI", I would suggest changing the subtitle from "patient and public involvement" to "Patient interviews" or "Patients' views".

Thank you, we really liked your suggestion "patients' views and changed the subtitle accordingly.

- Page 10, line 25: this paragraph provides a very generic description of the analysis of qualitative data. I would recommend you specify what method was used and provide an appropriate reference (e.g. thematic analysis, content analysis, framework analysis etc.)

We now specified the thematic analysis approached and cited the work published by Braun and Clarke (2012).

RESULTS

- Page 12, line 31 "predominantly rated as don't know". Please provide specific data. What do the authors mean by "predominantly"? 7/10 participants? 5/10? 8/10? Please be specific.

Thank you, this has been clarified as follows:

From these seven initial tests, PDI and SF-MPQ showed a median value of 3 (rated by 3 and 4 experts, respectively as "don't know") while the remaining tests showed a median rating of 2 ("useful").

Lastly (and perhaps not as important) please use “do not” rather than “don’t”.

Thank you, this has been corrected for all instances

- Please consider presenting results on the way participants ranked their top 3 tools in hierarchical order. This is described in the methods. For consistency it is important to present results of all the methods described. If the authors prefer not to present these results, please consider deleting it from the methods, whilst making sure the definition of consensus is still coherent with the methods, and results in the manuscript.

The results of the top 3 ranking are described in the results section. This was highlighted in yellow for clarity:

For the first round:

The ranking task placed the MIDAS first, followed by the HIT-6 and the NRS by the experts.

For the last round:

The revised ranking based on outcome measures from rounds one and two indicated that the MIDAS was the most useful tool, followed by the HIT-6 and the headache frequency. Headache diary, PDI and NPRS shared rank four.

- Table 1: I would leave a suggestion of adding some quantitative information about the levels of consensus against each tool presented in table 1. Like this, the reader can see at one glace how a particular tool did in the consensus exercise, and what comments were made by the expert panel.

Thank you, this is a great idea! We have added a column indicating the rank of each tool after the final round.

DISCUSSION

- Page 19, lines 41-42: please note that there are still come inconsistency between the terms “outcome” and “outcome measure”. “Associated symptoms” and “fear of attacks” are not outcome measures. These are outcomes. Please revise the conclusion as well for the same inconsistency. If these concepts are still confusing, a possible suggestion would be to use the term “measurement tool” instead of “outcome measure”.

Thank you, this was revised accordingly.

VERSION 3 – REVIEW

REVIEWER	Jim Odell Bournemouth University, UK
REVIEW RETURNED	01-Nov-2019

GENERAL COMMENTS	Minor comment, page 14. cut off levels for HIT6. It reads as though some reviewers think a 50% reduction in the HIT6 score is a suitable cutoff. e.g dropping from 60 to 30 This is virtually impossible with migraine let alone CM. So I may have misunderstood, it might be worth considering the clarity of the statement or explanation. Thank you for raising this issue of measurement instruments for non pharma/adjunctive studies

REVIEWER	Ana-Carolina Goncalves University of Southampton and Western Sussex Hospitals NHS Foundation Trust
REVIEW RETURNED	27-Nov-2019

GENERAL COMMENTS	Congratulations on this very interesting work and on the recent improvements to the manuscript. Please find some feedback below, which I hope can be useful to further improve the manuscript. SECTION ON STRENGTHS AND LIMITATIONS (below the abstract) 1) This section does not include any limitations. Only strengths. Please acknowledge important limitations of this work here, including the lack of patients and clinicians involved in the Delphi as key stakeholders. BACKGROUND 1) Much clearer. No further comments METHODS 1) The definition of consensus is now much clearer, thank you for addressing this comment. 2) Patient and public involvement – Thank you for looking into my guidance about what constitutes “patient and public involvement”. If patient and public involvement activities were not conducted, I would suggest simply deleting the section called “patient and public involvement” and only keeping the section named “Patients views” 3) Patient views - Please consider deleting the first sentence in this section “personal contacts with patients…”. Instead, a rationale is needed to having interviewed patients rather than include them in the Delphi alongside the other stakeholders. 4) More information is need on how were the patients recruited and what was the sampling strategy used. 5) The sentence “Only the factual contents of the interviews were considered” needs clarification. What do the authors mean by this? Otherwise consider deleting this sentence RESULTS 1) Consider using subheadings under the results section, in order to allow the reader to follow the information more easily. The following

	subheadings may be considered: Sample characterisation; Delphi round one; Delphi round two; Delphi round three; results from patient interviews. 2) Table 1 – I would suggest presenting the tools according to the final ranking order, so that the reader can easily find “the most important” tool on top of the table. 3) Table 3 – please avoid the use of closed questions in qualitative interviewing. Here some suggestions “what do you like/dislike about these tools?”; “what would a significant reduction in headache look like?” DISCUSSION 1) I would suggest starting the discussion with one paragraph summarising the study e.g. “The present study reports on an international Delphi survey, aiming to reach consensus on the measurement tools to be used in non-pharmacological intervention for migraine. Tools X and Y reached the definition of consensus and were ranked as the most relevant tools by the Delphi participants; conversely, patients valued Z and W”. 2) Please consider deleting the discussion about the “headache frequency” and “headache diary” being two separate tools or not. Multiple can measure the same outcome, but they are still different tools – I believe discussion about this is not necessary. 3) The paragraph starting with “Expert opinions varied widely…” is describing results rather than discussing them. Why is it important that experts’ opinions varied so much? Does that mean that experts will continue to choose using different tool and therefore the studies will not be comparable? Do we need new tools as none was considered “ideal”? 4) Reference to Wang et al is missing the year of publication. 5) Overall, I believe the discussion would benefit from being more focused and clearly stating: what do these results mean, why do they matter and what are the recommendations to future research. 6) As an additional reflection: Could it be controversial to refer to professionals “experts” and patients, simply as “patients”. Aren’t the interviewed patients’ experts in this field?

VERSION 3 – AUTHOR RESPONSE

Reviewer: 2
Reviewer Name: Jim Odell
Institution and Country: Bournemouth University, UK
Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
Minor comment, page 14. cut off levels for HIT6. It reads as though some reviewers think a 50% reduction in the HIT6 score is a suitable cutoff. e.g dropping from 60 to 30 This is virtually impossible with migraine let alone CM. So I may have misunderstood, it might be worth considering the clarity of the statement or explanation. Thank you for raising this issue of measurement instruments for non pharma/adjunctive studies

Author response:

Thank you for commenting on this. Indeed, this was suggested by one person. In the discussion, a more realistic value of a 2.5 to 6 point reduction is highlighted and the subsequent statement on the

50% reduction only focusses on headache frequency. This has now been clarified by stating: "*A ≥50% reduction of headache frequency is a target that is difficult to reach with non-pharmacological interventions, especially when provided as an adjunct to preventive medication and especially in chronic migraine"*

Reviewer: 3
Reviewer Name: Ana-Carolina Goncalves
Institution and Country: University of Southampton and Western Sussex Hospitals NHS Foundation Trust
Please state any competing interests or state 'None declared': none to declare

Please leave your comments for the authors below
Congratulations on this very interesting work and on the recent improvements to the manuscript. Please find some feedback below, which I hope can be useful to further improve the manuscript.

SECTION ON STRENGTHS AND LIMITATIONS (below the abstract)
1) This section does not include any limitations. Only strengths. Please acknowledge important limitations of this work here, including the lack of patients and clinicians involved in the Delphi as key stakeholders.

Author response:

A statement on the limitation that patients were not part of the Delphi process was included.

BACKGROUND
1) Much clearer. No further comments

METHODS
1) The definition of consensus is now much clearer, thank you for addressing this comment.
2) Patient and public involvement – Thank you for looking into my guidance about what constitutes "patient and public involvement". If patient and public involvement activities were not conducted, I would suggest simply deleting the section called "patient and public involvement" and only keeping the section named "Patients views"

Author response:

Thank you for the suggestion. However, the editor reminded us, that the section "patient and public involvement" is a journal requirement and we were asked to keep this section.

3) Patient views - Please consider deleting the first sentence in this section "personal contacts with patients…". Instead, a rationale is needed to having interviewed patients rather than include them in the Delphi alongside the other stakeholders.

Author response:

The first sentence was deleted. As a justification for treating patients differently from researchers, the following statement was added: "*Rather than including them in the Delphi process, patients were invited to take part in this study by being interviewed. It was anticipated that patients' views were*

more multifaceted and diverse and that the ranking tasks requested during the Delphi rounds would do their opinions not sufficient justice."

4) More information is need on how were the patients recruited and what was the sampling strategy used.

Author response:

The patients were a convenience sample of patients known to the researcher. This was clarified in the manuscript.

5) The sentence "Only the factual contents of the interviews were considered" needs clarification. What do the authors mean by this? Otherwise consider deleting this sentence

Author response:

The sentence was deleted.

RESULTS
1) Consider using subheadings under the results section, in order to allow the reader to follow the information more easily. The following subheadings may be considered: Sample characterisation; Delphi round one; Delphi round two; Delphi round three; results from patient interviews.

Author response:

Thank you for this suggestion, subheadings were added to the results section.

2) Table 1 – I would suggest presenting the tools according to the final ranking order, so that the reader can easily find "the most important" tool on top of the table.

Author response:

Thank you, table 1 was restructured.

3) Table 3 – please avoid the use of closed questions in qualitative interviewing. Here some suggestions "what do you like/dislike about these tools?"; "what would a significant reduction in headache look like?"

Author response:

Thank you. “What do you like about these tools?” was the correct question which was asked; the “what” was missing after copy & pasting the text and it was now inserted, again. The third question was asked the way it was reported and I am afraid I cannot change this retrospectively.

DISCUSSION

1) I would suggest starting the discussion with one paragraph summarising the study e.g. “The present study reports on an international Delphi survey, aiming to reach consensus on the measurement tools to be used in non-pharmacological intervention for migraine. Tools X and Y reached the definition of consensus and were ranked as the most relevant tools by the Delphi participants; conversely, patients valued Z and W”.

Author response:

Thank you. An introductory sentence was included as suggested.

2) Please consider deleting the discussion about the “headache frequency” and “headache diary” being two separate tools or not. Multiple can measure the same outcome, but they are still different tools – I believe discussion about this is not necessary.

Author response:

This section has been deleted.

3) The paragraph starting with “Expert opinions varied widely…” is describing results rather than discussing them. Why is it important that experts’ opinions varied so much? Does that mean that experts will continue to choose using different tool and therefore the studies will not be comparable? Do we need new tools as none was considered “ideal”?

Author response:

A sentence reflecting on this was added: “*To prevent researchers from using different tools in future research and thereby not allowing for the comparability of results, these limitations should be addressed by e.g. providing a version of the MIDAS only reflecting on the past 4 weeks rather than on the past 3 month.”*

4) Reference to Wang et al is missing the year of publication.

Author response:

Thank you for spotting this, the year was added.

5) Overall, I believe the discussion would benefit from being more focused and clearly stating: what do these results mean, why do they matter and what are the recommendations to future research.

Author response:

Thank you. We believe that the discussion clearly states that

"At this stage, there does not seem to be an ideal outcome measure. Neither does it seem to be clear which aspects of migraine (such as it's intensity, it's impact on a person's life, it's uncertainty,…) are the most important aspects to be measured."

The recommendation is stated as:

"*we promote the use of the MIDAS, the HIT-6, and headache frequency, and an outcome measure for quality of life (e.g., SF-36), which was preferred by patients and recently identified as a useful indicator for change"…Associated symptoms and fear of attacks should be considered as secondary outcomes.*

6) As an additional reflection: Could it be controversial to refer to professionals "experts" and patients, simply as "patients". Aren't the interviewed patients' experts in this field?

Author response:

Yes, patients are definitely experts for their symptoms. We have added the expression patient experts wherever it was possible in the text.