PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Comparative effectiveness of three versions of a stepped care model for insomnia differing in the amount of therapist support in internet-delivered treatment – study protocol for a pragmatic cluster randomised controlled trial (GET Sleep)
AUTHORS	Spiegelhalder, Kai; Baumeister, Harald; Al-Kamaly, Abdulwahab; Bader, Martina; Bauereiss, Natalie; Benz, Fee; Braun, Lina; Buntrock, Claudia; Burkhardt, Maike; Cuijpers, Pim; Domschke, Katharina; Dülsen, Patrick; Franke, Marvin; Frase, Lukas; Heber, Elena; Helm, Kathrin; Jentsch, Terry; Johann, Anna; Küchler, Ann-Marie; Kuhn, Michael; Lehr, Dirk; Maun, Andy; Morin, Charles M.; Moshagen, Morten; Richter, Kneginja; Schiel, Julian; Simon, Laura; Spille, Lukas; Weeß, Hans-Günter; Riemann, Dieter; Ebert, David

VERSION 1 – REVIEW

REVIEWER	David Ellard University of Warwick, Clinical Trials Unit
REVIEW RETURNED	13-Dec-2021

GENERAL COMMENTS	Thank you for inviting me to review this interesting protocol. This is a very ambitious randomised controlled trial with over 4000 participants recruited from general practices. I note this as an ambitious project because of the sample size and the fact that it is a complex four arm trial, which will be challenging. Generally the methodology is presented really well. I cannot however comment on the statistical set up in this trial as it is very complex (e.g. sample sizes and planned analyses) and beyond my expertise. I do have a number of comments. Firstly, I am somewhat surprised by the lack of patient and public involvement (PPI) in such a large trial. The outcome of which is not just important to clinicians! Overall the plan is to try and make life better for a large number of people who live with insomnia. Thus I feel patients and the public should have a voice in the project throughout. Having PPI working with the team would help to ensure that questions and interpretation are patient focussed and not just for a clinical audience. Qualitative interviews are mentioned but I feel they are skipped over and feel a little like an afterthought. In my opinion these are a really important part of the trial as it is no good having an intervention that you can deliver but that people don't like it. There is scant information on the analysis of qualitative data or any theoretical underpinning behind it. I note an economic evaluation is mentioned in the title and in the analysis section but it is not mentioned how this will be carried out

	(e.g. where is the data coming from?). I wish the authors well with their future work

REVIEWER	Helen Parsons University of Warwick, Division of Health Sciences
REVIEW RETURNED	16-Dec-2021

GENERAL COMMENTS	This paper described a large RCT for a complex intervention to aid patients with insomnia. The study design is ambitious, and so requires care when writing to ensure the reader understands. Presently, the protocol is not suitable for publication. I am uncertain if all issues can be addressed with clarification, but many issues can. The explanation of the interventions under study is confusing. I can find no explanation of what the “TAU” group receive. It is also unclear when participants are being randomised. As “step 1” is given under the intervention description, this implies that randomisation occurs before this step. However, from the description, the components under investigation are all in “step 2” and there is a possibility that participants will only require step 1. Furthermore, given that GPs can opt for participants to skip “step 1”, what is the purpose of including this step in the intervention? Surely randomising after the failure of “step 1” would allow for less dilution of treatment effects? Skipping “step 1” also complicated the timeline of follow up, as some participants will have not started “step 2” and others will be partly or fully completed four weeks post randomisation. It is also unclear why “step 3” is included in the study – this is for participants who fail step 2 and is applied equally to all randomisation groups. If this is to test for interactions between the randomisation groups and “step 3” then the study is very much underpowered. Of course, if this is not how the interventions work and participants are randomised at a different time, then this needs to be made clear The sample size calculation section is very badly written. The authors imply that an unbalanced design is used for reduced costs; which is confusing as efficiency is greatest with an equal allocation ratio (e.g. Julious, Steven A. "Sample sizes for clinical trials with normal data." Statistics in medicine 23.12 (2004): 1921-1986.). If this refers to the multiple comparators with control, this efficiency is usually gained from a larger control/TAU group (e.g. Machin, David, et al. Sample size tables for clinical studies. John Wiley & Sons, 2011.). Please add explanation on how this reduction in costs arise. Is it due to administration reasons, perhaps? Furthermore, I cannot replicate the sample size calculation as key parameters are missing and those which are there need to be justified. As a minimum, I would expect to see an explanation of the choice of cluster size, the ICC and correlation and number of observations included. An explanation of the unclustered sample size would also be useful as this is a complicated study and slight differences in rounding may make larger changes after applying the design effect. Other than this, the authors account for loss to follow up in their planned analysis, but lack any description of how this was accounted for in their sample size. Given this is likely to affect cluster sizes, considering drop out is critical at this stage. I am concerned about the effects of recruiting outside of the GP clusters on the model. Since the intervention is not delivered by

	group, it seems reasonable to assume that most of the subject correlation will be from patient similarity from that practice (e.g. levels of deprivation, etc). By grouping all self-referred participants together, I am not sure the correlation structure would be the same. The analysis as written is appropriate for a two parallel allocation groups; but needs amending to account for the four study groups. Throughout, I have assumed that pairwise comparisons are made (ie. Each group is tested against the three others), but this is not explicitly stated. Whilst accounting for type I error rates in multi armed studies does not have consensus, I expect to see at least some discussion of why no adjustments have been made. Especially given that the intervention arms of this study are essentially different dose regiments of the intervention, I would expect to see some control of the FWER (Wason, J.M.S., Stecher, L. & Mander, A.P. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done?. Trials 15, 364 (2014). https://doi.org/10.1186/1745-6215-15-364). The authors should also consider specifying which baseline variables will be included in their model, as cluster randomised studies often suffer from unbalanced demographics due to the unit of randomisation being different from the unit of analysis. Further details of the imputation process are also needed. For examples, if the imputed data is considered as a sensitivity or secondary analysis. The analysis of the subgroups should be explicitly stated in the analysis section with details of how this will be modelled The safety protocol section was poorly written or considered. It seems a simple step to ensure that their GP are notified that a participant is in distress/requires intervention, rather than just giving the participant information on how to contact their GP. It is disappointing that no patients have been included in the design so far. Given that there is no accepted minimum clinically important difference, this seems like an ideal situation where patient input would add great benefits. Furthermore, patient input on the acceptability of the interventions and length of the collected assessments would reassure me that the study will be successful. The authors should consider and justify their assessments. By my count, there are 17 outcomes collected at T2, plus the collection of sleep diaries and safety data. This seems like a high burden on participants, especially as the intervention at just step 2 is described as eight 60 minute sessions. Many of the assessments appear to be collecting very similar information. I also question the practicality and usefulness of giving yet another, unvalidated, questionnaire ("Dropout Questionnaire") to participants who do not complete the intervention. Minor comments The protocol includes sufficient detail to give the broad outline of the planned analyses, but the authors should consider developing/referring to a detailed statistical analysis plan to detail exactly what analyses are planned for the trial analysis and which are more exploratory. For example, the moderator and mediation analyses are very vague and sound like a potential "fishing" analysis (liable to be susceptible to p-hacking, etc); but are appropriate for exploratory analyses.

	Consider subheadings to separate efficacy, economic and qualitative analyses to aid the reader. The last paragraph in section 2.1 refers to noninferiority trials, this should be superiority trials The payment for taking part in a trial needs clarification. Presently, it seems that participants are rewarded only if they return assessments. To my understanding, payment should be irrespective if they return assessment as this could be perceived as (mild) inducement “Unauthorised access to the platform is not possible” is a very strong statement, especially given the recent log4j vulnerability (amongst others). I suspect that the authors mean that they have taken care to ensure robust security for the platform. I do not believe the study is “blinded with respect to group allocation and outcomes”. Other than asking for more explanation on how a statistician can be blinded to the data they are analysing, it will be obvious which allocation group is TAU because of the allocation ratio. Furthermore, given the plethora of data the study is collecting, it is possible that participant feedback will also unblind team members analysing (e.g. additional human guidance is not available for IG3)

REVIEWER	John Cape University College London, Clinical, Educational and Health Psychology
REVIEW RETURNED	23-Dec-2021

GENERAL COMMENTS	This is an ambitious and impressive planned study. The sample of 4,268 participants would indeed I believe be the largest ever clinical trial on insomnia and quite possibly the largest ever clinical trial including a psychological intervention for any condition. The background and rationale are well covered and argued, including the very low uptake of CBT for insomnia despite clinical guideline recommendations that it should be the first line treatment, the relevance of a treating insomnia within a stepped care model embedded in primary medical care and uncertainty about the optimal and most cost-effective level of support required for effective internet-based CBT-I treatment. A couple of general points about the protocol: 1. While the title is Effectiveness of a stepped care model, the design and power is focused on comparative effectiveness of the 3 different levels of support for internet-based CBT-I in step 2. With respect to stepped care, there is little focus on step 1 and no data to be collected on step 3 as far as I could see. This is not necessarily a problem, but it would be helpful for those doing literature searches on support for internet-based interventions if the title made clearer this focus 2. In my reading of the abstract and introduction, I had assumed that GPs would be doing the support for the internet CBT-I, so it was a bit of a surprise to see this was not the case when I got to the Methods. Could be others won’t read it in this way, but maybe the abstract could make clearer who supports the step 2 internet CBT-I Otherwise just a few specific points for clarification of methods that would help inform any replication and regarding the analysis:

	3. Are the main inclusion and exclusion criteria (ICD insomnia, exclusion of other sleep disorders, ICD mental health etc) assessed by the GP or from the screening questionnaire to participants? 4. At step 1, will a leaflet with the sleep education messages be given to patients as is common for psychoeducational interventions? If not, fine, but helpful to be clear in the protocol 5. Is movement from step 1 to step 2 entirely patient determined or is there a check in with the GP? And what is the process – does the patient get automatically sent a link after 4 weeks at step 1 to register? And do all patients get this? 6. What is duration of step 2 and is this standardized? My assumption is that the 12-week assessment point is planned to be at end of step 2 and before any patient would move on to step 3 7. What communication/feedback to GPs if any will there be from the step 2 CBT-I intervention either automated from the delivery software or from the psychologists providing the support? 8. For the move from step 2 to step 3, a criterion of ISI ≥ 15 is mentioned as guidance to GPs. Will GPs be informed of patients' ISI scores from their 12-week post step 2 research assessments or will GPs administer the ISI themselves to patients outside the research process? And whichever of these two takes place, given the ISI is the study primary outcome measure will any precautions be taken to blind patients as to their scores? 9. At step 3, it would be helpful to clarify what a sleep medicine specialist is in Germany and what kind of interventions they would carry out as I suspect this title does not travel and could mean different things in different countries 10. The Brief Fatigue Inventory is included in the Table but is not mentioned in the relevant Measures section of the text 11. The Study Design section includes as part of the text on CONSORT reporting that the study will be reported in accordance with the extensions for reporting ….. noninferiority trials. However, at no other point in the paper is a non-inferiority analysis mentioned. Please clarify. 12. What reliability parameter will be used to calculate the Reliable Change Index? Will this be pre-specified from one of the published Cronbach alpha estimates or calculated from the study data? And if clinical significant change is also to be calculated using the Jacobson and Truax approach (unclear from the text whether this is planned) what ISI clinical significance threshold will be used?

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: Dr. David Ellard, University of Warwick

Thank you for inviting me to review this interesting protocol. This is a very ambitious randomised controlled trial with over 4000 participants recruited from general practices. I note this as an ambitious project because of the sample size and the fact that it is a complex four arm trial, which will be challenging. Generally the methodology is presented really well. I cannot however comment on the statistical set up in this trial as it is very complex (e.g. sample sizes and planned analyses) and beyond my expertise.

Authors' Response: We thank the reviewer for the positive evaluation of our manuscript.

I do have a number of comments. Firstly, I am somewhat surprised by the lack of patient and public involvement (PPI) in such a large trial. The outcome of which is not just important to clinicians! Overall the plan is to try and make life better for a large number of people who live with insomnia. Thus I feel patients and the public should have a voice in the project throughout. Having PPI working with the

team would help to ensure that questions and interpretation are patient focussed and not just for a clinical audience.

Authors' Response: The insomnia researchers among the lead authors, Kai Spiegelhalder and Dieter Riemann, cooperate with two patient groups in several projects. In particular, representatives of the 'Allgemeiner Verband Chronische Schlafstörungen Deutschland e.V.' and 'Selbsthilfegruppe Ein- und Durchschlafstörungen am Pfalzklinikum Klingenmünster' regularly participate in meetings of these researchers. However, the current funding scheme (Innovationsfonds of the German Federal Joint Committee) does not specifically encourage or support a formal involvement of patient groups. Because of this, we have stated in the first submission of our study protocol that there is no patient involvement. However, we believe that it is more precise to state that 'Representatives of patients groups were not formally involved in the design of this study but will be involved in the discussion and dissemination of results. In addition, patients were involved in user-experience and usability testing of the platform for the internet intervention in order to ensure that the interface is user-friendly and adaptive to factors related to age, gender, and education'. Consequently, we have added this information in the manuscript (page 21, lines 11-15 of the redline version of the revised manuscript).

Qualitative interviews are mentioned but I feel they are skipped over and feel a little like an afterthought. In my opinion these are a really important part of the trial as it is no good having an intervention that you can deliver but that people don't like it. There is scant information on the analysis of qualitative data or any theoretical underpinning behind it.

Authors' Response: Following the advice of the reviewer, we have now provided more details on the collection and analysis of qualitative data in our study (pages 20-21).

I note an economic evaluation is mentioned in the title and in the analysis section but it is not mentioned how this will be carried out (e.g. where is the data coming from?).

Authors' Response: Please refer to sections 2.7.2. and 2.7.5. for a description of the data assessment for the health-economic evaluation. Health-care utilisation, patient and family expenditures and productivity losses due to absence from work or reduced efficiency during paid and unpaid work will be established with the Trimbos/iMTA questionnaire for costs associated with psychiatric illness (TiC-P), a retrospective self-report questionnaire covering the previous three months. A list of unit cost prices will be used to compute health care costs on a per-participant basis. In addition, secondary data from BARMER will be used to assess the validity of the TiC-P.

I wish the authors well with their future work

Authors' Response: Thank you very much!

Reviewer 2: Dr. Helen Parsons, University of Warwick

This paper described a large RCT for a complex intervention to aid patients with insomnia. The study design is ambitious, and so requires care when writing to ensure the reader understands. Presently, the protocol is not suitable for publication. I am uncertain if all issues can be addressed with clarification, but many issues can.

We thank the reviewer for the thoughtful evaluation of our manuscript and the constructive comments that helped us to improve our work.

The explanation of the interventions under study is confusing. I can find no explanation of what the "TAU" group receive.

Authors' Response: Thank you for this comment. We have now included an additional section in the manuscript describing the treatment in the TAU group (please refer to section 2.6. of the redline version of the revised manuscript, page 13, lines 1-8). Moreover, we have revised the description of the interventions in several places following the advice of the reviewers.

It is also unclear when participants are being randomised. As "step 1" is given under the intervention description, this implies that randomisation occurs before this step. However, from the description, the components under investigation are all in "step 2" and there is a possibility that participants will only require step 1. Furthermore, given that GPs can opt for participants to skip "step 1", what is the purpose of including this step in the intervention? Surely randomising after the failure of "step 1" would allow for less dilution of treatment effects? Skipping "step 1" also complicated the timeline of follow up, as some participants will have not started "step 2" and others will be partly or fully completed four weeks post randomisation. It is also unclear why "step 3" is included in the study – this is for participants who fail step 2 and is applied equally to all randomisation groups. If this is to test for interactions between the randomisation groups and "step 3" then the study is very much underpowered. Of course, if this is not how the interventions work and participants are randomised at a different time, then this needs to be made clear.

Authors' Response: Participants are not randomised in this trial. Instead, the study will use cluster randomisation of GPs which has been stated multiple times in the manuscript. Please note also that GPs are randomised in four arms which includes TAU as control condition. TAU does not include any step of the stepped care model. We hope that this addresses all of the concerns of the reviewer's comment.

The sample size calculation section is very badly written. The authors imply that an unbalanced design is used for reduced costs; which is confusing as efficiency is greatest with an equal allocation ratio (e.g. Julious, Steven A. "Sample sizes for clinical trials with normal data." Statistics in medicine 23.12 (2004): 1921-1986.). If this refers to the multiple comparators with control, this efficiency is usually gained from a larger control/TAU group (e.g. Machin, David, et al. Sample size tables for clinical studies. John Wiley & Sons, 2011.). Please add explanation on how this reduction in costs arise. Is it due to administration reasons, perhaps?

Authors' Response: The sample size calculation is based on the comparisons between the IGs and there is an equal allocation ratio for these three groups. Having an equal allocation ratio for all four groups would increase the number of GPs randomised into the TAU group and the number of patients receiving TAU by factor 3. In our view, this would be ethically problematic since we assume that TAU is inferior to the treatment in the IGs, and since we do not need this sample size for the TAU group to detect the expected differences between TAU and the IGs. In addition, a larger overall sample size is, of course, associated with increased costs. We have now added the information that the sample size calculation is based on the comparisons between the IGs in the revised version of the manuscript (page 18, line 3).

Furthermore, I cannot replicate the sample size calculation as key parameters are missing and those which are there need to be justified. As a minimum, I would expect to see an explanation of the choice of cluster size, the ICC and correlation and number of observations included. An explanation of the unclustered sample size would also be useful as this is a complicated study and slight differences in rounding may make larger changes after applying the design effect. Other than this, the authors account for loss to follow up in their planned analysis, but lack any description of how this was

accounted for in their sample size. Given this is likely to affect cluster sizes, considering drop out is critical at this stage.

Authors' Response: To our regret, we do not understand the reviewer's concern about missing key parameters of the sample size calculation. We have provided the following parameters: number of clusters: 320; cluster size: 9 ± 14 patients; ICC = 0.02; correlation of the outcome with the baseline assessment of r = 0.5; effect size d = 0.25; α = 0.05; (1-β) = 80%. Cluster size and standard deviation were determined based on experiences from prior clinical trials in varying digital and GP health care settings. To our knowledge, there is no systematic investigation of ICC values in insomnia-related primary care research. Because of this, the ICC of 0.02 was based on primary care studies in the field of depression (see Adams et al., J Clin Epidemiol 2004;57:785-94). Since correlations of post-treatment outcomes with baseline assessment are usually not reported in the insomnia field (see, e.g., a corresponding statement by Koffel et al., Sleep Med Rev 2015;19:6-16), chosing r = 0.5 was based on a recent own clinical trial (Johann et al., J Sleep Res 2020;29:e13102). Please note that the primary analysis of this trial will be conducted according to the intention-to-treat principle based on all patients with their original treatment allocation. Because of this, potential drop-outs are not accounted for in the sample size calculation.

I am concerned about the effects of recruiting outside of the GP clusters on the model. Since the intervention is not delivered by group, it seems reasonable to assume that most of the subject correlation will be from patient similarity from that practice (e.g. levels of deprivation, etc). By grouping all self-referred participants together, I am not sure the correlation structure would be the same.

Authors' Response: We thank the reviewer for raising this point. In our view, for reaching the envisioned sample size, it is of fundamental importance to include a group of GPs that uses telehealth consultations to treat patients that are recruited by online, print and broadcast media advertisements as well as postal mailings by the BARMER. While we agree with the general concern of the reviewer, we believe that there is no alternative to grouping these GPs together in one stratum for the stratified permuted block randomisation. We will discuss this as a study limitation.

The analysis as written is appropriate for a two parallel allocation groups; but needs amending to account for the four study groups. Throughout, I have assumed that pairwise comparisons are made (ie. Each group is tested against the three others), but this is not explicitly stated. Whilst accounting for type I error rates in multi armed studies does not have consensus, I expect to see at least some discussion of why no adjustments have been made. Especially given that the intervention arms of this study are essentially different dose regiments of the intervention, I would expect to see some control of the FWER (Wason, J.M.S., Stecher, L. & Mander, A.P. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done?. Trials 15, 364 (2014). https://doi.org/10.1186/1745-6215-15-364).

The authors should also consider specifying which baseline variables will be included in their model, as cluster randomised studies often suffer from unbalanced demographics due to the unit of randomisation being different from the unit of analysis. Further details of the imputation process are also needed. For examples, if the imputed data is considered as a sensitivity or secondary analysis.

Authors' Response: We thank the reviewer for these points. Indeed, pairwise comparisons between groups will be performed and this information is now included in the revised manuscript (page 18, lines 22-23). The alpha level of the main analyses will be adjusted using the Bonferroni-Holm procedure and this is now also mentioned in the manuscript (page 18, line 24-25). Concerning the inclusion of baseline variables in the model, we have now specified that baseline insomnia severity scores, age and gender will be included. However, please note that the stratification aims at minimising any imbalance in demographics. To our regret, we were not able to fully understand the

comment of the reviewer regarding the imputation process. With all due respect, we would like to ask the reviewer to explain the idea behind this comment in more detail.

The analysis of the subgroups should be explicitly stated in the analysis section with details of how this will be modelled.

Authors' Response: Thank you for this comment. The analyses will be conducted separately for each subgroup which is now explicitly stated in the revised manuscript (page 18, line 24).

The safety protocol section was poorly written or considered. It seems a simple step to ensure that their GP are notified that a participant is in distress/requires intervention, rather than just giving the participant information on how to contact their GP.

Authors' Response: We agree with the reviewer that it would be very simple to inform GPs about participants with suicidal ideation. However, there are two reasons why we did not include this in our safety protocol. First, this is an issue of data security and data protection. All patients would have to give their consent that the information is given to the responsible GP and this may negatively affect acceptance of the trials´ intervention. Second, as part of the safety protocol, we inform patients about several different available health services, and the GP may not be the most qualified person to consult for suicidal ideation. Please note, that the Ethics Committee of the Medical Center – University of Freiburg and the Ethics Committee of the State Chamber of Physicians ('Landesärztekammer Baden-Württemberg') have thoroughly checked and approved this safety protocol.

It is disappointing that no patients have been included in the design so far. Given that there is no accepted minimum clinically important difference, this seems like an ideal situation where patient input would add great benefits. Furthermore, patient input on the acceptability of the interventions and length of the collected assessments would reassure me that the study will be successful.

Authors' Response: The insomnia researchers among the lead authors, Kai Spiegelhalder and Dieter Riemann, cooperate with two patient groups in several projects. In particular, representatives of the 'Allgemeiner Verband Chronische Schlafstörungen Deutschland e.V.' and 'Selbsthilfegruppe Ein- und Durchschlafstörungen am Pfalzklinikum Klingenmünster' regularly participate in meetings of these researchers. However, the current funding scheme (Innovationsfonds of the German Federal Joint Committee) does not specifically encourage or support a formal involvement of patient groups. Because of this, we have stated in the first submission of our study protocol that there is no patient involvement. However, we believe that it is more precise to state that 'Representatives of patients groups were not formally involved in the design of this study but will be involved in the discussion and dissemination of results. In addition, patients were involved in user-experience and usability testing of the platform for the internet intervention in order to ensure that the interface is user-friendly and adaptive to factors related to age, gender, and education'. Consequently, we have added this information in the revised version of the manuscript (page 21, line 11-15).

The authors should consider and justify their assessments. By my count, there are 17 outcomes collected at T2, plus the collection of sleep diaries and safety data. This seems like a high burden on participants, especially as the intervention at just step 2 is described as eight 60 minute sessions. Many of the assessments appear to be collecting very similar information. I also question the practicality and usefulness of giving yet another, unvalidated, questionnaire ("Dropout Questionnaire") to participants who do not complete the intervention.

Authors' Response: We thank the reviewer for critically reviewing our assessment plan. Before the inclusion of the first patient, each of four research assistants at Ulm University have completed the

baseline questionnaires twice to determine the burden on participants. These individuals needed between 25 min 58 s and 1 hour 11 min 10 s to do so. In light of these data, we believe that the overall benefit for patients participating in this trial exceeds the burden of the assessments. Please note also that, in our view, there is no substantial overlap between different measures. In addition, we would like to emphasise that drop-outs are an important limitation of internet interventions for mental health (see e.g., Melville KM et al. Br J Clin Psychol 2010;49:455-71). Thus, identifying potential dropout reasons is a major aim of this trial, and, since there is no adequate validated measure, we would prefer to keep the self-developed questionnaire which will be validated in the current trial.

Minor comments: The protocol includes sufficient detail to give the broad outline of the planned analyses, but the authors should consider developing/referring to a detailed statistical analysis plan to detail exactly what analyses are planned for the trial analysis and which are more exploratory. For example, the moderator and mediation analyses are very vague and sound like a potential “fishing” analysis (liable to be susceptible to p-hacking, etc); but are appropriate for exploratory analyses.

Authors’ Response: The primary effectiveness analysis is described in section 2.10.1., and both the moderator and the mediator analyses were explicitly labelled as ‘exploratory’ (please refer to page 19). Thus, to our regret, it is not clear to us what additional information the reviewer would like to see in the manuscript.

Consider subheadings to separate efficacy, economic and qualitative analyses to aid the reader.

Authors’ Response: Thank you for this advice, we have now included subheadings in the section on the statistical analysis.

The last paragraph in section 2.1 refers to noninferiority trials, this should be superiority trials

Authors’ Response: We would like to apologise for this mistake and thank the reviewer for noticing this. The first author of the manuscript has recently received funding for a noninferiority trial and has confused the two studies when writing this sentence in the methods section. We have now removed the reference to the extension of the CONSORT 2010 statement for reporting noninferiority trials.

The payment for taking part in a trial needs clarification. Presently, it seems that participants are rewarded only if they return assessments. To my understanding, payment should be irrespective if they return assessment as this could be perceived as (mild) inducement.

Authors’ Response: Indeed, participants receive payment after completion of the online assessments. In our view, this is a common and legitimate strategy to increase adherence.

“Unauthorised access to the platform is not possible” is a very strong statement, especially given the recent log4j vulnerability (amongst others). I suspect that the authors mean that they have taken care to ensure robust security for the platform.

Authors’ Response: We agree with the reviewer and have re-written the corresponding sentence in the revised version of the manuscript (page 12, lines 14-15).

I do not believe the study is “blinded with respect to group allocation and outcomes”. Other than asking for more explanation on how a statistician can be blinded to the data they are analysing, it will be obvious which allocation group is TAU because of the allocation ratio. Furthermore, given the plethora of data the study is collecting, it is possible that participant feedback will also unblind team members analysing (e.g. additional human guidance is not available for IG3)

Authors' Response: We thank the reviewer for noticing this, and we agree. We have now removed the corresponding sentence from the revised manuscript.

Reviewer 3: Prof. John Cape, University College London

This is an ambitious and impressive planned study. The sample of 4,268 participants would indeed I believe be the largest ever clinical trial on insomnia and quite possibly the largest ever clinical trial including a psychological intervention for any condition. The background and rationale are well covered and argued, including the very low uptake of CBT for insomnia despite clinical guideline recommendations that it should be the first line treatment, the relevance of a treating insomnia within a stepped care model embedded in primary medical care and uncertainty about the optimal and most cost-effective level of support required for effective internet-based CBT-I treatment.

Authors' Response: We thank the reviewer for this positive evaluation of our trial protocol.

A couple of general points about the protocol:

1. While the title is Effectiveness of a stepped care model, the design and power is focused on comparative effectiveness of the 3 different levels of support for internet-based CBT-I in step 2. With respect to stepped care, there is little focus on step 1 and no data to be collected on step 3 as far as I could see. This is not necessarily a problem, but it would be helpful for those doing literature searches on support for internet-based interventions if the title made clearer this focus

Authors' Response: Following the advice of the reviewer, we have now changed the title of our manuscript to "Comparative effectiveness of three versions of a stepped care model for insomnia differing in the amount of therapist support in internet-delivered treatment – study protocol for a pragmatic cluster randomised controlled trial (GET Sleep)".

2. In my reading of the abstract and introduction, I had assumed that GPs would be doing the support for the internet CBT-I, so it was a bit of a surprise to see this was not the case when I got to the Methods. Could be others won't read it in this way, but maybe the abstract could make clearer who supports the step 2 internet CBT-I.

Authors' Response: Thank you for mentioning this, we have now re-written the corresponding sentence in the abstract to be more precise (page 3, line 7 of the redline version of the manuscript).

Otherwise just a few specific points for clarification of methods that would help inform any replication and regarding the analysis:

3. Are the main inclusion and exclusion criteria (ICD insomnia, exclusion of other sleep disorders, ICD mental health etc) assessed by the GP or from the screening questionnaire to participants?

Authors' Response: All inclusion and exclusion criteria are assessed by the responsible GPs, which is now more clearly stated in the revised manuscript (page 9, line 13).

4. At step 1, will a leaflet with the sleep education messages be given to patients as is common for psychoeducational interventions? If not, fine, but helpful to be clear in the protocol.

Authors' Response: Thank you for the opportunity to clarify this aspect of the study. There is no standardised leaflet that is used by the GPs in step 1 of the stepped care model. This is now made explicit in the revised version of the manuscript (page 11, lines 4-6).

5. Is movement from step 1 to step 2 entirely patient determined or is there a check in with the GP? And what is the process – does the patient get automatically sent a link after 4 weeks at step 1 to register? And do all patients get this?

Authors' Response: This is indeed correctly delineated by the reviewer and we agree that this could have been described more clearly in the manuscript. Consequently, we have re- written the corresponding sentence (page 11, lines 7-9).

6. What is duration of step 2 and is this standardized? My assumption is that the 12-week assessment point is planned to be at end of step 2 and before any patient would move on to step 3.

Authors' Response: Participants are instructed to complete one session of the internet intervention per week resulting in an overall duration of eight weeks for the step 2 treatment. However, participants are allowed to work through the sessions faster or slower accounting for interindividual differences in the therapeutic process (please refer to page 11, lines 22- 25). Thus, in some individuals, the T2 assessment after 12 weeks will be conducted before the end of step 2 treatment, and in some individuals, the T2 assessment will be conducted when participants are already in step 3 of the stepped care model.

7. What communication/feedback to GPs if any will there be from the step 2 CBT-I intervention either automated from the delivery software or from the psychologists providing the support?

Authors' Response: Thank you for pointing this out. The responsible e-coach of HelloBetter will send a report to the GP summarising step 2 treatment process and outcome including the recommendation about whether and by whom the treatment should be continued after step 2. This information is now added to the revised version of the manuscript (page 12, lines 19-23).

8. For the move from step 2 to step 3, a criterion of ISI ≥ 15 is mentioned as guidance to GPs. Will GPs be informed of patients' ISI scores from their 12-week post step 2 research assessments or will GPs administer the ISI themselves to patients outside the research process? And whichever of these two takes place, given the ISI is the study primary outcome measure will any precautions be taken to blind patients as to their scores?

Authors' Response: Thank for giving us the opportunity to clarify this. The GPs do not administer the ISI themselves. Instead, the report that the responsible e-coach of HelloBetter sends to the GP after step 2 treatment includes an ISI score based on a post-treatment ISI that participants fill in on the treatment platform of the GET.ON Institut für Online Gesundheitstrainings GmbH. This data is not part of the research process. We have now provided this information in the revised methods section of the manuscript (page 12, lines 19- 23). Patients do not receive the report of the e-coach and are thus blinded to their score. However, given that the ISI and guidelines for scoring/ interpretation are freely available on the internet, any precautions to blind patients to ISI scores are limited.

9. At step 3, it would be helpful to clarify what a sleep medicine specialist is in Germany and what kind of interventions they would carry out as I suspect this title does not travel and could mean different things in different countries.

Authors' Response: Following the advice of the reviewer, we have now included the information that the term 'sleep medicine specialist' is used in the manuscript to refer to medical doctors who have a

board-certification for sleep medicine. The intervention in step 3 is not standardised or controlled in this study, however, if the sleep medicine specialists adhere to the current insomnia guideline of the German Sleep Society, they would carry out face-to-face CBT-I or, if they believe that CBT-I has not been or cannot be sufficiently effective, they might also offer a pharmacological intervention.

10. The Brief Fatigue Inventory is included in the Table but is not mentioned in the relevant Measures section of the text

Authors' Response: We believe that the Brief Fatigue Inventory was mentioned under the subheading 'Potential treatment moderators and mediators' both in Table 1 and in the text (please refer to section 2.8.4. of the revised manuscript).

11. The Study Design section includes as part of the text on CONSORT reporting that the study will be reported in accordance with the extensions for reporting ….. noninferiority trials. However, at no other point in the paper is a non-inferiority analysis mentioned. Please clarify.

Authors' Response: We would like to apologise for this mistake and thank the reviewer for noticing this. The first author of the manuscript has recently received funding for a noninferiority trial and has confused the two studies when writing this sentence in the methods section. We have now removed the reference to the extension of the CONSORT 2010 statement for reporting noninferiority trials.

12. What reliability parameter will be used to calculate the Reliable Change Index? Will this be pre-specified from one of the published Cronbach alpha estimates or calculated from the study data? And if clinical significant change is also to be calculated using the Jacobson and Truax approach (unclear from the text whether this is planned) what ISI clinical significance threshold will be used?

Authors' Response: Thank you for this important comment. For calculating the Reliable Change Index, a pre-specified Cronbach's alpha of 0.92 will be used, based on a validation study in 410 primary care patients (Gagnon, Bélanger & Ivers, 2013; J Am Board Fam Med 2013;26:701-10). This information is now given in the revised version of the manuscript (page 19, lines 4-5). Clinical significance will be determined using Number Needed to Treat (NNT) analyses. This is now more clearly described on page 19, lines 1-2 of the revised manuscript.

VERSION 2 – REVIEW

REVIEWER	John Cape University College London, Clinical, Educational and Health Psychology
REVIEW RETURNED	19-Apr-2022

GENERAL COMMENTS	Thank you for addressing my previous comments so thoroughly and for the clarifying additions you have made the manuscript. Ideally, you might have considered using separate measures for the primary research outcome and for use by the e-coaches within the intervention rather than using the ISI for both. There are suggestions that repeated use of a measure may have an effect on the person completing the measure (whether interpreted as therapeutic via self-monitoring or as an artefact). But given that your main comparison is between the 3 intervention groups rather than with the TAU group (who unlike the intervention groups will only be adminstered the ISI for research assessments) I think this, while not ideal, is not a major issue. All the best with your recruitment.