PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	A cluster randomised controlled trial of a guided self-help mental health intervention in primary care
AUTHORS	Mathieson, Fiona; Stanley, James; Collings, Catherine (Sunny); Tester, Rachel; Dowell, Anthony

VERSION 1 - REVIEW

REVIEWER	Ava Schulz University of Zürich, Psychiatric Hospital Zürich, Switzerland
REVIEW RETURNED	28-May-2018

GENERAL COMMENTS	This two-arm cluster randomised trial investigated the effects of an ultra-brief intervention for mild to moderate health concerns of patients in general practice. 31 general practitioners were randomly assigned to provide either practice as usual or the novel brief intervention and recruited a total of 139 patients. ITT analyses show that the brief intervention did not improve outcome. The authors then discuss possible reasons for this null result and develop ideas for further development of the intervention. The topic under investigation has a broad public health impact and low-threshold interventions that are easily integrated into routine settings are in high demand. This study was rigorously conducted and used appropriate methods to investigate an important research question. I agree with the authors that “negative trials” such as this one merit publication (not only- but also- as an attempt to reduce publication bias) since there is a lot to learn from instances when intervention do not have the expected outcome. The results are valuable for the further development of interventions in primary care. I have, however, a few concerns that should be addressed before publication, which are listed below. My conclusion is that this a well conducted study that would be fit for publication in BMJ if the following concerns are addressed accordingly. Abstract: L38: It is unclear, what variables should be targeted by the intervention, “improve outcome” seems vague. Please specify. L21: “mean adjusted K10 difference = 1.68”, please include p-value or corresponding test statistics L25: “The UBI intervention did lead to better outcomes” – there is a “not” missing at this very crucial position in the text (Freudian slip? ;). Please correct, otherwise this looks as if the conclusions are not supported by the data in the slightest. p4,L39: In fact, results on the comparison between guided and unguided internet-based self-help are mixed; even though there is

	a tendency of superiority of guided treatment, there are studies showing that unguided interventions can be just as effective. L46,47: The current phrasing suggests a repetition with the "feasibility testing". Please rephrase. Methods: - general remark: I'd be interested in individual differences between GPs, have to controlled for specific GP effects, e.g. level of training, experience in psychological interventions, etc.? This seems especially relevant since one possible explanation of the null results may be that the GPs providing PAU were already providing high standard services. In a similar vein: Were medication effects controlled? Did you control for the number of patients treated by each GP, i.e. GP work load? p6, L15, L48: repetitive, please shorten - p8., L10: "guided" in this context is confusing, since it is mostly used in the context of webbased treatments. Please rephrase/clarify. p10, L13: Please add information on whether a difference of 3.2 points on the HADS is clinically relevant. p10. L54: It remains unclear why the assessment at 26 weeks was chosen as a primary outcome. p11, L19: Have you collected more detailed data on additional treatments? It would be interesting look into that and possibly conduct subgroup analyses. p12, L25: Please include p-values/test statistics when describing differences. p13, L44: This contradicts the abstracts, beyond the typo: The intervention did not only "not lead to an improved outcome", it fared actually worse than PAU. Please clearify. p15, L44: The authors point out an important possible explanation here: One reason for the null result could be that PAU received more extra care. I'd suggest elaborating this hypothesis more. In general, the manuscript would profit from a more thorough discussion of the reasons why the PAU was already so successful (which I think is one of the most interesting results of the study).

REVIEWER	Dr Caroline S Clarke University College London, UK
REVIEW RETURNED	04-Jun-2018

GENERAL COMMENTS	Very well written and clear paper. I have very few comments. Please note that I am a health economist who works on mental health trials (mainly in the UK), so my comments are from this perspective. Page 2: There is a "not" missing from "The UBI intervention did not lead to better outcomes than practice as usual" Perhaps could also include "brief intervention" as a keyword? Page 5: Typo: "while there was be clustering by practice" Page 6:

	The last paragraph is a repetition of the second half of the second paragraph. Page 7: “Randomisation was performed following individual GP consent as a single step, with randomisation conducted by the project biostatistician”. This is slightly confusing – does it mean that once the first GP from a practice had consented, then the practice was randomised? Supplementary table R3: It would be interesting to also know how many of the PAU patients (and UBI patients) were referred onwards to psychological or other counselling options, or to relevant community services (i.e. from description of PAU on page 8). Page 9: What is an academic mental health consumer? Is it an academic researcher on the team who also happens to use mental health services? Or something else? Page 11: “Additional treatments received during the trial (including medication and talking therapies) were analysed by study arm, based on self-report data collected at the 6 month follow-up. This descriptive analysis was not specified in the study protocol.” Was this analysis adjusted for baseline use of medication/talking therapies? Page 12: No comment is made on the fact that the two arms do not seem to be well balanced in some demographic variables, e.g. more younger people in UBI group. Perhaps this could be discussed, and possible impacts on the results discussed. General comments: There is not much mention of the fact that it seems that 60 GPs who received training did not recruit a single participant. Is this the case? How could this have been avoided? Also, for those who eventually recruited a patient a long time after being trained, were they offered a refresher session? It seems that the inclusion criteria limited recruitment a lot. Why did the funder insist on narrow inclusion criteria, and what would the authors have preferred to have done differently? Good points are made in the discussion about possible reasons for UBI not appearing more effective than PAU. What information was collected on the costs of the two arms? Could a cost-effectiveness analysis be done? As the outcome was the same for each arm, potentially a cost-minimisation analysis could be performed, i.e. if they are equally effective, then the cheaper option would be preferred. Some discussion on this would be important when considering the question of whether UBI should be rolled out or not.

REVIEWER	Juan Bellón El Palo Health Centre (SAS), redIAPP, IBIMA, Department of Public Health and Psychiatry, university of Málaga, Spain
REVIEW RETURNED	12-Jun-2018

GENERAL COMMENTS	Thanks for letting me review this paper. It aims to ascertain whether an ultra-brief intervention improves outcomes for patients in general practice with mild-to-moderate mental health. The research question is relevant and in general the methods used are appropriate. However, there are some points that should be clarified and improved: •It took 3 years in the recruitment and even then the minimum sample size required was not achieved. For researchers this is discouraging, but the worst thing is that it is difficult to draw conclusions from the data obtained. It cannot be concluded that it is a study with negative results because the lack of statistical power does not allow it. •In my opinion the two parallel arms of the trial are not the 'Ultra-Brief Intervention' (UBI) versus 'Practice as Usual' (PAU), but UBI + PAU vs PAU. The key question would be: how has the use of the PAU in both arms influenced the final results? It is therefore a post-randomization bias that should be adjusted. Another question related to the previous one would be: how the use of the UBI in the intervention group conditioned the use of the PAU in this same group? Although PAU resources could be used in both the intervention and control groups, GPs in the intervention group might be more reluctant to use them because they had the expectation that the UBI intervention would be effective. For example, in the supplementary table R3: patients who did not respond to the questions on medication status during trial were 20 and 9 for UBI and PAU group respectively, and for those who did answer there were 12 and 16 that started medication during trial respectively. In absolute terms, these data appear to be unbalanced between groups. We also do not know what type of medication was used and whether this use was adequate based on the diagnosis in each group, and there is much evidence that the appropriate use of antidepressants is effective in specific mental disorders. •In my opinion, the use of K-10 as a criterion for patient selection and primary outcome introduced a non-specific morbidity factor that diluted any positive effect of the UBI intervention. For example, the UBI may be effective in patients with depression or anxiety below the threshold, but not for moderate major depression or generalized anxiety disorder or for patients who suffer from the latter two disorders at the same time. Perhaps mixing all these patients in the same trial hid any possible positive effect of the UBI intervention. •Unless the GPs of New Zealand have a solid background as psychotherapists, the 2-hour training at UBI seems too short. For example, for the same patient with generalized anxiety disorder, the comparison regarding effectiveness between GPs with a 2-hour training and 3 sessions (30, 15 and 15 minutes respectively) and specialized psychotherapists performing at least 6-8 session

	of 45 minutes, a priori it would be in favor of the latter. As it seems to be deduced from the supplementary table R3 (Counselling sessions), if the GPs of the PAU group used more psychotherapists than those of the UBI group, this could condition the results. •It is not clear to me whether the authors have adequately controlled attrition biases. As reported in the Table 3, the results shown correspond to 70 patients from the UBI group and 69 from the PAU group, while 85 were randomized (80 used the UBI) and 75 respectively. Therefore, at least 10 and 6 of the UBI and PAU group respectively were not included in the analyses. In my opinion, these should also have been included in the analyses, maybe using multiple imputations. In addition, some data should be provided to support the MAR assumption. •The hierarchy in the data implies two levels. Patients nested in GPs and GPs nested in practices. The researchers decided to perform the randomization by GPs and therefore in the analyses they decide to adjust by GPs. In principle it seems an appropriate decision, however it would be convenient to calculate the intraclass correlation coefficients (ICC) at 'practice' level, and if both were relevant, then it should be adjusted for both, GPs and practice. If the ICC at practice level is not relevant, it should at least be showed in the results. •Following the CONSORT criteria for cluster RCTs, in the flowchart the figures of patients and clusters (GPs and practices) should be put in each step. •In the protocol, a priori only the analyses adjusted for the value of the respective dependent variable were cited; however, analyses adjusted for other variables measured at baseline were performed. It seems that there was a lack of balance in some variables measured at baseline, which is relatively common when randomization is done by cluster, especially when the number of clusters is small. In any case, as a sensitivity analysis, both types of analysis should be showed, adjusted for the baseline value of the dependent variable and also adjusted for the rest of the variables. It is not clear in the Table 3 and the figures 2-3 what were the adjustment variables.

REVIEWER	Jens Klotsche German Rheumatism Research Center Berlin, Germany
REVIEW RETURNED	17-Aug-2018

GENERAL COMMENTS	-The RCT was stopped before the planned end due to the stop of financial support. Therefore the initially planned sample size was not reached. oMore details about the planned and reached sample size should be provided. It is unclear how large the gap between the two numbers is. oThe study failed to show a significant difference between the two groups in the primary outcome. Is this a result of a lower statistical power?

	-The inclusion criteria seem to be arbitrary. The physician decides about the eligibility of the patient for the study. If the screening was positive, the patient was screened by the K10 for study inclusion. Why did you not provide consecutively the K10 to all patients to screen for study eligibility? -The authors should provide more details on attrition bias. The rate of patients with at least one follow-up visit is reported and seems to be comparable between the two groups. However, did you observe groups differences at specific time points. It is possible that the acceptance of an intervention is lower and the patients dropped out earlier in follow-up. -The two intervention groups are not balanced due to sex and age. Did you conduct sensitivity analyses by adjusting for both parameters? -Discussion,2nd paragraph: The trial was designed as superiority trial. It is not possible to get any conclusion about the non-inferiority of the two interventions even in case of the full sample size. The sample size of a non-inferiority trial has to be determined separately based on a non-inferiority margin. Usually, the sample size of a non-inferiority trial is larger than of a superiority trial.

REVIEWER	Ailish Hannigan Graduate Entry Medical School University of Limerick Ireland
REVIEW RETURNED	28-Aug-2018

GENERAL COMMENTS	This is a well-presented paper with no major statistical issues. As acknowledged by the authors, the results are unlikely to be generalisable to other settings/countries with less well-resourced practice as usual but the paper gives an interesting and relevant account of the challenges of carrying out pragmatic trials of mental health interventions in primary care including recruitment challenges and changes to practice as usual services during the trial. There was an improvement over time in both groups for the primary outcome with a larger mean improvement for the PAU group. The authors describe the improvement as reasonable - is there an accepted minimum clinically important change for K10 and how does the improvement in each group relate to this? From Table 3, at 8 weeks the mean difference in change in K10 between the two groups is close to zero. From 8 weeks onwards, the trend is for PAU to do better – in fact from Figure 2 mean K10 for UBI increases between 2 and 3 months. It raises the question as to what happens to the UBI group after the 5/6 week brief intervention. They revert to practice as usual? Further comment and discussion from the authors on this would be useful. It would also be useful to know more about the participating GPs in each group – age, gender, ethnicity, years' experience, and specialist training/interest in mental health– is this data available? There are two important typos – the conclusion in the abstract should read 'The UBI intervention did not lead to better outcomes..' and on page 13, line 46, the confidence interval is minus 1.18 to 4.85.

	It may be helpful for the reader to clarify that in Table 3 these are mean differences between groups of the change in outcomes over specified time periods.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Ava Schulz

Institution and Country: University of Zürich, Psychiatric Hospital Zürich, Switzerland Please state any competing interests or state 'None declared': none declared

Please leave your comments for the authors below This two-arm cluster randomised trial investigated the effects of an ultra-brief intervention for mild to moderate health concerns of patients in general practice. 31 general practitioners were randomly assigned to provide either practice as usual or the novel brief intervention and recruited a total of 139 patients. ITT analyses show that the brief intervention did not improve outcome. The authors then discuss possible reasons for this null result and develop ideas for further development of the intervention.

The topic under investigation has a broad public health impact and low-threshold interventions that are easily integrated into routine settings are in high demand. This study was rigorously conducted and used appropriate methods to investigate an important research question.

I agree with the authors that "negative trials" such as this one merit publication (not only- but also- as an attempt to reduce publication bias) since there is a lot to learn from instances when intervention do not have the expected outcome. The results are valuable for the further development of interventions in primary care. I have, however, a few concerns that should be addressed before publication, which are listed below. My conclusion is that this is a well conducted study that would be fit for publication in BMJ if the following concerns are addressed accordingly.

Thank you for these comments.

Abstract:

Point 1) L38: It is unclear, what variables should be targeted by the intervention, "improve outcome" seems vague. Please specify.

This has been amended to read 'Mental health outcomes'

Point 2) L21: "mean adjusted K10 difference = 1.68", please include p-value or corresponding test statistics

We have added the p-value for this comparison to the abstract (which already included the confidence interval for this difference)

Point 3) L25: "The UBI intervention did lead to better outcomes" – there is a "not" missing at this very crucial position in the text (Freudian slip? ;). Please correct, otherwise this looks as if the conclusions are not supported by the data in the slightest.

Thank you. This has been corrected.

Point 4) p4,L39: In fact, results on the comparison between guided and unguided internet-based self-help are mixed; even though there is a tendency of superiority of guided treatment, there are studies showing that unguided interventions can be just as effective.

We thank the reviewer for raising this point. While there has been commentary about this, overall the balance of evidence still seems to support our statement regarding guidance.

i.e. guidance is a beneficial feature of Internet-based interventions, although its effect is smaller than reported before when compared to unguided interventions. See for example (Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions—A systematic review. Internet Interventions. 2014 Oct 1;1(4):205-15.)

Point 5) L46,47: The current phrasing suggests a repetition with the “feasibility testing”. Please rephrase.

The second use of the word ‘feasibility’ has been removed from this sentence.

Methods:

Point 6) - general remark: I’d be interested in individual differences between GPs, have to controlled for specific GP effects, e.g. level of training, experience in psychological interventions, etc.? This seems especially relevant since one possible explanation of the null results may be that the GPs providing PAU were already providing high standard services. In a similar vein: Were medication effects controlled? Did you control for the number of patients treated by each GP, i.e. GP work load?

Unfortunately we did not collect such baseline data on GP past experience/training. It is possible that PAU already has a “high enough” standard of care that the intervention could not improve outcomes further (clinical comment here). We have commented on this point in the discussion.

On the second and third points: we did not formally control for medication effects (though these are summarised in the results/discussion for context) as this was not part of our formal analysis plan. Note also that medication data was only collected at final follow-up, which means these data were incomplete for many participants (only available for those who completed the 6 month assessment).

We also did not adjust for number of patients treated by each GP in the trial. We did not have sufficient data on the variability between GP workloads in order to perform this adjustment.

Point 7) p6, L15, L48: repetitive, please shorten

This has been fixed so as not to be repetitive

Point 8) - p8., L10: “guided” in this context is confusing, since it is mostly used in the context of webbased treatments. Please rephrase/clarify.

The term ‘guided’ has been replaced by ‘low intensity’

Point9) p10, L13: Please add information on whether a difference of 3.2 points on the HADS is clinically relevant.

Thank you for this comment. The work that has been done on MCIDs for the HADS measure suggest a change of at least 2 points on the total scale as a meaningful difference, though the available data is drawn from patients hospitalised with COPD. We have added a note to this effect in the discussion.

“The original sample size calculation also indicated that full recruitment would have achieved 80% power to detect a difference of 3.2 points on the HADS scale: this was a slightly bigger difference than the minimal clinically important difference cited in the literature (Puhan et al. 2008).”

Note that our study was powered to detect a difference of 4 points on the K10 scale: the statement about the power to detect differences on the HADS outcome was intended to illustrate power for this

secondary outcome, rather than the sample size having been set to detect a clinically meaningful difference on the HADS scale.

Point 10) p10. L54: It remains unclear why the assessment at 26 weeks was chosen as a primary outcome.

Conducting mental health research in both primary and secondary care settings can be challenging, particularly in regard to patient follow up. Many trials have a 3 month follow- up for their primary outcomes. We wished to continue data collection for as long as would be possible within the practical constraints of funding and patient retention. A six-month final follow-up was chosen to examine the long-term effects of the mental health intervention, and to ask whether any achieved outcomes would be sustainable.

Point 11) p11, L19: Have you collected more detailed data on additional treatments? It would be interesting look into that and possibly conduct subgroup analyses.

The only data we have on additional treatments is presented in Supplementary Table R3. We do not think it would be reasonable to conduct any subgroup analyses, given that none were planned in advance, and that we were not able to achieve our planned sample size for the main trial (which would limit what could be learned from a hypothesis-generating point of view).

Point12) p12, L25: Please include p-values/test statistics when describing differences.

Thank you for this comment. The aim of Table 1 is to describe the participant profiles and allow readers to examine potential imbalance at baseline. We have followed the CONSORT guidelines for presentation of this information: The CONSORT statement actively discourages the use of inferential statistics like p-values and confidence intervals in the context of describing the RCT cohort (see pages 16-17 of the referenced Explanatory document cited below.) Therefore we have not added hypothesis test results to the text or Table 1.

See: Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c869.

Point 13) p13, L44: This contradicts the abstracts, beyond the typo: The intervention did not only “not lead to an improved outcome”, it fared actually worse than PAU. Please clarify.

The section referred to by this statement (p13 L44 of the originally submitted paper) stated “the mean difference… favoured the PAU arm” which seems clear to us. The following lines under that point treated the result as indeterminate (from a confidence interval/hypothesis testing perspective). If this section is still unclear we are happy to update it further.

The discussion section (first line of discussion) seems to be akin to the comment raised here: we have altered this to read “The brief psychological treatment…did not lead to better outcomes than PAU in this pragmatic efficacy trial, with the point estimate favouring PAU over UBI.” which we hope addresses this point. We have also updated the start of the second paragraph in the discussion in light of this comment.

Point 14) p15, L44: The authors point out an important possible explanation here: One reason for the null result could be that PAU received more extra care. I’d suggest elaborating this hypothesis more. In general, the manuscript would profit from a more thorough discussion of the reasons why the PAU was already so successful (which I think is one of the most interesting results of the study).

Thank you – we have added additional text as follows;

For the last 10-20 years in many OECD jurisdictions there has been a focus on improving mental health care provision in primary care settings. In New Zealand this has taken the form of the

introduction of locally based Primary Mental Health Initiatives (PMHI), which have increased access to psychological services and provided opportunity for increased engagement (and remuneration) for General Practitioners to undertake mental health consultation work. (Dowell 2009) These opportunities were available to the PAU, and may partially explain the relative success of this 'control' arm in the study.

Reviewer: 2

Reviewer Name: Dr Caroline S Clarke

Institution and Country: University College London, UK Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below Very well written and clear paper. I have very few comments. Please note that I am a health economist who works on mental health trials (mainly in the UK), so my comments are from this perspective.

Page 2:

Point 1) There is a "not" missing from "The UBI intervention did not lead to better outcomes than practice as usual"

Thank you for this note: This has been corrected

Point2) Perhaps could also include "brief intervention" as a keyword?

We have amended 'Brief interventions' to "brief Intervention' as a key word

Point 3) Page 5: Typo: "while there was be clustering by practice"

This has been corrected

Point 4) Page 6: The last paragraph is a repetition of the second half of the second paragraph.

This has been corrected

Point 5) Page 7: "Randomisation was performed following individual GP consent as a single step, with randomisation conducted by the project biostatistician". This is slightly confusing – does it mean that once the first GP from a practice had consented, then the practice was randomised?

Thank you for this comment: we have rewritten this to clarify the randomisation process (and added one extra detail to the end of this sentence as follows)

"Practices were entered into the trial following consent from individual participating GPs in that practice. Randomisation of all consenting practices was conducted following this step by the project biostatistician (JS) using a computer-based randomisation following the above stratification profile."

Point 6) Supplementary table R3: It would be interesting to also know how many of the PAU patients (and UBI patients) were referred onwards to psychological or other counselling options, or to relevant community services (i.e. from description of PAU on page 8).

Please note that we have summarised the available information on additional treatment options in Supplementary Table R3 (extended GP consultations and referral to counselling, which includes both psychologists and counsellors). We did not ask participants to differentiate between these two types of talking therapists.

Point 7) Page 9: What is an academic mental health consumer? Is it an academic researcher on the team who also happens to use mental health services? Or something else?

We have added the definition that this as ‘an academic who is also a mental health service user and who conducts research from a service user perspective’.

Point 8) Page 11: “Additional treatments received during the trial (including medication and talking therapies) were analysed by study arm, based on self-report data collected at the 6 month follow-up. This descriptive analysis was not specified in the study protocol.” Was this analysis adjusted for baseline use of medication/talking therapies?

The analysis in Supplementary Table R3 is not adjusted for baseline use of medication/talking therapies: while we do not have data on past use of talking therapies, the “medication status during trial” part of supplementary table R3 includes a line reporting the proportions of participants who were on medication prior to starting the trial. We have a note at the end of the methods that this information was collected at the final 6 month follow-up (i.e. we did not collect data on medication at the time of the baseline data collection).

Point 9) Page 12: No comment is made on the fact that the two arms do not seem to be well balanced in some demographic variables, e.g. more younger people in UBI group. Perhaps this could be discussed, and possible impacts on the results discussed.

We agree that this would be useful to cover in the discussion and have added a section on this point (see below): This was already noted in the results (page 12) regarding what we would consider a slight imbalance (though there are no universal criteria as to how judge the magnitude of such imbalances). We note that the analysis of clinical outcomes was adjusted for these key demographic variables, which means that this imbalance should be accounted for in the results.

New note:

“The analyses presented here examined several arising issues that were not planned for at the start of the study. Firstly, there were imbalances on some demographic variables (gender and age group) between the two study arms. While this is sub-optimal, the analysis of primary and secondary outcomes adjusted for these and other sociodemographic factors, which means that these imbalances should be accounted for in the results.”

General comments:

Point 10) There is not much mention of the fact that it seems that 60 GPs who received training did not recruit a single participant. Is this the case? How could this have been avoided? Also, for those who eventually recruited a patient a long time after being trained, were they offered a refresher session?

As outlined in the manuscript (P17 L 50-53), Firstly, our recruitment was limited by specific entry criteria required by a funder (to allow access to treatments as part of the PAU group). An additional sentence has been added to the text to clarify this (p.21).

Point 11) It seems that the inclusion criteria limited recruitment a lot. Why did the funder insist on narrow inclusion criteria, and what would the authors have preferred to have done differently?

This paragraph now reads: Inclusion criteria were based on the access criteria of a local partner primary health organization (PHO) to psychological therapies. The criteria for access are specifically targeted at youth (defined as 18 to 24 years old), and, for individuals aged 25 years or older, patients with low income, or Māori or Pacific Island heritage.

Point 12) Good points are made in the discussion about possible reasons for UBI not appearing more effective than PAU. What information was collected on the costs of the two arms? Could a cost-effectiveness analysis be done? As the outcome was the same for each arm, potentially a cost-minimisation analysis could be performed, i.e. if they are equally effective, then the cheaper option would be preferred. Some discussion on this would be important when considering the question of whether UBI should be rolled out or not.

Within the resources and funding available for this study we were not able to perform a cost-effectiveness analysis.

Reviewer: 3

Reviewer Name: Juan Bellón

Institution and Country: El Palo Health Centre (SAS), redIAPP, IBIMA, Department of Public Health and Psychiatry, university of Málaga, Spain Please state any competing interests or state 'None declared': none

Please leave your comments for the authors below

Thanks for letting me review this paper. It aims to ascertain whether an ultra-brief intervention improves outcomes for patients in general practice with mild-to-moderate mental health. The research question is relevant and in general the methods used are appropriate. However, there are some points that should be clarified and improved:

Point 1) It took 3 years in the recruitment and even then the minimum sample size required was not achieved. For researchers this is discouraging, but the worst thing is that it is difficult to draw conclusions from the data obtained. It cannot be concluded that it is a study with negative results because the lack of statistical power does not allow it.

We agree and have acknowledged these points in the discussion. We have tried to be careful to describe the study results as "indeterminate" (that is we are not sure whether there really are positive or negative effects of the intervention). However, interpretation of the upper and lower bounds of the confidence intervals for the estimates (which do take into account lower power due to not reaching our intended sample size) suggest it is unlikely that the UBI arm has a substantial positive effect on outcomes relative to PAU. Reviewer 4 (point 7) commented on interpretation regarding non-inferiority, which we have further addressed below (as a follow-on from this comment).

Point 2) In my opinion the two parallel arms of the trial are not the 'Ultra-Brief Intervention' (UBI) versus 'Practice as Usual' (PAU), but UBI + PAU vs PAU. The key question would be: how has the use of the PAU in both arms influenced the final results? It is therefore a post-randomization bias that should be adjusted. Another question related to the previous one would be: how the use of the UBI in the intervention group conditioned the use of the PAU in this same group? Although PAU resources could be used in both the intervention and control groups, GPs in the intervention group might be more reluctant to use them because they had the expectation that the UBI intervention would be effective. For example, in the supplementary table R3: patients who did not respond to the questions on medication status during trial were 20 and 9 for UBI and PAU group respectively, and for those who did answer there were 12 and 16 that started medication during trial respectively. In absolute terms, these data appear to be unbalanced between groups.

Note that the "no response to question on medication status" is footnoted in Supplementary Table R3 as "Did not complete 6 month questionnaire and hence no data" – this imbalance in loss to follow-up is dealt with under reviewer point 5 below.

The reviewer is correct in stating that GPs in the intervention arm also had access to the PAU options. This was a pragmatic trial that did not want to disrupt usual management when required.

We have noted this in the manuscript.

Line 242 P 8

'The study protocol allowed for patients in either study arm to alter their treatment as needed (e.g. access other talking therapies, or commence mental health medications).'

We have added an additional sentence to make this clear prior to line 242:

'In New Zealand a stepped care approach to management guides the practitioner towards using the most appropriate therapy option for the severity of presentation. UBI was designed for mild to moderate presentations and in training GPs were comfortable with the use of the UBI approach for first line management'.

We also do not know what type of medication was used and whether this use was adequate based on the diagnosis in each group, and there is much evidence that the appropriate use of antidepressants is effective in specific mental disorders.

Regarding medication use: There is a considerable body of evidence that while medication may be useful in more severe and defined disorders, its effectiveness is diminished in less severe presentations. This has led to a more cautious approach to the use of medication in a New Zealand stepped care setting, for the mild / moderate presentations seen in the UBI trial. We consider the patient outcomes for both intervention and PAU groups to be in keeping with that expected from New Zealand primary care practice.

Please see:

Fournier JC, DeRubeis RJ, Hollon SD, Dimidjian S, Amsterdam JD, Shelton RC, Fawcett J. Antidepressant drug effects and depression severity: a patient-level meta-analysis. Jama. 2010 Jan 6;303(1):47-53.

The magnitude of benefit of antidepressant medication compared with placebo increases with severity of depression symptoms and may be minimal or non-existent, on average, in patients with mild or moderate symptoms.

Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R (January 2008). "Selective publication of antidepressant trials and its influence on apparent efficacy". The New England Journal of Medicine 358 (3): 252–60. doi:10.1056/NEJMsa065779

Point 3) In my opinion, the use of K-10 as a criterion for patient selection and primary outcome introduced a non-specific morbidity factor that diluted any positive effect of the UBI intervention. For example, the UBI may be effective in patients with depression or anxiety below the threshold, but not for moderate major depression or generalized anxiety disorder or for patients who suffer from the latter two disorders at the same time. Perhaps mixing all these patients in the same trial hid any possible positive effect of the UBI intervention.

This was a trial intended to replicate the pattern of presentation and management seen in New Zealand General Practice settings. The K-10 is commonly used as a screening tool to assess initial psychological symptom counts and it focuses on symptoms seen for the most common disorder in primary care – anxious depression (see reference below). While it is possible that UBI may perform better with some subgroups of patients, the aim of the trial was to assess its effectiveness in routine practice and presentations.

Goldberg, D. P., Reed, G. M., Robles, R., Minhas, F., Razzaque, B., Fortes, S., … Dowell, A. C., … Saxena, S. (2017). Screening for anxiety, depression, and anxious depression in primary care: A field study for ICD-11 PHC. Journal of Affective Disorders, 213, 199-206. doi: 10.1016/j.jad.2017.02.025

Point 4) Unless the GPs of New Zealand have a solid background as psychotherapists, the 2-hour training at UBI seems too short. For example, for the same patient with generalized anxiety disorder, the comparison regarding effectiveness between GPs with a 2-hour training and 3 sessions (30, 15 and 15 minutes respectively) and specialized psychotherapists performing at least 6-8 session of 45 minutes, a priori it would be in favor of the latter. As it seems to be deduced from the supplementary table R3 (Counselling sessions), if the GPs of the PAU group used more psychotherapists than those of the UBI group, this could condition the results.

As outlined above the presentation of common mental disorder symptoms in primary care differs from that in secondary care in terms of both severity, but also specificity of apparent diagnosis. For common presentations of anxious depression and for milder forms of GAD and Depression, clinical practice as undertaken by counsellors. or psychologists would be unlikely to have 6-8 sessions of 45 minutes duration. In our evaluation of the New Zealand primary mental health initiatives, counsellors and clinical psychologists had between 3-4 contacts with such patients (see reference below). UBI was designed to complement the existing mental health consulting skills of the GPs; the 2 hour training session was to provide a consultation framework within which to extend existing skills and hence we believe is appropriate. Our feasibility studies had suggested this amount of training was feasible and acceptable to GPs. Using the UBI tools and resources the GPs in the intervention group were encouraged to provide an initial management pathway (and were given the extra time and remuneration to do so) with the expectation this might reduce the need for subsequent referral to another mental health worker.

Dowell AC, Garrett S, Collings S, McBain L, McKinlay E, Stanley J. 2009.Evaluation of the Primary Mental Health Initiatives: Summary report 2008. Wellington: University of Otago and Ministry of Health. ISBN 978-0-478-31907-1 (print)ISBN 978-0-478-31908-8 (online)HP 4754

Point 5) It is not clear to me whether the authors have adequately controlled attrition biases. As reported in the Table 3, the results shown correspond to 70 patients from the UBI group and 69 from the PAU group, while 85 were randomized (80 used the UBI) and 75 respectively. Therefore, at least 10 and 6 of the UBI and PAU group respectively were not included in the analyses. In my opinion, these should also have been included in the analyses, maybe using multiple imputations. In addition, some data should be provided to support the MAR assumption.

Thank you for this comment – we have now conducted two sensitivity analyses using multiply imputed data to be able to include all recruited participants in the outcome analysis. Because the protocol did not specify this approach we have indicated that this is a “supplementary” analysis. Fuller detail on the methods used is given in the Supplementary Materials, and is outlined below. Because of the complexity of this analysis (both from a conduct and reporting point of view), we have restricted this additional analysis to the primary outcome (K10 score at 6 months post-baseline). For these analyses, we followed some of the suggestions implemented in Bell et al. (2013) and Sullivan et al. (2018).

Bell ML, Kenward MG, Fairclough DL, Horton NJ. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. BMJ. 2013;346:e8668.

Sullivan TR, White IR, Salter AB, Ryan P, Lee KJ. Should multiple imputation be the method of choice for handling missing data in randomized trials? Stat Methods Med Res. 2018;27(9):2610-26.

In the first analysis, we imputed follow-up outcome data for all participants with incomplete follow-up data. This means that “baseline data only” participants could be included and thus this sensitivity

analysis included all randomised participants (which is one of the concerns raised). This analysis makes the assumption that data are missing at random, conditional on observed data: for example, this analysis accounts for the fact that loss to follow up was more common in younger people, though it still assumes that the missing outcomes in the "baseline only" participants follow a similar pattern to the observed patterns in similar types of participants who participated in follow up (so this accounts for the fact that baseline-only participants tended to be younger).

In the second analysis, we assumed a different trajectory for participants who did not complete any post-baseline measures (under three different sets of assumptions, all assuming poorer outcomes for participants who did not participate in any follow-up). This set of results aims to address the potential for these people's outcomes to be missing not at random (MNAR) when assessing the potential impact on the intervention effect.

There was little difference between the effect sizes for the main results (as presented in Table 3) and these sensitivity analyses (judged relative to the MCID for the outcome, and also the statistical precision of the estimates as described in the confidence intervals): in the case of the MAR sensitivity analysis (imputed outcomes) this is likely because the assumptions are similar to the assumptions implicit in the main paper's linear mixed models (excepting that the new sensitivity analysis includes all randomised participants); and in the MNAR sensitivity analysis, the potential impact of any systematically different outcome trajectories for participants not completing any follow-up is limited by the absolute number of participants who were completely lost to follow up relative to participants included in the main reported analysis (15 / 85 in the UBI arm, and 6 / 75 in the PAU arm).

On the final point raised by the reviewer ("some data should be provided to support the MAR assumption"), we have added some commentary on whether the MAR assumption is reasonable: this is covered in the Supplementary Material in the section around the MAR/MNAR analysis. We

note here that the MAR assumption is intrinsically untestable in the quantitative sense within any given dataset, as it asks about whether (unobserved) outcomes in participants lost to follow up are systematically different to (observed) outcomes in those participants who remained in the trial. This information is never internally available within a given dataset (see e.g. Sullivan et al., 2018).

This combined material is covered in detail in the Supplementary Material, with a summary in the results (p. 18: not reproduced here as this takes up approximately one page in total) and a brief paragraph of implications in the discussion (p. 20) which states:

"We also examined the impact of analytical decisions on our primary outcome, particularly sensitivity analyses examining the potential impact of participants with no post-baseline data (excluded from the main analysis) on the reported intervention effect. There was more loss-to-follow-up observed in the UBI group than in the PAU group. These sensitivity analyses showed relatively little impact on our estimates under several sets of assumptions (Supplementary Methods and Results)."

Point 6) The hierarchy in the data implies two levels. Patients nested in GPs and GPs nested in practices. The researchers decided to perform the randomization by GPs and therefore in the analyses they decide to adjust by GPs. In principle it seems an appropriate decision, however it would be convenient to calculate the intraclass correlation coefficients (ICC) at 'practice' level, and if both were relevant, then it should be adjusted for both, GPs and practice. If the ICC at practice level is not relevant, it should at least be showed in the results.

Thank you for this comment: we have appended ICC values to Supplementary Table R4 for a "joint clustering" effect of having GPs nested within practices. We did not consider it possible to get truly independent estimates of "clustering by practice" since this would itself include clustering by GP as a sub-level. This multilevel clustering had minimal impact on the estimated ICCs for the K10 and HADS, though there appeared to be some increased clustering of patient outcomes on the WSAS and Health

Thermometer scales. This has been covered in more detail in the Supplementary Methods and Supplementary Table R4. (note that in updating this analysis we also updated some of the ICCs for the GP-clustering results, a few of which changed at the third decimal place).

We have added a note to this in the methods (bottom p. 17 - top p.18):

“We also examined clustering effects for GPs being nested within clustering by GP practice: this additional complexity (not implemented in our main analytical models) had little impact on ICCs for the K10 or HADS measures, though it did suggest slightly higher ICCs (greater clustering of outcomes than considering GPs alone) for the WSAS and Health Thermometer.”

Note that we did not consider re-running our primary analyses allowing for multi-level clustering as the random effects component of these main models are quite complex (as they already have patients repeated measurements nested within GPs), and more complex models could not be run with the addition of GP practice sitting above GPs in the random effects specification. This was not an issue for the ICCs reported in the supplementary material as these were calculated based on simplified mixed models (no repeat measurements per patient) and implemented in a different software library in R that could estimate the ICC (we have slightly expanded the notes in the supplementary methods regarding calculating the ICC).

Point 7) Following the CONSORT criteria for cluster RCTs, in the flowchart the figures of patients and clusters (GPs and practices) should be put in each step.

Thank you for this comment: we have amended the figure to reflect the clusters. In amending the figure it was very hard to squeeze in additional information on both numbers of GPs at each stage and number of practices: we have therefore added the number of GPs at each stage to the figure (as the most pertinent point to the flow of participants and the clusters formally considered in analysis) but have summarised the number of participating practices they represented in the text.

Point 8) In the protocol, a priori only the analyses adjusted for the value of the respective dependent variable were cited; however, analyses adjusted for other variables measured at baseline were performed. It seems that there was a lack of balance in some variables measured at baseline, which is relatively common when randomization is done by cluster, especially when the number of clusters is small. In any case, as a sensitivity analysis, both types of analysis should be showed, adjusted for the baseline value of the dependent variable and also adjusted for the rest of the variables. It is not clear in the Table 3 and the figures 2-3 what were the adjustment variables.

Thank you for this comment: the results in the main body of the paper are (as per methods) adjusted for baseline outcome scores and sociodemographic variables. We have clarified this in Table 3, and have also included the analysis adjusting only for baseline scores for the primary K10 outcome (as supplementary table R6). We have also clearly noted on P. 11 that the protocol stated we would only adjust for baseline scores:

“The original protocol stated that analyses would only be adjusted for baseline-values of each score: given some slight imbalance in sociodemographic characteristics it was decided to adjust for other baseline covariates in the main analyses. The originally planned analyses are presented in supplementary materials (overall patterns discussed in the body of the results).”

Reviewer: 4

Reviewer Name: Jens Klotsche

Institution and Country: German Rheumatism Research Center Berlin, Germany Please state any competing interests or state ‘None declared’: I do not have any conflict of interest.

Please leave your comments for the authors below

Point 1) The RCT was stopped before the planned end due to the stop of financial support. Therefore the initially planned sample size was not reached.

Yes that is correct. We added the word ‘planned’ to the following sentence in the discussion to emphasise this: ‘This meant we did not meet our planned sample size…’ (p.17, line 52). This is also discussed in the third paragraph of the discussion.

Point 2) More details about the planned and reached sample size should be provided. It is unclear how large the gap between the two numbers is.

Thank you for this comment. We have added this detail to the discussion to gather this information in one location (new text underlined below, page 17). We note that the planned sample size is presented on page 10 (methods) and the achieved sample size is in the results (page 13, and the recruitment flowchart Figure 1).

“We were unable to achieve full recruitment to match the pre-determined sample size: the study recruited 160 eligible participants across both study arms, against our target of 240 participants with complete data”.

Point 3) The study failed to show a significant difference between the two groups in the primary outcome. Is this a result of a lower statistical power?

The low recruitment (and hence lower power) reduced our ability to detect a difference between the groups: however, the confidence intervals for the estimates of differences reflect the achieved sample size, and hence we can make some interpretations of the likely “real” effect based on our study that suggest it is unlikely that there is a meaningful difference between the two study arms. A related issue was also raised under point 7, which is discussed below.

Point 4) The inclusion criteria seem to be arbitrary. The physician decides about the eligibility of the patient for the study. If the screening was positive, the patient was screened by the K10 for study inclusion. Why did you not provide consecutively the K10 to all patients to screen for study eligibility?

The study is a pragmatic trial which mirrors current clinical practice in primary care. The GP decides during the consultation whether from a clinical perspective the patient has mild / moderate mental health problem, and hence is suitable for inclusion in the study. The clinical decision was then followed by the K10 to ensure that the patient did not have a K10 indicating a higher level of mental stress than was suitable for the UBI intervention.

This study was not designed to offer a mental health intervention on the basis of those who achieved a certain K10 score per se (if screening all visitors the GP for mental distress), but to compare the outcomes of management for patients identified clinically as cases by the GP.

Point 5) The authors should provide more details on attrition bias. The rate of patients with at least one follow-up visit is reported and seems to be comparable between the two groups. However, did you observe groups differences at specific time points. It is possible that the acceptance of an intervention is lower and the patients dropped out earlier in follow-up.

Thank you for this comment: We have added sensitivity analyses that examine the impact of this loss to follow up on our primary outcome: these are discussed in response to Reviewer 3, point 5 above (summarised on p.18 of the manuscript and in more detail in the Supplementary Material, and commented in the discussion on p. 20).

We note that the attrition across the study is fully reported in the recruitment flowchart (Figure 1) which reports attrition in both study arms across all time points. Most of the loss to follow up was

between recruitment (baseline) and first follow up at two months, with more participants dropping out between baseline and two months in the UBI arm (n=15/85) than in the control arm (n=6/75). Only two patients were lost following this two month measurement in the UBI arm, and three in the PAU arm.

Point 6) The two intervention groups are not balanced due to sex and age. Did you conduct sensitivity analyses by adjusting for both parameters?

As per the methods section (p.11), the analyses presented in the body of the paper are adjusted for these two factors and other demographic factors. We have now clarified this in the results section and labelled the tables as to the adjustment variables. Please also see our response to Reviewer 3's point #8 about analysis as per the original analysis protocol where we only noted that we would adjust for baseline values of the score being analysed (this analysis is now included in Supplementary Table R6, and results summarised on p. 18 of the main paper).

Point 7) Discussion, 2nd paragraph: The trial was designed as superiority trial. It is not possible to get any conclusion about the non-inferiority of the two interventions even in case of the full sample size. The sample size of a non-inferiority trial has to be determined separately based on a non-inferiority margin. Usually, the sample size of a non-inferiority trial is larger than of a superiority trial.

Thank you for this comment. We have not claimed non-inferiority for the intervention on the basis of our results, but have rather stated that one cannot rule out inferiority based on this trial's results (taking the pre-defined difference of 4 points on the K10 as our measure of an important size difference). This is effectively saying that with the sample size and data we have we cannot reach a firm conclusion about whether UBI might be inferior to PAU. This is a standard interpretation of a wide confidence interval that captures a meaningful difference in outcomes (which is effectively a non-inferiority margin) and this kind of statement can (and perhaps should) be made from any study designed and planned as either a superiority or a non-inferiority trial.

The statistical issues in considering interpreting confidence intervals from a superiority trial with regards to non-inferiority statements are covered in Section IV of:

Committee for Proprietary Medicinal Products (CPMP). Points to consider on switching between superiority and non-inferiority. British Journal of Clinical Pharmacology. 2001;52(3):223-8.

Which specifically covers power (sample size) in IV.2.2:

"As indicated in IV.1.2, the results provided by the confidence interval supply a concrete assessment of the precision actually achieved by a clinical trial, superseding any calculations of power carried out before the trial was undertaken. The position of the lower end of the confidence interval relative to the agreed criterion of noninferiority provides the key information for making decisions about noninferiority."

Reviewer: 5

Reviewer Name: Ailish Hannigan

Institution and Country: Graduate Entry Medical School, University of Limerick, Ireland Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below This is a well-presented paper with no major statistical issues. As acknowledged by the authors, the results are unlikely to be generalisable to other settings/countries with less well-resourced practice as usual but the paper gives an interesting and relevant account of the challenges of carrying out pragmatic trials of mental health interventions in primary care including recruitment challenges and changes to practice as usual services during the trial.

Point 1) There was an improvement over time in both groups for the primary outcome with a larger mean improvement for the PAU group. The authors describe the improvement as reasonable - is there an accepted minimum clinically important change for K10 and how does the improvement in each group relate to this?

The mean changes in the study groups from baseline were 7.6 point change in the PAU group and 5.9 in the UBI group. From previous studies a clinically significant change has been estimated at between 6 and 7 points. See:

Rickwood DJ, Mazzer KR, Telford NR, Parker AG, Tanti CJ, McGorry PD. Changes in psychological distress and psychosocial functioning in young people visiting headspace centres for mental health problems. The Medical Journal of Australia. 2015 Jun 1;202(10):537-42.

Point 2) From Table 3, at 8 weeks the mean difference in change in K10 between the two groups is close to zero. From 8 weeks onwards, the trend is for PAU to do better – in fact from Figure 2 mean K10 for UBI increases between 2 and 3 months. It raises the question as to what happens to the UBI group after the 5/6 week brief intervention. They revert to practice as usual? Further comment and discussion from the authors on this would be useful.

Following the UBI intervention , patients revert to PAU, and will continue to receive available treatment options appropriate to their clinical status.

While we would agree that there is a tendency for PAU to do better after 2 months, we would suggest that the change in K10 score between two and three months cannot be described as an increase: the difference between the two measurements is very small, considered relative to the width of the confidence intervals (and hence the precision of those estimates).

Point 3) It would also be useful to know more about the participating GPs in each group – age, gender, ethnicity, years' experience, and specialist training/interest in mental health– is this data available?

Unfortunately we did not collect these data, and hence cannot present any results on this issue.

Point 4) There are two important typos – the conclusion in the abstract should read 'The UBI intervention did not lead to better outcomes..' and on page 13, line 46, the confidence interval is minus 1.18 to 4.85.

Thank you for spotting these: we have corrected them.

Point 5) It may be helpful for the reader to clarify that in Table 3 these are mean differences between groups of the change in outcomes over specified time periods.

Thank you for the comment: we have added this point to the title of the table. Note that because the linear mixed models adjust for baseline score (effectively forcing the initial outcomes measures to the same value in the two arms) the difference at any follow-up can equivalently be stated as either the mean difference at that time point, or the mean difference in changes since baseline.

VERSION 2 – REVIEW

REVIEWER	Ava Schulz University of Zurich
REVIEW RETURNED	12-Nov-2018

GENERAL COMMENTS	This RCT investigates the effects of a guided self-help mental health intervention in primary care. Results suggest that the intervention did not result in improved outcomes compared to PAU. This is the second round of reviews that I was able to contribute to. My main concerns with the manuscript in its original version were the lack of sensitivity analyses and a need for a more in-depth analysis of the null-results. All these concerns have been sufficiently addressed by the authors by including additional analyses and a more elaborated discussion. Further, a section to justify the authors' definition of clinically relevant change was added. Rather vague points in the methods section have been clarified and the discussion is now much more detailed. I especially appreciated the added section on p. 19 that put the project in the context of the local national health care system. Taken together, the manuscript has strongly benefitted from the revisions. In conclusion, I find the manuscript in its revised form fit for publication and comment the authors on their great work.

REVIEWER	Juan Bellón El Palo Health Centre (SAS), redIAPP, IBIMA, Department of Public Health and Psychiatry, university of Málaga, Spain
REVIEW RETURNED	06-Nov-2018

GENERAL COMMENTS	I would like to thank the authors for their effort and detail in answering the many questions that were raised by the 5 reviewers. The issue raised in the study is very pertinent and some of the questions I proposed have been explained and resolved satisfactorily. This trial is presented as a study with negative results; however, in my opinion two reasons still persist and limit the interpretation of the results as negative: 1) The sample size did not reach a minimum power to conclude as a negative result. Taking the K10 as an outcome and to obtain a power of 80% and alpha:0.05, with the real data of the study [power twomeans 29, diff (1.68) sd (6) m1 (22) m2 (12) rho (0.13)], 1292 patients would have been needed (STATA v15.2). With the 113 patients recruited in the study, the statistical power barely reached 10% 2) Possibly the patients in the control group received more potentially effective treatments (standard psychotherapy and pharmacological treatment) than those in the intervention group. This is logical since the GPs of the intervention group were confident that the UBI intervention could be effective, and a previous pilot study already demonstrated the feasibility and satisfaction of the GPs with the UBI. This is a post-randomization bias and the only way to control it is by adjusting in the analysis for these variables. However, the researchers did not collect information on these variables.

REVIEWER	Jens Klotsche German Rheumatism Research Center Berlin, Germany
REVIEW RETURNED	12-Nov-2018

GENERAL COMMENTS	The authors responded to my concerns adaquately. I do not have any further comment.

REVIEWER	Ailish Hannigan Graduate Entry Medical School University of Limerick Limerick Ireland
REVIEW RETURNED	23-Nov-2018

GENERAL COMMENTS	All my comments have been adequately addressed. It is surprising that background information wasn't collected on participating GPs - this should be added as a limitation. There is a sentence in the discussion 'Clinicians who participated in this study might be expected to be those who were motivated and skilled in supporting patients with mental health problems' but it needs to be clarified that no information on their background was collected and this is a major limitation.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 3

Reviewer Name: Juan Bellón

Institution and Country: El Palo Health Centre (SAS), redIAPP, IBIMA, Department of Public Health and Psychiatry, university of Málaga, Spain

I would like to thank the authors for their effort and detail in answering the many questions that were raised by the 5 reviewers. The issue raised in the study is very pertinent and some of the questions I proposed have been explained and resolved satisfactorily. This trial is presented as a study with negative results; however, in my opinion two reasons still persist and limit the interpretation of the results as negative:

1) The sample size did not reach a minimum power to conclude as a negative result. Taking the K10 as an outcome and to obtain a power of 80% and alpha:0.05, with the real data of the study [power twomeans 29, diff (1.68) sd (6) m1 (22) m2 (12) rho (0.13)], 1292 patients would have been needed (STATA v15.2). With the 113 patients recruited in the study, the statistical power barely reached 10%

We have followed the Editor’s comments (as given above) and have more carefully positioned our conclusions with respect to whether this is a “negative study” –these changes have been noted and described above.

We would like to note that the CONSORT statement specifically discourages post-hoc power calculations (especially when using the observed group difference from a study, as suggested,, rather than an a priori clinically meaningful difference), stating “There is little merit in a post hoc calculation of statistical power using the results of a trial; the power is then appropriately indicated by confidence intervals (see item 17).”

Moher et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869. doi: 10.1136/bmj.c86

A more detailed discussion is given in the reference cited by the CONSORT statement:

Goodman SN. The Use of Predicted Confidence Intervals When Planning Experiments and the Misuse of Power When Interpreting Results. Annals of Internal Medicine 1994;121(3) doi:10.7326/0003-4819-121-3-199408010-00008

2) Possibly the patients in the control group received more potentially effective treatments (standard psychotherapy and pharmacological treatment) than those in the intervention group. This is logical since the GPs of the intervention group were confident that the UBI intervention could be effective, and a previous pilot study already demonstrated the feasibility and satisfaction of the GPs with the UBI. This is a post-randomization bias and the only way to control it is by adjusting in the analysis for these variables. However, the researchers did not collect information on these variables.

Thank you for this note: we have added this point as a limitation in the discussion (second sentence below is new text on p. 21):

“Secondly, in this New Zealand context, the GPs in the PAU group had access to a sophisticated range of therapy options which included providing extended consultations themselves, as well as referring patients to psychological therapies such as counselling or CBT delivered by clinical psychologists (Dowell 2009). This introduces the possibility of post-randomisation bias in the control arm due to differential receipt of these other treatments: however, we did not collect details from patients on receipt of such treatments, and thus could not address this potential bias in our analyses.”

Reviewer: 4

Reviewer Name: Jens Klotsche

Institution and Country: German Rheumatism Research Center Berlin, Germany

The authors responded to my concerns adaquately. I do not have any further comment.

Thank you for the feedback.

Reviewer: 1

Reviewer Name: Ava Schulz

Institution and Country: University of Zurich

This RCT investigates the effects of a guided self-help mental health intervention in primary care. Results suggest that the intervention did not result in improved outcomes compared to PAU. This is the second round of reviews that I was able to contribute to. My main concerns with the manuscript in its original version were the lack of sensitivity analyses and a need for a more in-depth analysis of the null-results. All these concerns have been sufficiently addressed by the authors by including additional analyses and a more elaborated discussion. Further, a section to justify the authors' definition of clinically relevant change was added. Rather vague points in the methods section have been clarified and the discussion is now much more detailed. I especially appreciated the added section on p. 19 that put the project in the context of the local national health care system. Taken together, the manuscript has strongly benefitted from the revisions. In conclusion, I find the manuscript in its revised form fit for publication and comment the authors on their great work.

Thank you for these comments.

Reviewer: 5

Reviewer Name: Ailish Hannigan

Institution and Country: Graduate Entry Medical School, University of Limerick, Limerick, Ireland

All my comments have been adequately addressed. It is surprising that background information wasn't collected on participating GPs - this should be added as a limitation.

There is a sentence in the discussion 'Clinicians who participated in this study might be expected to be those who were motivated and skilled in supporting patients with mental health problems' but it needs to be clarified that no information on their background was collected and this is a major limitation.

Thank you for this suggestion: we have amended the text to note this as a limitation and make it clearer that this is a speculative point on which we have no internal study data.

New text is on page 22:

“We might expect that clinicians who participated in this study would be those who were motivated and skilled in supporting patients with mental health problems. This is a speculative point, as we did not collect this kind of data on clinician experience, which is a limitation of the study and needs to be considered when thinking about the generalisability of the current results to other settings.”