PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	COLLABORATIVE CARE MODEL FOR DEPRESSION IN RURAL NEPAL: A MIXED METHODS IMPLEMENTATION RESEARCH STUDY
AUTHORS	Rimal, Pragya; Choudhury, Nandini; Agrawal, Pawan; Basnet, Madhur; Bohara, Bhavendra; Citrin, David; Dhungana, Santosh; Gauchan, Bikash; Gupta, Priyanka; Gupta, Tula; Halliday, Scott; Kadayat, Bharat; Mahar, Ramesh; Maru, Duncan; Nguyen, Viet; Poudel, Sanjaya; Raut, Anant; Rawal, Janaki; Sapkota, Sabitri; Schwarz, Dan; Schwarz, Ryan; Shrestha, Srijana; Swar, Sikhar; Thapa, Aradhana; Thapa, Poshan; White, Rebecca; Acharya, Bibhav

VERSION 1 – REVIEW

REVIEWER	van Marwijk, Harm University of Brighton, Division of Primary Care and Public Health
REVIEW RETURNED	17-Feb-2021

GENERAL COMMENTS	Thank you for the opportunitiy to review this interesting and sympathetic paper. I liked it but have some critical reflections. First, it would benefit from using a checklist like the Standards for Reporting Implementation Studies (StaRI) Statement (BMJ). The first sentence of the abstract creates a slightly complicated contrast for me. 'Despite the high burden of depression and availability of evidence-based mental health service delivery models (e.g., the collaborative care model), patients in low-income countries lack access to adequate care.' To put these alongside even as a contrast is not so simple. Does it mean: 'Let us all embrace US ideas on healthcare and all will be well?' The US healthcare model, with its high levels of inequalilty, fragmentation, discontinuity, and costs, does not come naturally to me as the best exemplar system to export to poor countries (or to use as a chronic care comparator, any intervention will show effects compared to low levels of US long-term conditions usual care). There is some evidence of effectiveness, with small effect sizes (0.3) but these might not survive the transfer to other poorer countries (or be much better, the reader might want to see such as argument...). The CC model is also not an actual statistical or other 'model' as such but rather a clever clinical combination of all kinds of organisational and other improvement measures, ideas and interventions, but these tend to lack standardisation. Having been trained in CC by the original IMPACT team, I am clinically a big fan of CC but every new project seems to use a (slightly?) different version. A big question is also whether the effect estimates are derived from rather well-controlled settings (as in the IMPACT trial). I think a somewhat more critical discussion is

	needed of whether, if collaborative care models are so good, they have been implemented in the actual world of care across the world, and, if so, what elements work best, are sustainable etc (as in the Nigerian CC depression projects). There are a lot of studies but having tried to implement the model in several projects myself, it is a powerful idea that I like, but the 'common sense' element of it can make it actually hard to implement. 'Depression' can also be a medicalised lens in primary care, we found, for instance, that creates tensions in the clinical world (as in the UK). The PHQ was developed with a grant from Pfizer, as an example. A primary care tool free of pharma-input like the 4DSQ that also addresses the various other symptom dimensions relevant for primary care, such as distress and somatisation, seems a much better idea. Many problems in primary care are basically social or due to a lack of resources... A lot of problems are contextual 'distress' rather than 'depression'. Such medicalisation of distress can be an unwanted side effect of introducing these models in resource-poor settings. Would it increase their reliance on antidepressant medication or would simple behavioural interventions be enough (the problem-solving treatment kind)? . I agree that implementation and effectiveness data on real-world care delivery models are critical to help inform the expansion of access to quality mental healthcare. The effect data take centre stage in this paper, while (for real-world implementation), the sustainability should be: are they actually going to stay using this model? What elements work best? What would they need to keep doing it? The cohort data that now are central should have less emphasis. I must say I was surprised by the large amount of authors.

REVIEWER	Bachmann, Max Univeristy of East Anglia, Norwich Medical School, University of East Anglia
REVIEW RETURNED	24-Feb-2021

GENERAL COMMENTS	This paper usefully reports on the implementation of a complex intervention to improve management of depression in Nepal. It describes improvements in indicators of the quality of care, some qualitative perceptions of the intervention, and modifications of the intervention over time. This is a worthwhile contribution to knowledge and should be published. Major comments The paper also reports on the reduction in depressive symptoms in a minority of participants with moderate or severe screen-detect depressive symptoms at baseline. This is of some interest but, without a control group, cannot be interpreted as providing evidence that the intervention was effective, because of the problem of regression to the mean. It is well known that, in trials of participants with high levels of depressive symptoms, depression scores frequently decline to lower levels over time, even in control groups. This has been shown in trials using PHQ-9 score as outcome measure (see for example our paper https://doi.org/10.1016/j.jad.2020.12.123, Table 3). Regression of PHQ-9 score downwards is particularly likely in this study because only participants with high or moderate scores were followed up, and those with low or no depression symptoms were excluded. So the finding that 49% of participants followed up had >50%

	reduction in PHQ-9 could be largely or entirely due to chance alone. The authors do acknowledge that the absence of a control group was a limitation, but they do not elaborate on this fundamental problem. Therefore I recommend: 1/ This should not be called an implementation-effectiveness study (in the Title, Methods and elsewhere) because it cannot demonstrate effectiveness. 2/ The Discussion should consider regression to the mean as outlined above. In addition to noting that the reduction in PHQ-9 is similar to that seen in trials in high income countries, it should also consider how much reduction in PHQ-9 has been reported in the control groups of controlled trials that used PHQ-9 as outcome. 3/ The Conclusion (lines 569-570) should be changed from "…our findings suggest that an adapted CoCM enhanced providers' perception and delivery of mental healthcare and improved clinical outcomes in our setting", to "…our findings suggest that an adapted CoCM enhanced providers' perception and delivery of mental healthcare in our setting, and improved clinical outcomes in participants with high or moderate PHQ-9 scores were observed." Minor comments 4/ In Methods, more information should be provided about the screening procedure. It is only stated that 862 known hospital patients living near the hospital were tested with PHQ-9, but not how they were identified and selected. 5/ It seems strange that the prevalence of high and moderate symptoms of depression in a community based sample should be so high (72%). This should be discussed. 6/ Please provide some more detail about changes in PHQ-9, that is, 95% confidence interval for the 49% with >50% reduction score, and range, 25th and 75th centile for change in score (lines 370-375).

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Prof. Harm van Marwijk, University of Brighton, Brighton and Sussex Medical School

Comments to the Author:
Thank you for the opportunity to review this interesting and sympathetic paper. I liked it but have some critical reflections. First, it would benefit from using a checklist like the Standards for Reporting Implementation Studies (StaRI) Statement (BMJ).

- **Thank you for taking the time to review our paper and provide helpful feedback. We are including a STaRI checklist with this re-submission.**

The first sentence of the abstract creates a slightly complicated contrast for me. 'Despite the high burden of depression and availability of evidence-based mental health service delivery models (e.g., the collaborative care model), patients in low-income countries lack access to adequate care.' To put

these alongside even as a contrast is not so simple. Does it mean: 'Let us all embrace US ideas on healthcare and all will be well?' The US healthcare model, with its high levels of inequality, fragmentation, discontinuity, and costs, does not come naturally to me as the best exemplar system to export to poor countries (or to use as a chronic care comparator, any intervention will show effects compared to low levels of US long-term conditions usual care).

- **Thank you. We echo your critique of the US healthcare model and have edited the first sentence of the abstract to avoid any confusion that we are advocating for it as a panacea for complex global health challenges. The sentence in Line 88 now reads:**
    - "Despite carrying a disproportionately high burden of depression, patients in low-income countries lack access to effective care."

There is some evidence of effectiveness, with small effect sizes (0.3) but these might not survive the transfer to other poorer countries (or be much better, the reader might want to see such as argument...). The CC model is also not an actual statistical or other 'model' as such but rather a clever clinical combination of all kinds of organisational and other improvement measures, ideas and interventions, but these tend to lack standardisation. Having been trained in CC by the original IMPACT team, I am clinically a big fan of CC but every new project seems to use a (slightly?) different version. A big question is also whether the effect estimates are derived from rather well-controlled settings (as in the IMPACT trial). I think a somewhat more critical discussion is needed of whether, if collaborative care models are so good, they have been implemented in the actual world of care across the world, and, if so, what elements work best, are sustainable etc (as in the Nigerian CC depression projects). There are a lot of studies but having tried to implement the model in several projects myself, it is a powerful idea that I like, but the 'common sense' element of it can make it actually hard to implement.

- **Thank you for this feedback. We agree that a nuanced discussion of how several implementers have implemented the various components of the CoCM can provide an important contribution to the literature. While the current study is focused on just one site, we have added in a note and reference to one of our previously published articles, which uses the WHO Health Systems Framework to describe some of the modifications, challenges, and opportunities in implementing collaborative care in several low- and middle-income countries (LMICs) settings, along with strategies to address these. We have included the following (Line 204-205):**
    - "Most of the prior evidence for CoCM came from high-income countries. Meanwhile, several challenges need to be addressed in order to successfully implement CoCM in LMICs.[17]"

'Depression' can also be a medicalised lens in primary care, we found, for instance, that creates tensions in the clinical world (as in the UK). The PHQ was developed with a grant from Pfizer, as an example. A primary care tool free of pharma-input like the 4DSQ that also addresses the various other symptom dimensions relevant for primary care, such as distress and somatisation, seems a much better idea. Many problems in primary care are basically social or due to a lack of resources... A lot of problems are contextual 'distress' rather than 'depression'. Such medicalisation of distress can be an unwanted side effect of introducing these models in resource-poor settings. Would it increase their reliance on antidepressant medication or would simple behavioural interventions be enough (the problem-solving treatment kind)?

- **Thank you. Regarding the choice of tool to assess depression, we chose to use the PHQ-9 since it had already been validated in Nepal (reference 25 in the manuscript) using rigorous methods from social sciences and health sciences to compare against the gold standard of a full structured clinical interview for DSM (SCID) to help ensure that people with distress were not being diagnosed. The validation**

process incorporated local idioms of distress to ensure that terms that related to depression were selected while terms related to everyday distress that do not rise to the level of depression were not included. We used this validated version for our study. While we agree that the 4DSQ is an excellent tool, it has not yet been validated in Nepal. In addition, since it is a 50-item tool, it likely would have increased the amount of provider time needed for administration (compared to the 9-item PHQ-9), thereby limiting its feasibility in the study setting. We note the reason for using the PHQ-9 in lines 269-270:

  - **"**We used the PHQ-9 since it has been cross-culturally adapted and validated in Nepal and other low-resource settings as a provider-administered scale.[25]**"**

- **The risk of picking up distress is an important caution, especially in primary care settings. In addition to using a validated tool that employed interdisciplinary methods to preferentially screen for people with depression (rather than distress), our analysis cohort only included patients with moderate to severe depression at baseline, which further ensured that we did not include outcomes on those with mild symptoms or distress. In addition, the risk of over prescribing antidepressants is an important caution and we thank the reviewer for noting this. For patients with mild depression, our protocols recommended therapy from the counselors rather than antidepressants. For patients who were included in this analysis (those with moderate to severe depression), we still recommended use of therapy alongside antidepressant medications.**

I agree that implementation and effectiveness data on real-world care delivery models are critical to help inform the expansion of access to quality mental healthcare. The effect data take centre stage in this paper, while (for real-world implementation), the sustainability should be: are they actually going to stay using this model? What elements work best? What would they need to keep doing it? The cohort data that now are central should have less emphasis. I must say I was surprised by the large amount of authors.

- **Thank you. As noted above, we have revised the paper to focus more on the implementation data instead of the cohort data. With regard to the large number of co-authors, this accounts for contributions from our multi-country and multidisciplinary team that drew expertise from psychiatry, primary care, implementation science, electronic health records, and data analytics.**

Reviewer: 2
Prof. Max Bachmann, Univeristy of East Anglia

Comments to the Author:
This paper usefully reports on the implementation of a complex intervention to improve management of depression in Nepal. It describes improvements in indicators of the quality of care, some qualitative perceptions of the intervention, and modifications of the intervention over time. This is a worthwhile contribution to knowledge and should be published.

- **Thank you for your encouraging comments.**

Major comments

The paper also reports on the reduction in depressive symptoms in a minority of participants with moderate or severe screen-detect depressive symptoms at baseline. This is of some interest but, without a control group, cannot be interpreted as providing evidence that the intervention was effective, because of the problem of regression to the mean. It is well known that, in trials of

participants with high levels of depressive symptoms, depression scores frequently decline to lower levels over time, even in control groups. This has been shown in trials using PHQ-9 score as outcome measure (see for example our paper https://doi.org/10.1016/j.jad.2020.12.123, Table 3). Regression of PHQ-9 score downwards is particularly likely in this study because only participants with high or moderate scores were followed up, and those with low or no depression symptoms were excluded. So the finding that 49% of participants followed up had >50% reduction in PHQ-9 could be largely or entirely due to chance alone. The authors do acknowledge that the absence of a control group was a limitation, but they do not elaborate on this fundamental problem. Therefore I recommend:

- **Thank you. We have added in a note on regression to the mean as a limitation, as suggested. As your paper and Whiteford et al. (2013) suggest, regression to the mean is more likely to be seen in patients with lower severity of illness, i.e. those with mild depressive symptoms are likely to be having a bad time (rather than having serious depression), and may improve with no intervention at all. As noted in your paper, "Three-quarters of participants with moderate symptoms at baseline had minimal to mild symptoms (PHQ-9 score of < 9) at 6 months, sustained at 12 months. However, roughly one-third of participants with a history of depression, or moderately severe to severe symptoms at baseline, showed no improvement at 6 and 12 months. In the face of low treatment exposure, these findings are not surprising given a higher probability of spontaneous remission in people with milder depressive symptoms compared to those with severe symptoms (Whiteford et al., 2013)." As noted in both papers, regression to the mean is less likely in those with severe depression, which is unlikely to spontaneously resolve. Our study at least partially addressed this issue of regression to the mean by including only patients with moderate or high symptoms at baseline (PHQ-9 score ≥10) in outcome analysis. As such, we have included the following note in lines 558-563:**
  - **"**As is common in mental health studies, regression to the mean, whereby patients may experience spontaneous improvement in their symptoms without any intervention, may have contributed to the observed results. However, regression to the mean is more likely to be observed in those with milder depressive symptoms.[32] This was at least partially mitigated since we had excluded patients who were most likely to regress to the mean through our inclusion criteria for baseline PHQ>9."

1/ This should not be called an implementation-effectiveness study (in the Title, Methods and elsewhere) because it cannot demonstrate effectiveness.

- **Thank you for providing this valuable feedback. We have revised the manuscript to remove the term "effectiveness", as suggested.**

2/ The Discussion should consider regression to the mean as outlined above. In addition to noting - that the reduction in PHQ-9 is similar to that seen in trials in high income countries, it should also consider how much reduction in PHQ-9 has been reported in the control groups of controlled trials that used PHQ-9 as outcome.

- **As noted above, we have now included the role of regression to the mean in the discussion.**

3/ The Conclusion (lines 569-570) should be changed from "…our findings suggest that an adapted CoCM enhanced providers' perception and delivery of mental healthcare and improved clinical outcomes in our setting", to "…our findings suggest that an adapted CoCM enhanced providers' perception and delivery of mental healthcare in our setting, and improved clinical outcomes in participants with high or moderate PHQ-9 scores were observed."

- **Thank you, we have modified this as suggested. This now reads as follows in lines 566-568:**

- **"…**our findings suggest that an adapted CoCM enhanced providers' perception and delivery of mental healthcare in our setting, and we observed improved clinical outcomes in patients with moderate or severe depression."

Minor comments

4/ In Methods, more information should be provided about the screening procedure. It is only stated that 862 known hospital patients living near the hospital were tested with PHQ-9, but not how they were identified and selected.

- **Thank you for pointing this out – we have clarified that patients were screened at the hospital based on PCPs' judgment using mhGAP protocols since universal screening was infeasible in the study context. The section on screening now says the following (lines 280-286):**
  - **"**Inclusion criteria were: …(ii) receiving care through CoCM at the study site primary care clinic and living in the hospital's immediate catchment area, (iii) assessed at least once with PHQ-9 during the study period, and (iv) having moderate or severe unipolar depression at baseline (i.e., PHQ-9 score ≥10). Local stakeholders noted that universal screening was infeasible and to facilitate real-world implementation, PCPs used clinical judgment based on the case descriptions in mhGAP protocols to decide which patients to screen for depression.[3]

5/ It seems strange that the prevalence of high and moderate symptoms of depression in a community based sample should be so high (72%). This should be discussed.

- **We have clarified (as in the point above) that this was a facility-based sample (not community based) as patients who presented for care in the clinic were preferentially screened. We have also noted under "Study site" that this area did not have access to any psychiatric services. As such, the community burden of untreated mental illness was high. We have also added in a note about this in the discussion, as suggested.**
  - Lines 189-190: **"**Access to mental healthcare is limited in this region, and the nearest psychiatrist is 14 hours away by road."
  - Lines 516-523: "As mental health services became available in our primary care system, a programmatically meaningful number of patients (862 over two years) were assessed using PHQ-9 in a setting where patients previously had limited or no access to quality mental healthcare. This likely included patients who previously sought tertiary care or did not seek care because of barriers like distance and stigma. PCPs at the facility screened patients with suspected mental illness, if their presentation was similar to the cases described in mhGAP training. These factors may explain the large proportion (73%) of patients with high or moderate symptoms of depression in our study."

6/ Please provide some more detail about changes in PHQ-9, that is, 95% confidence interval for the 49% with >50% reduction score, and range, 25th and 75th centile for change in score (lines 370-375).

- **Thank you. We have provided the suggested details. The section now reads (lines 495-500):**
  - **"**Of the 201 patients in the analysis cohort, 99 (49%, 95% CI: 42%, 56%) demonstrated substantial clinical response, i.e., their most recent PHQ-9 score was at least 50% lower than their baseline score. A subset of these patients (n=25, 12% of the cohort) showed remission in depression, i.e., their PHQ-9 score dropped to below 5 at their most recent follow-up (95% CI: 8%, 17%). The median change in PHQ-9 score in this cohort was -7 points [Q1: -9, Q3: -2] which was both statistically significant ($p<0.0001$) at $\alpha=0.05$ and clinically meaningful."

VERSION 2 – REVIEW

REVIEWER	van Marwijk, Harm University of Brighton, Division of Primary Care and Public Health
REVIEW RETURNED	10-May-2021

GENERAL COMMENTS	Nice rebuttal, well done

REVIEWER	Bachmann, Max Univeristy of East Anglia, Norwich Medical School, University of East Anglia
REVIEW RETURNED	12-May-2021

GENERAL COMMENTS	The authors have thoroughly revised the manuscript as recommended by this reviewer. It provides a detailed description of the process and experience of the implementation process, and of changes in patients' depression outcomes. The limitations are clearly discussed. I believe it is now suitable for publication without further revision.