PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	A Longitudinal Evaluation of a Countywide Alternative to the Quality and Outcomes Framework in UK General Practice, Aimed at Improving Person Centred Coordinated Care.
AUTHORS	Close, James; Fosh, Ben; Wheat, Hannah; Horrell, Jane; Lee, William; Byng, Richard; Bainbridge, Michael; Blackwell, Richard; Witts, Louise; Hall, Louise; Lloyd, Helen

VERSION 1 - REVIEW

REVIEWER	Mark Ashworth King's College London, UK I chaired the QOF 'Technical Working Group', 2017-18 for NHSE. This was the group responsible for making recommendations to NHSE on the future of QOF.
REVIEW RETURNED	18-Feb-2019

GENERAL COMMENTS	Thank you for asking me to review this paper. Comments: 1) Introduction: perhaps the opening sentence of para 3 (Line 99) could be amended: "In response to such criticisms, both the NHS Chief Executive and the General Practitioners Committee (GPC) Chairman have backed the removal of QOF." NHSE has decided to continue QOF in modified format so this opening sentence now looks rather politicised. True, their were important voices calling for removal; but there was also a counterargument which prevailed (in spite, of course, of the move away from QOF in Scotland). The same applies to the concluding 'Implications for the Future' (Line 414). 2) The Introduction provides a good summary of the limitations of the QOF. But it fails to provide a broad summary of the proposed alternative. A series of documents are cited. But the 'positive vision' for leaving QOF is not conveyed. Some summary of the essence and guiding principles of the 3 alternatives cited would help (Symphony Vanguard, Village Agents, Health Connections Mendip). Confusingly, some of the detail is in Box 1 and this lists 7 schemes: 4 with the label 'Test and Learn', one of which is called 'Test and Learn - South Somerset Vanguard' (is this the same as 'Symphony Vanguard' mentioned in the Introduction; one of which is called 'Village Agents Service (I presume the same as 'Village Agents' mentioned in the Introduction). In other words, the description of the QOF alternative needs to be rationalised so that it is consistent; and it needs to be clear (what was the vision - at the moment it comes across as a series of rather disconnected schemes?). Earlier in the Introduction, the helpful concept of P3C was introduced; how do these disparate initiatives relate to P3C.

	3) Introduction, line 124. It doesn't sound sufficiently neutral when researchers describe the scheme they are involved in with language such as: "There was a genuine passion and commitment to improving P3C". 4) Typos and minor errors: there are quite a few. Thus in the Methods pg7, one of the headings states 'Practioner' instead of 'Practitioner' (line 176). The P3C-OCT tool is discussed on pg 7 where the authors state that it "broadly correspond to five domains" - and then they list 6 domains, not 5, after this statement (line 189). Also a references to "SPSS" (line 230) when the authors mean, "SPQS". 5) At times, the authors refer to under performance which then turns out to be non-significant. In other words, it wasnt 'under performance'. For example, Line 280: "This suggests that SPQS practices were underperforming against the control group at time 1 (e.g. a score of 5.8 versus 6.2; p=0.64), whereas later in the evaluation, at time-point 2, this situation had been reversed (6.7 versus 6.2; p=0.41) – although these are both non-significant.". Much better to state that there was no significant difference between practices both before and after the intervention (then give the values). Similarly in the Discussion, the authors state (Line 387): "Whilst this evaluation did not assess costs on healthcare ..., a recent US-based review found large (albeit not statistically significant) average healthcare savings with interventions that have parallels to the models being deployed in Somerset". In fact, the finding was not so much 'large savings' and 'non-significant savings'. The interpretation of these findings in the Discussion does not appear impartial. 6) Multiple testing really should have a lower P value of 0.05 for a test of significance. Thus in Table 2, 6 sub-domains are tested. One has P value 0.03 and one P value of 0.00. I would not count the first P value as significant with multiple testing. Also, to be precise, there is no such thing as P=0.00. It should be P<0.001. 7) Results: 'Discretion from QOF and time saving' (Line 283) to 'Retention of QOF savings' (end Line 337). Almost all this data presented is qualitative. This was not part of this study. A further paper will be presented on qualitative findings. Without a clear description of qualitative methodology, the selection of quotes could be seen as 'cherry picking'. My own interpretation is that all qualitative data should be removed, as per the Methods description. Instead, the quantitative data, rather buried away in the Supplemental section, pg21, should be presented as this describes the perceived time savings. 8) Limitations of the study (Line 405): this section really ought to offer a more robust analysis of why the Practitioner Experience Survey showed no difference in intervention practices; whereas practitioners reported time savings in supplemental questions. Some attempt is made to answer this question (Lines 410-3) but more detail is required. How might this change have been detected? These instruments were internally designed - why did they not capture the key feature of the intervention (time freed up for patient centred care)? 9) References: having found non-significant patient experience and practitioner experience scores, the authors should reference

	the 3D study by Chris Salisbury, the largest RCT to date to explore multimorbidity (and also with negative findings but with a detailed analysis of why findings were negative). 10) Response to Referee comments pg36. There has been a good response to my own Referee comments (one of which I stand corrected and clearly I made an error myself - Line 33; my apologies). In particular, there is now a clearer description of why the 18 non-SPQS practices in Somerset were not used as control practices; also there is a stronger justification of the validity of the 3 selected survey instruments; also the addition of 'effect sizes' to the Results as presented. I welcome the positive response of the authors.

REVIEWER	Lindsay Forbes University of Kent, UK
REVIEW RETURNED	03-Mar-2019

GENERAL COMMENTS	This paper is improved. The abstract could set out the results more clearly, emphasising that the authors did not find any differences in patient or professional experience between intervention and control in the first sentence of the results. Focusing on the time savings, which was not the main element of data collection, is inappropriate. The order of the conclusions should be changed. I don't think 'discretion from QOF' is easy to understand. In the main text methods, I think it would be helpful to understand what the range of scores on the P3C measures are and what low and a high scores are and mean in terms of actual patient care. I am still a little unclear about what was actually different about the care delivered in the SPQS practices ; I find Box 1 difficult to understand - I have no idea what a 'symphony hub system' or a 'virtual hub' or 'reconnection to the community' are, for example. I think the authors should attempt to describe in concrete terms what was actually happening differently in the participating practices. Rather than have a patient and public involvement section, which looks a bit token, I suggest that what the patients/public actually did is incorporated in the methods in the relevant places. In the methods it is quite difficult to find out how data collection was carried out - I suggest this is put in a separate section from the data collection tools. There is still a bit of discussion in the methods and results. These should be presented without commentary on the validity of methods or meaning of results. Main text results - In the first section of the results, I think the authors mean ' the responses of the two groups are compared in table 1', rather than 'the two groups of responses are compared in table 1. The education columns of table 1 are bit odd - I am not sure what 'college' is in this context. It would be useful to know what a score on the P3c-EQ of 23-24 actually means. Is that high or low? I am not sure what 'admissible' means in this context. From line 283 to line 337 is all results for which the methods section provides insufficient data about the nature of the questions. I also still think much of the data in this part of the

	section should be included in the planned qualitative paper because they distract from the key message from the quantitative findings: SPQS was not associated with any detectable improvement or deterioration in validated measures patient or professional experience, or in hospital admissions over a relatively short time after implementation, but there was some evidence of organisational change for the better. There are several unexplained or doubly explained abbreviations and a few typos.

VERSION 1 – AUTHOR RESPONSE

Please leave your comments for the authors below

Thank you for asking me to review this paper. Comments:

1) Introduction: perhaps the opening sentence of para 3 (Line 99) could be amended: "In response to such criticisms, both the NHS Chief Executive and the General Practitioners Committee (GPC) Chairman have backed the removal of QOF." NHSE has decided to continue QOF in modified format so this opening sentence now looks rather politicised. True, their were important voices calling for removal; but there was also a counterargument which prevailed (in spite, of course, of the move away from QOF in Scotland). The same applies to the concluding 'Implications for the Future' (Line 414).

Fair point. These lines have been amended.

In the first instance, the change has been fairly minor, with the insertion of "… Chairman have previously backed the removal of QOF". This minor alteration does, however, set the political context for the time SPQS was started (which is important for the introduction)

In contrast, larger changes have been made to the discussion. This now reflects an up-to-date political context, with two recent news references.

"Whilst previous calls for the removal of QOF in England [52] have not been reiterated, recent policy has seen moves towards a reformed, streamlined version of QOF [53,54]. With QOF continuing to evolve, lessons from SPQS have…"

2) The Introduction provides a good summary of the limitations of the QOF. But it fails to provide a broad summary of the proposed alternative. A series of documents are cited. But the 'positive vision' for leaving QOF is not conveyed. Some summary of the essence and guiding principles of the 3 alternatives cited would help (Symphony Vanguard, Village Agents, Health Connections Mendip). Confusingly, some of the detail is in Box 1 and this lists 7 schemes: 4 with the label 'Test and Learn', one of which is called 'Test and Learn - South Somerset Vanguard' (is this the same as 'Symphony Vanguard' mentioned in the Introduction; one of which is called 'Village Agents Service (I presume the same as 'Village Agents' mentioned in the Introduction). In other words, the description of the QOF alternative needs to be rationalised so that it is consistent; and it needs to be clear (what was the vision - at the moment it comes across as a series of rather disconnected schemes?). Earlier in the

Introduction, the helpful concept of P3C was introduced; how do these disparate initiatives relate to P3C.

In our opinion, the implementation of SPQS was somewhat patchwork, and the vision not well laid out in the contract. In many ways, SPQS implementation was a series of “rather disconnected schemes”. Hence the lack of conveying a consistent, positive vision in our introduction.

Nonetheless, we have made amends to improve clarity in the introduction (lines 114-122). The ‘shared vision’ of these schemes have been stated. The schemes have been consistently names. Box 1 has been rearranged to make the link between the 3 ‘Test & Learn’ initiatives clearer.

3) Introduction, line 124. It doesn't sound sufficiently neutral when researchers describe the scheme they are involved in with language such as: "There was a genuine passion and commitment to improving P3C".

This sentence fragment has been deleted.

4) Typos and minor errors: there are quite a few. Thus in the Methods pg7, one of the headings states 'Practioner' instead of 'Practitioner' (line 176). The P3C-OCT tool is discussed on pg 7 where the authors state that it "broadly correspond to five domains" - and then they list 6 domains, not 5, after this statement (line 189). Also a references to "SPSS" (line 230) when the authors mean, "SPQS".

Thanks. Sorted.

5) At times, the authors refer to under performance which then turns out to be non-significant. In other words, it wasnt 'under performance'. For example, Line 280: "This

suggests that SPQS practices were underperforming against the control group at time 1 (e.g. a score of 5.8 versus 6.2; p=0.64), whereas later in the evaluation, at time-point 2, this situation had been reversed (6.7 versus 6.2; p=0.41) – although these are both non-significant.". Much better to state that there was no significant difference between practices both before and after the intervention (then give the values).

This has been altered as suggested.” Aggregate results for the P3C-OCT revealed that control practices had an aggregate score of 6.2 on the P3C-OCT, with no significant difference between SPQS and control practices either before (a score of 5.8 versus 6.2; p=0.64) or after (6.7 versus 6.2; p=0.41) the intervention.“

Similarly in the Discussion, the authors state (Line 387): "Whilst this evaluation did not assess costs on healthcare ..., a recent US-based review found large (albeit not statistically significant) average healthcare savings with interventions that have parallels to the models being deployed in Somerset". In fact, the finding was not so much 'large savings' and 'non-significant savings'. The interpretation of these findings in the Discussion does not appear impartial.

This sentence has been deleted, and is not necessary for publication

6) Multiple testing really should have a lower P value of 0.05 for a test of significance. Thus in Table 2, 6 sub-domains are tested. One has P value 0.03 and one P value of 0.00. I would not count the first P value as significant with multiple testing. Also, to be precise, there is no such thing as P=0.00. It should be P<0.001.

Both changes have been made, with significance now being tested with Bonferroni adjustment, and a line added in table to reflect this change.

7) Results: 'Discretion from QOF and time saving' (Line 283) to 'Retention of QOF savings' (end Line 337). Almost all this data presented is qualitative. This was not part of this study. A further paper will be presented on qualitative findings. Without a clear description of qualitative methodology, the selection of quotes could be seen as 'cherry picking'. My own interpretation is that all qualitative data should be removed, as per the Methods description. Instead, the quantitative data, rather buried away in the Supplemental section, pg21, should be presented as this describes the perceived time savings.

These lines (296-307) had been shortened from previous version, but have now been deleted.

I'm a bit confused about the "Instead, the quantitative data, rather buried away in the Supplemental section, pg21, should be presented as this describes the perceived time savings."

Is the reviewer referring to figure 2? In the final (published version), this figure should appear interlaced with main text (e.g. it won't be buried away in supplementary info).

8) Limitations of the study (Line 405): this section really ought to offer a more robust analysis of why the Practitioner Experience Survey showed no difference in intervention practices; whereas practitioners reported time savings in supplemental questions. Some attempt is made to answer this question (Lines 410-3) but more detail is required. How might this change have been detected? These instruments were internally designed - why did they not capture the key feature of the intervention (time freed up for patient centred care)?

I think there is some misunderstanding over instruments. The questions about time savings were presented in figure 2, and these were prepended to the P3C-OCT, because they were considered an organisational issue. We made no hypothesis that these time savings would improve practitioner experiences (because they just reflected a shift in workload, not an improvement).

To reflect this statement, this has actually been clarified much earlier in the paper (lines 323-4): “In this manner, the time savings leveraged from QOF were not hypothesised to lead to an improvement of experiences for practitioners, but instead a shift in workload.”

The time freed up from lack of SPQS was simple leveraged for greater involvement in P3C. This was established via results of P3C-OCT. Any benefits to practitioner experience would have been due to working in a more holistic/person-centred manner, rather than due to reduced workload.

9) References: having found non-significant patient experience and practitioner experience scores, the authors should reference the 3D study by Chris Salisbury, the largest RCT to date to explore multimorbidity (and also with negative findings but with a detailed analysis of why findings were negative).

I have added a comment and reference to this paper at lines 394-399.

10) Response to Referee comments pg36. There has been a good response to my own Referee comments (one of which I stand corrected and clearly I made an error myself – Line 33; my apologies). In particular, there is now a clearer description of why the 18 non-SPQS practices in Somerset were not used as control practices; also there is a stronger justification of the validity of the 3 selected survey instruments; also the addition of ‘effect sizes’ to the Results as presented. I welcome the positive response of the authors.

Thanks for the positive responses; the suggestions have all made a positive contribution to the paper. Thanks.

Reviewer: 2

Reviewer Name: Lindsay Forbes

Institution and Country: University of Kent, UK

Please state any competing interests or state ‘None declared’: None Declared

Please leave your comments for the authors below

This paper is improved.

The abstract could set out the results more clearly, emphasising that the authors did not find any differences in patient or professional experience between intervention and control in the first sentence of the results.

The sentence about patient/practitioner experience has been moved as suggested, to the first line of the results section.

Focusing on the time savings, which was not the main element of data collection, is inappropriate. The order of the conclusions should be changed. I don't think 'discretion from QOF' is easy to understand.

To be honest, I haven't come across anyone else who has struggled with this terminology.

Discretion: the freedom to decide what should be done in a particular situation.

To me, this seems unambiguous. Especially when it comes after the phrase "deincentivisation of QOF".

However, I have changed the offending phrase to "Discretion from QOF incentives". Hopefully this is clearer.

Also note that the abstract word limit is 300 words, so it is challenging to fully elucidate these sorts of concepts in an abstract.

In the main text methods, I think it would be helpful to understand what the range of scores on the P3C measures are and what low and a high scores are and mean in terms of actual patient care.

Score ranges, maximum and meaning have been indicated in the methods section.

I am still a little unclear about what was actually different about the care delivered in the SPQS practices ; I find Box 1 difficult to understand - I have no idea what a 'symphony hub system' or a 'virtual hub' or 'reconnection to the community' are, for example. I think the authors should attempt to describe in concrete terms what was actually happening differently in the participating practices.

This raises some similar points to the previous reviewer, and changes have been made as above to be more clearly indicate the over-arching goals of the schemes, and how these relate to the sub-schemes.

Again, I think the usage of the word 'hub' is completely unambiguous.

Hub: The effective centre of an activity, region, or network.

Nonetheless, I have reworded this section to help with clarity. “South Somerset Symphony Vanguard: A symphony “hub” system located at Yeovil District Hospital, where complex patients receive extra support from Health Coaches/Key Workers at the Symphony hub service, although they with complex patients remaining remain under management of GP practice”

With “virtual hub”, I think this phrase also makes sense to most people. Nonetheless, I have added the caveat of a “multidisciplinary team moving between practice”

Please note that due to word constraints, it is difficult to provide much more information about these services. In fact, some of the information has had to be put in a box to side-step these length problems. Also note that references are provided to the schemes themselves, where more details can be found. Also, our supplementary information includes the analysis of the STPs, which details the different involvement of practices, and highlights the patchwork nature of the SPQS implementation.

Rather than have a patient and public involvement section, which looks a bit token, I suggest that what the patients/public actually did is incorporated in the methods in the relevant places.

I agree, but this was included due to the formatting rules of BMJ Open.

In the methods it is quite difficult to find out how data collection was carried out - I suggest this is put in a separate section from the data collection tools.

I have added a “data collection” section after the tools, lines 218-229

There is still a bit of discussion in the methods and results. These should be presented without commentary on the validity of methods or meaning of results.

Main text results - In the first section of the results, I think the authors mean ' the responses of the two groups are compared in table 1', rather than 'the two groups of responses are compared in table 1.

Changed.

The education columns of table 1 are bit odd - I am not sure what 'college' is in this context. It would be useful to know what a score on the P3c-EQ of 23-24 actually means. Is that high or low?

I am not sure what 'admissible' means in this context.

I added sentence about score reflecting positive experiences of care.

“College” was originally qualified like this in the demographic questionnaire:

“College education/vocational training (3-4 years of post O-level or GCSE)”

I have amended the table to say “College/Vocational”. I realise this is still slightly ambiguous, but given the relative non-importance of this issue, I don’t think further discussion is necessary in the paper. However, I would be happy to remove this data, if requested. It has no impact on the findings.

Admissible has been qualified with “(i.e. complete and timely)”

From line 283 to line 337 is all results for which the methods section provides insufficient data about the nature of the questions. I also still think much of the data in this part of the section should be included in the planned qualitative paper because they distract from the key message from the quantitative findings: SPQS was not associated with any detectable improvement or deterioration in validated measures patient or professional experience, or in hospital admissions over a relatively short time after implementation, but there was some evidence of organisational change for the better.

This has been removed at the request of the first reviewer also, see above.

There are several unexplained or doubly explained abbreviations and a few typos.

BMJ open papers seem to have abbreviation explanations in both the abstract, and then repeated in the full text – I am just repeating this policy. I have, however, checked for further redundancy, and picked up a couple of instances, which have been corrected.

Please do tell me of any other typos that I may have missed.

VERSION 2 – REVIEW

REVIEWER	Mark Ashworth King's College London
REVIEW RETURNED	09-Apr-2019

GENERAL COMMENTS	Thank you for asking me to review this paper once more. I agree that the recent changes have resulted in considerable improvement. I believe that each of the Reviewer comments have been addressed either fully or partially. This is encouraging. Comments:

	1) There are still several typos. This really isn't acceptable in a 2nd revision. For example, Line 92: 'practices and ractice data'; Line 120 reads, 'The Quality Outcomes Framework (QOF)' - it should be 'Quality and Outcomes Framework'. Line 138 reads: 'reduces consultations to 'box-ticking' exercise' - it should read, 'to a 'box-ticking' exercise. And so on. There are many more. It just seems rather sloppy at this stage not to have checked language/wording. 2) I am concerned about something I had not questioned in previous reviews. The Results are now more clearly presented. But that has revealed another difficulty. The comparison used three instruments: patient experience (P3C-EQ), staff experience (P3C-practitioner) and organisational data (PC3-OCT). The first two instruments showed no significant difference between intervention and control practices. The positive findings emerged from the third instrument, PC3-OCT. However, this instrument was administered in different ways to intervention and control practices (Line 244): "Each SPQS practice was requested to complete the P3C-OCT at two time points (from Feb-Aug 2016 and Dec 2016-Mar 2017). In contrast, control practices only completed the P3C-OCT once (at Time 2)." Turning to the Results (Line 311), the scores improved significantly in intervention practices: "...an increase (0.9; p=0.034) in aggregate scores on the P3C-OCT between T1 (5.8) to T2 (6.7)". However, this is an uncontrolled observation. The control practices did not have the opportunity to demonstrate 'improvement. In fact, the mean score in control practices was not significantly different to the mean score in intervention practices. As a result, a more cautious interpretation of the Results is that 3 questionnaires were administered, and there were NO significant differences between intervention and control practices based on responses to all 3 questionnaires. It would be fair to say that intervention practices demonstrated 'significant improvements' (Line 316); but since it was not tested in control practices, we do not know if there were 'significant improvements' in control practices too. This needs to be clearly stated as a limitation. And the Results presented more cautiously. 3) The real positive findings of this study emerge from additional questions asked to the intervention practices alone. These practices were asked additional questions on the P3C-OCT questionnaire which related to the intervention itself. Respondents gave largely positive responses (Lines 333-354). Further benefits are listed in the para beginning Line 332, which described details reported by intervention practices of benefits arising from time saving through not having to complete QOF requirements. These findings form the core findings of this paper. I therefore think this paper should be presented in two parts. Firstly, the negative statistical findings from all three instruments. Secondly, the reported benefits of the intervention as elicited by 'additional questions' using a validated scale, the P3C-OCT. These 'additional questions' (Line 333) were only offered to intervention practices. But reported benefits were plausibly linked to the intervention, such as time saving, more Federation-level working, etc. However, these positive findings need clearer presentation. Firstly there were quantitative findings from the 'additional questions', summarised in Figure 2 and, as far as I can see, in Supplementary File 1 (pg24). The chart in Supplementary File 1 contains a lot of quantitative detail which could usefully be

	summarised in the Results. Secondly, there were qualitative findings reported in Freetext responses (Line 346). If these findings are to be reported, the qualitative methodology should be stated in the Methods; or the authors could simply give examples of freetext responses which support the quantitative data. 4) Supplementary File 1 appears to name individual practices. I may be mistaken, but if so, the identifiers should be removed.

REVIEWER	Lindsay Forbes University of Kent, UK
REVIEW RETURNED	22-Apr-2019

GENERAL COMMENTS	The manuscript is much improved. I have a couple of minor comments about the abstract that would improve it. I would suggest that in the 'setting' section, 'we evaluated the impact on care' is rather too general, rather it should be something like 'the impact on processes and outcomes of care' . In the results, I think that 'care delivery was altered via stronger federation-level agreements and informal networks etc' is speculative and should be labelled as such, or omitted from the abstract. There are still many typos, and a few inconsistent tenses, that I am sure subeditors will sort out.

VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name: Mark Ashworth

Institution and Country: King's College London

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below

Thank you for asking me to review this paper once more. I agree that the recent changes have resulted in considerable improvement. I believe that each of the Reviewer comments have been addressed either fully or partially. This is encouraging.

Comments:

1) There are still several typos. This really isn't acceptable in a 2nd revision. For example, Line 92: 'practices and ractice data'; Line 120 reads, 'The Quality Outcomes Framework (QOF)' - it should be 'Quality and Outcomes Framework'. Line 138 reads: 'reduces consultations to ‘box-ticking’ exercise' - it should read, 'to a 'box-ticking' exercise. And so on. There are many more. It just seems rather sloppy at this stage not to have checked language/wording.

These have been altered, thankyou.

However, it has to be pointed out that some of these points are rather pedantic. For instance, both disambiguations of QOF (i.e. Quality AND outcomes framework Vs. Quality Outcomes Framework) are in common usage, with the more fluent 'Quality Outcome Framework' becoming more common, as evidence by recent NHS digital usage, eg

https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/general-practice-data-hub/quality-outcomes-framework-qof

2) I am concerned about something I had not questioned in previous reviews. The Results are now more clearly presented. But that has revealed another difficulty. The comparison used three instruments: patient experience (P3C-EQ), staff experience (P3C-practitioner) and organisational data (PC3-OCT). The first two instruments showed no significant difference between intervention and control practices. The positive findings emerged from the third instrument, PC3-OCT. However, this instrument was administered in different ways to intervention and control practices (Line 244): "Each SPQS practice was requested to complete the P3C-OCT at two time points (from Feb-Aug 2016 and Dec 2016-Mar 2017). In contrast, control practices only completed the P3C-OCT once (at Time 2)." Turning to the Results (Line 311), the scores improved significantly in intervention practices: "...an increase (0.9; p=0.034) in

aggregate scores on the P3C-OCT between T1 (5.8) to T2 (6.7)". However, this is an uncontrolled observation. The control practices did not have the opportunity to demonstrate 'improvement. In fact, the mean score in control practices was not significantly different to the mean score in intervention practices. As a result, a more cautious interpretation of the Results is that 3 questionnaires were administered, and there were NO significant differences between intervention and control practices based on responses to all 3 questionnaires. It would be fair to say that intervention practices demonstrated 'significant improvements' (Line 316); but since it was not tested in control practices, we do not know if there were 'significant improvements' in control practices too. This needs to be clearly stated as a limitation. And the Results presented more cautiously.

This point has already been addressed in previous versions of the MS. We had already made changes in the results section to address the criticism that "there were NO significant differences between intervention and control practices". At lines 333-335, the paper already states "Aggregate results for the P3C-OCT revealed that control practices had an aggregate score of 6.2 on the P3C-OCT, with no significant difference between SPQS and control practices either before (a score of 5.8 versus 6.2; p=0.64) or after (6.7 versus 6.2; p=0.41) the intervention."

However, we have now added a line to the 'limitations' sections of the discussion, reflecting these suggestions (line 456-9).

3) The real positive findings of this study emerge from additional questions asked to the intervention practices alone. These practices were asked additional questions on the P3C-OCT questionnaire which related to the intervention itself. Respondents gave largely positive responses (Lines 333-354). Further benefits are listed in the para beginning Line 332, which described details reported by

intervention practices of benefits arising from time saving through not having to complete QOF requirements. These findings form the core findings of this paper.

I therefore think this paper should be presented in two parts. Firstly, the negative statistical findings from all three instruments. Secondly, the reported benefits of the intervention as elicited by 'additional questions' using a validated scale, the P3C-OCT. These 'additional questions' (Line 333) were only offered to intervention practices. But reported benefits were plausibly linked to the intervention, such as time saving, more Federation-level working, etc. However, these positive findings need clearer presentation

First, I disagree that these results (3 short questions about time savings)are “the core findings of the paper”. Whilst interesting, these were a secondary (and minor) question, which are inherently of a subjective nature.

Secondly, I disagree that “reported benefits were plausibly linked to the intervention, such as time saving, more Federation-level working, etc.” These questions were specifically asking about time savings from “removal of QOF”. The nature of the question specifically delimits that the time-savings are directly from QOF (and not from e.g. more federation level working).

These findings are very clearly displayed in both graph and narrative form. I do not see how they could be displayed more clearly.

Firstly there were quantitative findings from the 'additional questions', summarised in Figure 2 and, as far as I can see, in Supplementary File 1 (pg24). The chart in Supplementary File 1 contains a lot of quantitative detail which could usefully be summarised in the Results.

Figure 2 (questions about time savings) have no relation to Supplementary File 1 (results of the STPs), so I’m not quite sure how to address this point…

Nonetheless, it is a reasonable suggestion to provide totals for each “activity” mentioned in the STPs/Supplementary File 1.

However, we are limited to a total of five tables/figures with BMJ Open, so a new table cannot be inserted into the publication. Instead, I have updated Supplementary Figure 1 with a new column, showing a summary of the results.

Secondly, there were qualitative findings reported in Freetext responses (Line 346). If these findings are to be reported, the qualitative methodology should be stated in the Methods; or the authors could simply give examples of freetext responses which support the quantitative data.

As has already been discussed on the previous two rounds of review, this section has already been substantially altered. There is now only minimal freetext responses, which do indeed “simply give examples of freetext responses which support the quantitative data”: As stated in the paper, the free

text response boxes confirmed the plans of the STPs (see introduction and Supplementary File 1). Thus, for this point, we are already doing as suggested by the reviewer (e.g. only citing as an example of the data in the STPs)

4) Supplementary File 1 appears to name individual practices. I may be mistaken, but if so, the identifiers should be removed.

We are technically OK with this under permissions previously sought from practices, but the reviewer is probably correct to be cautious. Identifiers have been removed. I have updated the supplementary table.

Reviewer: 2

Reviewer Name: Lindsay Forbes

Institution and Country: University of Kent, UK

Please state any competing interests or state ‘None declared’: None

Please leave your comments for the authors below

The manuscript is much improved. I have a couple of minor comments about the abstract that would improve it. I would suggest that in the 'setting' section, 'we evaluated the impact on care' is rather too general, rather it should be something like 'the impact on processes and outcomes of care' .

Done.

In the results, I think that 'care delivery was altered via stronger federation-level agreements and informal networks etc' is speculative and should be labelled as such, or omitted from the abstract.

The wording is now more cautious.

There are still many typos, and a few inconsistent tenses, that I am sure subeditors will sort out.

Did another proof read.

VERSION 3 - REVIEW

REVIEWER	Mark Ashworth King's College London, UK
REVIEW RETURNED	03-May-2019

GENERAL COMMENTS	Thank you for asking me to review (re-review) this paper. Comments: 1) There are still small typos. This is disappointing after so many revisions. For example, the paper title (pg 3) misses out the word, 'and' which appears in other versions of the title (true, a relatively small point). 2) The relatively small changes made to the paper have resulted in more cautious presentation of the findings which is good. However, I wonder if the changes in this respect are sufficient? The Abstract illustrates the difficulty. The Abstract states: "The evaluation used matched data from 55 SPQS practices and 17 regional control practices for three survey instruments.". It would therefore be reasonable for the reader to assume that all the Results are derived from 'matched data'. Of the three instruments administered, there were NO differences between intervention and control practices - this should be explicit and clearly presented in the Abstract. The only statistically significant finding was: "significant increase in P3C (one of the three survey instruments) oriented organisational processes". This is an unmatched finding and should also be explicitly stated in the Abstract. The P3C instrument was not administered on a second occasion to matched practices and we therefore do not know if there were similar 'significant increases' in P3C in matched practices. In other words, the only statistically significant finding was derived from un-matched data. I do agree with the authors who stated in response to my Review: "This point has already been addressed in previous versions of the MS. We had already made changes in the results section to address the criticism that “there were NO significant differences between intervention and control practices”". Indeed, I agree that they have responded which is very positive. However, I still think they have not gone far enough, particularly in the Abstract which has the potential to mislead. Provided that claims of benefit are presented more cautiously, I would think the paper would tell a clearer story.

VERSION 3 – AUTHOR RESPONSE

We have made the minor changes mentioned above.

However, the problem with the abstract, (as always), is word constraints. Therefore, further elaboration is not possible (without a loss of clarity elsewhere). However, I have removed the word "matched" and added the words "SPQS practice data" to make a clearer presentation of where the positive findings are. This approach retains brevity, whilst also stopping any readers from over-inferring the results.