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)	Process evaluation of the Data-driven Quality Improvement in Primary Care (DQIP) trial: Case-study evaluation of adoption and maintenance of a complex intervention to reduce high-risk primary care prescribing
AUTHORS	Grant, Aileen; Dreischulte, Tobias; Guthrie, Bruce

VERSION 1 - REVIEW

REVIEWER	Martin Gulliford King's College London, UK
REVIEW RETURNED	12-Dec-2016

GENERAL COMMENTS	1. Ethics: The study has been approved by an ethics committee and participants gave consent to publication of anonymised data. However, as I understood the paper, the data are not anonymised. For example 'One GP in Orosay was highly motivated and engaged, but one felt they had no obligation to participate despite having received the initial payment'. Unless I mis-understood, this does not appear to be anonymised. I suggest replacing the GP locations with numbers or letters, throughout. Apart from this I thought this was a very good paper that could be published. 2. In the final sentence of the introduction, the aim of the study could be clearly stated. 3. The conclusions of the study do not appear to be generalisable beyond the present sample: 'the four case-study practices which immediately implemented the DQIP intervention ... In contrast, the six practices in which implementation was problematic were more variable' . The text of the paper would benefit from revision throughout so that general readers can draw generalisable conclusions. 4. Limitations do not appear to be acknowledged. Personally, I am not sure that sampling responders and non-responders is a good idea. These could represent extremes of (random/Normal) variation in primary care and may not provide stable differences from which we can draw firm conclusions. Regression to the mean could result. This could be one limitation.

REVIEWER	Barbara Clyne Royal College of Surgeons in Ireland Ireland
REVIEW RETURNED	13-Dec-2016

GENERAL COMMENTS	Many thanks for asking me to review this paper. The authors are to be commended on a significant body of research in the DQIP trial and associated process evaluation. This paper seems to form part of a series of papers reporting the data from the process evaluation. Given the size of the process evaluation this is an appropriate dissemination plan but as described below, I sometimes found it difficult to evaluate this paper where reference is made to other unpublished work. This current paper requires some further clarification for the reader to appraise it as an individual paper. 1. This paper reports a multiple/comparative case study, conducted within the context of a mixed methods process evaluation. While the title highlights this, the first part of the methods section (page 5) is a little convoluted in delivering this. I think the methods should describe this design before giving the context of the process evaluation. 2. Qualitative data collection: “the GP most involved in the DQIP work, a GP less involved…” – how was this defined? 3. Could the topic guide be included as an appendix, or a table highlighting the types of questions/areas addressed – seeing as the topic guide was revised over time? 4. Small point – was data fully anonysmised or pseudoanonymised? Are the quote IDs the real names of where the practices are? Maybe you could clarify this here. 5. Quantitative data collection: Reference to the unpublished paper is unhelpful here. Essentially, from my reading, quantitative data from the process evaluation was used to sample cases, and then to classify the case study practices in terms of reach etc. This needs to be made more explicit, page 7/8 ‘mixed methods’ should be amalgamated with this section 6. Analysis: Analysis section describes how the researchers were essentially ‘blind’ in terms of interpreting the data until the trial results were known. However, page 5 (case study sampling) describes how cases were sampled based on their reductions in high risk prescribing @ 4 months post intervention so essentially the researchers knew by proxy which practices were implementers/non-implementers. This seems contradictory to me. 7. Analysis: “A coding frame was developed inductively from field-notes, initial interviews and topic guides, framework[15] and logic model” – isn’t this then a combination of inductive and deductive methods seeing as the framework and logic model identified a priori themes? 8. Analysis: While NPT is an appropriate theoretical lens for exploring potential implementation and sustainability of the intervention, not enough detail is presented in the results section and I don’t think it is appropriate to describe it in the discussion. The methods and the results sections need to provide more clarity as to the process of how the findings from the initial coding/thematic analysis were mapped onto the four constructs and 16 sub-constructs of the NPT framework. A table could be used to convey this? 9. Results: the word legitimate/legitimation is used throughout the results without any definition as to what constitutes legitimation 10. Results: refer to the tool not working well in some practices, was it piloted? Is this a limitation 11. Results: Table 2 is ordered from the least to most successful adopters/implementers – is that dictated by reach? For example, Hirta delivered to 45% and had effectiveness of 75% but Taransay is

	'the most successful' with delivery of 33% and effectiveness 59%.Please clarify the assumptions this gradation are based on. 12. Discussion: Reference to unpublished papers again doesn't enhance the readers understanding of this paper and its contributions to the literature. 13. Discussion: While Pincer is the closest trial and evaluation to DQIP, this is not the only process evaluation of an RCT in primary care. There is a lot of learning presented here in terms of implementation and implications for future research in this area that the authors have not giving sufficient consideration to in the context of other research. 14. Discussion: the discussion section does not present strengths and weaknesses of the study. Minor comments • Abstract & intro: Define DQIP • Abstract: Setting – add country • Page 5, line 111 should read "to participate and to published anonymised data".

REVIEWER	Tony Avery University of Nottingham, UK Bruce Guthrie is a co-applicant on an NIHR Programme Grant for Applied Research that I am leading (NIHR PGfAR - RP-PG-1214-20012).
REVIEW RETURNED	29-Dec-2016

GENERAL COMMENTS	This is an important process evaluation because it relates to a study that reports an intervention that has been shown to substantially reduce high-risk prescribing. The insights into what worked and why will be useful for others planning similar interventions. Overall the approach to sampling seems reasonable in this study. Although the authors state that the sampling was purposive, did they really plan from the start to look at four practices with initially successful implementation and six practices with delayed implementation, or had they planned for five of each? This is not a critically important issue and I don't think the paper needs to be changed. The authors state that: 'Sampling was additionally structured to recruit practices starting the trial at different times, and to ensure a mix of small and large practices (<5000 and ≥5000 registered patients)'. I don't have data on the size of general practices in Fife, but practices with 5000 patients would be below average in England. Consider whether 'small and large' is the appropriate description here. The approach taken to qualitative data collection seems reasonable with around four people being interviewed in each practice, and the lead GP being interviewed 3-6 months later. Ideally, the research team would have conducted interviews in the early stages of the implementation, as well as six months after the implementation. Also, given that the authors found maintenance of the effects of the intervention at 12 months it would have been useful to have had some interviews at or beyond this point.

	The approach taken to analysis seems appropriate, including the use of Normalisation Process Theory (I thought this worked well). The main results are based on the analysis of a substantial volume of qualitative data. Overall, the results provide important insights into the success of the intervention. For the practices assumed to be 'initial non-adopters' there are insights that are important and confirm the findings of other studies, and common sense, such as the need to convince practitioners that the intervention is genuinely important, and the need for adequate administrative and healthcare professional time to implement the intervention. It would be interesting to know the views of the authors on the sustainability of this sort of intervention given the comment that they work was sometimes/often done out of hours: "…we ended up doing it in the evenings; it was after six o'clock, so it was sitting here at half past six, seven o'clock in the evening ploughing through patients' notes." Overall, the paper is well written and I think it will make a valuable contribution to the literature.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1 Comment 4. "Limitations do not appear to be acknowledged. Personally, I am not sure that sampling responders and non-responders is a good idea. These could represent extremes of (random/Normal) variation in primary care and may not provide stable differences from which we can draw firm conclusions. Regression to the mean could result. This could be one limitation."
Reviewer 2 Comment 14. "Discussion: the discussion section does not present strengths and weaknesses of the study.
RESPONSE
We note that reviewer 3 (comment 1) believed the sampling of practices to be reasonable. We agree that regression to the mean can be a consequence of sampling outliers, but we do not believe this to be the case in our study. Regression to the mean is a particular problem if outliers are sampled at a single time point, whereas we sampled on the basis of a visual assessment of a time series. Regression to the mean should also apply to both the low and high outliers, in which case we would expect that practices selected as clear 'initial responders' would turn out not be so successful by trial end. However, as figure 2 makes clear, the four initial responders remain at the upper end of the distribution of response which is not consistent with regression to the mean. Additionally, the qualitative judgement of final effectiveness (made blind to the quantitative data at trial end) is broadly consistent with the observed quantitative distribution of change in prescribing shown in figure 2, which is again not consistent with a regression to the mean explanation. We do accept that regression to the mean may still be present and now acknowledge it in the limitations.
A strengths and limitations section has been added starting line 465
"A strength of the study overall is that we were sufficiently resourced to develop and published a pre-planned and pre-specified study[18] and carry out a well-resourced and rigorous process evaluation of a third of all practices included in the trial. [19, 20] A strength of this analysis is the attention to context, illustrating its influence on intervention implementation including that context is not fixed, for example in relation to variable staffing over time. The sampling method we chose has advantages and disadvantages. Using run charts showing change in the targeted high-risk prescribing, we chose to sample six practices which did not appear to have implemented the intervention, and four practices

which had. We felt that purposive sampling for heterogeneity in this way would help us more clearly understand the barriers and facilitators to trial delivery and intervention implementation. We recognise that sampling from the two ends of the distribution of initial response may limit generalisability to a wider population, but this is mitigated by the fact that we included one third of practices participating in the trial in the process evaluation. In principle, it also risks regression to the mean where outliers from both ends of the distribution would be expected to become more similar to each other because being an outlier at one time point may be due to chance variation. We do not believe this was the case here, because the four initial responders remained amongst the most effective at trial end, and the two initial non-responders which delivered the largest reductions in high-risk prescribing by trial end were both judged qualitatively (blind to final trial results) to have significantly implemented after a delay."

Reviewer 1 Comment 3. "The conclusions of the study do not appear to be generalisable beyond the present sample: 'the four case-study practices which immediately implemented the DQIP intervention ... In contrast, the six practices in which implementation was problematic were more variable'. The text of the paper would benefit from revision throughout so that general readers can draw generalisable conclusions."

Reviewer 2 Comment 13. "Discussion: While Pincer is the closest trial and evaluation to DQIP, this is not the only process evaluation of an RCT in primary care. There is a lot of learning presented here in terms of implementation and implications for future research in this area that the authors have not giving sufficient consideration to in the context of other research."

RESPONSE

Generalisation from qualitative research is often problematic because of small sample sizes, although this process evaluation is larger than most since it includes in-depth examination of 10 of the 33 participating practices (interviews with multiple professionals including key informant interviews at more than one time point, with additional observational and quantitative data). We have added a statement to the start of the discussion section (line 357) to emphasise the sample included in this paper was a third of the trial sample which we believe adds weight to the findings "(30% of trial practices)." The issue of generalisability has been further addressed in the new limitations section starting at line 465 (described above), and additional text starting at line 441 which includes a broader discussion of the literature.

"These findings are likely to be generalizable beyond this immediate study and context. For example, Kennedy et al found that the Whole System Informing Self-management Engagement (WISE) intervention was not routinely adopted by practices because it was not perceived as relevant or legitimate activity or a priority for general practice, and it did not fit within existing work. These findings align with the findings of this study that lack of coherence and cognitive participation were important barriers to implementation, and therefore may be important targets for intervention.[26] Likewise, Berendesen et al found the BeweegKuur (combined lifestyle intervention) was not implemented according to protocol and had poor sustainability because patient's expectations of the intervention were not met and tailoring to general practice and patient contexts was required. [27] Moore et al have also shown how implementation of an exercise intervention varied by patient characteristics and context. [28] Like this study, the process evaluation of the implementation of the ESTEEM trial comparing GP versus nurse telephone triage of same-day appointment requests highlighted the importance of context on intervention implementation, and suggested that allowing autonomy to deliver the intervention to suit practice and patient contexts was likely important for effective implementation. [29]

Our published paper examining professional perceptions of the key components of the intervention delivered to practices (education, financial incentives and the informatics tool) found that all components were 'active' although at different stages of recruitment and adoption 4[20]. This paper adds that, within a clear definition of the work practices were expected to do with patients, allowing them the freedom to tailor implementation and develop their own processes to suit their context was important for effective implementation. The paper also shows that characteristics associated with implementation varied with context. Having a single motivated individual GP in small practices appeared to be sufficient for effective implementation, but shared vision and joint working additionally

appeared important in larger practices."
Reviewer 1 Comment 1. "Ethics: The study has been approved by an ethics committee and participants gave consent to publication of anonymised data. However, as I understood the paper, the data are not anonymised. For example 'One GP in Orosay was highly motivated and engaged, but one felt they had no obligation to participate despite having received the initial payment'. Unless I mis-understood, this does not appear to be anonymised. I suggest replacing the GP locations with numbers or letters, throughout."
Reviewer 2 Comment 4. "Small point – was data fully anonysmised or pseudoanonymised? Are the quote IDs the real names of where the practices are? Maybe you could clarify this here."
RESPONSE
We have clarified that the practice names used are pseudonyms (they are the names of uninhabited small Scottish islands), which are the same as used in the other two process evaluation papers published or in submission which facilitates reading across all three papers.
Line 133-134, new text in bold "All data has been anonymised, including by using pseudonyms for practices."
Reviewer 1 Comment 2. "In the final sentence of the introduction, the aim of the study could be clearly stated."
RESPONSE
The final sentence of the background section has been re-written to more clearly state our aims. Starting line 96 (new text in bold) "The aim of this analysis is to examine how different practices responded to the intervention delivered to them by the research team in terms of their adoption of the work, their reorganisation to deliver the intended change in care to patients, and whether implementation was sustained over time."
Reviewer 2 Comment 1. "This paper reports a multiple/comparative case study, conducted within the context of a mixed methods process evaluation. While the title highlights this, the first part of the methods section (page 5) is a little convoluted in delivering this. I think the methods should describe this design before giving the context of the process evaluation."

RESPONSE
The design sentence has been moved to the start of this section. Page 5 line 101. Moved text in bold. "The design was a mixed-method comparative case-study, with general practices the unit of analysis.[16, 17] The overall design and methods have been described previously in the published protocol.[18] In brief, we used a mixed-methods parallel process evaluation to examine the implementation of selected processes and their associations with change in high-risk prescribing at practice level."

Reviewer 2 Comment 2. "Qualitative data collection: "the GP most involved in the DQIP work, a GP less involved…" – how was this defined?"

RESPONSE
Definitions have been added starting line 125. New text in bold. "In each practice, the GP most involved in the DQIP work (leading the review work), a GP less involved (who may have been involved in the review work to some degree)…"

Reviewer 2 Comment 3. "Could the topic guide be included as an appendix, or a table highlighting the types of questions/areas addressed – seeing as the topic guide was revised over time?"
RESPONSE
We have attached topic guide 1 and the generic topic guide 2. Topic guide 2 was edited before the second interview with the GP most involved (GP 1 in each practice) based on their first interview responses and those of the practice manager, pharmacist and the less involved GP (GP2).
Reviewer 2 Comment 5. "Quantitative data collection: Reference to the unpublished paper is unhelpful here. Essentially, from my reading, quantitative data from the process evaluation was used

to sample cases, and then to classify the case study practices in terms of reach etc. This needs to be made more explicit, page 7/8 'mixed methods' should be amalgamated with this section."
Reviewer 2 Comment 12. "Discussion: Reference to unpublished papers again doesn't enhance the readers understanding of this paper and its contributions to the literature."
RESPONSE
The quantitative paper already has a revision under consideration at another journal and is likely to be published before this paper. We will update the reference before publication, and have also provided more information about the quantitative data used. The quantitative data collection section (starting line 135) has been amended as follows (new text is in bold).
"Data from the DQIP IT tool was used to sample practices using run-charts to visually assess change in performance in the first four months of implementation, but was not used alongside the qualitative analysis reported here which was otherwise blind to any quantitative process or outcome data. Data from the same source was also used after qualitative analysis was complete to measure and categorise reach, delivery to the patient and maintenance. Reach was measured as the percentage of patients flagged as needing a review who had received at least one review, delivery to the patient as the percentage of flagged patients who had further action in response to initial review (eg a medication change, or an invitation to consult), and maintenance as reach in the final 24 weeks of the intervention period. Effectiveness was defined as the relative change in the mean rate of high-risk prescribing trial primary outcome measure between baseline (the 48 weeks before the intervention) and the final 24 weeks of the intervention. Quantitative data was only used to explore whether the qualitative judgements made about implementation were consistent with observed data on reach, delivery, maintenance and effectiveness. Associations between quantitative practice-level process and effectiveness data will be reported separately [19]."
Reviewer 2 Comment 6. "Analysis: Analysis section describes how the researchers were essentially 'blind' in terms of interpreting the data until the trial results were known. However, page 5 (case study sampling) describes how cases were sampled based on their reductions in high risk prescribing @ 4 months post intervention so essentially the researchers knew by proxy which practices were implementers/non-implementers. This seems contradictory to me."
RESPONSE
In order to sample the case study practices, the qualitative researcher used run charts of change over time in the primary outcome approximately four months after each practice started the intervention. As the methods state, the sampling is therefore based on a visual assessment of change in prescribing as a proxy for successful initial implementation. That is quite distinct from knowing whether the main trial analysis showed that the intervention was effective, which the qualitative researcher was blinded to until qualitative analysis was complete. As already described, we have amended the quantitative data collection text starting at line 136 to make this clearer.
"Data from the DQIP IT tool was used to sample practices using run-charts to visually assess change in performance in the first four months of implementation, but was not used alongside the qualitative analysis reported here which was otherwise blind to any quantitative process or outcome data."
Reviewer 2 Comment 7. "Analysis: "A coding frame was developed inductively from field-notes, initial interviews and topic guides, framework[15] and logic model" – isn't this then a combination of inductive and deductive methods seeing as the framework and logic model identified a priori themes?"
RESPONSE
We have amended the text to reflect this (line 171). New text has been added in bold. "A coding frame was developed inductively and deductively from field-notes, initial interviews and topic guides, framework[15] and logic model…"
Reviewer 2 Comment 8. "Analysis: While NPT is an appropriate theoretical lens for exploring potential implementation and sustainability of the intervention, not enough detail is presented in the results section and I don't think it is appropriate to describe it in the discussion. The methods and the results sections need to provide more clarity as to the process of how the findings from the initial coding/thematic analysis were mapped onto the four constructs and 16 sub-constructs of the NPT framework. A table could be used to convey this?"

RESPONSE
We agree more clarity could be provided and have added some sentences starting at line 180. New text in bold below. We have also included a table used in data collection and analysis which shows our DQIP specific definitions for the NPT constructs.
“Analysis drew on Normalisation Process Theory (NPT), which focuses attention on how interventions become integrated, embedded and routinized into social contexts.[21] AG analysed the data collected in the study twice, inductively and deductively as described above and deductively based on Normalisation Process Theory. Each NPT construct was defined specifically for DQIP and these definitions can be found in table 3. Coding charts and memos were developed for each NPT construct and higher order constructs (coherence, cognitive participation, collective action and reflective monitoring). NPT interpretation and coding reliability was established through a workshop with NPT experts. The data was explored for negative cases.”
Reviewer 2 Comment 9. “Results: the word legitimate/legitimation is used throughout the results without any definition as to what constitutes legitimation”
RESPONSE
Legitimation is now defined in table 3 described in the previous response.
Reviewer 2 Comment 10. “Results: refer to the tool not working well in some practices, was it piloted? Is this a limitation?”
RESPONSE
We believe that it is a finding of a technical barrier to intervention implementation rather than a limitation of the process evaluation.
Reviewer 2 Comment 11. “Results: Table 2 is ordered from the least to most successful adopters/implementers – is that dictated by reach? For example, Hirta delivered to 45% and had effectiveness of 75% but Taransay is ‘the most successful’ with delivery of 33% and effectiveness 59%. Please clarify the assumptions this gradation are based on.”
RESPONSE
As stated above, the quantitative data in the table is provided to contextualise the qualitative judgements made about implementation success in each practice (those qualitative judgements were made blind to this trial-end quantitative data). We have added footnotes to tables 1 and 2 to clarify this.
“a. Ordered from top to bottom in terms of the practices judged from qualitative analysis to have been the least (top) to most (bottom) successful implementers”

Reviewer 2 Minor comment 1. “Abstract & intro: Define DQIP” –
RESPONSE
New text in bold
Line 22 “Explore how different practices responded to the Data driven Quality Improvement in Primary care (DQIP) intervention in terms of their adoption of the work…”
Reviewer 2 Minor comment 2. Abstract: Setting – add country
RESPONSE
New text in bold. “Ten (30%) primary care practices participating in the trial. Scotland, United Kingdom.”
Reviewer 2 Minor comment 3. Page 5, line 111 should read “to participate and to published anonymised data”.
RESPONSE
Unchanged because this is what is written in the text at this line.
Reviewer 3 Comment 1. “Overall the approach to sampling seems reasonable in this study. Although the authors state that the sampling was purposive, did they really plan from the start to look at four practices with initially successful implementation and six practices with delayed implementation, or had they planned for five of each? This is not a critically important issue and I don't think the paper needs to be changed.”
RESPONSE

The sampling frame had three pre-specified purposive elements (initial implementation, list-size and time of starting the intervention), but was also pre-specified in the published protocol to vary with initial analysis. The first eight case studies were balanced (four responders, four non-responders) but initial analysis showed more variation within the non-responders than the responders, and we therefore chose to recruit two non-responders as the final pair.
Reviewer 3 Comment 2. "The authors state that: 'Sampling was additionally structured to recruit practices starting the trial at different times, and to ensure a mix of small and large practices (<5000 and ≥5000 registered patients)'. I don't have data on the size of general practices in Fife, but practices with 5000 patients would be below average in England. Consider whether 'small and large' is the appropriate description here."
RESPONSE
Reflecting greater rurality, median listsizes in Scotland at the time of the trial were ~5200. We recognise that this is not the case in England, and have changed the text at line 118. New text in bold. "Sampling was additionally structured to recruit practices starting the trial at different times, and to ensure a mix of smaller and larger practices (<5000 and ≥5000 registered patients),…"
Reviewer 3 Comment 3. "It would be interesting to know the views of the authors on the sustainability of this sort of intervention given the comment that they work was sometimes/often done out of hours: "…we ended up doing it in the evenings; it was after six o'clock, so it was sitting here at half past six, seven o'clock in the evening ploughing through patients' notes."

RESPONSE
When the informatics tool was first switched on, approximately half of the patients requiring review across the entire one year duration of the trial were flagged, meaning that there was an initial high burden of work (interview 1 with the lead GP took place about this time), however, once this was completed, sustaining regular review was much less work. The key to sustainability is reducing new initiation to reduce future review burden. None of the practices considered the work burdensome, consistent with the degree of effort being considered worth it and/or this type of work being normal rather than particularly burdensome.

VERSION 2 – REVIEW

REVIEWER	Barbara Clyne Royal College of Surgeons in Ireland
REVIEW RETURNED	09-Feb-2017

GENERAL COMMENTS	The authors have addressed my concerns. Overall, the paper is well written and I think it will make a valuable contribution to the literature

REVIEWER	Anthony Avery University of Nottingham, UK I am chief investigator on an NIHR Programme Grant stating 1/3/17 and Bruce Guthrie is a co-investigator.
REVIEW RETURNED	12-Feb-2017

GENERAL COMMENTS	I am happy with the changes the authors have made to this paper.