

## 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

|                            |  |
|----------------------------|--|
| <b>TITLE (PROVISIONAL)</b> | Does sector matter for the quality of care services? A secondary analysis of social care services regulated by the Care Inspectorate in Scotland |
| <b>AUTHORS</b>             | Bach-Mortensen, Anders; Montgomery, Paul   |

### VERSION 1 – REVIEW

|                        |   |
|------------------------|---|
| <b>REVIEWER</b>        | Prof David Challis<br>PSSRU, University of Manchester, UK |
| <b>REVIEW RETURNED</b> | 04-Apr-2018   |

|                         |   |
|-------------------------|---|
| <b>GENERAL COMMENTS</b> | <p>The legend of tables 6 and 7 would benefit from the same format - 6 has public and third first; 7 has private first. The sign of the coefficient will change but interpretation is easier with constant format.</p> <p>A minor comment about the generalisability and size of Scotland and its data should be added.</p> <p>p14 line 46 should read "receive"</p> <p>p16 line line 15 should read "effects"</p> <p>p16 line 30 should read "consistent with"</p> <p>p16 line 40 should read "effects"</p> <p>There are a number of similar modest typo/drafting errors in the Discussion and Conclusions which should be addressed. The paper is of relevance and interest and with minor revision is worthy of publication.</p> |
|-------------------------|---|

|                        |   |
|------------------------|---|
| <b>REVIEWER</b>        | Geir Hiller Holom<br>Department of Health Management and Health Economics,<br>Institute of Health and Society, University of Oslo, Norway |
| <b>REVIEW RETURNED</b> | 16-May-2018   |

|                         |  |
|-------------------------|--|
| <b>GENERAL COMMENTS</b> | <p>General comments</p> <p>The paper's attempt to generalize findings across a wide range of very different social services is heroic. Unfortunately, the paper lacks a satisfactory level of detail.</p> <p>In general you need to justify better that 1) the services included are comparable across private, public and third sector (are they really providing the same services? What kind of Day Care of Children', what kind of 'Housing Support service', what kind of 'Adoption Service'?) and 2) the patients or users of the services (case-mix) are comparable (possible case-mix differences are briefly introduced by the authors in the Introduction section). This key and most important part of the paper is highly unsatisfactory. You may be comparing apples to pears and suffer from a serious</p> |
|-------------------------|--|

|  |   |
|--|---|
|  | <p>bias without a more detailed description of the services and (adjustment of) the users of the services.</p> <p>I would like to see more discussion of relevant previous international studies on the topic. The social care services, such as 'Day Care of Children', 'Housing Support service' and 'Adoption Service', are very different from specialized health services.</p> <p>The paper's lack of detail level is evident in the Discussion section and in the Conclusion section.</p> <p>Detailed comments</p> <p>The paper's readability needs to be improved. The recommended word limit is exceeded, as is the number of tables and figures (There are actually 7, but grouped into 4). The graphical presentation of the figures has significant room for improvement. The appendix is comprehensive, but appears confusing and lacks detailed citations in the text.</p> <p>Abstract</p> <ul style="list-style-type: none"> <li>- The abstract gives the impression of that the paper also assesses health services, while the paper is about social services/care services.</li> <li>- In the abstract and throughout the paper I would consider the use of 'private for-profit organisations', instead of 'private organisations'. In Table 1, the term for-profit is used. As also described in table 1, third sector organisations are often private...</li> </ul> <p>Strength and limitation</p> <ul style="list-style-type: none"> <li>- 'Only care service types that were implemented by both private, public and third sector organisations were included in the analysis.' This is not a strength, but allows comparison. 'There were few observations for organisations with very low quality levels.' The technical implications around this can be better explained to the reader.</li> </ul> <p>Introduction</p> <ul style="list-style-type: none"> <li>- The introduction section is too long and has too much theoretical focus for an empirical paper in a medical journal.</li> <li>- Relatively much attention is given to the payment and commissioning structures, but little information is presented about the process and differences among the three sectors. Since the paper's focus is about quality, the most important information should be limited and may be summarized in a box/table.</li> <li>- Table 1 should be removed or moved to the Appendix. I would not prioritize this table as it adds little value.</li> </ul> <p>Past research</p> <ul style="list-style-type: none"> <li>- This section should be removed completely. The most important and relevant existing literature should be included in the Introduction section (as background) or in the Discussion section.</li> <li>- In this section, and throughout the article, I would be careful to 1) uncritically compare papers with different quality metrics and 2) to compare health services to social care services. For example, mortality rates following surgery are very different</li> </ul> |
|--|---|

|  |   |
|--|---|
|  | <p>from various metrics used by organizations providing child minding, children day-care services, support services, care homes for the elderly, and housing.</p> <ul style="list-style-type: none"> <li>- The literature included in the paper is confusing and not directly relevant. This is also evident in the discussion section. Although there is little existing literature does this section have clear shortcomings.</li> <li>- (Note: In Holom and Hagen (2017), there were readmissions following surgery at PFPs, but the PFP patients were readmitted to a public or private non-profit hospital).</li> </ul> <p>Methods</p> <ul style="list-style-type: none"> <li>- 'The CI data is available for download on the official website and updated monthly. This belongs to the Data section.</li> <li>- 'The main purpose of the analysis is to test whether care organisations from different sectors vary in quality-related outcome'. This is no new information and should be removed.</li> <li>- 'CI' should be defined even though it is clear for the reader what this refers to.</li> </ul> <p>Explanatory variables</p> <ul style="list-style-type: none"> <li>- 'We exclude care service types which have no sector variation (e.g. child minding).' Which services were included/excluded?</li> <li>- 'We also ran the models on the largest sub-categories of services to explore if the results were generalisable to all included service types'. Which services were included in the sub-categories?</li> </ul> <p>Dependent variable</p> <ul style="list-style-type: none"> <li>- 'The dependent variables of interest are the four quality domains, as measured by the CI.' Please refer to the table.</li> </ul> <p>Descriptive statistics</p> <ul style="list-style-type: none"> <li>- 'The dataset includes information on 13,311 care organisations based in Scotland.' This is already included in the data section.</li> </ul> <p>Predicted probabilities</p> <ul style="list-style-type: none"> <li>- 7 figures are too many. The graphical presentation should be improved. Table 4 lacks citation in the text.</li> </ul> <p>Discussion</p> <ul style="list-style-type: none"> <li>- The present version does not, in a satisfactory way, explain the differences observed or at least speculate around potential explanations. Neither are the results, in a satisfactory way, discussed and compared to existing literature.</li> </ul> <p>Conclusions</p> <p>'Although the results are consistent to theory and past research, more research is needed to explore what exact characteristics (e.g. salaries) account for the variation in quality across sectors.' Do the authors believe that salaries can explain the variation across the different social services? How?</p> <p>Thank you for the opportunity to review this research.</p> |
|--|---|

## VERSION 1 – AUTHOR RESPONSE

### Reviewer 1

The legend of tables 6 and 7 would benefit from the same format - 6 has public and third first; 7 has private first. The sign of the coefficient will change but interpretation is easier with constant format. A minor comment about the generalisability and size of Scotland and its data should be added.

p14 line 46 should read "receive"

p16 line line 15 should read "effects"

p16 line 30 should read "consistent with"

p16 line 40 should read "effects"

There are a number of similar modest typo/drafting errors in the Discussion and Conclusions which should be addressed. The paper is of relevance and interest and with minor revision is worthy of publication.

We thank the reviewer for these comments. We have amended the text as suggested by the comments above. Also, we've added a section discussing the generalisability of the data considering the Scottish context on p 16, section 2.

### Reviewer 2

#### General comments

The paper's attempt to generalize findings across a wide range of very different social services is heroic. Unfortunately, the paper lacks a satisfactory level of detail.

In general you need to justify better that 1) the services included are comparable across private, public and third sector (are they really providing the same services? What kind of 'Day Care of Children', what kind of 'Housing Support service', what kind of 'Adoption Service?') and 2) the patients or users of the services (case-mix) are comparable (possible case-mix differences are briefly introduced by the authors in the Introduction section). This key and most important part of the paper is highly unsatisfactory. You may be comparing apples to pears and suffer from a serious bias without a more detailed description of the services and (adjustment of) the users of the services.

I would like to see more discussion of relevant previous international studies on the topic. The social care services, such as 'Day Care of Children', 'Housing Support service' and 'Adoption Service', are very different from specialized health services.

The paper's lack of detail level is evident in the Discussion section and in the Conclusion section.

We thank the reviewer for these helpful and rigorous comments. To address the main comments regarding the comparability of the included service types and their clients, we've made several efforts to provide clarification and justification of the heterogeneity of the data.

First, we've included a full description of the included services in Appendix A, which is referred to on p 5. This table outlines how the Care Inspectorate defines each service type.

Second, it's important to note that the service types included in the analysis are treated as they are classified in the data-set. That is, the analysis is conducted on the assumption that the definition of services provided by the Care Inspectorate is consistent for organisations across different sectors, which is in line with the fact that the Care Inspectorate follow the same inspection procedure regardless of sector. We've clarified this assumption in the methodology section on p 7.

Last, we've added a section in the discussion which addresses the concern that service types and clients may differ systematically across sector. To address this issue, we've controlled for all factors allowed for by the data, and re-run the analyses on all major service types to find the same pattern as when they are aggregated. However, we acknowledge that this is an issue deserving further attention going forward. The section now reads as follows:

"A more substantial limitation of the analysis is that service types may differ systematically across sector. For example, it may be that nursing homes or the clients served by FPOs are inherently different from third and public sector providers. To address this issue, we controlled for all organisational characteristics allowed for by the data (e.g. service types, client group, and number of employees), and ran the statistical models individually on all service types, which showed the same substantive pattern for all included services. However, considering that past research have identified differences in clients served by for-profit and non-profit hospitals [41], further research is needed to explore client variation across sector for social care services. "(p 14)

To the best of our knowledge, there is very limited research on this topic in the context of social care settings. The only study we know of is that of Barron and West (2017), which we've already included in the manuscript.

#### Detailed comments

The paper's readability needs to be improved. The recommended word limit is exceeded, as is the number of tables and figures (There are actually 7, but grouped into 4). The graphical presentation of the figures has significant room for improvement. The appendix is comprehensive, but appears confusing and lacks detailed citations in the text.

We've redesigned all figures to improve clarity and comprehension and cited the appendix in more detail throughout the text.

#### Abstract

- The abstract gives the impression of that the paper also assesses health services, while the paper is about social services/care services. <sup>[1]</sup><sub>SEP</sub>

To clarify this, we've repeatedly specified that the sample is 'social care organisations'.

- In the abstract and throughout the paper I would consider the use of 'private for-profit organisations', instead of 'private organisations'. In Table 1, the term for-profit is used. As also described in table 1, third sector organisations are often private... <sup>[1]</sup><sub>SEP</sub>

We thank the reviewer for this comment. Throughout the paper, we've exchanged 'private' with 'for-profit'.

#### Strength and limitation <sup>[1]</sup><sub>SEP</sub>

- 'Only care service types that were implemented by both private, public and third sector organisations were included in the analysis.' This is not a strength, but allows comparison.

We've removed this bullet point from the section.

'There were few observations for organisations with very low quality levels.' The technical implications around this can be better explained to the reader.

We thank the reviewer for this comment. We've rewritten the bullet point to: "There were few observations for organisations with very low quality levels, making comparison across sector difficult in these categories."

#### Introduction

- The introduction section is too long and has too much theoretical focus for an empirical paper in a medical journal. [SEP]

We thank the reviewer for this comment. While it is true that the paper is empirical, we believe that one of the flaws of existing research on this topic is the failure to clarify how sector can be thought to matter to service-related outcomes, as this is indeed the fundamental assumption behind any attempt to investigate sector differences. Therefore, we would very much prefer to keep some theoretical discussion, so that the distinction of sector is clear to the reader. However, we acknowledge that the section could be written more concisely and have therefore revised it heavily.

- Relatively much attention is given to the payment and commissioning structures, but little information is presented about the process and differences among the three sectors. Since the paper's focus is about quality, the most important information should be limited and may be summarized in a box/table. [SEP]

As mentioned above, we have heavily modified the introduction, hopefully allowing for a more concise background to the topic. There is unfortunately little research (to our knowledge) that addresses how the commissioning process differ for different providers in the context of social care organisations, which is something we're planning to explore going forward.

- Table 1 should be removed or moved to the Appendix. I would not prioritize this table as it adds little value. [SEP]

Related to the point of theory, we would very much prefer to keep the table in the text, so that the distinction of sector is clearly presented to the reader to enable clarity and definition. We've reduced our discussion of theory in the text, allowing the table to add important and significant information.

#### Past research [SEP]

- This section should be removed completely. The most important and relevant existing literature should be included in the Introduction section (as background) or in the Discussion section. [SEP]

To address this comment, we have heavily adapted this section and merged it with the introduction.

- In this section, and throughout the article, I would be careful to 1) uncritically compare papers with different quality metrics and 2) to compare health services to social care services. For example, mortality rates following surgery are very different from various metrics used by organizations providing child minding, children day-care services, support services, care homes for the elderly, and housing. [SEP]

We thank the reviewer for this comment. We believe it's important to summarise the literature on this topic and considering its heterogeneity, we've found it necessary to include literature from healthcare settings. However, as mentioned above we have made substantial modification to this section, which now mainly discusses the overview of reviews, which, although claiming to investigate healthcare services, also includes reviews on nursing homes.

- The literature included in the paper is confusing and not directly relevant. This is also evident in the discussion section. Although there is little existing literature does this section have clear shortcomings. [SEP]

See comment above.

- (Note: In Holom and Hagen (2017), there were readmissions following surgery at PFPs, but the PFP patients were readmitted to a public or private non-profit hospital).

We apologise for this misunderstanding.

Methods

- 'The CI data is available for download on the official website and updated monthly. This belongs to the Data section.'

We have moved this sentence to the data section.

- 'The main purpose of the analysis is to test whether care organisations from different sectors vary in quality-related outcome'. This is no new information and should be removed.

We have modified as suggested.

- 'CI' should be defined even though it is clear for the reader what this refers to.

We have modified as suggested, and clarified the acronym on page 5.

Explanatory variables

- 'We exclude care service types which have no sector variation (e.g. child minding). Which services were included/excluded?'

We thank the reviewer for this comment. We have now specified all service types with limited sector variation on p 7. These include: adoption services, adult placement services, child minding, fostering services, nurse agencies, and offender accommodation services.

- 'We also ran the models on the largest sub-categories of services to explore if the results were generalisable to all included service types'. Which services were included in the sub-categories?'

We thank the reviewer for this comment and have clarified what service types on which we re-ran the analysis on p 7. These include: home services, children day care services, housing support services, and support services.

Dependent variable

- 'The dependent variables of interest are the four quality domains, as measured by the CI.' Please refer to the table.

We have amended the text as suggested.

Descriptive statistics

- 'The dataset includes information on 13,311 care organisations based in Scotland.' This is already included in the data section.

We have deleted this sentence.

Predicted probabilities

- 7 figures are too many. The graphical presentation should be improved. Table 4 lacks citation in the text.

As addressed above, we've redesigned the figures to improve clarity and comprehension. Table 4 is cited and described on p 8.

#### Discussion

- The present version does not, in a satisfactory way, explain the differences observed or at least speculate around potential explanations. Neither are the results, in a satisfactory way, discussed and compared to existing literature.

To address this comment, we've modified the discussion to clarify: (1) How the findings correspond to existing literature on the same topic (i.e. Barron and West (2017)). (2) The concern of heterogeneity of the included services and the clients they serve. (3) That the associations found in the analysis, although being robust to rigorous control and sub-analyses, cannot be thought of as causal and further research on *mechanisms* are warranted going forward. (4) That the findings apply to context of the data (i.e. Scotland), and that further research is needed to explore if the findings can be replicated to other contexts both geographically and for different service types (e.g. healthcare).

#### Conclusions

'Although the results are consistent to theory and past research, more research is needed to explore what exact characteristics (e.g. salaries) account for the variation in quality across sectors.' Do the authors believe that salaries can explain the variation across the different social services? How?

As discussed in the limitation session, the reported results essentially constitute robust correlations, which cannot be thought of as causal. This comment is simply meant to point out the the fact that there may be many unobserved factors (such as wages/salaries) potentially confounding the associations identified by the analysis.

Thank you for the opportunity to review this research.

### VERSION 2 – REVIEW

|                        |  |
|------------------------|--|
| <b>REVIEWER</b>        | Geir Hiller Holom<br>University of Oslo, Faculty of Medicine |
| <b>REVIEW RETURNED</b> | 17-Jul-2018  |

|                         |  |
|-------------------------|--|
| <b>GENERAL COMMENTS</b> | <p>This manuscript has been modestly improved. I remain concerned that you may be comparing apples to pears and suffer from a serious bias without adjusting for the users of the services.</p> <p>Also, the services are still, in general, very broadly grouped, making trustworthy comparisons very difficult.</p> <p>Regarding readability, the paper has improved but there are still too many tables and the graphical presentations do not add much to the paper.</p> |
|-------------------------|--|

|                        |  |
|------------------------|--|
| <b>REVIEWER</b>        | Joe Nolan<br>Northern Kentucky University, USA |
| <b>REVIEW RETURNED</b> | 05-Aug-2018                                    |

|                                |   |
|--------------------------------|---|
| <p><b>GENERAL COMMENTS</b></p> | <p>Please note that I reviewed ONLY the statistical methods/results in this manuscript. I recommend major revision because an overly complex statistical model has been used, when it appears that a much better model, involving fewer limitations, is available. Change of model should result in results that are much easier for the reader to understand what's going on, while likely still supporting overall conclusions in a similar way. Please see attachment regarding the statistical review.</p> <p>Statistical Review</p> <p>Choice of Model<br/>         Logistic regression is an appropriate method here. However, given the lack of poor quality observations, a simplification to category 5 and 6 ('good quality') and category 1, 2, 3, and 4 ('not good quality') would make a lot of sense here. Similar mechanism of combining would be appropriate for medium and high risk. This would both simplify the reporting as well as the statistics, since it would allow for much more straight-forward binary logistic regression models to be used.</p> <p>Reporting of Predicted Probabilities<br/>         Generally, predicted probabilities should be preferred to odds ratios in any case, as they provide the reader with a much better understanding of the chance of a positive/negative event.</p> <p>Tables 3-5<br/>         I find these wholly inappropriate. There are two essential issues:<br/>         1. It is completely unclear what the 240 indicated p-values are actually testing. What is meant by change "over five levels of quality"???? What is represented by the 1,2,3,4,5 at the top of the table? Furthermore, there is no correction for multiplicity of testing. Therefore one would EXPECT around 10 significant results even in a reality where there is nothing to be found.<br/>         2. The tables report only point estimates. Confidence intervals should be used to ascertain, even for statistically significant results, whether or not they may be practically important.</p> <p>Table 6</p> <ul style="list-style-type: none"> <li>• See comments above, particularly as regards multiplicity.</li> </ul> <p>Figures 1-3<br/>         Figures 1-3 are important (and far more useful than Tables 3-5). I make the following notes:<br/>         1. Does the vertical axis represent probability? Or percent? Use one or the other and be consistent.<br/>         2. What is being estimated by the percentages is somewhat difficult to discern, and some of the markings are overlapping result in some things not coming out of the picture. The use of binary logistic regression model will simplify this tremendously, as you will only need ONE demarcation for "high". "not-high" would be inferred as the opposite. Therefore only two colors would be needed (one for Quality of Care, the other for Quality of Staff, in the first picture). Or perhaps, change the horizontal axis to the item and let three colors represent the different sectors, for one simplified graphic.<br/>         3. The confidence intervals provided by the bars are an absolute necessity and I'm glad they are included.<br/>         Both the Tables and Figures will become far easier to interpret/understand if quality levels are collapsed into a binary variable and binary logistic regression is used. In addition to those</p> |
|--------------------------------|---|

|  |   |
|--|---|
|  | <p>things, for the binary models the significance tests for Odds Ratio (for “ordsec” in each model) should be reported.</p> <p>Summary</p> <p>Simplification of the model to avoid the impact of some of the greatest limitations, as well as to simplify your reported, is highly recommended. I think you will still have viable results here, with the overall outcomes of your paper not changing a whole lot. Ultimately, I think the understanding of readers will also be substantially aided by the simplification as well.</p> |
|--|---|

## VERSION 2 – AUTHOR RESPONSE

### Author Response Letter

#### Reviewer 2:

This manuscript has been modestly improved. I remain concerned that you may be comparing apples to pears and suffer from a serious bias without adjusting for the users of the services.

We were curious about this reviewer comment. While we understand the concern of heterogeneity, we believe the manuscript offers an important, worthwhile and rigorous description and analysis of existing care services in Scotland. The heterogeneity is intentional and a strength of the analysis. For example, the consistency of our findings suggest that the direction of effects is similar across all included service types, although further research should be undertaken to replicate our results. Of course, we acknowledge that the results are not without flaws – as we also discuss in the manuscript – but believe that the novelty and consistency of our results are worthwhile to disseminate, in a number of ways. It contributes to a small, but growing body of research.

Finally on this point, we agree that it would be interesting to adjust for the type of service users across sector, but this information is not collected by the Care Inspectorate, and thus such adjustment is unfortunately not possible.

Also, the services are still, in general, very broadly grouped, making trustworthy comparisons very difficult.

See above, and note, we are only able to categorise the services in the way that they are grouped by the Care Inspectorate. We believe the consistency of the findings suggests that they are in fact likely to be trustworthy.

Regarding readability, the paper has improved but there are still too many tables and the graphical presentations do not add much to the paper.

The information provided in the tables was completed so as to adhere to the STROBE reporting guidelines (as required for BMJ Open publications). However, by merging tables 4-5, we have now reduced the number of tables to 6 and have also merged all the information from the figures into one (see Figure 1).

### Reviewer: 3

#### Choice of Model

Logistic regression is an appropriate method here. However, given the lack of poor quality observations, a simplification to category 5 and 6 ('good quality') and category 1, 2, 3, and 4 ('not good quality') would make a lot of sense here. Similar mechanism of combining would be appropriate for medium and high risk. This would both simplify the reporting as well as the statistics, since it would allow for much more straight-forward binary logistic regression models to be used.

We thank the reviewer for these comments. We put a lot of thought into how to operationalise the data and conduct the analyses in a rigorous manner. We therefore completely understand and acknowledge the appeal in dichotomising the data into two categories. We do, however, worry that doing so will send a misleading signal about the performance of *for-profit* organisations. Our findings show that although 'sector' appears to be a relatively consistent predictor of quality and risk of services, the relationship is characterised by non-linearity. Specifically, we don't find a lot of variation across sector on the probability of organisations performing 'poorly'. That is, we do not find that *for-profit* organisations are performing poorly (in the sense of performing below acceptable thresholds), but rather that public and third sector organisations are more likely to perform highly, whereas *for-profit* organisations are more likely to perform within 'adequate' threshold. Therefore, merging the 'poor' and 'adequate' categories into 'poor' may dissolve the quite important nuance of 'poor' and 'adequate'.

Further, it's important to note that the we have already merged the 6-point scale into three categories for more intuitive interpretation, i.e. 'good' (5+6), 'adequate' (3+4), and 'poor' (1 and 2) (see Figure 1 and Tables 5-6 in the manuscript).

We chose generalised ordered regression due to the ordering of the data, and also due to the fact that the proportional odds assumption was violated across all quality domains and in the risk outcome (which is a central assumption for ordered logistic regression). However, it is worth mentioning that we ran the same analyses using ordered logistic regression and multinomial regression models, which led to substantially similar results.

#### Reporting of Predicted Probabilities

Generally, predicted probabilities should be preferred to odds ratios in any case, as they provide the reader with a much better understanding of the chance of a positive/negative event.

We thank the reviewer for the comment, which summarises the rationale we followed when designing the analysis.

#### Tables 3-5

I find these wholly inappropriate. There are two essential issues:

It is completely unclear what the 240 indicated p-values are actually testing. What is meant by change "over five levels of quality"???? What is represented by the 1,2,3,4,5 at the top of the table? Furthermore, there is no correction for multiplicity of testing. Therefore one would EXPECT around 10 significant results even in a reality where there is nothing to be found. <sup>[1]</sup> <sub>[SEP]</sub>

The tables report only point estimates. Confidence intervals should be used to ascertain, even for

statistically significant results, whether or not they may be practically important. [SEP]

Thank you. It's important to note that these comments apply to the tables reported in our appendix, which are meant as complementary to the analysis. We have heavily revised Appendix A to avoid further confusion. Now all tables in both the appendix and manuscript include Bonferroni adjusted confidence intervals.

Table 6

- See comments above, particularly as regards multiplicity. Figures 1-3

Figures 1-3 are important (and far more useful than Tables 3-5). I make the following notes:

Does the vertical axis represent probability? Or percent? Use one or the other and be consistent. [SEP]

The vertical axis represents predicted probabilities in percentages and is now clearly labelled.

What is being estimated by the percentages is somewhat difficult to discern, and some of the markings are [SEP] overlapping result in some things not coming out of the picture. The use of binary logistic regression model will simplify this tremendously, as you will only need ONE demarcation for "high". "not-high" would be inferred as the opposite. Therefore only two colors would be needed (one for Quality of Care, the other for Quality of Staff, in the first picture). Or perhaps, change the horizontal axis to the item and let three colors represent the different sectors, for one simplified graphic. [SEP]

Thanks for this helpful comment. We have followed the suggestion to merge all information from the graphs into one simplified Figure (see Figure 1).

The confidence intervals provided by the bars are an absolute necessity and I'm glad they are included. [SEP]

Both the Tables and Figures will become far easier to interpret/understand if quality levels are collapsed into a binary variable and binary logistic regression is used. In addition to those things, for the binary models the significance tests for Odds Ratio (for "ordsec" in each model) should be reported.

Please see our first comment. The significance tests for the ORs and corresponding CIs for the complaint and requirement outcomes can be found in Table A4 in Appendix A.

Summary

Simplification of the model to avoid the impact of some of the greatest limitations, as well as to simplify your reported, is highly recommended. I think you will still have viable results here, with the overall outcomes of your paper not changing a whole lot. Ultimately, I think the understanding of readers will also be substantially aided by the simplification as well.

We acknowledge that demarcating the data will allow for an easier interpretation, but as addressed above, we believe that demarcating the data any further will result in a too simplified model, which will fail to meaningfully portray the relationship between the dependent and independent variables.

**Additional comments by Reviewer 3**

The reference to Tables 3-5 is a reference to the three tables in the appendix, beginning on page 27. On my initial read, I did not pick up on the "appendix" piece, as with several figures at the end it did not surprise me that some tables were at the end as well. I suppose it goes without saying at this point that it would be a good idea not to have multiple tables in the paper/appendix that have the same numbers!

We apologise for the confusion regarding the numbering of the tables. We have now renamed the tables in the appendix, so that the distinction should be clear.

As far as results go, Tables 3-5 in the appendix imply inference to a population, while tables 3-5 in the main paper are simply descriptive of the sample. Table 6, however, is in the main paper and has the same issues as the three tables in the appendix. Also, Figures 1-3 in the paper (which are connected/similar, I think, to the information in Tables 3-5 in the appendix) are much more useful in terms of population inference, since they include CI's. If the items in the appendix are to be included, they should be changed so that confidence intervals for the probabilities, not point estimates, are provided. This would eliminate a need for p-values in those tables, but those aren't particularly useful to begin with. An adjustment for multiplicity should also be made when producing the CI's.

We thank the review for this comment. All tables both in the manuscript and appendix now include Bonferroni corrected confidence intervals. This adjustment did, however, not lead to major changes in the results.

I hope this is helpful.

### VERSION 3 – REVIEW

|                        |   |
|------------------------|---|
| <b>REVIEWER</b>        | Geir Hiller Holom<br>University of Oslo |
| <b>REVIEW RETURNED</b> | 22-Oct-2018                             |

|                         |  |
|-------------------------|--|
| <b>GENERAL COMMENTS</b> | <p>Following the authors' first submission, I provided several general and detailed comments. I remain concerned about my crucial general comment:</p> <p>The paper's attempt to generalize findings across a wide range of very different social services is heroic. Unfortunately, the paper lacks a satisfactory level of detail.</p> <p>In general you need to justify better that 1) the services included are comparable across private, public and third sector (are they really providing the same services? What kind of 'Day Care of Children', what kind of 'Housing Support service', what kind of 'Adoption Service'?) and 2) the patients or users of the services (case-mix) are comparable (possible case-mix differences are briefly introduced by the authors in the Introduction section). This key and most important part of the paper is highly unsatisfactory. You may be comparing apples to pears and suffer from a serious bias without a more detailed description of the services and (adjustment of) the users of the services.</p> |
|-------------------------|--|

|  |   |
|--|---|
|  | <p>Following two rounds of review, I cannot see major changes or efforts in meeting these concerns. The author’s last response that “(...) The heterogeneity is intentional and a strength of the analysis. For example, the consistency of our findings suggest that the direction of effects is similar across all included service types, although further research should be undertaken to replicate our results.” is not about or addressing this concern.</p> <p>Unfortunately, I still find this crucial point highly unsatisfactory. When looking at previous literature in the field, it is likely that services and clients differ across sectors. This leaves the paper imprecise and reduces its scientific value. In case the data collected from the Care Inspectorate does not allow for such level of detail, the paper should aim accordingly.</p> <p>Thank you for the opportunity to review this research.</p> |
|--|---|

|                        |   |
|------------------------|---|
| <b>REVIEWER</b>        | Joe Nolan<br>Northern Kentucky University |
| <b>REVIEW RETURNED</b> | 05-Oct-2018                               |

|                         |   |
|-------------------------|---|
| <b>GENERAL COMMENTS</b> | Regarding the choice of model, the rationale provided by the authors for splitting into three categories rather than two is acceptable. Revisions to include CI's and adjust for multiplicity make this a substantially better paper, as do the changes made to simplify the figures. I find the statistics in this paper now meet the quality needed for acceptance. |
|-------------------------|---|

### VERSION 3 – AUTHOR RESPONSE

#### **Reviewer #2**

Following the authors’ first submission, I provided several general and detailed comments.

I remain concerned about my crucial general comment:

The paper’s attempt to generalize findings across a wide range of very different social services is heroic. Unfortunately, the paper lacks a satisfactory level of detail.

In general you need to justify better that 1) the services included are comparable across private, public and third sector (are they really providing the same services? What kind of 'Day Care of Children', what kind of 'Housing Support service', what kind of 'Adoption Service?') and 2) the patients or users of the services (case-mix) are comparable (possible case-mix differences are briefly introduced by the authors in the Introduction section). This key and most important part of the paper is highly unsatisfactory. You may be comparing apples to pears and suffer from a serious bias without a more detailed description of the services and (adjustment of) the users of the services.

Following two rounds of review, I cannot see major changes or efforts in meeting these concerns. The author’s last response that “(...) The heterogeneity is intentional and a strength of the analysis. For

example, the consistency of our findings suggest that the direction of effects is similar across all included service types, although further research should be undertaken to replicate our results.” is not about or addressing this concern.

Unfortunately, I still find this crucial point highly unsatisfactory. When looking at previous literature in the field, it is likely that services and clients differ across sectors. This leaves the paper imprecise and reduces its scientific value. In case the data collected from the Care Inspectorate does not allow for such level of detail, the paper should aim accordingly.

### **Author response:**

Thank you for these comments. We yet again acknowledge and agree to the fact that controlling for individual level data on clients served would be appropriate for our analysis. We understand that you have found client variation yourself among patients undergoing certain types of surgery in for-profit and non-profit hospitals in Norway (<https://bmjopen.bmj.com/content/bmjopen/8/6/e019780.full.pdf>). However, for existing research on social care services, we are not aware of any studies that have included individual level client data when investigating performance or quality across sectors, which makes it unclear what literature you’re referring to in the comment: “*When looking at previous literature in the field, it is likely that services and clients differ across sectors.*”. That is not to say it would not be worthwhile analysis to do (given data availability), as it is indeed an intuitive and theoretically sound control variable to include. However, it is important to note that our research question do not revolve around clients, and we would also question the applicability of client variation among patients receiving certain types of surgery in Norwegian hospitals to social care organisations in Scotland (assuming that’s the ‘previous literature’ you’re referring to). To clarify this issue, we’ve added the following clarification in the discussion section:

“A more substantial limitation of the analysis is that service types may differ systematically across sectors. For example, it may be that clients served by FPOs are inherently different from those served by third and public sector providers. To address this issue, we controlled for all organisational characteristics allowed for by the data (e.g. service types, years of registration, and number of employees), and ran the models separately on all major service types, showing the same substantive pattern as reported in the main results. Although all models adjusted for the types of clients served by the social care providers (e.g. children, older people, dementia patients), the data collected by the Care Inspectorate do not include individual level information on clients served, and we could thus not investigate client variation in terms of, for example, socio-economic status across sectors. To the best of our knowledge, there is no existing research investigating client variation across sectors for social care services [3, 40, 41], which is important to explore in future research.”(p14-15)

In terms of the included service types and how different types of providers are distributed among them, it is vital that we note (again) that we only included service types with observations across all sectors. To make this clearer, we’ve included table A5 in the appendix, which shows the raw cross tabulations of how the quality and risk domains are distributed among different service types and across sectors (showing the same substantive pattern as when aggregated). Additionally, we’ve included figures B1-B5 in appendix C, which display the predicted probabilities for the 4 biggest service types to be rated of poor, adequate, or high quality, and of low, medium, and high risk across all quality and risk domains (also showing a similar direction as the main results). We’ve clarified our sub-analysis in the results section:

“We reran the analyses on the biggest sub-groups of services (i.e. care homes, support services, child care services, and housing support services), which can be found in in table A5 in Appendix A and figures B1-B5 in Appendix C. Most of these sub-analyses resulted in slightly reduced effects of sector and with less clear differences between PSOs and TSOs. However, both the raw cross tabulations and the predicted probabilities generally display the same substantive pattern and direction of effects with FPOs being more likely to be classified as of adequate quality and of higher risk relative to TSOs and PSOs, which are more likely to be rated as high quality and of lower risk.” (p13)

Also, as we noted for the first round of revisions, table A1 includes a full description of all included service types, and table A2 shows how different service types are distributed among sectors. These can both be found in appendix A.

We hope these efforts clarify your concerns.

Thank you for the opportunity to review this research.

### **Reviewer #3**

Regarding the choice of model, the rationale provided by the authors for splitting into three categories rather than two is acceptable. Revisions to include CI's and adjust for multiplicity make this a substantially better paper, as do the changes made to simplify the figures. I find the statistics in this paper now meet the quality needed for acceptance.

### **Author response:**

We thank the reviewer for the very helpful comments, which improved the models and statistical reporting of the paper.