PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The reliability of hospital scores for the Cancer Patient Experience Survey: analysis of publicly reported patient survey data
AUTHORS	Abel, Gary; Gomez-Cano, Mayam; Pham, Tra My; Lyratzopoulos, Georgios

VERSION 1 – REVIEW

REVIEWER	Chris Graham Picker Institute Europe Picker Institute Europe provides survey management and implementation services for national and local organisations, including numerous national surveys similar to the cancer patient experience survey (CPES). We do not have this involvement in CPES itself and I do not consider that I have a material conflict of interest.
REVIEW RETURNED	11-Feb-2019

GENERAL COMMENTS	Thank you for the opportunity to review this interesting and well written paper, which I enjoyed reading. The manuscript addresses a simple but important question: how reliable are estimates from England's national cancer patient experience survey (CPES), and are they suitable for high-stakes institutional performance assessment. This question is investigated by using published survey results to investigate between hospital and within hospital variation, and reliability is calculated as the proportion of variation in hospital level mean scores arising from true variation between hospitals. The majority of estimates are found to have low (<0.7) reliability and strategies for addressing this are proposed. Overall, the manuscript is of good quality and the methodology is appropriate and clearly described. There were some areas that could be improved or clarified but most are minor. Where there larger issues, they primarily include points that would benefit from further expansion - including describing the current use of the survey in the introduction, and investigating the benefits and risks of changing the sample size in the results and discussion. I have no substantive concerns about the methodology, analysis, or interpretation. For convenience, I've split my comments into 'major' and 'minor' sections below. I hope that you will find them to be constructive and helpful. Major issues

	1. p4, 17-21 - Although I don't disagree with the conclusions it expresses, the opening paragraph of the introduction could be stronger and better evidenced. The opening sentence would benefit from a supporting citation or rewording, particularly because "the quality improvement movement" is not unequivocal on the subject of measurement. The sources cited for the second sentence are examples of measures rather than evidence that the number has increased or that public reporting has become the norm. It would also be worth making clear the point that the trend towards public reporting and performance measurement is international. 2. p4 - introduction - as the purpose of the paper is to discuss the suitability of CPES for high stakes comparisons, it is odd that only one sentence is devoted to describing the survey and its current use. It would be helpful - especially for readers outside of England - to have a description of how the survey is conducted and how the results are used, and this should be more specific than the current statements about it being 'reported publicly and used by healthcare improvement teams'. 3. p6, 53 - p7, 5 - The authors have clearly sought to keep the paper concise, and it is commendably focused. I would, however, have liked to see this paragraph developed further. It's not clear why only a fourfold increase in sample size or an 80% threshold are reported (other, perhaps, than to reflect the maximum available sample per year?) - but it would be interesting to know what proportion of estimates would be reliable for different levels of increase in sample size. Even a simple chart showing the proportion of estimates meeting the 0.7 reliability threshold for different sample sizes or multiples of the current sample size would be helpful in letting readers better understand the likely trade-off between cost and data quality. 4. p8, 42-58 - arguably this paragraph oversimplifies the options for increasing the sample size. Moving the sampling window from 3 to 12 months would not simply increase the size of the population to sample from - it would also influences the composition of that population. For example, patients with more aggressive cancers should be comparatively underrepresented in a 12 month vs a 3 month sample due to different rates of death. Patients with regular but infrequent appointments - eg those with annual check ups - will also be overrepresented in a 12 month vs a 3 month sample. These kind of changes have implications for the comparability of data over time, and to the extent that those comparisons are an aim of the survey it means that taking the steps required to support high stakes performance comparisons might be contrary to other aims. 5. p9, 4-5 - I agree with the use of the phrase 'high stakes' here, but it is never explained and may be unfamiliar to some readers. it would be helpful for a definition to be added either in the introduction or in the conclusions so that readers are clear on the kinds of comparisons that should be considered unsafe in the case of poor data reliability. Minor issues 1. p1, 47-48 - ethical approval declaration - it would be appropriate to state whether the survey itself was subject to ethical review.

	2. p2, 32 - abstract - participants - it would be helpful to state the number of patients participating in the survey for the benefit of readers unfamiliar with CPES. 3. p2, 49 - typo - 'some hospital' should read 'some hospitals' 4. p2 - 53-4 (and elsewhere) - "The English Patient Experience Survey represents a globally unique source...". Does it? Scotland & Wales have very similar surveys, so I don't think is accurate. 5. p3 - 16-17 - very minor point, but "measures" would be more appropriate than "items" here - as the point the authors are making is that low reliability is a consequence of both items and trust-level sample sizes. 6. p4 - 23-24 - The text here essentially refers to the definition of health service quality offered by Darzi in the NHS Next Stage Review, so citing that report or the accompanying article in the Lancet would be appropriate. Parallels to the 'triple aim' promoted by IHI in the United States may also help to demonstrate the international relevance of this. 7. p4 - 34-35 = typo missing 'of' in "prior examination [of] the..." 8. p4, 58-59 - another very minor point - but the statement that patients "were known to be alive on the day of survey mail out" is inaccurate, because the patient list is checked for recorded deaths. The statement "were not known to have died prior to the survey mail out" would be more accurate, but possibly harder for readers to follow. A fuller description would be to say that "The patient list was checked against national records prior to mailout, and any patients recorded as having died were removed" - but perhaps this is too much detail when the authors already cite the full details of the survey administration. 9. The results section is inconsistent in use of decimals after percentages - the first has 1 dp, others 0 dp 10. p5, 53-60 - from the description of patient involvement it's not at all clear what the role of patient involvement was in this specific analysis. Did the advisory group consider this analysis or its findings? 11. p6, 39-43 - another very minor point. The first sentence here ("Given that reliability depends...") rightly implies that the method of calculating reliability means that the independent impact of number of responses, between hospital variance, and percent positive responses on reliability are predictable. The next two sentences almost imply that these relationships have been discovered in the course of analysing the results, which I don't think was the intention. Similarly the passive phrasing ("Hospitals which tended to have lower reliability also had smaller sample sizes") could be clearer in helping non-statistical readers understand that smaller samples directly cause lower reliability. 12. p6, 39 - The text here says that "reliability depends on the sample size". It would be more accurate to say that it depends on the number of responses to the given item in the given hospital, but I understand that the reason for wording it like this is to

	highlight the importance of the sorting of the hospital axis in figure 3. I think this might be easier for readers to follow if the sorting of the hospital axis was mentioned in the text as well as in the legend of figure 3. 13. p6, 51-2 - where items are cited by number, it would be helpful to state which items they are. It would also be helpful to include the wording of each question in the supplementary tables. 14. Figures 1 and 2 - I infer that the x (hospital) axis is supported by total number of respondents within hospital. It would be worth stating this - particularly for figure 2 where it is less obvious (but where there does appear to be some association between hospital size, as measured by the number of respondents, and positivity of patients for certain questions) 15. Figure 3 - I wonder whether moving the y and x axis labels to sit between the scatter plots and main grid would make it more obvious for readers that they share categories and orders? Referring to the sorting in the labels would also help - eg "Item, in order of variance" and "Hospital, in order of participant sample size" instead of just "Item" and "Hospital"?

REVIEWER	AlanM. Zaslavsky Harvard University, USA
REVIEW RETURNED	13-Feb-2019

GENERAL COMMENTS	I was pleased to find that this paper implemented Spearman-Brown reliability estimates on these hospital survey data, since this is a statistical approach that works well and is (and should be) widely implemented in this context. Nonetheless I have some suggestions which, if implemented, might increase the value of this paper. (1) You didn;t explain why the analysis was applied to unadjusted rather than casemix-adjusted scores. Was this because you only had access to publicly reported data? Care for cancer is very heterogeneous, involving a variety of treatment modalities and degrees of burden on patients; furthermore the outcomes are not always successful. So I would expect casemix effects to be potentially quite substantial; the author's conjecture that they would tend to reduce reliability may well be correct. Even without clinical data (which might have been unavailable) you probably have a fair amount of information that could be used in these models; even the screeners for some sections might loosely represent the treatments received (or suffered) by the patient. If there is any way you could get the necessary data, substituting adjusted results would improve this manuscript. Unfortunately your results for this specific survey and reporting exercise cannot be relied upon if you can't incorporate or simulate the casemix adjustment,and you should make this clear to your readers. (2) More information about the items on the survey would be helpful. Are all reports on single survey items, or are there any composites used in reporting? (These might be more reliable than single items, although there is no prior guarantee. (3) I have some issues with the discussion of what are sometimes called "topped-out" items, that is, those for which responses are almost 100% favorable. I agree that such items are a sign of

	success and universally high scores might indicate that it is no longer useful to report them. However, there may be some value to those items that is masked by the appearance of large sampling variance and consequent low reliability due to your analysis conducted on the logistic scale, whose transformation from the probability scale stretches out values near the extremes. Thus a large variance on the logistic scale might represent very little variability in probabilities. For example, it sounds impressively bad that two hospitals are statistically indistinguishable when their odds ratio is 2 on an item. However this sounds less serious when we learn that one has a score of 98% and the other, 99%. In general the normal approximations we apply to make comparisons don't work very well at these extremes. (See for example Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Statistical science. 2001 May 1:101-17.) Although this comparison of topped-out hospitals might be irrelevant and unreliable, we still benefit from the evidence that both hospitals are better than one with a score of 70%. This point should be considered when assessing the value of a score for which some hospitals but not all are "topped out". (4) You might mention that improving the survey response rate would be another way to improve reliability. (5) You might point out that the optimal design for a survey that puts equal importance on every hospital is an equal sample size for each hospital. A proportional (equal sampling rate) design reflects the popular misconception that it takes a bigger sample to estimate a rate in a bigger population. (6) the big tables at the end should be cut off and moved to a supplement, preferably in machine-readable format.

REVIEWER	Richard Wagland University of Southampton
REVIEW RETURNED	23-Feb-2019

GENERAL COMMENTS	This is an important study of publicly reported NCPES data, which has implications for the way in which it should be reported (i.e. in the interests of transparency), and recommendations for sampling that should be considered.

VERSION 1 – AUTHOR RESPONSE

Reply to Reviewer's comments, manuscript: Manuscript ID bmjopen-2019-029037: "The reliability of hospital scores for the Cancer Patient Experience Survey: analysis of publicly reported patient survey data"

Dear BMJ Open

thank you for the thorough review of our paper and the insightful and constructive comments offered by the two Reviewers. We have considered and addressed all comments as outlined below.

Reviewer 1: Chris Graham

Major issues	
The manuscript addresses a simple but important question: how reliable are estimates from England's national cancer patient experience survey (CPES), and are they suitable for high-stakes institutional performance assessment. This question is investigated by using published survey results to investigate between hospital and within hospital variation, and reliability is calculated as the proportion of variation in hospital level mean scores arising from true variation between hospitals. The majority of estimates are found to have low (<0.7) reliability and strategies for addressing this are proposed. Overall, the manuscript is of good quality and the methodology is appropriate and clearly described. There were some areas that could be improved or clarified but most are minor. Where there larger issues, they primarily include points that would benefit from further expansion - including describing the current use of the survey in the introduction, and investigating the benefits and risks of changing the sample size in the results and discussion. I have no substantive concerns about the methodology, analysis, or interpretation. For convenience, I've split my comments into 'major' and 'minor' sections below. I hope that you will find them to be constructive and helpful.	We would like to thank the Reviewer for their overall positive comments.
1. p4, 17-21 - Although I don't disagree with the conclusions it expresses, the opening paragraph of the introduction could be stronger and better evidenced. The opening sentence would benefit from a supporting citation or rewording, particularly because "the quality improvement movement" is not unequivocal on the subject of measurement. The sources cited for the second sentence are examples of measures rather than evidence that the number has increased or that public reporting has become the norm. It would also be worth making clear the point that the trend towards public reporting and performance measurement is international.	Thank you, we have added reference to a key publication on measurement of quality in health care, as suggested (Raleigh VS, Foot C Getting the measure of quality. Opportunities and challenges. The King's Fund 2010. ISBN: 978 1 85717 590 5. https://www.kingsfund.org.uk/sites/default/files/Getting-the-measure-of-quality-Veena-Raleigh-Catherine-Foot-The-Kings-Fund-January-2010.pdf). We have additionally edited the first sentence to indicate the international relevance of the matter.

2. p4 - introduction - as the purpose of the paper is to discuss the suitability of CPES for high stakes comparisons, it is odd that only one sentence is devoted to describing the survey and its current use. It would be helpful - especially for readers outside of England - to have a description of how the survey is conducted and how the results are used, and this should be more specific than the current statements about it being 'reported publicly and used mngby healthcare improvement teams'.	Thank you, we have added the following text to address this comment, within the first paragraph of 'Data' in 'Methods'. **'The (English) National Cancer Patient Experience Survey 2016 survey questionnaire is the sixth iteration of the survey first undertaken in 2010. It includes many evaluative questions covering the experience of diagnosis, diagnostic testing, shared-decision-making, specialist nursing, inpatient care**, **anti-cancer treatment (surgery, radiotherapy, chemotherapy), hospital discharge and care in the community, together with an overall item for overall satisfaction with care. Survey results are reported publicly for each English hospital to drive local quality improvements, to assist commissioners and providers of cancer care; and to inform the work of the various charities and stakeholder groups supporting cancer patients.** **The survey was mailed to all adult patients (aged 16 and over) discharged from a National Health Service hospital after inpatient or a day case cancer-related treatment during April-June 2016 after vital status checks at survey mail-out (between 3-5 months after the sampling period).'**
3. p6, 53 - p7, 5 - The authors have clearly sought to keep the paper concise, and it is commendably focused. I would, however, have liked to see this paragraph developed further. It's not clear why only a fourfold increase in sample size or an 80% threshold are reported (other, perhaps, than to reflect the maximum available sample per year?) - but it would be interesting to know what proportion of estimates would be reliable for different levels of increase in sample size. Even a simple chart showing the proportion of estimates meeting the 0.7 reliability threshold for different sample sizes or multiples of the current sample size would be helpful in letting readers better understand the likely trade-off between cost and data quality.	We have taken on board the Reviewer's comment and have expanded this paragraph as well as including the suggested graph. The paragraph now reads (bold fonts denoting new text): "Overall the reliability of hospitals scores for the survey is low. Of the 7,377 individual hospital-question pairs, only 35 % reached a reliability of 0.7, the minimum generally considered to be useful. As it is possible to improve reliability by increasing the sample size **for a given hospital** we calculated how many multiples of the current sample size would be required to reach a reliability of 0.7 for each question **(Figure 4) It would be possible to increase the percentage of hospital scores reaching a**

	reliability of 0.7 to 60% by doubling the individual hospital sample size. Further increases lead to smaller gains, though 80% of the individual hospital scores would have achieved a reliability of 0.7 or more with 4 times as much data (**which represents the upper limit of what could be achieved within a single year of data collection, though could also be achieved by aggregating over longer time periods**)."
4. p8, 42-58 - arguably this paragraph oversimplifies the options for increasing the sample size. Moving the sampling window from 3 to 12 months would not simply increase the size of the population to sample from - it would also influences the composition of that population. For example, patients with more aggressive cancers should be comparatively underrepresented in a 12 month vs a 3 month sample due to different rates of death. Patients with regular but infrequent appointments - eg those with annual check ups - will also be overrepresented in a 12 month vs a 3 month sample. These kind of changes have implications for the comparability of data over time, and to the extent that those comparisons are an aim of the survey it means that taking the steps required to support high stakes performance comparisons might be contrary to other aims.	We appreciate this point very much having previously made it ourselves in a previous publication considering the responders to the survey. We now make the point explicitly in the paper adding the following text to the paragraph "**However, changing the length of the sampling window will likely impact the composition of responders as this is dictated by variation in treatment modalities, early mortality and non-response, the effect of which will depend on the sampling window ([reference]). This in turn may impact the ability to compare results to those from previous years**." Where 'reference' is Abel GA, Saunders CL, Lyratzopoulos G. Post-sampling mortality and non-response patterns in the English Cancer Patient Experience Survey: Implications for epidemiological studies based on surveys of cancer patients. Cancer Epidemiol. 2016;41:34-41. doi: 10.1016/j.canep.2015.12.010.
5. p9, 4-5 - I agree with the use of the phrase 'high stakes' here, but it is never explained and may be unfamiliar to some readers. it would be helpful for a definition to be added either in the introduction or in the conclusions so that readers are clear on the kinds of comparisons that should be considered unsafe in the case of poor data reliability.	We agree and we have accordingly added the following text within the 5th paragraph of Methods: **'High stakes purposes have important consequences for an individual or organisation (i.e. when attached to a financial incentive, publicly reported league**

	tables or an outcome measure in a research study) and therefore require high measure reliability. Reliability can takes values from 0 to 1. Values <0.70 are considered to represent low reliability, whereas values ≥0.90 represent high reliability, required for 'high-stake' purposes; in-between values are considered to represent adequate reliability.'
Minor issues	
1. p1, 47-48 - ethical approval declaration - it would be appropriate to state whether the survey itself was subject to ethical review.	Thank you, we have added: **"The actual survey was conducted by the survey providers after obtaining section 251 approval of the NHS Act 2006 and Health Service (Control of Patient Information) Regulations 2002."**
2. p2, 32 - abstract - participants - it would be helpful to state the number of patients participating in the survey for the benefit of readers unfamiliar with CPES.	We have added the number of patients included in the analysis in the Abstract. (Please note that this is slightly lower than the number of responders due to the exclusion of two hospitals with fewer than 20 responders).
3. p2, 49 - typo - 'some hospital' should read 'some hospitals'	We have corrected. Thanks.
4. p2 - 53-4 (and elsewhere) - "The English Patient Experience Survey represents a globally unique source...". Does it? Scotland & Wales have very similar surveys, so I don't think is accurate.	Because this sentence forms part of the Conclusion of the Abstract, we do not feel we can elaborate further. However, for the Editor's and the Reviewer's attention, we do indeed believe that the English CPES is a globally unique resource, if one consider: -The multiple years of data available since 2010 (approximately 420,000 patient surveys, the largest such resource anywhere in the world at present) -The fact that some of the survey waves have been linked to cancer registration data -The open access to unlinked data -The number of publications that have arisen from the survey thus far (far greater than any other cancer patient experience survey)

	Our statement does not aim to denigrate the importance of other national surveys, particularly in other UK countries.
5. p3 - 16-17 - very minor point, but "measures" would be more appropriate than "items" here - as the point the authors are making is that low reliability is a consequence of both items and trust-level sample sizes.	We agree with the Reviewer; we have now changed the term “items” to “**survey scores**” to be consistent with the language used in the rest of the paper.
6. p4 - 23-24 - The text here essentially refers to the definition of health service quality offered by Darzi in the NHS Next Stage Review, so citing that report or the accompanying article in the Lancet would be appropriate. Parallels to the 'triple aim' promoted by IHI in the United States may also help to demonstrate the international relevance of this.	Thank you very much, and we now cite the following paper by Ara Darzi as suggested. We have opted for the Lancet publication as more easily retrievable by readers. Ara Darzi, Quality and the NHS Next Stage Review, The Lancet, Volume 371, Issue 9624, 2008, Pages 1563-1564, ISSN 0140-6736, https://doi.org/10.1016/S0140-6736(08)60672-8. (http://www.sciencedirect.com/science/article/pii/S0140673608606728)
7. p4 - 34-35 = typo missing 'of' in "prior examination [of] the..."	Corrected, thank you.
8. p4, 58-59 - another very minor point - but the statement that patients "were known to be alive on the day of survey mail out" is inaccurate, because the patient list is checked for recorded deaths. The statement "were not known to have died prior to the survey mail out" would be more accurate, but possibly harder for readers to follow. A fuller description would be to say that "The patient list was checked against national	Thank you, to address the comment we have changed the text to “**were not known to have died prior to the survey mail out**”.

records prior to mailout, and any patients recorded as having died were removed" - but perhaps this is too much detail when the authors already cite the full details of the survey administration.	
9. The results section is inconsistent in use of decimals after percentages - the first has 1 dp, others 0 dp	We now consistently report all percentage values in the text to 0 dp (i.e. excluding any decimal points).
10. p5, 53-60 - from the description of patient involvement it's not at all clear what the role of patient involvement was in this specific analysis. Did the advisory group consider this analysis or its findings?	Annually the authors meet with an advisory group to discuss the approaches and findings, and those were presented in the last meeting.
11. p6, 39-43 - another very minor point. The first sentence here ("Given that reliability depends...") rightly implies that the method of calculating reliability means that the independent impact of number of responses, between hospital variance, and percent positive responses on reliability are predictable. The next two sentences almost imply that these relationships have been discovered in the course of analysing the results, which I don't think was the intention. Similarly the passive phrasing ("Hospitals which tended to have lower reliability also had smaller sample sizes") could be clearer in helping non-statistical readers understand that smaller samples directly cause lower reliability.	We have prefaced the second sentence in this paragraph with “Consistent with this” to make it clear we are not claiming this as a discovery of our making. However, we would like to keep the following phrasing as is. The reason for this is that a small sample size does not guarantee poor reliability. It is possible that a large variance between hospitals would counteract it and no one factor should be considered in isolation. We are also aware this is a results section and so we are reporting our findings rather than interpreting them at this point.
12. p6, 39 - The text here says that "reliability depends on the sample size". It would be more accurate to say that it depends on the number of responses to the given item in the given hospital, but I understand that the reason for wording it like this is to highlight the importance of the sorting of the hospital axis in figure 3. I think this might be easier for readers to follow if the sorting of the hospital axis was mentioned in the text as well as in the legend of figure 3.	We appreciate this suggestion and have added the details to the paper. As the ordering applies to Figures 1, 2, and 3 we make it clear in the first paragraph of results with the following text “**Our findings are displayed in three figures each comprising a grid of hospitals by questions. Hospitals are ordered according to the number of responders and questions are ordered according to the between hospital variance**.”
13. p6, 51-2 - where items are cited by number, it would be helpful to state which items they are. It would also be helpful to include the wording of each question in the supplementary tables.	The wordings of the questions are very long and it is not feasible to incorporate this text into the tables. Furthermore, although the survey instruments are freely accessible (as referenced) by the readers via the survey provider’s website, we have no permission to reproduce them.

14. Figures 1 and 2 - I infer that the x (hospital) axis is supported by total number of respondents within hospital. It would be worth stating this - particularly for figure 2 where it is less obvious (but where there does appear to be some association between hospital size, as measured by the number of respondents, and positivity of patients for certain questions)	Thank you, please see response to point 12 (above)
15. Figure 3 - I wonder whether moving the y and x axis labels to sit between the scatter plots and main grid would make it more obvious for readers that they share categories and orders? Referring to the sorting in the labels would also help - eg "Item, in order of variance" and "Hospital, in order of participant sample size" instead of just "Item" and "Hospital"?	Thank you for this suggestion. We did try moving the labels as you suggested, however, this actually made it less clear that the axes were shared as they acted as a barrier between the graphs. We have however, taken on board your suggestion and referred to the sorting in the labels.

Reviewer 2: Alan M. Zaslavsky

I was pleased to find that this paper implemented Spearman-Brown reliability estimates on these hospital survey data, since this is a statistical approach that works well and is (and should be) widely implemented in this context. Nonetheless I have some suggestions which, if implemented, might increase the value of this paper.	Thank you very much for your positive assessment and for your constructive comments.

(1) You didn't explain why the analysis was applied to unadjusted rather than casemix-adjusted scores. Was this because you only had access to publicly reported data? Care for cancer is very heterogeneous, involving a variety of treatment modalities and degrees of burden on patients; furthermore the outcomes are not always successful. So I would expect casemix effects to be potentially quite substantial; the author's conjecture that they would tend to reduce reliability may well be correct. Even without clinical data (which might have been unavailable) you probably have a fair amount of information that could be used in these models; even the screeners for some sections might loosely represent the treatments received (or suffered) by the patient. If there is any way you could get the necessary data, substituting adjusted results would improve this manuscript. Unfortunately your results for this specific survey and reporting exercise cannot be relied upon if you can't incorporate or simulate the casemix adjustment, and you should make this clear to your readers.	We very much appreciate the point made that case mix may be very important with regards to the reliability of hospital scores for this survey. However, we have not been able to address this due to not having the data to do so, as the Reviewer rightly suspected. The only data we have available to us is the publicly reported scores. We can get access to individual patient data for this survey, but due to data governance issues it is not made available with hospital identifiers (or even pseudo-identifiers). For this reason, we cannot make any changes to the paper in response to this comment, though we would, in an ideal world, like to do so. Additionally, in the 3rd paragraph of the Discussion, we indeed acknowledge this issue as a limitation as indeed suggested: '**Its main limitation is that our analysis does not take into account the influence of patient case-mix. Certain patient groups have systematic tendencies towards reporting positive experiences compared to others[22, 23] and for this reason the results of the survey are reported in both adjusted and unadjusted form**.'
(2) More information about the items on the survey would be helpful. Are all reports on single survey items, or are there any composites used in reporting? (These might be more reliable than single items, although there is no prior guarantee.	Thank you, and we have added relevant text as outlined in our reply to Reviewer 1, comment 2 (above). And indeed, all reports relate to single survey items (no composites are used in the reporting of this survey).

(3) I have some issues with the discussion of what are sometimes called "topped-out" items, that is, those for which responses are almost 100% favorable. I agree that such items are a sign of success and universally high scores might indicate that it is no longer useful to report them. However, there may be some value to those items that is masked by the appearance of large sampling variance and consequent low reliability due to your analysis conducted on the logistic scale, whose transformation from the probability scale stretches out values near the extremes. Thus a large variance on the logistic scale might represent very little variability in probabilities. For example, it sounds impressively bad that two hospitals are statistically indistinguishable when their odds ratio is 2 on an item. However this sounds less serious when we learn that one has a score of 98% and the other, 99%. In general the normal approximations we apply to make comparisons don't work very well at these extremes. (See for example Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Statistical science. 2001 May 1:101-17.) Although this comparison of topped-out hospitals might be irrelevant and unreliable, we still benefit from the evidence that both hospitals are better than one with a score of 70%. This point should be considered when assessing the value of a score for which some hospitals but not all are "topped out".	We agree with much of what is said. However we feel there is a limit to what we can discuss on this topic in a short space whilst maintaining accessibility to the less statistically minded. To take your example of an odds ratio of 2 sounding impressive, but the difference between 98 and 99% being less meaningful. Another way to consider this is that twice as many people in one hospital experience the poor outcome than the other. The significance of a poor experience being experienced by only 1 or 2% of patients will depend on what the experience item is. Given there is potentially a lot to unpick here, and our discussion is already rather long, we would rather not open this subject up in this paper, although we fully appreciate its importance. In response to the final issue of it being useful to know that all hospitals are performing above a minimum threshold, we have addressed this by adding the following text to the fifth paragraph of discussion "**Furthermore, it can also be useful to know that all hospitals are performing above a target level even though we may not be able to distinguish between them**."
(4) You might mention that improving the survey response rate would be another way to improve reliability.	We have addressed this point by adding the following text to the end of the discussion "**Similarly, improvements in response rate can also increase the sample size, in turn improving reliability. The scope for improvement in this survey may be limited due to it already having a high response rate, but for other surveys that may not be true**."

(5) You might point out that the optimal design for a survey that puts equal importance on every hospital is an equal sample size for each hospital. A proportional (equal sampling rate) design reflects the popular misconception that it takes a bigger sample to estimate a rate in a bigger population.	We agree with the point the Reviewer has made and have added the suggested text into the middle of the final paragraph of the Discussion section.
(6) the big tables at the end should be cut off and moved to a supplement, preferably in machine-readable format.	We have indeed supplied them as Excel spreadsheets. The erroneous appearance in the pdf version is a result of the file conversion process for purposes of review. These tables will not appear in the final published manuscript, only as an online supplementary file.

VERSION 2 – REVIEW

REVIEWER	Chris Graham Picker Institute Europe, United Kingdom My organisation receives fees for designing and implementing patient surveys similar to (but not including) the survey used in this analysis. This does not amount to a material conflict of interest.
REVIEW RETURNED	17-Apr-2019

GENERAL COMMENTS	I am grateful to the authors for their clear and thorough response to reviewer comments. From this response and the revised manuscript, it is clear that the vast majority of points raised by both reviewers have been addressed with changes that help to improve the manuscript. The additional chart (figure 4) is particularly helpful in demonstrating the diminishing returns associated with increasing the survey's sample size. Where the authors have chosen not to incorporate suggested changes, the reasons given for this are sensible and well justified. Overall, this is a strong paper and I have no further comments to add.

VERSION 2 – AUTHOR RESPONSE

Many thanks for the last set of comments on our paper. We have updated the Strengths and Limitations section in line with the editor's comments. The reviewers comments do not make any suggestions for changes and so these are the only changes that have been made. We hope this addresses any remaining issues with our manuscript and we look forwards to progressing with the publication of this paper.