Author’s response to reviews

Title: Reliability of patient-reported complications following hip or knee arthroplasty procedures

Authors:

Sung Mu Heo (brian.sm.heo@gmail.com)

Justine Naylor (Justine.Naylor@sswhs.nsw.gov.au)

Ian Harris (ian.harris@unsw.edu.au)

Timothy Churches (timothy.churches@unsw.edu.au)

Version: 2 Date: 10 Nov 2018

Author’s response to reviews:

Reviewer 1:

Major issues:

Many cases of egregious lack of references for big, generic statements that need validation. There are numerous examples which need to be rectified. Some of the big ones:

We have revised our manuscript with substantiating evidence as outlined in the comments below. The references we have added are as follows:


a) The very first sentence of the manuscript. How do the authors know that complications awareness is a major impact on patients' decisions to pursue the surgery? From our experience, it's often not related at all.

This comment has been removed and the opening paragraph has been written as follows:

“Patient reported outcome measures (PROMs) are a reliable tool for understanding patients' perceived outcomes, and are typically implemented via telephone or mail follow-up via a centralised registry [1]. Complications are often included in such registries as these data can be analysed to inform best practice and reduce the rates of these complications and their associated economic and social burden [2,5].”

The following reference was included:

b) The first sentence of the second paragraph. I would argue that contacting patients directly from a registry is not inexpensive, and is not time-saving. Also has issues with the point of reference. Sure, it's cheaper for the clinic to have someone else (the registry) call the patient, but is it cheaper for society and for the patient?

The statements regarding cost and time saving have been removed and replaced with an explanation on the economic and social burden of surgical complications in the end of the first paragraph as follows:

“Complications are often included in such registries as these data can be analysed to inform best practice and reduce the rates of these complications and their associated economic and social burden [2,5].”


c) The "surprising finding" paragraph in Discussion. There are dozens of papers that describe pain expectations after TKA/THA which the authors don't cite at all that get at this issue in great detail. It's not a surprise.
We have removed “surprising finding” from the sentence. It now reads as follows: “This study observed high rates of false negative results for unexpected pain as seen by Visser et al and Franneby et al.”

d) "Patients are seldom bothered by minor differences in leg length and often overlook these as long as they have improved function." How do the authors know this? This sentence seemingly states fact and explanation without attributing either. The following 2 sentences in the paragraph have similar issues as well.

Statements regarding leg length discrepancy and infection have been revised and substantiated with evidence, as follows:

“Patients with minor leg-length discrepancies may have few symptoms which may explain the under-reporting by patients, whereas surgeons place great importance on leg length due to its possible detrimental outcomes [22]. This study showed poor patient specificity in identifying SSI, a finding that was also observed in a study by Zellmer et al., who suggested that may be improved with validated infection education material [23].”


There is a huge limitation that isn't discussed at all. How do we know how the registry asks the patient about their complications? There is a sentence in the manuscript that exemplifies this near the top of page 6. Providers were more likely to report neuropathy and patients were more likely to report numbness. Well, of course! Numbness is a word that's not medically exclusive and would be known to patients with lower literacy, while neuropathy is not a familiar layman's
word. If the registry callers do not explain the words, but simply asks if the patient had this complication yes/no, then that's a huge failing. Ultimately, this manuscript needs to explain this issue, preferably by issuing the whole phone script.

Callers use a standard data collection form for complications as per protocol. Patients are asked if they experienced a complication other than reoperation or readmission, and if they answer “yes”, they are openly asked to specify, and their responses were recorded from a list of complications on the form. Patients are not asked if they specifically experienced complications as such as neuropathy or paraesthesia which are more medically exclusive terminology and are only recorded if patients report the specific complication unprompted. This has been added in the methods section as follows:

“If patients responded Yes to complications not requiring readmission, patients were further asked to specify their complication with an open ended question as to not prompt the patient. Callers then recorded the complications on the standard data collection form which consists of the following; surgical site infection (SSI) requiring oral antibiotics, SSI requiring intravenous antibiotics, deep vein thrombosis (DVT) index leg, DVT other leg, DVT both legs, pulmonary embolism (PE), dislocation, joint stiffness, bladder infection or urinary retention, fracture, unexpected pain, cardiac, stroke, leg length discrepancy, joint or lower limb swelling, paraesthesia or numbness, cellulitis, neuropathy, muscle weakness, respiratory infection, other, and unknown.”

Deleted in the discussion: “The current study has found similar results and suggests that clarity in what defines a complication may decrease the observed disagreement. For example, neuropathy and numbness are related complications, but neuropathy was reported by more surgeons and patients were more likely to report numbness. Further, muscle weakness encompasses anything from slight difficulty in movement to complete immobilization.”

Replaced with: “ACORN callers similarly enquire if the patients have experienced a complication and then ask the patient to specify details, without prompting for specific types of complication. Although this provides a window into patients' experience and recollection of post-operative complications, it does not provide information on the extent or severity. For example, muscle weakness can refer to slight difficulty in movement to complete immobilisation. Clinical registries such as ACORN may benefit from further enquiring about severity of complications.”
The 6-month follow-up timeframe is quite vague. It needs to be specifically defined, such as 150-210 days postoperative or similar. If ACORN doesn't define it, then it should be a discussed limitation.

It is defined in the ACORN report as 6 months +/- 1 month. This has been added to the Background section as follows:

“The Arthroplasty Clinical Outcomes Registry, National (ACORN) is an Australian orthopaedic registry collecting clinical and patient-reported outcomes by telephone interview following elective primary and revision total hip arthroplasty (THA) and total knee arthroplasty (TKA) at six months (+/- one month) post-operatively.”

Similarly, the authors admit that 3 of the 6 surgeons actually saw most of their follow-ups by 8 weeks, which equates to roughly 2 months or 60 days, which is well below the common 90-day cutoff for postoperative complications. This isn't addressed at all and really throws the entire paper into question. Why should we accept a paper about complications if it doesn't include the standard window for complications? Why should we accept a paper that compares surgeon-reported complications and patient-reported complications when the reports are at different times? By my read of the paper, it's highly possible and easy that discrepancies could both be right: a surgeon doesn't report a complication at the 8-week follow-up, but the patient has one at 3 months and then reports it at the 6-month registry interview.

For this reason, we undertook the subgroup analyses by follow-up time and found no significant differences, as shown in table 4. The surgeon data were grouped into <6 weeks, 6-8 weeks, 3-5 months, 6 months and 6-12 months and was compared to the 6-month ACORN data. The surgeons’ records were accepted as gold standard, as mentioned in the methods, which reflect the real-life surgeon awareness of post-operative complications. This study compared ACORN data to the gold standard, it did not aim to test the validity of the gold standard. This has been addressed in the Methods section as follows:

“The surgeon follow-up times were categorised into <6 weeks, 6-8 weeks, 3-5 months, 6 months and 6-12 months, and were individually compared to the 6-month ACORN data.”
The paragraph addressing this issue was revised in the Discussion section as follows:

“Patients followed up outside the six-month mark were not excluded because in clinical practice not all patients are reviewed at the same time point by surgeons. This was addressed by including subgroup analyses in which patients reviewed at six months post-surgery can be compared to those followed up at different times, which showed no significant differences. Further, this study accepted the surgeons' records as gold standard as it reflects the real-life surgeon awareness of patient complications. Studies have discussed that although surgeons may have a better idea about what constitutes “true” medical and surgical complications, only the patients have the complete picture of adverse events [8, 11, 27]. Surgeons may also be susceptible to overlooking minor complications and keeping incomplete or inaccurate records, and alternative care sought from other health services for complications will not have been captured [14]. However, using patient recollection as the gold standard has its own set of problems, as it may be subject to recall bias. Both sources have limitations and this study suggests that agreement values (percentage agreement) may be more appropriate in assessing the accuracy of PRC.”

If I read the last paragraph of Results correctly, the issues listed above may be moot, but without showing actual data or any mention in the Discussion. If so, some of those are truly surprising and should get paragraphs of discussion and displayed, rather than just pushed away.

We have addressed differences in follow-up time by adding a subgroup analysis shown in table 4 which showed no significant differences.

The findings of the authors essentially repeat all of their citations and aren't surprising. They need to do a better job justifying the publication of yet another paper with this general finding.

We have sought to specifically examine the data captured in the ACORN registry for clinical use, which has not been done previously. Furthermore, our study is focused on total hip and knee arthroplasties. The aim of this study has been clarified in the background section as follows:
“This study aims to assess the reliability of PRC following THA and TKA recorded in ACORN compared to surgeons' medical records by calculating sensitivity, specificity, PPV, NPV and agreement values.”

Need much more information on the physician's records that were extracted. Were they EMRs or paper notes? A mix? How much investigation occurred? Who did the extraction: surgeon, nurse, medical assistant, someone else? There is a general dearth of necessary information like this in the Methods…

The data was gathered using the identical complications section of the ACORN questionnaire form as mentioned in the Methods section by the lead author. We have added the information suggested in the Methods as follows:

“Information on post-discharge complications for each patient was abstracted by the lead author from the electronic medical records maintained by each surgeon in their practice. Abstraction was repeated for a subset to confirm the reliability of the data collection. The items abstracted were the same as those collected from each patient at the six-month post-operative follow-up interview using the identical questionnaire used by ACORN. Patients without record of follow-up review by surgeons were excluded and a substitute patient from the randomised list of patients was substituted. Data collection continued until at least 50 patients with follow-up had been collected for each surgeon. Where more time was available, data from additional randomly-chosen patient records were collected from each surgeon’s records.”

Abstract: Should mention that ACORN is an Australian registry since it's not obvious from the title and this journal is not Australian-specific or orthopaedics-specific. Should also be bluntly mentioned in the manuscript as well. It wasn't obvious to me.

This has been included, as follows:

“Methods: This was a retrospective descriptive study of 364 patients who had completed their six-month follow-up review questionnaire in the Arthroplasty Clinical Outcomes Registry, National (ACORN), an Australian orthopaedic registry. Patient-reported complications following total hip arthroplasty (THA) and total knee arthroplasty (TKA) were compared to surgeon-
reported complications recorded in their electronic medical records at their various follow-up appointments.”

Abstract: The wording implies that all patients were seen at 6-month follow-up. However, the full text reveals that this wasn't the case for half the surgeons, and frankly should be changed throughout. Please rephrase the abstract accordingly.

We did not intend to imply that all patients were followed up at 6 months by their surgeons. We have clarified this in the abstract, as follows:

“Patient-reported complications following total hip arthroplasty (THA) and total knee arthroplasty (TKA) were compared to surgeon-reported complications recorded in their electronic medical records at their various follow-up appointments.”

We have also clarified this in the methods, as follows: “The surgeon follow-up times were categorised into <6 weeks, 6-8 weeks, 3-5 months, 6 months and 6-12 months, and were individually compared to the 6-month ACORN data.”

Abstract: I don't think the authors need to mention "2 x 2 tables." Too detailed and they don't actually show them in the manuscript itself.

This has been removed.

Abstract: Doesn't mention the actual procedures in question except for Key words. Need to explicitly state "total knee arthroplasty" and "total hip arthroplasty." The word "total" or the acronyms "THA" and "TKA" should be used throughout the manuscript to distinguish from other types of hip and knee arthroplasty.

This has been added, as follows:
Abstract:

“Patient-reported complications following total hip arthroplasty (THA) and total knee arthroplasty (TKA) were compared to surgeon-reported complications recorded in their electronic medical records at their various follow-up appointments.”

Background

“The Arthroplasty Clinical Outcomes Registry, National (ACORN) is an Australian orthopaedic registry collecting clinical and patient-reported outcomes by telephone interview following elective primary and revision total hip arthroplasty (THA) and total knee arthroplasty (TKA) at six months (+/-one month) post-operatively.”

The utility of Kappa versus % agreement discussion is interesting and I generally agree with the authors. That said, the authors use a fairly blanket decisiveness that I feel is not warranted by the size/power of their data, and I don't think it contributes much on the whole. They may want to think about limiting their conclusions on this topic or excise from the manuscript.

We have limited our conclusion in the Discussion section as follows:

“Although Cohen's kappa accounts for chance agreement, the literature has noted that the assumptions made about rater independence may overestimate chance agreement, thereby underestimating the agreement value [25, 26]. The implication for his study is that because of the low incidence of many of the complications, the kappa statistics are not robust to very small changes in agreement.”

Reviewer 2

REQUESTED REVISIONS:

Minor Issues
Abstract - there are numerous typing/spelling errors. For example...what is the 'medical les' is 'les' a typo? Specificity is incorrectly typed; non-significant differences in the Results Abstract section - please check and revise throughout.

This would appear to be a problem when viewing the files generated by the BMC publication system – the missing letters are the ligatures “ff”, “fl” and “fi”. The manuscript was submitted as a LaTeX file using the approved BMC LaTeX template, and the LaTeX correctly generated a PDF file on the authors’ system, a copy of which was uploaded to the BMC publication system along with the other files, for reference. If the problem persists, the BMC editors and/or technical staff may be able to assist. In the author’s experience, the Adobe Acrobat viewer often renders such PDF files correctly. If the reviewer has been given a Word file, then the problem is in the conversion to Word, which is carried out by the BMC publication system as far as we are aware. We can confirm that there are no selling errors or missing ligatures etc in the manuscript as submitted.

Major Issues

Abstract - I am not sure the results section is sufficient. This should be revised when the paper is considered as a whole but I would anticipate a difference in reliability based on the type of complication reported. To analyse a PE with a superficial wound infection as equal complications feels somewhat inappropriate. I think the analysis should be stratified by severity of complication. This will change the abstract's results section.

The analysis was stratified not by severity of complication, but category type as shown in table 1. We have clarified that different complications showed varying results in the abstract as follows:

“Values varied depending on the type and category of complication.”

Abstract - "Analyses performed categorizing patients by surgeon, joint operated and time between surgery and follow-up revealed non-significant differences between these groups" - I am not sure whether this level of detail is really needed in this paper. Answer the main question
first and then consider this secondary questions. Ultimately the results are most probably underpowered therefore the non-significant differences may be attributed to type 2 statistical error and therefore erroneous. Please consider.

This sentence has been removed from the abstract. However, the analysis was included to assess whether there were differences in results when differences in follow-up period, joint and surgeon were accounted for, and was therefore a necessary part of analyses. Table 4 has been added to address this.

Abstract - Conclusion - this is the correct conclusion. The sensitivity detects how good a test is (PROMS in this instance) for detecting a complication when one exists. The specificity is how good a test is for detecting when one does not exist….the results (based on the data), indicate that patients are good at detecting the absence of a complication (0.98) but not good a reporting the presence (0.14). HOWEVER, I think the data should be analysed by type of complication rather than complication per se. I suspect that this will radically change the results and interpretation. I would recommend the authors consider this, even if just by analysing my major and minor complication. This is important for surveillance work and long-term monitoring. This will also help contextualise the recommendation on 'greater attention to the clarity of the questions asked' as we will have a much better idea on what questions should be asked then.

Although complications were not grouped by severity, they were grouped by category into thromboembolic events, infections, joint problems, medical complications and subjective complications as seen in table 1 and the results of these analyses are shown in table 3.

Background - Page 1 line 54 - Please provide some indication on what is 'relatively low' in complication rates. This is important to provide the reader with context.

“Complications following hip and knee arthroplasty are relatively low” has been deleted and revised in the Background section as follows: “The ACORN 2016 annual report reported post-discharge complication rates of 0.055% to 5.4% for THA and 0.0% to 13.0% for TKA [6].”

Page 2 line 12-15 - Please provide data on sensitivity and specificity so that the reader can start to get an idea on what the comparative data is compared to this paper's results.

Specific data has been added along with PPV, NPV and agreement values found in existing literature in the Background section as follows:

“Overall, these studies have indicated high negative predictive values (NPV, 95.0% to 98.2%) but low positive predictive values (PPV, 26.0% to 83.3%) and varying levels of concordance (56.4% to 97.2%) and agreement (11.0% to 100.0%). One study on bone marrow transplant patients showed varying sensitivity (52.9% to 100.0%) and specificity (75.4% to 100.0%) dependent on complication type [13]. Another study monitoring patients' ability to report surgical site infections reported high sensitivity (83.3%) and specificity (93.7% to 98.1%) [16].”

Page 2 line 24 - The reasoning for doing this study is not clear. Please tell the reader explicitly why we need this analysis and what this analysis provides over previous studies [3-15 from the paper's reference list].

The purpose of this study was to validate ACORN complication data, and therefore patient-reported complications as a whole, and this has been clarified as the aim in the Background section. It is also important for future, similar data collection post arthroplasty. Clarified as follows:

“This study aims to assess the reliability of PRC following THA and TKA recorded in ACORN compared to surgeons' medical records by calculating sensitivity, specificity, PPV, NPV and agreement values.”

Page 3 line 41 - Please provide a sample size calculation to provide a justification for this cohort of 'approximately 50' per surgeon. There is limited clarity on this and a risk of selection bias. How were the 6 surgeon's selected? This would be useful to know to aid the transparency on the cohort's identification.
Details of the sample size calculations have been added. The six surgeons with the highest volume of procedures in ACORN were selected as mentioned in the methods section. The process of patient selection has been included in the Methods. Further subgroup analyses shown in table 4 were done to assess for any sample size differences between surgeons and were found to be non-significant. Sample size calculation was explained in the Methods section as follows:

“Sample size calculation

Sample size calculation for statistical power to detect a Cohen's kappa agreement statistic in the range 0.4 to 0.7 with a mean complication prevalence of 10% (range 1.0 to 15.0%) and with a standard alpha parameter of 0.05, was undertaken using the kappaSize package for R, yielding a minimum sample size of 300 patients [20]. To obtain an even distribution of patients, data from at least 50 patients from each surgeon who had most recently completed their six-month follow-up at April 2016 were acquired. An R script was used to randomly sample from all patients for each surgeon from 2015 from January to October inclusive, and the search was extended backward into 2014 if the patient volume was insufficient for some surgeons.”

Page 3 line 58 - I do not think that the subgroup analyses are helpful. I think analysing by joint (THA or TKA may be useful if the authors think that the pathway and process is different for these joints (they are not in my clinical experience from a surveillance perspective)). The time between surgery and follow-up should not be substantially difference and if it is, I would suggest the authors consider whether this should be an exclusion criteria. It is hard to ascertain whether the 'attending surgeon' should be an important factor as we have insufficient information on them and the process to which they work. I think analysing by type of complication would be much more valuable and would be more likely to provide differential results. Analysing by re-admission, reoperation and other complication would be a start. I would also suggest re-analysing by minor versus major complications. Page 3 lines 9 to 19 highlights a number of complications which the authors can make a clinical judgement on the severity of complication. This may be grouped into two or three different categorises. Please consider analysing the PPV, NPV, sensitivity and specificity by such groupings. I think this will provide a more insightful outcome.

“Attending” has been omitted. As discussed in the limitations, the subgroup analysis was performed to overcome the differences in follow-up periods. In practice, not all patients are followed up by surgeons at the 6-month mark as they were for ACORN. The subgroup analyses showed insignificant differences (table 4).
Complications were categorised into thromboembolic events, infections, joint problems, medical complications and subjective complications as seen in table 1, and analysed as such results of which are shown in table 3.

Results - Page 4 line 6 - I do not think that this is really a random sampling process at all. The researchers could select the time-scale, the surgeon's involved and potentially, as not clearly reported, the participants actually included. This needs clarifying.

Patients were randomly sampled. The surgeons selected were a convenience sample from surgeons contributing their data to the ACORN registry. The time-scale was set to nominally match that used by the ACORN registry, and is not a random sample. Details of the sample size calculation have been added, explaining the sampling process in the Methods section, as follows:

“Sample size calculation

Sample size calculation for statistical power to detect a Cohen's kappa agreement statistic in the range 0.4 to 0.7 with a mean complication prevalence of 10% (range 1.0 to 15.0%) and with a standard alpha parameter of 0.05, was undertaken using the kappaSize package for R, yielding a minimum sample size of 300 patients [20]. To obtain an even distribution of patients, data from at least 50 patients from each surgeon who had most recently completed their six-month follow-up at April 2016 were acquired. An R script was used to randomly sample from all patients for each surgeon from 2015 from January to October inclusive, and the search was extended backward into 2014 if the patient volume was insufficient for some surgeons.”

Results - Page 4, line 7 - how can a participant (potentially 24) be included if they did not review a surgeon review which formed the basis of the comparison to PROM. Please consider the reporting and/or eligibility criteria.

The 24 patients without record of follow-up with a surgeon have been excluded in our analyses. Have added to the Methods section: “Patients without record of follow-up review by surgeons were excluded from analyses.”
Results - Page 4 Line 12 - I would strongly suggest excluding Surgeons A, B and E as the follow-up time and therefore recall error is potentially different to Surgeons C, D and F…this is a major confounder which massively impacts on the findings. Please re-consider this approach and identify alternative surgeons or include the cohort from Surgeon C, D and F to reach sufficient participants to meet the sample size. Acknowledging this as Limitation 2 (Page 6, Line 60) is insufficient - something needed to be done to prevent this from happening and I think not recruiting from Surgeon A, B and E would be advantageous in this instance given the impact this could have on the study's conclusions.

All surgeons were included in the analyses as surgeon preference of follow-up periods were unknown to the authors prior to their selection. We have now added the subgroup analyses shown in table 4 which show no significant differences in the accuracy of the data between patients when grouped by follow-up time. Further, we have discussed that this study accepted the surgeons’ records as gold standard, as mentioned in the methods, to reflect real-life surgeon awareness of post-operative complications in the final paragraph of the discussion section as follows:

“Patients followed up outside the six-month mark were not excluded because in clinical practice not all patients are reviewed at the same time point by surgeons. This was addressed by including subgroup analyses in which patients reviewed at six months post-surgery can be compared to those followed up at different times, which showed no significant differences. Further, this study accepted the surgeons' records as gold standard as it reflects the real-life surgeon awareness of patient complications. Studies have discussed that although surgeons may have a better idea about what constitutes “true” medical and surgical complications, only the patients have the complete picture of adverse events [8, 11, 27]. Surgeons may also be susceptible to overlooking minor complications and keeping incomplete or inaccurate records, and alternative care sought from other health services for complications will not have been captured [14]. However, using patient recollection as the gold standard has its own set of problems, as it may be subject to recall bias. Both sources have limitations and this study suggests that agreement values (percentage agreement) may be more appropriate in assessing the accuracy of PRC.”

Results - Page 4 line 30-40 - I am cautious about the interpretation of the results because of the grouping issues as stated above.
The non-significant differences from the subgroup analyses indicate that follow-up time discrepancies do not threaten the validity of the conclusions. We have addressed the reviewer’s comments and believe our interpretation of the results is appropriate.

Results - Page 4 Line 40-47 - I am cautious about the interpretation until a power calculation has been presented, as stated previously.

Sample size calculation has been provided in the Methods section as follows:

“Sample size calculation

A sample size with statistical power to detect a kappa agreement statistic in the range 0.4 to 0.7 with a mean complications prevalence of 10% (range 1.0-15.0%) with a standard alpha parameter of 0.05 was calculated to be a total of 300 patients. To obtain an even distribution of patients, data from at least 50 patients from each surgeon who had most recently completed their six-month follow-up at April 2016 were acquired. An R script was used to randomly sample from all patients for each surgeon from 2015 from January to October inclusive, and the search was extended backward into 2014 if the patient volume was insufficient for some surgeons.”

Discussion - I am concerned about the interpretation because of the analysis approach adopted. Please find comments above.

The non-significant differences from the subgroup analyses indicate that follow-up time discrepancies do not threaten the validity of the conclusions, and this study aimed to compare ACORN data to surgeons’ records, having already accepted it as the gold standard. Other concerns regarding sampling and complication categorization have been addressed by providing the sample size calculation and referring to the categorization in table 1.

Discussion - the interpretation is reasonable however I would recommend the authors reflect on the qualitative literature on reporting, expectations and outcomes which will help support their assumptions on patient recovery and complications in Page 5 and 6.
The manuscript has been corrected according to reviewers’ comments to include fewer assumptions in the discussion which are replaced by comments substantiated with evidence as follows:

Deleted: “This may be because patients are seldom bothered by minor differences in leg length and often overlook these as long as they have improved function, whereas surgeons regard unequal leg lengths as a sign of poor technique. Regarding infection, the mean patient age for this study was 68 years, with most of the patients in an age range which is likely to be prescribed multiple medications. To these patients, antibiotics may become “just another pill” and may possibly account for their under-reporting of low level wound infection.”

Replaced with: “Patients with minor leg-length discrepancies may have few symptoms which may explain the under-reporting by patients, whereas surgeons place great importance on leg length due to its possible detrimental outcomes [22]. This study showed poor patient specificity in identifying SSI, a finding that was also observed in a study by Zellmer et al., who suggested that may be improved with validated infection education material [23].”

Discussion - Limitations (Page 6 Line 53-60) - I do not think the authors made a sufficient account for the major limitation in this study which is around the 'gold standard' comparison as 'surgeon records' (Page 3 line 53). This is major as there may be issues of poor reporting of medical complications, particularly minor complications which may have cause some discrepancy in reporting of complications. Furthermore, what is the risk of non-reporting complications to the surgeon's hospital but self-managing or seeking alternative medical care such as from a community physician or other orthopaedic surgeon? Again, this may provide a discrepancy which should be addressed in the paper's limitation section but also provides a reason for categorising analyses by different levels of severity of complication. Lesser complications may not require hospital consultation or may have greater risk of poor recall than more memorable major complications which need urgent medical attention.

We have noted the potential susceptibility of surgeons underreporting minor complications and presentation to alternative medical care in the discussion as follows:
“Surgeons may also be susceptible to overlooking minor complications and keeping incomplete or inaccurate records, and alternative care sought from other health services for complications will not have been captured [14].”

As mentioned above, we have categorised complications by type not severity, as shown in table 1.

We agree that both surgeon and patient reporting have limitations which raise the question of which data source to use as gold standard. We have amended the discussion to suggest that agreement values are more appropriate than sensitivity, specificity, PPV and NPV, as follows:

“However, using patient recollection as the gold standard has its own set of problems, as it may be subject to recall bias. Both sources have limitations and this study suggests that agreement values (percentage agreement) may be more appropriate in assessing the accuracy of PRC.”

Conclusion - I am cautious due to the methodological issues I have raised earlier. The authors need to re-format the data and then re-analyse. They may also need to half their sample due to issues in their choice of surgeons. The authors may also need to consider how the data was collected. These are fundamental issues.

We have discussed why we have included all surgeons in our analysis despite the differences in time: to reflect real-life surgeon awareness of post-operative complications. The surgeons' records were accepted to be the gold standard in our study but have discussed that the literature has shown that it is not a perfect gold standard. This is addressed in the final paragraph of the discussion, as follows:

“Patients followed up outside the six-month mark were not excluded because in clinical practice not all patients are reviewed at the same time point by surgeons. This was addressed by including subgroup analyses in which patients reviewed at six months post-surgery can be compared to those followed up at different times, which showed no significant differences. Further, this study accepted the surgeons' records as gold standard as it reflects the real-life surgeon awareness of patient complications. Studies have discussed that although surgeons may have a better idea about what constitutes “true” medical and surgical complications, only the patients have
the complete picture of adverse events [8, 11, 27]. Surgeons may also be susceptible to overlooking minor complications and keeping incomplete or inaccurate records, and alternative care sought from other health services for complications will not have been captured [14]. However, using patient recollection as the gold standard has its own set of problems, as it may be subject to recall bias. Both sources have limitations and this study suggests that agreement values (percentage agreement) may be more appropriate in assessing the accuracy of PRC.”

We have also clarified how the data was collected in the Methods section as follows:

“Information on post-discharge complications for each patient was abstracted by the lead author from the electronic medical records maintained by each surgeon in their practice. Abstraction was repeated for a subset to confirm the reliability of the data collection. The items abstracted were the same as those collected from each patient at the six-month post-operative follow-up interview using the identical questionnaire used by ACORN. Patients without record of follow-up review by surgeons were excluded and a substitute patient from the randomised list of patients was substituted. Data collection continued until at least 50 patients with follow-up had been collected for each surgeon. Where more time was available, data from additional randomly-chosen patient records were collected from each surgeons' records.”