PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (see an example) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below. Some articles will have been accepted based in part or entirely on reviews undertaken for other BMJ Group journals. These will be reproduced where possible.

ARTICLE DETAILS

TITLE (PROVISIONAL)	PATIENT- AND SURGEON-ADJUSTED CONTROL CHARTS FOR MONITORING INDIVIDUAL PERFORMANCE
AUTHORS	Maruthappu, Mahiben; Carty, Matthew; Lipsitz, Stuart; Wright, John; Orgill, Dennis; Duclos, Antoine

VERSION 1 - REVIEW

REVIEWER	Cassandra Villegas, MD MPH Department of Surgery University of Arizona Tucson, AZ United States of America
REVIEW RETURNED	28-Oct-2013

GENERAL COMMENTS	Please do a thorough check of references cited. There are numerous examples of citations that are grossly inaccurate. For instance (but not limited to), line 46 of page 14 "...been used as a robust indicator of learning and outcome [17, 30, 31, 32, 33, 45]" refers to reference 45, but there is no reference 45 in the manuscript. Or it is referring to Peersman, et al. for "Prolonged operative time..." which is misnumbered in the reference section (between references 33 and 34 on page 20). Also, line 8 of page 15 "...substitute for surgical experience in prior published studies [17, 18, 19, 39]." refers to reference 39 but only 37 references are listed. THE STUDY The authors have tackled a very germane topic in surgical research, namely that of adjusted outcomes in evaluating surgeon performance. Specifically, the goal was to provide a proof-of-concept investigation for the use of case-mix and surgeon-experience adjusted control charts as a method for benchmarking surgeon performance on TKRs. The authors used a large database of TKR's to compare adjusted operative time curves based on either case-mix alone or case mix with surgeon experience. The study employed robust statistical methodology for which the authors should be commended. Not only did they control for case-mix by regression analysis but they also controlled for clustering (using generalized estimating) as well as recognizing that adjusted operative times may not be linear. Ultimately, the authors concluded that the incorporation of surgeon experience in addition to case-mix significantly changed the

evaluation of surgeon performance.

COMMENTS ON RESULTS & CONCLUSIONS
1) Table 1 reports overall patient demographics and comorbidities. Though not required but helpful would be an additional two columns that compared those demographics between the training and testing datasets. Even a cursory sentence indicating that no differences (assuming that there were none) could be found in the two populations would be sufficient...

2) Please include the 30.3 min standard deviation time to the reported mean operative time of 109 min in the first paragraph of the Results section in order to give the reader some sense of operative time variability.

3) To their credit, the authors recognize and address the fact that years of experience may not correlate with the number of TKR cases. However, this may not be sufficient. This can be seen in Table 1 if one calculates the number of cases/year. At its most generous, those 10 cases were in just one year, resulting in 10 cases/year. At the other end of the spectrum, one surgeon performed 1,871 cases. If those cases were performed by the longest practicing surgeon (35 years), then she or he would have done approx 53 cases per year. This indicates that there is substantial variability in the number of cases done per year by each surgeon, significantly undermining the validity of using years of practice as a proxy for surgeon experience. One way to address this limitation would be to limit the study to those surgeons with similar cases/year (surgeons must have performed a minimum of 20 cases/year), though this would admittedly require reanalysis of the data. Alternatively, one could report in the supplemental section the number of TKR's/year for each surgeon to better allow readers to properly evaluate the use of years of experience as a proxy for the number of TKR's performed.

4) Figure 1 deserves more clarification. If this is being read correctly, then this is the reduction in the calculated adjusted operative time and not a plot of a surgeon's actual operative time. If so, then that should be made explicit in line 19 of page 12 of the results section (ie "a substantial decline in the patient-risk adjusted operative time..."). In addition, the model would suggest that with sufficient experience (5th hash mark on the x-axis), that a surgeon could expect a decrease in operative time of nearly 120 minutes. Please comment on the validity of this model in light of the fact that the actual mean operative time was only 109 minutes and that 95% of them were done <170 minutes (based on a SD of 30 min). Does this overstate the impact of surgeon experience on operative time?

5) Why these comorbidities? While these would certainly be expected to affect patient outcomes such as mortality, surgical site infection, length of stay, need for re-operation, etc., the ability of these comorbidities to affect operative time itself is less clear. Other comorbidities such as Factor V Leiden deficiency, anti-coagulation use, revision of prior TKR, or even BMI would seem more relevant to impacting operative time than a patient's COPD status. Was the selection of these comorbidities based on previous studies affecting operative time or instead other patient outcomes? Perhaps based on statistical inclusion using forward, backward, or step-wise regression? Please clarify. Also, a small table of the case-mix model

with beta coefficients added to the supplemental section would help readers to judge the fit of your modeling.

6) One of the limitations with databases that cover a substantial time period (15 years) is the potential for changes in operative technique/technology. Please indicate whether the four surgeons from the testing dataset collectively practiced over the 15 years of the dataset and thus remove that confounder or perhaps whether surgical practice/technology for TKR has remained relatively unchanged during that period.

7) This study does not address the topic of resident/trainee involvement in cases and operative times. Numerous studies have found that trainee involvement prolongs operative time (PMID 22365858, 22476835, 23520071). Given that this study was at an academic tertiary care center, to what degree was resident involvement accounted for in this study and how does it impact/confound the results?

8) As the authors correctly pointed out, larger datasets combined with more accurate statistical methodology has allowed for more comprehensive and granular detail on surgical outcomes (ie we can now control for both patient and surgeon level characteristics on operative time). All who have operated recognize the benefits of having a surgical team that is fluent in the procedure and can anticipate surgical needs. Is it possible that some of the operative time variation may be accounted for by surgical staff experience? Specifically, did the four surgeons (A-D) use the same surgical staff?

As a side note, it is quite likely that specific surgical staff was not captured in the original dataset. However, if it was and the authors did wish to incorporate it into their modeling, you might find hierarchical modeling to be more robust than GEE when dealing with multi-level/dependent data.

GENERAL COMMENTS
Overall, the paper is well done, especially since it addresses the challenging and multi-factorial subject of surgeon evaluation. There are some reference issues that are mildly distracting and the paper would benefit from reporting some underlying data, but overall the paper merits publication with minor revisions.

REVIEWER	Gary Collins University of Oxford
REVIEW RETURNED	26-Nov-2013

GENERAL COMMENTS	This is a fascinating paper taking quite a novel approach to evaluate surgeon performance. The resulting graphs are interesting and clear and the manuscript (generally) clearly written. My only criticism is whilst the methods per se are clear, I am struggling to join the dots and find the description rater abstract and I'm fairly sure most readers will struggle to grapple with what is actually being done. If an independent group want to replicate this then I think some more 'hand-holding' is required. The paper would benefit from being slightly more transparent and simplistic in their description.

VERSION 1 – AUTHOR RESPONSE

REVIEWER 1 COMMENTS:

Reviewer Name Cassandra Villegas, MD MPH
Institution and Country Department of Surgery
University of Arizona
Tucson, AZ
United States of America
Please state any competing interests or state ‘None declared’: None declared.
Please do a thorough check of references cited. There are numerous examples of citations that are grossly inaccurate. For instance (but not limited to), line 46 of page 14 "...been used as a robust indicator of learning and outcome [17, 30, 31, 32, 33, 45]" refers to reference 45, but there is no reference 45 in the manuscript. Or it is referring to Peersman, et al. for "Prolonged operative time..." which is misnumbered in the reference section (between references 33 and 34 on page 20).

Also, line 8 of page 15 "...substitute for surgical experience in prior published studies [17, 18, 19, 39]." refers to reference 39 but only 37 references are listed.

>>> We have reviewed and appropriately amended the references in accordance with the reviewer’s suggestions.

THE STUDY
The authors have tackled a very germane topic in surgical research, namely that of adjusted outcomes in evaluating surgeon performance. Specifically, the goal was to provide a proof-of-concept investigation for the use of case-mix and surgeon-experience adjusted control charts as a method for benchmarking surgeon performance on TKRs.

The authors used a large database of TKR's to compare adjusted operative time curves based on either case-mix alone or case mix with surgeon experience.

The study employed robust statistical methodology for which the authors should be commended. Not only did they control for case-mix by regression analysis but they also controlled for clustering (using generalized estimating) as well as recognizing that adjusted operative times may not be linear.

Ultimately, the authors concluded that the incorporation of surgeon experience in addition to case-mix significantly changed the evaluation of surgeon performance.

COMMENTS ON RESULTS & CONCLUSIONS
1) Table 1 reports overall patient demographics and comorbidities. Though not required but helpful would be an additional two columns that compared those demographics between the training and testing datasets. Even a cursory sentence indicating that no differences (assuming that there were none) could be found in the two populations would be sufficient...

>>>We have updated Table 1, comparing patient demographics in the training and testing datasets.

2) Please include the 30.3 min standard deviation time to the reported mean operative time of 109 min in the first paragraph of the Results section in order to give the reader some sense of operative time variability.

>>> We have included the standard deviation time in the first paragraph of the results as suggested.

3) To their credit, the authors recognize and address the fact that years of experience may not correlate with the number of TKR cases. However, this may not be sufficient. This can be seen in Table 1 if one calculates the number of cases/year. At its most generous, those 10 cases were in just one year, resulting in 10 cases/year. At the other end of the spectrum, one surgeon performed 1,871 cases. If those cases were performed by the longest practicing surgeon (35 years), then she or he would have done approx 53 cases per year. This indicates that there is substantial variability in the number of cases done per year by each surgeon, significantly undermining the validity of using years of practice as a proxy for surgeon experience. One way to address this limitation would be to limit the study to those surgeons with similar cases/year (surgeons must have performed a minimum of 20 cases/year), though this would admittedly require reanalysis of the data. Alternatively, one could report in the supplemental section the number of TKR's/year for each surgeon to better allow readers to properly evaluate the use of years of experience as a proxy for the number of TKR's performed.

>>> We have introduced a new table (Table 2), summarising the number of TKRs performed by each surgeon in the testing and training datasets. To maintain anonymity we have not stated the number of years over which each surgeon performed the TKRs.

4) Figure 1 deserves more clarification. If this is being read correctly, then this is the reduction in the calculated adjusted operative time and not a plot of a surgeon's actual operative time. If so, then that should be made explicit in line 19 of page 12 of the results section (ie "a substantial decline in the patient-risk adjusted operative time..."). In addition, the model would suggest that with sufficient experience (5th hash mark on the x-axis), that a surgeon could expect a decrease in operative time of nearly 120 minutes. Please comment on the validity of this model in light of the fact that the actual mean operative time was only 109 minutes and that 95% of them were done <170 minutes (based on a SD of 30 min). Does this overstate the impact of surgeon experience on operative time?

>>> We have amended the phrasing in the results to “substantial decline in risk-adjusted operative time” as suggested by the reviewer. With regards to the second point, we respectfully posit that that the issue raised by the reviewer regarding anticipated operative time reflects a misunderstanding of the data presented. Our analysis suggests that increasing surgeon experience can contribute to a substantial reduction in operative time over the course of one’s operative career; this improvement can yield as much as 120 minutes in improved operative efficiency when comparing mean operative time at the beginning of one’s career to mean operative time at the stage of maximal efficiency. The

reported mean operative time for the entire study population does not fully illustrate the markedly higher mean operative time witnessed in the neophyte stages of operative maturity – rather, this is more fully represented in the performance curve that better reflects average operative time at various levels of surgeon experience. The opportunity for improvement referenced in the manuscript describes the maximal efficiency differential that separates the performance curve apex and nadir rather than an improvement goal relative to the population mean.

5) Why these comorbidities? While these would certainly be expected to affect patient outcomes such as mortality, surgical site infection, length of stay, need for re-operation, etc., the ability of these comorbidities to affect operative time itself is less clear. Other comorbidities such as Factor V Leiden deficiency, anti-coagulation use, revision of prior TKR, or even BMI would seem more relevant to impacting operative time than a patient's COPD status. Was the selection of these comorbidities based on previous studies affecting operative time or instead other patient outcomes? Perhaps based on statistical inclusion using forward, backward, or step-wise regression? Please clarify. Also, a small table of the case-mix model with beta coefficients added to the supplemental section would help readers to judge the fit of your modeling.

>>> We included all available co-morbidities in our dataset into the case-mix adjustment model. Other co-morbidities such as Factor V Leiden deficiency could not be included as they were not recorded at the time of the operation. We have updated the methods section of the manuscript to state this.

6) One of the limitations with databases that cover a substantial time period (15 years) is the potential for changes in operative technique/technology. Please indicate whether the four surgeons from the testing dataset collectively practiced over the 15 years of the dataset and thus remove that confounder or perhaps whether surgical practice/technology for TKR has remained relatively unchanged during that period.

>>> We acknowledge that our database covers a substantial time period and that operative technique/technology may have changed; we have now included this as a limitation of the study.

7) This study does not address the topic of resident/trainee involvement in cases and operative times. Numerous studies have found that trainee involvement prolongs operative time (PMID 22365858, 22476835, 23520071). Given that this study was at an academic tertiary care center, to what degree was resident involvement accounted for in this study and how does it impact/confound the results?

>>> Our group is currently evaluating the impact of teamwork and resident involvement operative performance in TKR. These factors were not adjusted for in our analyses and have now been included as a limitation of the study in the discussion.

8) As the authors correctly pointed out, larger datasets combined with more accurate statistical methodology has allowed for more comprehensive and granular detail on surgical outcomes (ie we can now control for both patient and surgeon level characteristics on operative time). All who have operated recognize the benefits of having a surgical team that is fluent in the procedure and can anticipate surgical needs. Is it possible that some of the operative time variation may be accounted for by surgical staff experience? Specifically, did the four surgeons (A-D) use the same surgical staff?

As a side note, it is quite likely that specific surgical staff was not captured in the original dataset. However, if it was and the authors did wish to incorporate it into their modeling, you might find hierarchical modeling to be more robust than GEE when dealing with multi-level/dependent data.

>>> As suggested by the reviewer, unfortunately data on specific surgical staff was not captured in the original dataset; we therefore cannot answer this question with accuracy.

REVIEWER 2 COMMENTS

Reviewer Name Gary Collins
Institution and Country University of Oxford
Please state any competing interests or state ‘None declared’: None declared

this is a fascinating paper taking quite a novel approach to evaluate surgeon performance. The resulting graphs are interesting and clear and the manuscript (generally) clearly written.

My only criticism is whilst the methods per se are clear, I am struggling to join the dots and find the description rater abstract and I'm fairly sure most readers will struggle to grapple with what is actually being done. If an independent group want to replicate this then I think some more 'hand-holding' is required.

The paper would benefit from being slightly more transparent and simplistic in their description.