PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Impact of patient characteristics on the Canadian Patient Experiences Survey – Inpatient Care survey- analysis from an academic tertiary care centre
AUTHORS	Rothwell, Diana; Al Zayadi, Amal; Sundaresan, Sudhir; Ramsay, Tim; Forster, Alan; Rubens, F

VERSION 1 – REVIEW

REVIEWER	Nabil Natafgi University of Maryland, Baltimore, USA
REVIEW RETURNED	01-Feb-2018

GENERAL COMMENTS	Abstract: 1. First sentence in the abstract: not all 4 global questions are P<0.001 for physical health (recommend hospital is 0.018). Similarly, education & rate experience (p=0.007)... Strengths and Limitations: 1. Avoid decisive terms such as "conclusively". The study design and methodology may not allow for such definitive conclusions. 2. Basically, you summarized the conclusions or stated implications for practice rather than strengths and limitations. I am fine with this approach, but I am not sure if this is what the journal is asking for. Strength, for example, maybe you inclusion of other satisfaction domains or other predictors such as perception of health not reported earlier in the Canadian context. Weakness would address study limitations. Introduction 1. Topbox scoring: (a) Kemp et al (ref 20) seem to have collapsed experience into 10 vs. 0-9. Can you discuss how you came to your choice of 10-9 vs. 0-8? What did other studies do? what is their and your rationale? (b) it seems this may be more of a 'methods' discussion as opposed to 'introduction'. Methods: 1. Was the survey solely administered in English or was it translated to French, as well? 2. What was the response rate? 3. Did all CPES-IC surveys match with admin data? If not, what did you do with mismatches or unmatched cases? 4. How did you deal with missing data? 5. Your reported Elixhauser measure seems to have produced <0. Can you discuss why is there negative values and what does that mean? 6. Clarify which tool did you use to identify PSIs using ICD codes... Is it the one by AHRQ? If so, which which PSIs? any? Some are

	surgical others are medical, did you use the different PSIs for the different admitting departments? 7. Page 8 of 47 Line 50: I think this should read "There are 7 admitting departments" 8. Can you discuss the impact of missing data from one surgical +one medical division on selection bias? How does those 2 divisions compare to others in terms of (1) study outcomes (i.e. rate exp., recommend hosp, etc.) and (2) patient characteristics + perceived health. 9. Composite domains: did you collapse the questions into the said domains? or where they previously determined by the survey creators and/or published in literature? what is the correlation between some of those domains (e.g. rate experience & rate hospital; or recommend hospital and rate hospital). Someone might argue they are measuring the same thing. 10. Can you comment briefly on the validity and reliability of the CPES-IC as a tool? was it validated by anyone before? Results: 1. Check reported numbers of respondents and clarify differences between text (2989) and Table 1 (2935). 2. What is the rationale behind excluding maternity, rehab, and mental health admissions? Also, this discussion belongs to the methods, where you discussed exclusion of psychiatry and ophthalmology. 3. Page 10 of 47 Line 36: " The institution consists of the patients from [...]" may fit better when discussing the setting in methods. 4. Page 10 of 47 Line 43: age and discharge disposition are statistically significant 5. Is there a composite / single patient experience measure for CPES-IC? 6. Page 10 Line 54: be careful with statements that may imply causality (that cannot be established here). 7. Page 11 Line 26: Figures 1 thru 4 shows measures for rate experience, recommend hospital, rate hospital, and overall helped. 8. For pairwise comparisons (figures 1-4), what are you adjusting for? all the variables presented in the multivariate models? 9. Page 11 Lines 29-36: confusing. You discuss the unadjusted pairwise comparison btw Surg & Med for rate experience being not significant. And then (this not significant) difference disappear after adjustments. Discussion: 1. Page 13 Line 26: Even though it is a single payer system, the private nature of providers still does not generate enough competition to attract patients? 2.There are no federal or provincial provisions (or anticipated provisions) for linking payment to some performance measures (including pt experiences)? 3. Page 14 Line 20: Your findings showed that admit urgent and LOS were not significant predictors (as opposed to Ref 20). What do you think are the reasons? Is there a difference in the case-mix of the studies? Or is this because you adjusted for some additional variables? Did you consider running the same models Kemp did, and see if you get comparable results to them, in your dataset? If so, then you may want to know which covariates in multivariate models removed the significance. Is it the physical and mental health? This would be an interesting finding, I think. 4. Page 15 Line 25: "The analysis highlights the differences in adjusted and unadjusted [...]"/ Physical and mental health are not

	demographics. 5. Page 15 Line 31: Elixclass (comorbidity) was not stat. significant in your multivariate models. Although it is very surprising finding, I am not sure you can come to your conclusions here with that finding. Again I am surprised about the Elixclass not being signifcant and I am not sure if it is related to the way you modified that variable. Also, what was your rationale for the classification you used (i.e. <0, 0, 1-5, >13). I think this argument needs to be based on theoretical/clinical grounds supported by stats from your database. The 19% of people having 6+ commodities seem a little high and primarily coming from medicine. 6. I agree that physical and mental health as predictors of experience is an important finding. Also reinforcing the importance of factoring differences between case-mix among diff depts is important. I also agree with your statement that this recognition should enhances engagement of staff facing challenges of pts with chronic conditions. Yet, I think this shouldn't give depts caring for pts with chronic conditions a pass on lower scores for patient experience, simply because of the case-mix they serve. Rather, leadership should provide more resources and / or opportunities for those depts to address this challenge and reach scores comparable to other departments, albeit the differences in case-mix. 7. Page 16: It is not clear which ones are the patient care domains (I am assuming the one with doctors communication, nurses communication, etc.). If so, I am not sure I necessarily understand why they were excluded. I see that they can be correlated with other predictors, but I do not think in a way more than some of your current predictors correlate with each other, anyhow. So maybe look at the correlations quantitatively and present those numbers (e.g. use VIF or other appropriate statistic). Another way looking at could be treating them as a separate set of analyses where they are the outcomes predicted by the same predictors you currently have in your presented models. Tables: 1. Table 1: Check the "n" in the first row for Total and different departments. For example mental health adds up to 2938; also the four depts adds up to 2,896 2. Did you consider collapsing some of the variables, especially in cells where you have too few n (e.g. education, age, elixclass)? 3. Are the Race categories mutually exclusive? 4. In the Tables legends you have ALC, I did not see that in the tables 5. Table 1: "ED isit" should read "ED visit" 6. It would be helpful to define the disposition categories in the methods. What does home-setting include? what does another health facility include? Were there any death cases? 7. How did you code LOS for the multivariable analyses? Did you leave it as a count variable? If so what does the (>3 days) mean? Or did you change it to a binary variable (>3 days vs. < 3 days) 8. Did you consider looking at the ED visit within 7 days as an outcome predicted by patient experience? With that, I am not sure if you should include ED visit w/in 7 days as a predictor of patient experience (even if it wasn't statistically significant). Same argument may apply for discharge disposition. 9. Be consistent in the ordering of the variables in the tables for ease of comparison across Tables. Also, you may want to merge Tables 2-4 in a single table as they have the same predictors. 10. Table 3: Physical Health - Poor - remove the third figure after the

	point 11. Since you chose topbox scoring for experiences, did you consider similar scoring for self-rated health... i.e. coding excellent vs. other? or excellent + V. good vs. others? 12. I am not necessarily recommending doing so, but did you consider collapsing education and / or age categories? or using age as continuous variable? 13. Can you include the results of the univariate analyses in comparison to the multivariate comparison to better see the effect of adjustments Figures 1. Although it's results might be a little surprising, I like figure 5 as it may have some implications for practice and you do mention that in your discussion. Can you just clarify which variables entered in this analysis? Is does this adjust for some of the demographics you were adjusting for earlier? Other comments 1. Is it possible to include a figure that shows the distribution of your dependent variable (patient experiences) prior to switching it to a binary variable. You may chose to not eventually include that in the final paper, but it is worth looking at to understand the raw distribution.

REVIEWER	Rachel Foskett-Tharby University of Birmingham, UK NHS England, UK
REVIEW RETURNED	12-Feb-2018

GENERAL COMMENTS	This was an interesting, well drafted paper. A couple of minor points: Page 5 lines 47-54 and page 6 lines 3-24: I found this section a little confusing with the classification of the additional questions into groups and then domains. I wonder if this should be revised to make it clearer when you are describing the additional questions in the Canadian survey. Page 8: methods - it would be useful to have a little more information about the Elixhauser comorbidity measure and why this was chosen rather than any other one. The reference to this (line 40) needs updating.

REVIEWER	Vincent S. Staggs, PhD Children's Mercy Kansas City niversity of Missouri--Kansas City Kansas City, MO USA
REVIEW RETURNED	29-Mar-2018

GENERAL COMMENTS	Thank you for the opportunity to review this work. I found the paper well-written overall, and the study seems thoughtful and well done. As a biostatistician I'll focus on the statistical aspects. I hope the authors find these comments helpful. 1. Readers may be interested in seeing the correlation matrix, or at least the correlations between the dependent measures and the quantitative explanatory variables. No need to include p-values with the correlations (which would exacerbate the multiple testing issue). 2a. Clustering. With only two campuses I would simply include campus as a fixed (not random) effect; I'm not sure that two campuses from a single hospital can provide a meaningful estimate of between-hospital variance in the population of hospitals, and a small number of clusters can be problematic in mixed models (see Kenward & Roger, 1997).

	2b. The more important source of clustering to consider is unit/ward. It sounds like patients came from eight units total, and I would recommend including in each model a random intercept for unit (in addition to the categorical fixed effect for department type) and, given the small number of clusters, using the Kenward-Roger degree of freedom method. 3. Did the authors examine multicollinearity among the explanatory variables in the logistic regression models (e.g., VIFs)? It seems likely that some variables are correlated (e.g., ICU with LOS and patient safety event, age with physical health) so it seems prudent to check this. 4. Multiple testing. The authors note using the Bonferroni adjustment for pairwise comparisons, presumably for the department comparisons shown in the Figures. But with roughly 200 hypothesis tests carried out, multiple testing remains something of a concern. The Bonferroni method is conservative and may severely reduce power if applied across all tests (although the large sample size will help). One option to reduce the number of tests would be combining a couple of dependent measures if they're highly correlated (as I would guess some are), or just choosing one measure from those highly correlated. Another would be treating physical health, mental health, Elixclass score, and age as quantitative (not categorical) variables. It would also help (a little) to collapse across some categories for education and/or race. The authors also might consider the Benjamini-Hochberg False Discovery Rate (FDR) method, which is a nice, less conservative, alternative to Bonferroni adjustment. 5a. For the analyses shown in the Figures, please explain under Statistical Analyses where the unadjusted and adjusted predicted percent topbox numbers come from, and how the pairwise tests were carried out with these variables. 5b. I'm not sure how much the figures themselves add; perhaps it would suffice to show the effects of adjustment by reporting these numbers in a table. 6. In describing the logistic regression analyses (P10) the authors mention only likelihood ratio tests. However, the LRT p-values do not generally match the OR p-values in the tables (e.g., in Table 3 the LRT p-value is very different from the OR p-value for Any PSI), so I assume the latter must be Wald or profile likelihood test p-values. In any case, it would be helpful to clarify the description of tests in the Statistical Analyses section. 7. In reporting results of the logistic regression models I would include the C-statistic/AUC for each. Also, please label the second column in the regression tables. 8. On a minor note, there's no need to log-transform income if converting to deciles. Log-transformed income values will fall in exactly the same deciles as the untransformed values, as assignment to deciles is determined by ranks and log is a strictly monotone (and thus rank-preserving) transformation. 9. On P10, first sentence, something like this might be clearer: "After dichotomizing each of the four overall care questions [(a) … (b) … (c) … (d) …] based on the 'topbox' response criteria defined above, we fit a separate logistic regression model for each question to model the odds of topbox response as a function of [explanatory variables]." 10. The authors may have addressed this, but I'm curious how they classified patients who were admitted in one unit but ended up in another. Or perhaps this was rare. 11. It would be helpful to describe the Key Driver Analysis in more detail under Statistical Anaylses. On first encountering "vertical

	separation of the quadrants" I didn't follow. 12. The non-linear effects of age (with highest ratings in the middle and lower ratings for youngest and oldest patients) may be worth some discussion. 13. I found the writing good overall but would recommend having someone unfamiliar with the study read and edit the paper. There are places where additional clarity would be beneficial.

REVIEWER	Hilde Hestad Iversen Norwegian Institute of Public Health, Norway
REVIEW RETURNED	15-Apr-2018

GENERAL COMMENTS	Impact of patient characteristics on the Canadian Patient Experiences Survey – Inpatient Care survey- analysis from an academic tertiary care centre The objective of this article is to determine the role of patient demographics, care domains and self-perceived health status in the analysis and interpretation of results from the Canadian Patient Experience Survey-Inpatient Care (CPES-IC). Hospital patients were randomly sampled post-discharge. Logistic regression models were developed to analyze topbox scoring on questions of global care. It is concluded that caution should be exercised in using patient-satisfaction surveys to compare performance between different health-care provision entities, as differences could be explained by variation in patient mix rather than variation in performance. The manuscript addresses an important issue. I agree that even if case-mix adjustment often have a small impact on hospital ratings, the rankings of some hospitals may be substantially affected and it can lead to reductions in the bias in comparisons between hospitals. Subgroups of patients who have the same experiences may still provide different responses because some subgroups of patients may be more generous than others in providing positive responses, while others are more critical. This might be related to for example sociodemographic factors, not quality. However, the aim of the current study is not clear to me. Is it to be able to compare different units or departments at the same hospital or to compare hospitals? Also the reason why potential adjustments is an important area of interest when measuring patient experiences and the underlying rationale behind this should be explained more thoroughly. It is not obvious how the current results should be used in the follow-up work at the hospital. On page 7 we are told that the overall objective was to compare the value of the self-reported background variables with covariates from a hospital database, in the development of a statistical model to predict topbox scoring in the four survey questions related to overall care. Please elaborate more on this aim, is the aim to explore or compare? The study focus mostly on global ratings, except for the key driver analysis, often skewed towards the positive end of the scale and showing less variation than specific experience items. Any reflections on this subject? In the introduction we are given information on how the Canadian Patient Experience Survey – Inpatient Care (CPES-IC) differ from the HCAHPS in the US. Some if this information, especially the elaboration on specific items but also on top-box ratings, should be included in the methods section.

	The modification of the HCAHPS survey was developed through a collaboration between different parties, but how many items were changed? Only the items measuring the new topics, or some of the others as well? The authors mention in the discussion section that the newly developed questions added should have been testet, this is very important to explore the validity and reliability of the items. Have the other domains/questions from the HCAHPS been validated in in the Canadian context? We are told that the patient experience survey are routinely administered in four provinces in Canada, however, that there is limited familiarity in the assessment of patient experience and the use of such surveys. How are these results reported and responded to today? Which institutions carry out or are responsible for the surveys? Are the current results compared? Are they case-mix adjusted and weighted, or is this work the first effort to explore a possible case-mix model? Again, the aim here might just be to compare departments at the current hospital? The abstract tells us that the participants of the study were randomly sampled post-discharge. I did not find more information on the administration of the survey in the method section, either on the number of patients that were invited to participate or the response rate. How did the subjects respond, electronically or on paper? How were they contacted? Where did they fill out the questionnaire, at home or at the hospital? Were the hospital responsible for the survey? We are not informed about the response rate or given any information on non-response. Do the authors have any reflections regarding the representativeness for inpatients in general? I am a bit uncertain on the choice of analyses because I did not fully comprehend the aim of the study (see my previous comment). Please explain more thoroughly why these specific analyses were chosen. Multi-level regression, for example, gives a more robust and complete picture.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Nabil Natafgi
Institution and Country: University of Maryland, Baltimore, USA
Please state any competing interests: None declared

Please leave your comments for the authors below

Abstract:
1. First sentence in the abstract: not all 4 global questions are P<0.001 for physical health (recommend hospital is 0.018). Similarly, education & rate experience (p=0.007)...

We assume the reviewer means “First sentence in Results section of abstract”. We have changed this to $p<0.05$.

Strengths and Limitations:

1. Avoid decisive terms such as "conclusively". The study design and methodology may not allow for such definitive conclusions.
2. Basically, you summarized the conclusions or stated implications for practice rather than strengths and limitations. I am fine with this approach, but I am not sure if this is what the journal is asking for. Strength, for example, maybe you inclusion of other satisfaction domains or other predictors such as perception of health not reported earlier in the Canadian context. Weakness would address study limitations.

As noted above, the section on Strengths and Weaknesses has been completely re-written.

Introduction
1. Topbox scoring: (a) Kemp et al (ref 20) seem to have collapsed experience into 10 vs. 0-9. Can you discuss how you came to your choice of 10-9 vs. 0-8? What did other studies do? what is their and your rationale? (b) it seems this may be more of a 'methods' discussion as opposed to 'introduction'.

As far as we are aware, the Kemp study is the only one in which “10” was considered topbox, whereas all of the rest have used 9-10. The latter is the accepted definition utilized by Canadian Institute for Health Information (CIHI) and the Centers for Medicare and Medicaid (CMS). Further, in order to compare our institution’s results with the literature, it is standard practice to define topbox in this manner (e.g. Sacks GD et al, JAMA Surg 2015;150(9):858-64 // Jha AK et al NEJM 2008;359(18):1921-31)

Methods:
1. Was the survey solely administered in English or was it translated to French, as well?

The patient experience survey is administered in the language of choice (French or English).

2. What was the response rate?

The overall response rate for the survey in the time period indicated was 43%. This has been added to the manuscript.

3. Did all CPES-IC surveys match with admin data? If not, what did you do with mismatches or unmatched cases?

All surveys were matched to administrative data and there were no unmatched cases.

4. How did you deal with missing data?

There was essentially no missing data in the administrative database. There were 6/2989 cases (0.2%) in which the discharge disposition was not reported and 1/2989 (0.03%) missing data on

postop complication incidence, marital status and agegroup. For the Patient Mix Adjuster questions, the number of missing entries was <0.1%. Imputation was not used in either case.

5. Your reported Elixhauser measure seems to have produced <0. Can you discuss why is there negative values and what does that mean?

One of the authors (AJF) has previously published on the modified form of the Elixhauser Comorbidity Measure (van Walraven et al, Med Care 2009;47(6):626-33) through which a single score can be derived for each patient. In this study, a series of 30 patient co-morbidities were identified for each patient (e.g. CHF, metastatic cancer). All were tested as univariates for their association for in-hospital mortality. Ultimately, in the final multivariate analysis, many co-morbidities (e.g. obesity and the obesity "paradox") were actually determined to be protective of mortality. To derive the score, integer values were given for each co-morbidity – those associated with increased mortality had a positive score whereas those "protective" had a negative score. As the final score is derived through summation of these co-morbidity scores, there are some patients who would have negative scores. The same calculations would be evident if using the Charlson score (Charlson ME Pompei P Ales KL et al. J Chron Dis. 1987;40:373-383). As discussed below, we have treated the score as a continuous variable.

6. Clarify which tool did you use to identify PSIs using ICD codes... Is it the one by AHRQ? If so, which which PSIs? any? Some are surgical others are medical, did you use the different PSIs for the different admitting departments?

Different PSIs were not used by different departments. We are comfortable with the excellent reference we have provided in the text (Southern DA et al, Med Care 2016) which summarizes the process of identification and its validation. Note that the final author on the current paper (AJF) is one of the authors on this communication.

7. Page 8 of 47 Line 50: I think this should read "There are 7 admitting departments"

Thank you for picking up this oversight.

8. Can you discuss the impact of missing data from one surgical +one medical division on selection bias? How does those 2 divisions compare to others in terms of (1) study outcomes (i.e. rate exp., recommend hosp, etc.) and (2) patient characteristics + perceived health.

Though part of The Ottawa Hospital, due to a different financing envelope from the Ministry, the divisions of cardiology and cardiac surgery at the University of Ottawa Heart Institute have a different data collection system and contract with the survey vendor (NRC Picker). Notably they do not have access to the individual survey data at the patient level such that the data cannot be linked back to the clinical administrative database. Overall results with regards to global domains in patient experience are generally excellent on these services, however no further extrapolations could be determined due to inability to link to physician or service.

9*. Composite domains: did you collapse the questions into the said domains? or were they previously determined by the survey creators and/or published in literature? what is the correlation between some of those domains (e.g. rate experience & rate hospital; or recommend hospital and rate hospital). Someone might argue they are measuring the same thing.*

Most of the composite domains from the Canadian survey are identical to the HCAHPS survey composite domains, the latter of which have been validated extensively in the literature (e.g. communication with nurses, communication with doctors). We would like to clarify that “rate experience” and “rate hospital” are not domains, but represent key global questions, whereas domains represent the average of a series of questions. We agree that global questions may be correlated and may measure the same thing to some degree, however, by convention, all are measured in the survey and all are of interest for varying reasons to each institution.

10. Can you comment briefly on the validity and reliability of the CPES-IC as a tool? was it validated by anyone before?

We would refer the reviewer to the CIHI web site for more information on the CPES-IC. The Canadian survey contains identical questions as the HCAPS survey, as well as additional context-specific questions that are not found in the HCAPS survey (which were not analyzed in this paper). Regardless, the global questions (rate experience, recommend hospital etc.) are identical to those in the HCAPS survey and the latter have been tested thoroughly for validity and reliability in the US. For further information on the Canadian survey, we refer the reviewer to;

https://www.cihi.ca/en/cpes_ic_procedure_20140501_en.pdf

Results:
1. Check reported numbers of respondents and clarify differences between text (2989) and Table 1 (2935).

We thank the reviewer for picking up this discrepancy. The number in the text has been corrected. We have removed patients who were not in any of the four target departments. The table has been corrected.

2. What is the rationale behind excluding maternity, rehab, and mental health admissions? Also, this discussion belongs to the methods, where you discussed exclusion of psychiatry and ophthalmology.

A different survey is used for maternity and mental health. Rehabilitation is a specific department (not under medicine) and is primarily outpatient with <5 admissions. Due to the small number, it was excluded. This statement was moved to the methods.

3. Page 10 of 47 Line 36: " The institution consists of the patients from [...]" may fit better when discussing the setting in methods.

This has been moved as suggested.

4. Page 10 of 47 Line 43: age and discharge disposition are statistically significant

We thank the reviewer for identifying this and this has been corrected.

5. Is there a composite / single patient experience measure for CPES-IC?

There is no composite/single measure for patient experience although many studies have used the answers to the global questions (rate experience, recommend hospital etc.) as singularly important.

6. Page 10 Line 54: be careful with statements that may imply causality (that cannot be established here).

The wording has been changed to reflect “association”.

7. Page 11 Line 26: Figures 1 thru 4 shows measures for rate experience, recommend hospital, rate hospital, and overall helped.

We thank the reviewer for picking up this discrepancy. This has been corrected.

8. For pairwise comparisons (figures 1-4), what are you adjusting for? all the variables presented in the multivariate models?

The reviewer is correct: adjustment was completed using all of the variables in the multivariable model.

9. Page 11 Lines 29-36: confusing. You discuss the unadjusted pairwise comparison btw Surg & Med for rate experience being not significant. And then (this not significant) difference disappear after adjustments.

Although the unadjusted results were not statistically significantly different, the p value was 0.054 and thus we felt it appropriate to indicate that there was a strong trend supporting a difference between the unadjusted results in these two departments.

Discussion:
1. Page 13 Line 26: Even though it is a single payer system, the private nature of providers still does not generate enough competition to attract patients?

The reviewer is correct. The Canadian system is centralized in nature. As an example, for a catchment area of 1.5-2 million people, there is only one very large cardiac surgery unit. For a similar population in the US, there might be 5-8 units, each competing for patients.

2.There are no federal or provincial provisions (or anticipated provisions) for linking payment to some performance measures (including pt experiences)?

At this time, there are no provisions for hospital reimbursement based upon performance measures other than waiting list numbers and wait duration. However, we are aware that the respective ministries are very interested in implementing a linkage in the near future.

3. Page 14 Line 20: Your findings showed that admit urgent and LOS were not significant predictors (as opposed to Ref 20). What do you think are the reasons? Is there a difference in the case-mix of the studies? Or is this because you adjusted for some additional variables? Did you consider running the same models Kemp did, and see if you get comparable results to them, in your dataset? If so, then you may want to know which covariates in multivariate models removed the significance. Is it the physical and mental health? This would be an interesting finding, I think.

As indicated in the text, the Kemp survey did not correct for the two factors that we consistently found to be the strongest covariates (patient-perceived physical and mental health status) therefore we disagree with their methodology. Further, this is the only manuscript we have found where topbox was solely indicated by a score of 10, as opposed to 9-10. Therefore we would not be able to compare to their results if we ran the same analysis.

4. Page 15 Line 25: "The analysis highlights the differences in adjusted and unadjusted [...]"/ Physical and mental health are not demographics.

We have added the words “and other” after the word “demographic”.

5. Page 15 Line 31: Elixclass (comorbidity) was not stat. significant in your multivariate models. Although it is very surprising finding, I am not sure you can come to your conclusions here with that finding. Again I am surprised about the Elixclass not being significant and I am not sure if it is related to the way you modified that variable. Also, what was your rationale for the classification you used (i.e. <0, 0, 1-5, >13). I think this argument needs to be based on theoretical/clinical grounds supported by stats from your database. The 19% of people having 6+ commodities seem a little high and primarily coming from medicine.

We thank the reviewer for these comments. In this revision, we have treated the Elixscore as a continuous variable. Further, due to the skewness of the distribution, the Elixscore was log-transformed. This has not significantly changed the results in the analysis.

Of note, an Elixscore of “6” does not indicate 6+ comorbidities, but rather that the sum of the weight of comorbidities is 6. For example, metastatic solid tumor as an isolated comorbidity has an Elixhauser value of +12.

6. I agree that physical and mental health as predictors of experience is an important finding. Also reinforcing the importance of factoring differences between case-mix among diff depts is important. I also agree with your statement that this recognition should enhances engagement of staff facing challenges of pts with chronic conditions. Yet, I think this shouldn't give depts caring for pts with chronic conditions a pass on lower scores for patient experience, simply because of the case-mix they serve. Rather, leadership should provide more resources and / or opportunities for those depts to address this challenge and reach scores comparable to other departments, albeit the differences in case-mix.

We agree with the reviewer’s comments on this issue.

7. Page 16: It is not clear which ones are the patient care domains (I am assuming the one with doctors communication, nurses communication, etc.). If so, I am not sure I necessarily understand why they were excluded. I see that they can be correlated with other predictors, but I do not think in a way more than some of your current predictors correlate with each other, anyhow. So maybe look at the correlations quantitatively and present those numbers (e.g. use VIF or other appropriate statistic). Another way looking at could be treating them as a separate set of analyses where they are the outcomes predicted by the same predictors you currently have in your presented models.

We agree with the reviewer, however due to manuscript space, we had to limit the analysis to the global questions. The objective of the paper was to establish that global questions cannot be interpreted without adjustment with covariates reflecting patient perceptions and demographics. We believe that it is very likely that “adjustment” is also necessary in the patient care domain questions.

Tables

1. Table 1: Check the "n" in the first row for Total and different departments. For example mental health adds up to 2938; also the four depts adds up to 2,896

The first row has been corrected. In some situations, there were missing data in PMA (<1%)and thus to totals do not always add up. A comment to this effect has been added to the table.

2. Did you consider collapsing some of the variables, especially in cells where you have too few n (e.g. education, age, elixclass)?

3. Are the Race categories mutually exclusive?

As indicated above, we have re-analyzed the data using the Elixhauser score as a log-transformed continuous variable. Age is already significant in all of the questions, whereas education is significant in three of four questions, so it is not clear what the benefit would be to use a different means of classification. Finally education is a self-reported item on the survey so it may not be accurate to

collapse some of the answers. However, we have already collapsed some of the age items based on geographic localization (e.g. Filipino – Southeast Asian – Korean – Japanese = Oriental)

4. In the Tables legends you have ALC, I did not see that in the tables

We thank the reviewer for picking up this oversight. This has been removed.

5. Table 1: "ED isit" should read "ED visit"

This has been corrected

6. It would be helpful to define the categories in the methods. What does home-setting include? what does another health facility include? Were there any death cases?

Patients who died prior to the patient's discharge were excluded. The definition of discharge disposition categories has now been included in the methods.

"Discharge disposition was divided into three categories: 1) Discharged to the patient's home without support services 2) Discharged home or to a home-setting with support services (e.g. senior's lodge, attendant care, home care, meals on wheels, homemaking etc.) 3) Discharged to another health care facility (e.g. continuing care, acute care inpatient) or other (palliative care/hospice, addiction treatment etc.)"

7. How did you code LOS for the multivariable analyses? Did you leave it as a count variable? If so what does the (>3 days) mean? Or did you change it to a binary variable (>3 days vs. < 3 days)

LOS was a binary variable (< or > 3 days).

8. Did you consider looking at the ED visit within 7 days as an outcome predicted by patient experience? With that, I am not sure if you should include ED visit w/in 7 days as a predictor of patient experience (even if it wasn't statistically significant). Same argument may apply for discharge disposition.

Both of these covariates were tested and the results are indicated in Tables 2-5. ED visit within 7 days was not significant in any of the four questions. Discharge disposition was significant for "recommend this hospital" and "rate this hospital". We did not test ED visit within 7 days as an outcome; though this is an interesting question, it wouldn't be related to our primary objectives. We do still believe that there are strong arguments to support that having to come back to hospital within 7 days of discharge, would "colour" one's perspective on the hospital stay and the experience, as the survey is done several weeks after discharge. We also strongly feel that discharge disposition would impact the patient experience.

9. Be consistent in the ordering of the variables in the tables for ease of comparison across Tables. Also, you may want to merge Tables 2-4 in a single table as they have the same predictors.

We have tried to merge these tables but we found the product unwieldy and difficult to understand. We are open to changing to this recommendation at the editor's request.

10. Table 3: Physical Health - Poor - remove the third figure after the point

This has been corrected.

11. Since you chose topbox scoring for experiences, did you consider similar scoring for self-rated health... i.e. coding excellent vs. other? or excellent + V. good vs. others?

We did not consider this. As the variable is already significant, it is not clear how this would impact the results.

12. I am not necessarily recommending doing so, but did you consider collapsing education and / or age categories? or using age as continuous variable?

This was addressed above.

13. Can you include the results of the univariate analyses in comparison to the multivariate comparison to better see the effect of adjustments

An additional column has been added to provide this information.

Figures
1. Although it's results might be a little surprising, I like figure 5 as it may have some implications for practice and you do mention that in your discussion. Can you just clarify which variables entered in this analysis? Is does this adjust for some of the demographics you were adjusting for earlier?

The domains (e.g. doctor communication, nurse communication) were adjusted for all of the same covariates as the global questions. As the reviewer has perceived, the driver analysis provides the clinician the opportunity to identify areas that have high yield of focus to improve the results on a global measure such as overall experience.

Other comments
1. Is it possible to include a figure that shows the distribution of your dependent variable (patient experiences) prior to switching it to a binary variable. You may chose to not eventually include that in

the final paper, but it is worth looking at to understand the raw distribution.

We agree that this would be of interest, but we do not feel it would necessarily add to the study and due to length constraints we have not included this.

Reviewer: 2
Reviewer Name: Rachel Foskett-Tharby
Institution and Country: University of Birmingham, UK, NHS England, UK
Please state any competing interests: Senior Policy Lead at NHS England with a focus upon the use of financial incentives in healthcare.

Please leave your comments for the authors below

This was an interesting, well drafted paper. A couple of minor points:
Page 5 lines 47-54 and page 6 lines 3-24: I found this section a little confusing with the classification of the additional questions into groups and then domains. I wonder if this should be revised to make it clearer when you are describing the additional questions in the Canadian survey.

We have made revisions to this section to render it more readable.

Page 8: methods - it would be useful to have a little more information about the Elixhauser comorbidity measure and why this was chosen rather than any other one. The reference to this (line 40) needs updating.

This section has been revised. The Elixhauser score was utilized as one of the authors (AJF) was the initial investigator in the validation of this score and the administrative data was established to collect the 30 binary covariates used in its derivation. It is anticipated that use of the Carlson score would give identical results but the administrative database is not designed to determine this and this would require a major restructuring.

Reviewer: 3
Reviewer Name: Vincent S. Staggs, PhD
Institution and Country: Children's Mercy Kansas City, University of Missouri--Kansas City, Kansas City, MO, USA
Please state any competing interests: None declared

Please leave your comments for the authors below

Thank you for the opportunity to review this work. I found the paper well-written overall, and the study seems thoughtful and well done. As a biostatistician I'll focus on the statistical aspects. I hope the authors find these comments helpful.
1. Readers may be interested in seeing the correlation matrix, or at least the correlations between the dependent measures and the quantitative explanatory variables. No need to include p-values with the correlations (which would exacerbate the multiple testing issue).

We thank the reviewer for the suggestion, however we feel that including the correlation matrix in the manuscript would make the paper too long for most readers. Regardless, if the editor feels that it is important we would be willing to submit it as an online appendix.

2a. Clustering. With only two campuses I would simply include campus as a fixed (not random) effect; I'm not sure that two campuses from a single hospital can provide a meaningful estimate of between-hospital variance in the population of hospitals, and a small number of clusters can be problematic in mixed models (see Kenward & Roger, 1997).

We have re-designed the analysis to include campus as a fixed effect.

2b. The more important source of clustering to consider is unit/ward. It sounds like patients came from eight units total, and I would recommend including in each model a random intercept for unit (in addition to the categorical fixed effect for department type) and, given the small number of clusters, using the Kenward-Roger degree of freedom method.

There are actually 30 different units in the two hospitals. Further, it is not uncommon for patients to be moved between units – for example from the ICU to the ward, or from an acute care ward to a chronic ward. The dataset only provides the ward of discharge.

3. Did the authors examine multicollinearity among the explanatory variables in the logistic regression models (e.g., VIFs)? It seems likely that some variables are correlated (e.g., ICU with LOS and patient safety event, age with physical health) so it seems prudent to check this.

The reviewer raises and interesting question regarding multicollinearity in logistic regression modelling. This remains a controversial topic and I refer the reviewer to a recent discourse in Statalist with comments from one of the tenured members.

https://www.statalist.org/forums/forum/general-stata-discussion/general/1398913-multicollinearity-in-binary-logistic-regression

On review of the standard errors, they do not appear to be excessive.

4. Multiple testing. The authors note using the Bonferroni adjustment for pairwise comparisons, presumably for the department comparisons shown in the Figures. But with roughly 200 hypothesis tests carried out, multiple testing remains something of a concern. The Bonferroni method is conservative and may severely reduce power if applied across all tests (although the large sample size will help). One option to reduce the number of tests would be combining a couple of dependent measures if they're highly correlated (as I would guess some are), or just choosing one measure from those highly correlated. Another would be treating physical health, mental health, Elixclass score, and age as quantitative (not categorical) variables. It would also help (a little) to collapse across some categories for education and/or race. The authors also might consider the Benjamini-Hochberg False Discovery Rate (FDR) method, which is a nice, less conservative, alternative to Bonferroni adjustment.

We thank the reviewer for thoughtfully pointing out that we weren't entirely clear on how we adjusted for multiple testing. As indicated above, Elixclass has been converted to a continuous variable which would somewhat decrease the reviewer's concerns. Further, we used a modified more conservative multiple comparison test in Stata (mcompare(*method* adjustall)).

5a. For the analyses shown in the Figures, please explain under Statistical Analyses where the unadjusted and adjusted predicted percent topbox numbers come from, and how the pairwise tests were carried out with these variables.
5b. I'm not sure how much the figures themselves add; perhaps it would suffice to show the effects of adjustment by reporting these numbers in a table.

The following was added to the Statistical Analysis section:

"Predicted marginal means after modelling with and without adjustment using the covariates were plotted with 95% CI."

6. In describing the logistic regression analyses (P10) the authors mention only likelihood ratio tests. However, the LRT p-values do not generally match the OR p-values in the tables (e.g., in Table 3 the LRT p-value is very different from the OR p-value for Any PSI), so I assume the latter must be Wald or profile likelihood test p-values. In any case, it would be helpful to clarify the description of tests in the Statistical Analyses section.

The following was changed in the Statistical Analyses section:

"The association of each covariate was assessed using likelihood-ratio chi square testing."

7. In reporting results of the logistic regression models I would include the C-statistic/AUC for each. Also, please label the second column in the regression tables.

The AUC for each of the four global questions has been added to the legends of each table. We are not clear on the second request and this may be due to confusion on the formatting of the table – the table length is greater than one page and thus the columns are not labelled on the second page.

8. *On a minor note, there's no need to log-transform income if converting to deciles. Log-transformed income values will fall in exactly the same deciles as the untransformed values, as assignment to deciles is determined by ranks and log is a strictly monotone (and thus rank-preserving) transformation.*

We have repeated the analysis without log-transforming the data. We thank the reviewer for this suggestion. Of note, there were no differences in the outcomes.

9. On P10, first sentence, something like this might be clearer: "After dichotomizing each of the four overall care questions [(a) … (b) … (c) … (d) …] based on the 'topbox' response criteria defined

above, we fit a separate logistic regression model for each question to model the odds of topbox response as a function of [explanatory variables]."

We have re-written this as suggested by the reviewer.

10. The authors may have addressed this, but I'm curious how they classified patients who were admitted in one unit but ended up in another. Or perhaps this was rare.

This was not rare and we do not have this data, thus there is no way that we could determine this. This was one of the reasons we could not use unit as either a fixed or random covariate.

11. It would be helpful to describe the Key Driver Analysis in more detail under Statistical Anaylses. On first encountering "vertical separation of the quadrants" I didn't follow.

We have re-written this section.

12. The non-linear effects of age (with highest ratings in the middle and lower ratings for youngest and oldest patients) may be worth some discussion.

We have added a comment to address this in the discussion

13. I found the writing good overall but would recommend having someone unfamiliar with the study read and edit the paper. There are places where additional clarity would be beneficial.

We thank the reviewer for their constructive comments and we have substantially edited the paper for readability.

Reviewer: 4
Reviewer Name: Hilde Hestad Iversen
Institution and Country: Norwegian Institute of Public Health, Norway
The aim of the current study is not clear to me. Is it to be able to compare different units or departments at the same hospital or to compare hospitals? Also the reason why potential adjustments is an important area of interest when measuring patient experiences and the underlying rationale behind this should be explained more thoroughly. ….It is not obvious how the current results should be used in the follow-up work at the hospital. On page 7 we are told that the overall objective was to compare the value of the self-reported background variables with covariates from a hospital database, in the development of a statistical model to predict topbox scoring in the four survey questions related to overall care. Please elaborate more on this aim, is the aim to explore or compare?

We have enhanced the introduction and discussion to better rationalize the goal of the manuscript.

The study focus mostly on global ratings, except for the key driver analysis, often skewed towards the positive end of the scale and showing less variation than specific experience items. Any reflections on this subject?

We agree that we did focus on global ratings as opposed to specific composite domains such as doctor communication. In particular, hospitals are driven to essentially compete on these global questions (e.g. overall experience) and they are far more important from a corporate point of view. There is no reason however to believe that the rationale and methodologies used to adjust for these answers using covariates we have identified, would be the same if one was measuring and comparing the composite domain questions such as communication with nurses. Due to space constraints, we elected to focus on the global questions.

In the introduction we are given information on how the Canadian Patient Experience Survey – Inpatient Care (CPES-IC) differ from the HCAHPS in the US. Some if this information, especially the elaboration on specific items but also on top-box ratings, should be included in the methods section.

As described above, some of this has been elaborated upon.

The modification of the HCAHPS survey was developed through a collaboration between different parties, but how many items were changed? Only the items measuring the new topics, or some of the others as well? The authors mention in the discussion section that the newly developed questions added should have been testet, this is very important to explore the validity and reliability of the items. Have the other domains/questions from the HCAHPS been validated in in the Canadian context?

These are very good questions. We have elaborated more (as described above) on the differences between HCAHPS. We are not aware of previous validation of the HCAHPS in the Canadian context and we have confirmed this with the key researchers at CIHI (personal communication).

We are told that the patient experience survey are routinely administered in four provinces in Canada, however, that there is limited familiarity in the assessment of patient experience and the use of such surveys. How are these results reported and responded to today? Which institutions carry out or are responsible for the surveys? Are the current results compared? Are they case-mix adjusted and weighted, or is this work the first effort to explore a possible case-mix model? Again, the aim here might just be to compare departments at the current hospital?

Although the same survey is used nationally, different vendors distribute the survey and analyze the data in each province. Results are given back to the Institute as well as reported centrally to CIHI. At this point, no inter-institutional comparisons are done. Our report supports the rationale that should this happen, one must correct for case mix. We believe our results support that adjustment is essential whether comparing institutions, departments or divisions.

The abstract tells us that the participants of the study were randomly sampled post-discharge. I did not find more information on the administration of the survey in the method section, either on the number of patients that were invited to participate or the response rate. How did the subjects respond, electronically or on paper? How were they contacted? Where did they fill out the questionnaire, at

home or at the hospital? Were the hospital responsible for the survey? We are not informed about the response rate or given any information on non-response. Do the authors have any reflections regarding the representativeness for inpatients in general?

We have added details regarding the survey administration and the general response rate. The vendor (in this case NRC Health, Toronto, Canada) distributes the survey. For the majority of hospitals, they complete the analysis whereas at The Ottawa Hospital, the raw data is returned to the Data Warehouse where linkage with the administrative database is completed. We have an upcoming publication (Rubens Chen Ramsay Forster Wells Sundaresan. The Development of a positive deviancy strategy to identify excellence in patient experience. European Journal for Person Centered Healthcare, 2018, in press) looking at the differences between responders and non-responders to patient experience surveys from patients discharged from a Department of Surgery. Non-responders were more likely to be single and young with lower income and a lower comorbidity status. They were also more likely to have been admitted emergently, with longer length of stays and less likely to be discharged home. As we have now indicated in the limitations, there is some evidence that non-responders may be slightly different from responders affecting the generalization of the results.

I am a bit uncertain on the choice of analyses because I did not fully comprehend the aim of the study (see my previous comment). Please explain more thoroughly why these specific analyses were chosen. Multi-level regression, for example, gives a more robust and complete picture.

As discussed above, we believe we have enhanced the description of the objectives to better clarify these points. We thank the reviewer for emphasizing this important point.

FORMATTING AMENDMENTS (if any)
Required amendments will be listed here; please include these changes in your revised version:
1.Supplementary file citation
- We have noticed that you have uploaded the file "CPES-IC Survey" under 'supplementary file'. However, we can't see any citation for this file within the main text. If this file needs to be published as supplementary file, please cite it as 'supplementary file' in the main text. Otherwise, please change its file designation to 'Supplementary file for editors only'.

We have uploaded the full file as a supplement to the manuscript

VERSION 2 – REVIEW

REVIEWER	Nabil Natafgi University of Maryland School of Pharmacy, USA
REVIEW RETURNED	17-May-2018

GENERAL COMMENTS	Thank you for addressing the comments raised by all reviewers. I think the paper is shaping well and is clearer. Please find below a few additional thoughts for your consideration: 1. I think I am still confused with the total (n) numbers reported, particularly in Table 1. For instance, in the revised version you

	indicate that 2896 patients responded to the survey. Looking at Table 1, you report that same number (2896) in the total column/row. However, the sum of some variables are still more than the total number of respondents. For example, mental health adds up to 2898. Admit adds up to 2948 and age adds up to 2947. I can see how some variables can add up to less than 2896 (for missing) but not the other way around. Am I missing something here? 2. For figures 1 thru 4, I think it will be helpful to add the following to the footnote: "adjustment was completed using all pf the variables in the multivariable model". 3. Indicate in methods that patients who died prior to discharge were excluded from analyses. 4. I feel adding distribution of each of the 4 global scores may help the reader get a sense of "generosity" of patients in rating inpatient care in the study context. You can include this as an online-only supplementary. Alternatively, you can add basic descriptives of for each of the four global measures to Table 1 (as binary 9-10 vs. 1-8). 5. In the "results" section, "Campus site was found to be a factor as a random effect in rate hospital"; I think this should now read fixed effect? Also, now it is significant for rate experience. 6. In the discussion, towards the end you mention "few of the covariates from the administrative database were significant in models describing perceptions of excellence in individual questions of overall care (length of stay, ICU stay, marital status)". Looking at your tables, it looks like LOS and ICU was not significant for any of the four global measures and is Marital status is only significant for one of them. Finally, I would suggest to revise your discussion section to elaborate a little more on some tangible implications and recommendations for institutions like yours or others with similar context/setting.

REVIEWER	Vincent Staggs Children's Mercy Kansas City University of Missouri-Kansas City
REVIEW RETURNED	16-May-2018

GENERAL COMMENTS	The authors have thoughtfully addressed my concerns. My only additional comments are (1) a bit more information about the Stata modeling in the manuscript would be needed for the study to be fully repeated, (2) the section describing the departments excluded and included could be a bit clearer, and (3) inclusion of site in the model could be mentioned in the text itself. These are very easy edits not requiring further review by me.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Nabil Natafgi
Institution and Country: University of Maryland School of Pharmacy, USA
Please state any competing interests: None declared

1. I think I am still confused with the total (n) numbers reported, particularly in Table 1. For instance, in the revised version you indicate that 2896 patients responded to the survey. Looking at Table 1, you report that same number (2896) in the total column/row. However, the sum of some variables are still

more than the total number of respondents. For example, mental health adds up to 2898. Admit adds up to 2948 and age adds up to 2947. I can see how some variables can add up to less than 2896 (for missing) but not the other way around. Am I missing something here?

Thank you to the reviewer for recognizing this. We have incorrectly reported the total. We derived it by adding the numbers for Surgery (1699), Medicine (1023), Family Medicine (79) and Obs/Gyn (95). However we reviewed the original dataset and there are n=2989 eligible observations.

Differences in the numbers for each group (e.g. mental health at 2898) reflect missing data for that particular measure/survey question.

2. For figures 1 thru 4, I think it will be helpful to add the following to the footnote: "adjustment was completed using all pf the variables in the multivariable model".

This has been added as suggested.

3. Indicate in methods that patients who died prior to discharge were excluded from analyses.

This has been added as suggested.

4. I feel adding distribution of each of the 4 global scores may help the reader get a sense of "generosity" of patients in rating inpatient care in the study context. You can include this as an online-only supplementary. Alternatively, you can add basic descriptives of for each of the four global measures to Table 1 (as binary 9-10 vs. 1-8).

We have added this to Table I.

5. In the "results" section, "Campus site was found to be a factor as a random effect in rate hospital"; I think this should now read fixed effect? Also, now it is significant for rate experience.

As suggested, we have corrected this wording regarding random and fixed effects. However we note that in Table 2, the p value for Campus was 0.332, thus it is not significant.

6. In the discussion, towards the end you mention "few of the covariates from the administrative database were significant in models describing perceptions of excellence in individual questions of overall care (length of stay, ICU stay, marital status)". Looking at your tables, it looks like LOS and ICU was not significant for any of the

four global measures and is Marital status is only significant for one of them.

We are sorry we were not more clear. The items in parentheses were reflecting covariates in the administrative data that were NOT significant. We have therefore removed the parentheses.

Finally, I would suggest to revise your discussion section to elaborate a little more on some tangible implications and recommendations for institutions like yours or others with similar context/setting.

We have added the following to the final paragraph of the discussion:

Health care institutions must incorporate patient demographics and self-reported aspects of perceived health into the analysis of patient experience data to properly interpret and compare this information particularly when comparing departments and units within the institution.

Reviewer: 3
Reviewer Name: Vincent Staggs
Institution and Country: Children's Mercy Kansas City
University of Missouri-Kansas City
Please state any competing interests: None declared

Please leave your comments for the authors below

The authors have thoughtfully addressed my concerns. My only additional comments are (1) a bit more information about the Stata modeling in the manuscript would be needed for the study to be fully repeated, (2) the section describing the departments excluded and included could be a bit clearer, and (3) inclusion of site in the model could be mentioned in the text itself. These are very easy edits not requiring further review by me.

We have modified the methods to allow for better reproducibility of the analysis in Stata. We have edited the wording on the departments included and excluded. We did not add a line indicating that campus is now being analyzed as a fixed effect as it is evident from the Tables that it is being treated as such. The reviewers were provided with earlier iterations of the text in which it was treated as a random effect, however all references to this have been removed and thus we believe that it will be perceived as any other fixed effect in the interpretation of the analysis by the reader.

VERSION 3 – REVIEW

REVIEWER	Nabil Natafgi University of Maryland, Baltimore, USA
REVIEW RETURNED	16-Jul-2018

GENERAL COMMENTS	None - my previous comments were addressed.