PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Counterintuitive Results From Observational Data: A Case Study and Discussion
AUTHORS	Doty, Erik; Stone, David; McCague, Ned; Celi, Leo Anthony

VERSION 1 - REVIEW

REVIEWER	Peter Smith University of Southampton, UK
REVIEW RETURNED	03-Dec-2018

GENERAL COMMENTS	This paper presents a framework to validate counterintuitive results in retrospective studies, and a study of the relationship between perceived pain in the ICU and three outcomes (30-day mortality, 1-year mortality and hospital length of stay), where counterintuitive results were obtained. However, the title only refers to the former, while most of the abstract and the first part of the paper focus on the latter. Assuming the framework is meant to be the main contribution of the paper (and I think this would be a valuable contribution), then I recommend that the paper is substantially rewritten with the development of the framework first and the study used as a case study to demonstrate the application of the framework. Such a paper (and its abstract) would not follow the format used to report the results of a study, and it is an editorial decision whether such a paper would be suitable for publication on BMJ Open. I would also consider adding a second case study where the framework helps identify problems with the study rather than concluding that the counterintuitive results are correct findings. Producing a comprehensive version of Figure 1 should be a key part in the framework. Therefore, for your study more, details of the other (1917 – 844 – 23 = 1050) patients excluded from the final cohort should be provided in this figure. You include non-CABG surgery in the exclusion criteria. Does this mean that you excluded patients who had both CABG and non-CABG surgery, and therefore they are some of the 1050 patients? Also, I would include in Figure 1 a box with the 1889 patients used in the sensitivity analysis. “Unconditional logistic regression with Fisher’s optimization” is SAS-speak. As far as I can tell, you are fitting logistic regression models using maximum likelihood estimation. If this is not the

	case, then you need to provide a reference and justification for using an alternative method. Also, did you consider modelling the log of length of stay, and did you use the residuals from the linear regression model to check the normality assumption? On page 6, you refer to ordinal regression, but all you are doing here is discretising the risk factor, not changing the dependent variable to be ordinal. The description of the method and results of the falsification hypothesis testing is unclear. Please confirm that nausea was regressed on pain and, after controlling for the potential confounding variables, pain was non-significant, and that this was the case for all four measures of pain. One source error, which I am not sure you have fully investigated for the study (or which category of Table 3 it falls in), is selection/nonresponse bias. You state on page 9 that 'Sampling bias was also minimal'. However, you excluded all patients with missing data on the confounding variables and those that had no pain scores recorded. I suggest you consider including multiple imputation of the missing values of the confounding variables and/or the missing pain scores, as well coding all patients with no pain score as having no pain, in the sensitivity analyses. Note there is a contradiction between you exclusion criteria ("missing data on confounding variable(s)") and the Strengths ("Large sample size with complete covariate data").

REVIEWER	Martine J. Barons University of Warwick, UK
REVIEW RETURNED	03-Dec-2018

GENERAL COMMENTS	For BMJO: It is harder to review papers sent with tables and figures separated from the relevant text and from their captions. Analyzing Counterintuitive Data This paper seeks for uncover the relationship between outcomes of length of hospital stay, 30-day mortality and 1-year mortality and patient-reported pain post CABG surgery for a subset of relatively healthy patients who were able to be and were extubated within 24 hours after postoperative ICU admission. Since this modelling produces a counterintuitive result, the authors also offer some thoughts on how to check for errors in the data and contextual checks which might alert researchers to when the counterintuitive nature of the modelling result might be due to errors. Page 2 line 25 please clarify what you mean by confounding variables Page 2 line 33, please add confidence intervals for the LOS Page 4 line 25, please list the variables considered to be confounders Page 4 line 32, Figure 1 needs to be improved so that losses at every stage are reported; currently you have 1050 patients excluded for a non-specified reason. Page 5 line 3 you mention later the potential unreliability of the self-reported scores, but not how this might affect results.

	Page 5 line 11 were they any differences observed between reported pain scores for patients prescribed different medications? Page 5 line 35, please give details of the e-score as for the OASIS P Page 6 line 45, what was the mean (DS) frequency of pain assessment? Page 7 line 32 @The results were similar…' please expand. Page 8 line 48 Emphasis that you do not claim the subjective experience of pain causes positive outcomes and therefore these results should not be misunderstood as a recommendation to reduce pain relief. Page 9 line 28 Oral analgesic Page 10 line 16 Is the range of pain reported similar to that reported in larger cohorts or other conditions, such as the ones you cite? Page 11 use 'counterintuitive finding' throughout Page 18-19 Please provide a separate table for each model, which includes the OR / p-values as appropriate for all the variables in the model so their relative importance can be assessed. Page 18 T1 please label what is in brackets; in the case of gender and in hospital I assume this is (%). Total what? Narcotics during stay? Please clarify Page 20 T3 and Page 21 T4: these are very specific to the specific case of CABG in places (e.g. lab error) and should be more general. I would prefer as 3-column table telling me the name of the possible problem, what to check for, what to do to fix it. The important role of STROBE, CONSORT and other checklists in promoting reproducibility should be included in the discussion of counterintuitive findings. The discussion section would benefit from more headings and subheadings. I would suggest that it is not the data that was counterintuitive, but the results of the data modelling. No mention is made of the possible difficulties that may arise from violation of the modelling assumptions, not how these were checked in your analysis of the CABG case.

REVIEWER	Hamada Ahmed Hamada Faculty o physical therapy, Cairo University
REVIEW RETURNED	04-Dec-2018

GENERAL COMMENTS	Thanks for the opportunity to review this Retrospective study. title: add study design in your title. Abstract: 1.Please define abbreviations beforehand and avoid them in the abstract. Results: 1. add statistical analysis section. 2. need more data in logestic and multiple regression such as (which type of logestic were used (ordinal logestic or binominal logestic), where is the χ2 and R2 values?, where is B coefficients?. general results needs more details and reorganized.

VERSION 1 – AUTHOR RESPONSE

In response to reviewer 1, Dr. Peter Smith:

This paper presents a framework to validate counterintuitive results in retrospective studies, and a study of the relationship between perceived pain in the ICU and three outcomes (30-day mortality, 1-year mortality and hospital length of stay), where counterintuitive results were obtained. However, the title only refers to the former, while most of the abstract and the first part of the paper focus on the latter. Assuming the framework is meant to be the main contribution of the paper (and I think this would be a valuable contribution), then I recommend that the paper is substantially rewritten with the development of the framework first and the study used as a case study to demonstrate the application of the framework. Such a paper (and its abstract) would not follow the format used to report the results of a study, and it is an editorial decision whether such a paper would be suitable for publication on BMJ Open. I would also consider adding a second case study where the framework helps identify problems with the study rather than concluding that the counterintuitive results are correct findings.

We appreciate and understand Dr. Smith's point in which there appears to be a disconnect between our title, the abstract and first part of the paper. Rather than completely restructure the paper, we believe that a simper and perhaps more effective solution to the reviewer's point can be achieved by first changing the title to "Counterintuitive results from big data: A case study and discussion" so that while the reader will understand the overall concept of the paper, they will also know to expect the sequence of the CABG pain case study first, and the discussion of counterintuitive data later. We will adjust the abstract so that it is better balanced in this regard . Thank you for pointing out this issue- your input makes this a better paper.

We also feel that we have raised similar issues in the discussion, such as the research on lower hemoglobin transfusion triggers, so that another new case is unnecessary for us to present. It was those paradoxical CABG pain-outcome findings that made us think more deeply about the issue of counterintuitive data, making us realize that it will become an increasingly important issue as more and more clinical data is analyzed and utilized in this newly digitized medical world of progressively bigger data. We hope to raise clinicians' awareness about this as a possible issue as well as to present a possible framework for them to approach such findings.

The description of the method and results of the falsification hypothesis testing is unclear. Please confirm that nausea was regressed on pain and, after controlling for the potential confounding variables, pain was non-significant, and that this was the case for all four measures of pain.

We reworded the description in order to be more clear. The methodology for falsification hypothesis testing was published in JAMA by Prasad and Jena, and is listed as reference #14. The general concept is to test a hypothesis that is known to be unlikely to influence the outcomes, in our case nausea, against the outcomes of interest, in our case mortality and length of stay. We also tested delirium, which is known to have effects on our outcomes of interest. The results, in our case negative for nausea and positive for delirium, are designed to add validity to our observed outcomes with regards to pain levels, and support that they are less likely to be due to chance as the cohort follows other known facts.

Thank you for the suggestion, we hope this response clears up any confusion in this regard.

One source error, which I am not sure you have fully investigated for the study (or which category of Table 3 it falls in), is selection/nonresponse bias. You state on page 9 that 'Sampling bias was also minimal'. However, you excluded all patients with missing data on the confounding variables and those that had no pain scores recorded. I suggest you consider including multiple imputation of the missing values of the confounding variables and/or the missing pain scores, as well coding all patients with no pain score as having no pain, in the sensitivity analyses. Note there is a contradiction between

you exclusion criteria (“missing data on confounding variable(s)”) and the Strengths (“Large sample size with complete covariate data”).

While there was some missing data, it was minimal: Of the 1917 CABG patients, only 28 subject who had missing data. We have updated figure 1 to better reflect this. While we agree that including multiple imputations would have been beneficial, we did not believe it to be necessary as the number with missing data was so relatively small. We have also clarified the contradiction you mentioned.

In Response to reviewer 2, Dr. Martine J. Barons

Page 2 line 25 please clarify what you mean by confounding variables

Page 4 line 25, please list the variables considered to be confounders

We have defined our confounders in the covariate subheading of the methods section. We did refer to confounders several times prior to defining them. However due to word count constraints, particularly in the abstract, we are unable to additionally define them outside of the covariate section

Page 18-19 Please provide a separate table for each model, which includes the OR / p-values as appropriate for all the variables in the model so their relative importance can be assessed.

While we also understand the value of having the complete picture, we didn’t want to overcrowd the paper an excessive number of tables; in fact, including separate tables for each model would yield a minimum of 24 different tables in this paper. We selected what we believed to be the most appropriate information to convey our message. We believe the most effective solution would be to include the complete statistical analysis output as an online supplemental file to this paper. We believe this will help maximize the readability of our paper, but also allow readers access to additional statistical data if they desire it.

In response to reviewer 3, Dr. Hamada Ahmed Hamada

1.Please define abbreviations beforehand and avoid them in the abstract.

Results:

1. add statistical analysis section.

2. need more data in logestic and multiple regression such as (which type of logestic were used (ordinal logestic or binominal logestic), where is the $\chi 2$ and R2 values?, where is B coefficients?.

general results needs more details and reorganized.

Thank you for reviewing our paper and for your suggestions. We have revised the abstract accordingly. Additionally, we do have a statistical analysis section in our methods. We have retitled this to “Statistical Analysis” from “Analysis” for improved clarity.

As for the data from the regression models: We tried to include the most pertinent data to avoid overwhelming readers with an excessive amount of number and increase readability. Due to the fact that we have a plethora of statistical data from the various models, we thought it would be more efficient to include the entire output from all our models as an online supplemental file.

VERSION 2 – REVIEW

REVIEWER	Peter Smith University of Southampton, UK
REVIEW RETURNED	30-Jan-2019

GENERAL COMMENTS	By addressing most of the reviewers' comments, including mine, you have improved the paper. However, I still believe there is further work to be done before it would be acceptable for publication. While I still think that my structure would increase the readership of the paper, I am now happy that the research questions and study objectives are clearly defined, and the abstract is accurate, balanced and complete. However, I do not think the title is appropriate: your case study does not use 'big data' and as an alternative I suggest 'Counterintuitive Results from Observational Data: A Case Study and Discussion'. Figure 1 is much more informative now. I was pleased to learn that only 28 patients had missing data. However, it would be interesting to know how many of these were extubated within 24 hours and therefore were excluded from the primary analysis. I would also mention somewhere in the main paper that 22 patients died during hospitalization, and possibly add this information to Figure 1. The supplementary material is a great addition to the paper, but I hope it will be made more readable if the paper is accepted for publication. Furthermore, I still think you should consider whether or not it might be more appropriate to model the log of length of stay. Your understanding of a falsification hypothesis testing is odds with mine. Furthermore, I could not find the statement '[a] distinct and highly unlikely hypothesis is tested against the outcomes of interest' in Prasad and Jena (2013). My understanding is that you take an outcome that is highly unlikely to be related to the 'intervention of interest', in your case pain, and test whether or not is it. This is in agreement with the definition in Prasad and Jena (2013): 'A falsification hypothesis is a claim, distinct from the one being tested, that researchers believe is highly unlikely to be causally related to the intervention in question.' This is then followed with a 'for instance' concerning PPI use and pneumonia, where tests for associations between PPI use and soft tissue infraction and myocardial infarction are proposed. Therefore, in your case study, you need to find outcomes that should not be causally related to pain. Given that I hope you will be able to address my concerns above, below are a few minor comments and suggestions which I think would improve the paper. Abstract Objectives: I would retain 'To explore …'. Design: 'The case study is a …' Conclusion: '… incomplete or has not been firmly based on empirical evidence'.

	Table 1: replace ‘oasis’ with ‘OASIS’, ‘e_score’ with ‘E_score’ (twice) and ‘elixhauser’ with ‘Elixhauser’. Page 10, line 11 (in pdf): ‘Case Study’. Page 11, line 5: ‘Also, patients …’. Page 11, line 47: ‘… confounder and were unable to control completely for … database.’ Page 12, line 11: ‘... and Examples’.

REVIEWER	Martine J. Barons University of Warwick UK
REVIEW RETURNED	01-Feb-2019

GENERAL COMMENTS	The authors have addressed the points raised by the reviewers

REVIEWER	Hamada Ahmed Hamada Cairo University, Egypt
REVIEW RETURNED	26-Jan-2019

GENERAL COMMENTS	please don't putting output of spss as it is. please revise all result section

VERSION 2 – AUTHOR RESPONSE

-In response to reviewer 1, Dr. Smith:

We appreciate your concerns and have altered the title of our paper as suggested. We have also made all the minor edits suggested.

We have added the mortality numbers to the results section of our paper. We looked into adding the mortality information to Figure One, but found that this added visual clutter to the extent that it might then falsely appear that we were excluding patients who died in our primary analyses. Unfortunately, we also do not know the intubation length of those with the missing data.

Regarding your concern about modelling the log of length of stay initially: This is something we also had considered when first running our models, however we and our statistician felt the data was sufficiently normal so that this was not necessary.

We have altered our falsification testing to better conform with Prasad and Jena's methodology. We re-ran the analysis looking for an association between pain and nausea and continued to find no association

We have edited the supplemental file to improved readability by eliminating some unnecessary tables, as well as adding headings and a table of contents

Thank you again for reviewing our paper.

-In response to reviewer 3, Dr. Hamada:

We appreciate you taking your time to again review our paper. We feel that we have included the most pertinent information in our results sections that doesn't overwhelm the readers with an excessive amount of numbers. We believe including the statistical output would be beneficial to those who wanted additional information. We have further edited the supplemental file to make it easier to navigate.

VERSION 3 - REVIEW

REVIEWER	Peter Smith University of Southampton, UK
REVIEW RETURNED	11-Mar-2019

GENERAL COMMENTS	I now think that this paper is acceptable for publication, although a few minor changes to the text are required: Page 2, line 50: '… incomplete or has not been firmly …'. Page 5, line 32: 'Nausea was also tested against our exposure for use …'. Page 9, line 55: '… hospital and our …'. Page 10, line 11: '… erroneous due to a selection effect'? Page 12, line 45: '… database.'