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)	The impact of the implementation of a fast-track on emergency department length of stay and quality of care indicators in the Champagne-Ardenne region: a before-after study
AUTHORS	Chrusciel, Jan; Fontaine, Xavier; Devillard, Arnaud; Cordonnier, Aurélien; Kanagaratnam, Lukshe; Laplanche, David; Sanchez, Stéphane

VERSION 1 - REVIEW

REVIEWER	Li Chao-Jui Kaohsiung Chang Gung Memorial Hospital, Taiwan
REVIEW RETURNED	11-Sep-2018

GENERAL COMMENTS	The statistical methods are too complicated to me. Please review it by specialist statistician.

REVIEWER	Anna Marie Chang Thomas Jefferson University, Philadelphia PA USA
REVIEW RETURNED	20-Sep-2018

GENERAL COMMENTS	Thank you for your submission of “Impact of the implementation fo a fast-track on emergency department length of stay and quality of care indicators” This is an interesting paper that adds to the body of literature that split track and having a separate area of minor illnesses appears to decrease overall LOS. there should not be new results presented in the discussion section. The discussion is also muddled by the lack of focus on hypothesizing on how fast track decreases LOS. there are references to case management, readmissions, and other topics that are relevant to ED crowding, but may or may not relate to fast track. Strengths and limitations section I am not clear as to how more hospitals in the control group would alter the impact of the results of your study? including more hospitals in general may be helpful Abstract: I am unclear as to why you use your primary outcome as proportion of patients with >4h LOS and access block, versus just median or mean LOS. Would separate out the sentences re: access block and readmissions.

Conclusion: do you mean <4 hours and is it < or <?
Here you mention the connection to quality of care indicators—what do you mean by this

The introduction is a bit long specifically in its flow of why length of stay matters/is a quality indicator.
Intervention:
Were all new 12 rooms used for fast track?
Who determined
In your paper, you report LWBS as well as 7 and 30 day readmission. Why is the LWBS not reported in abstract?
In the methods you describe lwbs, 7 and 30 day readmission as the quality of care indicators?
What are the "clinically significant predictors of LOS" included in the model?

Results:
I do not think table 2 is necessary, perhaps a sentence re: the increase in FTE is all that is necessary.

Page 8, line 29: what do you mean than outpatients?
Unclear to me where the subgroup analysis came from as it was not described in the methods section.

Table 3: is this LOS for all patients? I would like to see LOS for admitted vs discharged patients and vs fast track area patients. Presumably the discharged patients would see more difference at the intervention hospital?
Why does table 3 separate out only patients with injuries?

I would like to see numbers of patients who went to fast track versus the main ED vs the new fast track area. If the numbers are high with not suitable staffing patterns then it would not change LOS

Discussion:
Are you focusing on patients with < or >4 hours LOS again?
"extension" here is confusing, do you mean adding beds?
Page 13, line 12: these are new data being presented for differing LOS for different ICD 10 codes. This was not discussed at all during results section.
Line 37: the DEED II trial seems misplaced here as I think the readmissions piece has muddled up the data.
Page 14, again presenting new data re: neoplasms and diseases
Unclear to me what case management has to do with LOS and the focus of having a fast track area in the ED.
Why would readmissions decrease LOS? And again, presenting new data in discussion section.

Supplementary:
Page 26
Line 8: Not sure what "dependent in every day life" means
Line 28: define "major wound"
Page 27 line 40: what is urgent treatment injections

REVIEWER	Janos Sandor Department of Preventive Medicine Faculty of Public Health University of Debrecen Hungary
REVIEW RETURNED	28-Nov-2018

GENERAL COMMENTS

Thank you very much for the opportunity to review your manuscript. I was asked to evaluate the applied statistical methods.

Regarding descriptive statistics:
1.
Time of the intervention is not reported in the text.
2.
It is written that PS classification by data was not possible in all records. Sometimes expert opinion was the base of categorization. It is not declared how frequent this secondary approach was. Further, why was not applied a statistical method to estimate missing data (e.g.: multiple imputation)?
3.
Continuous variables (age, LOS) have obviously not normal distribution (page 8, lines 27-60). The mean and SD seems to be not proper summary measures to describe the distribution of observed data.
4.
The structure of table 3 is proper. But the p values are reported in a not consistent manner: sometimes with 3 decimals, sometimes with 4 decimals. The usual way with 3 decimals seems to appropriate in this table. Footnotes have to be added to the table on (a) exact dates of the first and the second periods, and on (b) the name of test to calculate the p-values (hopefully, a test which was able to evaluate the not normally distributed data).
5. There are disease specific descriptive statistics in the Results section (page 8, lines 42-60), but this analysis is not mentioned in the Methods section.

Regarding regression modeling:
1.
The difference-in-difference analysis by logistic regression modelling with interaction term for time and exposure is proper method to answer the study questions.
2.
Definition of the regression model (page 6; line 40) is good. Cited references are proper.
3.
Table 4 on results of modeling has appropriate structure. It contains p-values reported with 3 decimals and 4 decimals. It is to be corrected. Furthermore, reporting p-values and 95% confidence intervals are redundant. Considering that p-values are not informative about the size of the effect, and due to the big numbers analyzed the statistical significant effects are not necessarily clinically important, reporting only the 95% confidence interval is informative enough
4.
Results of regression models are reported in table 4 needs some explanation, since the table does not contain the odds ratios for the period and for the location; we can see only the calculated measures for interaction term. Why? If the tested model was applied without inserting location and period as explanatory variables then the model is not appropriate. If these variables were included in the model then the results for them should be added to the table.
5.

	The regression model contained 7-day readmission as variable to control for the quality of care. There are 3 indicators used for quality of care description (7-day readmission, 30-day readmission, number of patients leaving without being seen) in the paper. It is not written explicitly what was the reason for using 7-day readmission only in modeling. 6. It is written in page 10-line 3 that exponentiated difference-in-difference was estimated, and the reported odds ratio is the same shown in table 4. It is not mentioned in the Methods that the parameters had transformed before modeling to correct the lack of normality, and it is not declared in the title of table 4 that some variables were transformed. 7. Criteria of applying difference-in-difference analysis are not evaluated properly. (a) The similar before-intervention trends in the studied hospitals are not checked properly. Statistical evaluation of before-intervention-trends in both hospitals needed (the graphic presentation is not enough.) (b) The criterion of common shocks is violated as it is acknowledged by authors (page 15, lines 36-45). Without demonstrating that this problem does not jeopardize the validity, results are not convincing and are not able to answer the study questions.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Li Chao-Jui

Institution and Country: Kaohsiung Chang Gung Memorial Hospital, Taiwan

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below

The statistical methods are too complicated to me. Please review it by specialist statistician.

Thank you for asking for a specialized review. Our method aims to take into account patient-level confounders and natural trends over time. Numerous publications assign causal effects to public health interventions whereas in fact the observed modifications were the consequence of regional trends and would have occurred even without the interventions. Our design controls for such trends as well as for patient-related confounders.

Reviewer: 2

Reviewer Name: Anna Marie Chang

Institution and Country: Thomas Jefferson University, Philadelphia PA USA

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below

Thank you for your submission of “Impact of the implementation fo a fast-track on emergency department length of stay and quality of care indicators”

This is an interesting paper that adds to the body of literature that split track and having a separate area of minor illnesses appears to decrease overall LOS.

Thank you for your helpful review.

there should not be new results presented in the discussion section. The discussion is also muddled by the lack of focus on hypothesizing on how fast track decreases LOS. there are references to case management, readmissions, and other topics that are relevant to ED crowding, but may or may not relate to fast track.

The decrease in length of stay observed after the implementation of a fast-track can be explained by several mechanisms, which are not mutually exclusive: decreased crowding due to rapid patient discharge, floorplan modifications allowing faster patient transfers, or physician and nurse role adjustments. These mechanisms could act in a synergic manner, by enabling the creation of patient “streams”.

The problem of the impact of the fast track on other ED patients needs to be considered. Studies have shown that carefully designed fast-track systems usually do not adversely affect other patients.

We added these points to the discussion with the relevant references.

Strengths and limitations section

I am not clear as to how more hospitals in the control group would alter the impact of the results of your study? including more hospitals in general may be helpful

We agree that by including more hospitals, we would be more confident that our findings are not limited to our specific setting but can be generalized to other hospitals with a similar profile. To clarify this, we changed “more hospitals in the control group” to “more hospitals”.

Abstract:

I am unclear as to why you use your primary outcome as proportion of patients with >4h LOS and access block, versus just median or mean LOS.

Would separate out the sentences re: access block and readmissions.

In the multivariable analysis, we used categorical variables because they usually require fewer modelling assumptions. As the model is a key part of the article, we think it is important to keep the results of the multivariable model in the abstract. We checked that we also added the mean LOS. We rephrased the abstract to make it easier to read and within the 300 words limit.

Conclusion: do you mean <4 hours and is it < or <?

Here you mention the connection to quality of care indicators—what do you mean by this

We can confirm that we meant ≥ (superior or equal to) 4 hours. Overall, the intervention resulted in a decrease in the proportion of stays that were longer than or equal to four hours. This is consistent with an improvement in patient flow which can be attributed to the intervention. However, patients that were hospitalized after the emergency department did not benefit from the intervention. Their increased length of stay, however, could be due to other factors (hospital-level bed availability).

The introduction is a bit long specifically in its flow of why length of stay matters/is a quality indicator.

Following your comment, we shortened the introduction.

We wanted to ensure the readers understood that our point of view is patient-centered even though the article uses indicators such as length of stay which are often used from a health economics/hospital management point of view.

Intervention:

Were all new 12 rooms used for fast track?

The fast-track was a dedicated area with 6 consultation spaces. We added a precision in the revised manuscript.

Who determined

In your paper, you report LWBS as well as 7 and 30 day readmission. Why is the LWBS not reported in abstract?

We added the number of patients LWBS in the abstract.

In the methods you describe lwbs, 7 and 30 day readmission as the quality of care indicators?

Yes. At the time of writing they were considered quality of care indicators. We added a sentence stating that readmissions data need to be interpreted very carefully and in conjunction with other quality of care indicators. At the end of the discussion, we also insist that further studies should include more quality of care indicators.

What are the “clinically significant predictors of LOS” included in the model?

The aim of this sentence was to introduce a description of the variables included in the model. Following your comment, we removed this sentence.

Results:

I do not think table 2 is necessary, perhaps a sentence re: the increase in FTE is all that is necessary.

We removed table 2 and replaced it by a sentence.

Page 8, line 29: what do you mean than outpatients? Unclear to me where the subgroup analysis came from as it was not described in the methods section.

We deleted this sentence.

Table 3: is this LOS for all patients?

The LOS for all patients is reported on the left side of the table. We added “for all patients” to clarify.

I would like to see LOS for admitted vs discharged patients and vs fast track area patients. Presumably the discharged patients would see more difference at the intervention hospital?

This is correct. The data is shown in Table 2 of the revised manuscript. Patients discharged from the intervention hospital decreased their mean length of stay by 22 minutes (from 230 to 208 minutes). Patients discharged from the control hospital had a stable length of stay (from 203 to 201 minutes).

Therefore, the difference between both periods was highest in the intervention group for discharged patients.

Why does table 3 separate out only patients with injuries?

It seemed some of these patients (with light injuries) could be representative of patients of the fast-track. However, because this could be confusing we deleted this part of the table.

I would like to see numbers of patients who went to fast track versus the main ED vs the new fast track area. If the numbers are high with not suitable staffing patterns then it would not change LOS

Patients who went through the fast-track were not indicated in our database. However, we know which patients were admitted to the hospital after staying in the emergency department. The number of patients that were admitted to the hospital after the emergency department remained stable (from 14,795 to 14,864 ; +0.4%). On the other side, the number of patients who were not hospitalized rose from 38,971 to 43,100 (+10.6%) (Table 2 of revised manuscript). This can be attributed to new patients who were managed in the fast-track (fast-track patients are rarely hospitalized after their ED stay).

Discussion:

Are you focusing on patients with < or >4 hours LOS again?

“extension” here is confusing, do you mean adding beds?

Yes, we replaced “the extension” by “the addition of new beds”.

Page 13, line 12: these are new data being presented for differing LOS for different ICD 10 codes. This was not discussed at all during results section.

We now present these data in the Results section.

Line 37: the DEED II trial seems misplaced here as I think the readmissions piece has muddled up the data.

We removed the DEED II trial.

Page 14, again presenting new data re: neoplasms and diseases

Unclear to me what case management has to do with LOS and the focus of having a fast track area in the ED.

This could decrease the input component to the emergency department, which could prevent crowding which in turn reduces LOS. However, as is does no seem entirely clear, we removed this sentence.

Why would readmissions decrease LOS?

For readmissions that take place a short time after the first admission, it could be due to knowledge of the patient's recent history which means that less history-taking and complementary examinations are needed, making the emergency department stay shorter.

And again, presenting new data in discussion section.

As these data can be seen in the tables, we removed them from the Discussion.

Supplementary:

Page 26

Line 8: Not sure what "dependent in every day life" means

We changed this to "Disabled patients or patients with reduced autonomy"

Line 28: define "major wound"

We changed this to "The patient needs a wound suture with expected duration of > 20 min"

Page 27 line 40: what is urgent treatment injections

We changed this line to "Urgent local drugs administration (e.g. intracavernous phenylephrine)"

Thank you for your helpful review.

Reviewer: 3

Reviewer Name: Janos Sandor

Institution and Country: Department of Preventive Medicine, Faculty of Public Health, University of Debrecen, Hungary

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below

Thank you very much for the opportunity to review your manuscript. I was asked to evaluate the applied statistical methods.

Thank you for your very precise review.

Regarding descriptive statistics:

1.

Time of the intervention is not reported in the text.

We added the time of the intervention in the Methods section.

2.

It is written that PS classification by data was not possible in all records. Sometimes expert opinion was the base of categorization. It is not declared how frequent this secondary approach was.

This was the case in 11.9 % of cases. We added this in the methods section of the revised manuscript.

Further, why was not applied a statistical method to estimate missing data (e.g.: multiple imputation)?

We agree that multiple imputation is a good method for treating missing data. However, it was difficult for us to carry out multiple imputation on this large dataset.

3.

Continuous variables (age, LOS) have obviously not normal distribution (page 8, lines 27-60). The mean and SD seems to be not proper summary measures to describe the distribution of observed data.

We agree that the median can be a good summary statistic for asymetric data. However, it is important for us to use the mean, because by multiplying LOS by the number of patients we can estimate total bed utilization. This is a meaningful information for hospital managers.

4.

The structure of table 3 is proper. But the p values are reported in a not consistent manner: sometimes with 3 decimals, sometimes with 4 decimals. The usual way with 3 decimals seems to appropriate in this table.

We changed the p values from 4 decimals to 3 decimals.

Footnotes have to be added to the table on (a) exact dates of the first and the second periods, and on (b) the name of test to calculate the p-values (hopefully, a test which was able to evaluate the not normally distributed data).

We added the footnotes for exact dates of both periods. We also added footnotes for the names of tests to calculate p-values. NB: Because the number of patients in our study was very high, the Central Limit Theorem suggests we could use these tests. This can be verified using a bootstrap estimation of the sample mean's distribution.

5. There are disease specific descriptive statistics in the Results section (page 8, lines 42-60), but this analysis is not mentioned in the Methods section.

We added a mention of this analysis in the Methods section.

Regarding regression modeling:

1.

The difference-in-difference analysis by logistic regression modelling with interaction term for time and exposure is proper method to answer the study questions.

2.

Definition of the regression model (page 6; line 40) is good. Cited references are proper.

3.

Table 4 on results of modeling has appropriate structure. It contains p-values reported with 3 decimals and 4 decimals. It is to be corrected. Furthermore, reporting p-values and 95% confidence intervals are redundant. Considering that p-values are not informative about the size of the effect, and due to the big numbers analyzed the statistical significant effects are not necessarily clinically important, reporting only the 95% confidence interval is informative enough

We removed the p-values.

4.

Results of regression models are reported in table 4 needs some explanation, since the table does not contain the odds ratios for the period and for the location; we can see only the calculated measures for interaction term. Why?

People often misinterpret the meaning of main effects in the presence of interactions. In this case, we were interested in the intervention effect in the center where the intervention was implemented. This was modelled with the interaction term, which was the quantity of interest.

We were less interested in the effect of a particular center, which is why we did not report these coefficients.

If the tested model was applied without inserting location and period as explanatory variables then the model is not appropriate. If these variables were included in the model then the results for them should be added to the table.

These variables were included in the model simultaneously with the interaction term. We added the results in the table.

5.

The regression model contained 7-day readmission as variable to control for the quality of care. There are 3 indicators used for quality of care description (7-day readmission, 30-day readmission, number of patients leaving without being seen) in the paper. It is not written explicitly what was the reason for using 7-day readmission only in modeling.

Only the 7-day readmissions variable was kept in the model because early readmissions could be more relevant than late readmissions from the hospital’s point of view. The nature of 7-day readmissions could be different from 30-day readmissions. For example:

Graham KL, Auerbach AD, Schnipper JL, et al. Preventability of Early Versus Late Hospital Readmissions in a National Cohort of General Medicine Patients. Ann Intern Med 2018;168:766–74. doi:10.7326/M17-1724.

We added this reference to the article.

6.

It is written in page 10-line 3 that exponentiated difference-in-difference was estimated, and the reported odds ratio is the same shown in table 4. It is not mentioned in the Methods that the parameters had transformed before modeling to correct the lack of normality, and it is not declared in the title of table 4 that some variables were transformed.

We added a sentence to explain this in the Methods section : "To facilitate modelling, length of stay was transformed in a binary variable using thresholds classically found in the litterature."

Reference: Khanna S, Boyle J, Good N, et al. New emergency department quality measure: from access block to National Emergency Access Target compliance. Emerg Med Australas EMA 2013;25:565–72. doi:10.1111/1742-6723.12139

We verified that all variables in the table were correctly identified as categorical.

7.

Criteria of applying difference-in-difference analysis are not evaluated properly. (a) The similar before-intervention trends in the studied hospitals are not checked properly. Statistical evaluation of before-intervention-trends in both hospitals needed (the graphic presentation is not enough.)

Thank you for raising this point.

Random fluctuations could make the common trends test indicate a significant effect, whereas from the organizational point of view, no major changes took place in the hospitals. This could be the reason why common trends are rarely tested in the litterature. As applied researchers, we are most confident when we measure statistical indicators of the effect of interventions backed by theoretical and empirical evidence. We cannot provide a better evaluation of the common trend hypothesis.

(b) The criterion of common shocks is violated as it is acknowledged by authors (page 15, lines 36-45). Without demonstrating that this problem does not jeopardize the validity, results are not convincing and are not able to answer the study questions.

We moved this sentence to the end of the discussion. We are not aware of any epidemic that could have affected the population served by one hospital without affecting the other during the study period. Although the design we used relies on assumptions about the intervention and control hospitals which are difficult to assess, we believe it conveys more information than a before-after study restricted to the intervention hospital.

Thank you again for reviewing our paper.

VERSION 2 – REVIEW

REVIEWER	János Sándor University of Debrecen, Faculty of Public Health, Department of Preventive Medicine, Hungary
REVIEW RETURNED	10-Feb-2019

GENERAL COMMENTS	(MY QUESTIONS AND MY COMMENTS ON THE AUTHORS' ANSWERS ARE IN ITALICS-CAPITALS.) Regarding descriptive statistics: 1. TIME OF THE INTERVENTION IS NOT REPORTED IN THE TEXT. We added the time of the intervention in the Methods section. PROPER MODIFICATION 2. IT IS WRITTEN THAT PS CLASSIFICATION BY DATA WAS NOT POSSIBLE IN ALL RECORDS. SOMETIMES EXPERT OPINION WAS THE BASE OF CATEGORIZATION. IT IS NOT DECLARED HOW FREQUENT THIS SECONDARY APPROACH WAS. This was the case in 11.9 % of cases. We added this in the methods section of the revised manuscript. PROPER MODIFICATION, BUT I MISS THE DISCUSSION OF THE INFLUENCE OF THIS VALIDITY PROBLEM ON THE FINAL RESULTS. FURTHER, WHY WAS NOT APPLIED A STATISTICAL METHOD TO ESTIMATE MISSING DATA (E.G.: MULTIPLE IMPUTATION)? We agree that multiple imputation is a good method for treating missing data. However, it was difficult for us to carry out multiple imputation on this large dataset. IT SHOULD BE ACKNOWLEDGED AMONG LIMITATIONS. 3. CONTINUOUS VARIABLES (AGE, LOS) HAVE OBVIOUSLY NOT NORMAL DISTRIBUTION (PAGE 8, LINES 27-60). THE MEAN AND SD SEEMS TO BE NOT PROPER SUMMARY MEASURES TO DESCRIBE THE DISTRIBUTION OF OBSERVED DATA. We agree that the median can be a good summary statistic for asymetric data. However, it is important for us to use the mean, because by multiplying LOS by the number of patients we can estimate total bed utilization. This is a meaningful information for hospital managers. IT IS NOT A MAJOR ISSUE, BUT… ALTHOUGH, MEAN VALUE AND STANDARD DEVIATION CAN BE CALCULATED FOR

	VARIABLE WITH NON-NORMAL DISTRIBUTION, BUT THESE HAVE NO MEANING AT ALL. IN DESCRIPTION, THE MEDIAN WITH INTER-QUARTILE RANGE CAN BE CORRECT AND INFORMATIVE. 4. THE STRUCTURE OF TABLE 3 IS PROPER. BUT THE P VALUES ARE REPORTED IN A NOT CONSISTENT MANNER: SOMETIMES WITH 3 DECIMALS, SOMETIMES WITH 4 DECIMALS. THE USUAL WAY WITH 3 DECIMALS SEEMS TO APPROPRIATE IN THIS TABLE. We changed the p values from 4 decimals to 3 decimals. PROPER MODIFICATION FOOTNOTES HAVE TO BE ADDED TO THE TABLE ON (A) EXACT DATES OF THE FIRST AND THE SECOND PERIODS, AND ON (B) THE NAME OF TEST TO CALCULATE THE P-VALUES (HOPEFULLY, A TEST WHICH WAS ABLE TO EVALUATE THE NOT NORMALLY DISTRIBUTED DATA). We added the footnotes for exact dates of both periods. We also added footnotes for the names of tests to calculate p-values. NB: Because the number of patients in our study was very high, the Central Limit Theorem suggests we could use these tests. This can be verified using a bootstrap estimation of the sample mean's distribution. REGARDING FOOTNOTES: PROPER MODIFICATION REGARDING STATISTICAL TESTS: THIS EXPLANATION SHOULD BE WRITTEN NOT ONLY FOR REVIEWERS. IT IS MORE IMPORTANT TO BE PRESENTED FOR READERS! 5. THERE ARE DISEASE SPECIFIC DESCRIPTIVE STATISTICS IN THE RESULTS SECTION (PAGE 8, LINES 42-60), BUT THIS ANALYSIS IS NOT MENTIONED IN THE METHODS SECTION. We added a mention of this analysis in the Methods section. PROPER MODIFICATION Regarding regression modeling: 1. The difference-in-difference analysis by logistic regression modelling with interaction term for time and exposure is proper method to answer the study questions. 2. Definition of the regression model (page 6; line 40) is good. Cited references are proper. 3. TABLE 4 ON RESULTS OF MODELING HAS APPROPRIATE STRUCTURE. IT CONTAINS P-VALUES REPORTED WITH 3 DECIMALS AND 4 DECIMALS. IT IS TO BE CORRECTED. FURTHERMORE, REPORTING P-VALUES AND 95% CONFIDENCE INTERVALS ARE REDUNDANT. CONSIDERING THAT P-VALUES ARE NOT INFORMATIVE ABOUT THE SIZE OF THE EFFECT, AND DUE TO THE BIG NUMBERS ANALYZED THE STATISTICAL SIGNIFICANT EFFECTS ARE NOT NECESSARILY CLINICALLY IMPORTANT, REPORTING ONLY THE 95% CONFIDENCE INTERVAL IS INFORMATIVE ENOUGH We removed the p-values. PROPER MODIFICATION 4. RESULTS OF REGRESSION MODELS ARE REPORTED IN TABLE 4 NEEDS SOME EXPLANATION, SINCE THE TABLE DOES NOT CONTAIN THE ODDS RATIOS FOR THE PERIOD

	AND FOR THE LOCATION; WE CAN SEE ONLY THE CALCULATED MEASURES FOR INTERACTION TERM. WHY? People often misinterpret the meaning of main effects in the presence of interactions. In this case, we were interested in the intervention effect in the center where the intervention was implemented. This was modelled with the interaction term, which was the quantity of interest. We were less interested in the effect of a particular center, which is why we did not report these coefficients. If the tested model was applied without inserting location and period as explanatory variables then the model is not appropriate. If these variables were included in the model then the results for them should be added to the table. These variables were included in the model simultaneously with the interaction term. We added the results in the table. REGARDING LOCATION: THE LOCATION IS SIGNIFICANT FACTOR FOR BOTH MODEL PRESENTED IN TABLE 4 (ACCORDING TO THE ORIGINAL VERSION). IT MEANS THAT THERE WERE SIGNIFICANT DIFFERENCES BETWEEN INTERVENTION AND CONTROL HOSPITALS. IT NEEDS EXPLICIT DISCUSSION WHETHER THE CONTROL HOSPITALS WERE USEFUL FOR THE DIFFERENCE-IN-DIFFERENCE ANALYSIS. 5. THE REGRESSION MODEL CONTAINED 7-DAY READMISSION AS VARIABLE TO CONTROL FOR THE QUALITY OF CARE. THERE ARE 3 INDICATORS USED FOR QUALITY OF CARE DESCRIPTION (7-DAY READMISSION, 30-DAY READMISSION, NUMBER OF PATIENTS LEAVING WITHOUT BEING SEEN) IN THE PAPER. IT IS NOT WRITTEN EXPLICITLY WHAT WAS THE REASON FOR USING 7-DAY READMISSION ONLY IN MODELING. Only the 7-day readmissions variable was kept in the model because early readmissions could be more relevant than late readmissions from the hospital's point of view. The nature of 7-day readmissions could be different from 30-day readmissions. For example: Graham KL, Auerbach AD, Schnipper JL, et al. Preventability of Early Versus Late Hospital Readmissions in a National Cohort of General Medicine Patients. Ann Intern Med 2018;168:766–74. doi:10.7326/M17-1724. We added this reference to the article. THIS ANSWER IS NOT STATISTICAL - IT CAN BE PROPER MODIFICATION 6. IT IS WRITTEN IN PAGE 10-LINE 3 THAT EXPONENTIATED DIFFERENCE-IN-DIFFERENCE WAS ESTIMATED, AND THE REPORTED ODDS RATIO IS THE SAME SHOWN IN TABLE 4. IT IS NOT MENTIONED IN THE METHODS THAT THE PARAMETERS HAD TRANSFORMED BEFORE MODELING TO CORRECT THE LACK OF NORMALITY, AND IT IS NOT DECLARED IN THE TITLE OF TABLE 4 THAT SOME VARIABLES WERE TRANSFORMED. We added a sentence to explain this in the Methods section : "To facilitate modelling, length of stay was transformed in a binary variable using thresholds classically found in the litterature." Reference: Khanna S, Boyle J, Good N, et al. New emergency department quality measure: from access block to National Emergency Access Target compliance. Emerg Med Australas EMA 2013;25:565–72. doi:10.1111/1742-6723.12139

	We verified that all variables in the table were correctly identified as categorical. PROPER MODIFICATION 7. CRITERIA OF APPLYING DIFFERENCE-IN-DIFFERENCE ANALYSIS ARE NOT EVALUATED PROPERLY. (A) THE SIMILAR BEFORE-INTERVENTION TRENDS IN THE STUDIED HOSPITALS ARE NOT CHECKED PROPERLY. STATISTICAL EVALUATION OF BEFORE-INTERVENTION-TRENDS IN BOTH HOSPITALS NEEDED (THE GRAPHIC PRESENTATION IS NOT ENOUGH.) Thank you for raising this point. Random fluctuations could make the common trends test indicate a significant effect, whereas from the organizational point of view, no major changes took place in the hospitals. This could be the reason why common trends are rarely tested in the litterature. As applied researchers, we are most confident when we measure statistical indicators of the effect of interventions backed by theoretical and empirical evidence. We cannot provide a better evaluation of the common trend hypothesis. (B) THE CRITERION OF COMMON SHOCKS IS VIOLATED AS IT IS ACKNOWLEDGED BY AUTHORS (PAGE 15, LINES 36-45). WITHOUT DEMONSTRATING THAT THIS PROBLEM DOES NOT JEOPARDIZE THE VALIDITY, RESULTS ARE NOT CONVINCING AND ARE NOT ABLE TO ANSWER THE STUDY QUESTIONS. We moved this sentence to the end of the discussion. We are not aware of any epidemic that could have affected the population served by one hospital without affecting the other during the study period. Although the design we used relies on assumptions about the intervention and control hospitals which are difficult to assess, we believe it conveys more information than a before-after study restricted to the intervention hospital. THE DIFFERENCE-IN-DIFFERENCE ANALYSIS IS ABLE TO PRODUCE MORE CONVINCING EVIDENCES THAN THE BEFORE-AFTER ANALYSIS WHEN THE CRITERIA FOR ITS APPLICATION ARE MET. IF THESE CRITERIA ARE NOT MET OR AUTHORS HAVE NO DATA TO DEMONSTRATE THAT THOSE ARE MET, THEN THIS STATISTICAL APPROACH IS NOT ESTABLISHED. ACCORDING TO THE ANSWERS: (A) THE SIMILAR BEFORE-INTERVENTION TREND IS NOT MET (LOCATION SPECIFIC ORs INSERTED INTO TABLE 4) (B) THE COMMON SHOCK IS VIOLATED AS IT IS ACKNOWLEDGED BY AUTHORS. AUTHORS COULD NOT ANSWER THE VALIDITY RELATED QUESTIONS. UNTIL THEY CAN DO IT, THE CONCLUSIONS ARE NOT CONVINCINGLY SUPPORTED BY THE PRESENTED RESULTS.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 3

Reviewer Name: János Sándor

Institution and Country: University of Debrecen, Faculty of Public Health, Department of Preventive Medicine, Hungary

Please state any competing interests or state ‘None declared’: None declared.

Please leave your comments for the authors below

(MY QUESTIONS AND MY COMMENTS ON THE AUTHORS' ANSWERS ARE IN ITALICS-CAPITALS.)

Regarding descriptive statistics:

1.

TIME OF THE INTERVENTION IS NOT REPORTED IN THE TEXT.

We added the time of the intervention in the Methods section.

PROPER MODIFICATION

2.

IT IS WRITTEN THAT PS CLASSIFICATION BY DATA WAS NOT POSSIBLE IN ALL RECORDS. SOMETIMES EXPERT OPINION WAS THE BASE OF CATEGORIZATION. IT IS NOT DECLARED HOW FREQUENT THIS SECONDARY APPROACH WAS.

This was the case in 11.9 % of cases. We added this in the methods section of the revised manuscript.

PROPER MODIFICATION, BUT I MISS THE DISCUSSION OF THE INFLUENCE OF THIS VALIDITY PROBLEM ON THE FINAL RESULTS.

FURTHER, WHY WAS NOT APPLIED A STATISTICAL METHOD TO ESTIMATE MISSING DATA (E.G.: MULTIPLE IMPUTATION)?

We added a paragraph regarding missing data in the Discussion.

We agree that multiple imputation is a good method for treating missing data. However, it was difficult for us to carry out multiple imputation on this large dataset.

IT SHOULD BE ACKNOWLEDGED AMONG LIMITATIONS.

We added this to the limitations in the Discussion.

3.

CONTINUOUS VARIABLES (AGE, LOS) HAVE OBVIOUSLY NOT NORMAL DISTRIBUTION (PAGE 8, LINES 27-60). THE MEAN AND SD SEEMS TO BE NOT PROPER SUMMARY MEASURES TO DESCRIBE THE DISTRIBUTION OF OBSERVED DATA.

We agree that the median can be a good summary statistic for asymetric data. However, it is important for us to use the mean, because by multiplying LOS by the number of patients we can estimate total bed utilization. This is a meaningful information for hospital managers.

IT IS NOT A MAJOR ISSUE, BUT… ALTHOUGH, MEAN VALUE AND STANDARD DEVIATION CAN BE CALCULATED FOR VARIABLE WITH NON-NORMAL DISTRIBUTION, BUT THESE HAVE NO MEANING AT ALL. IN DESCRIPTION, THE MEDIAN WITH INTER-QUARTILE RANGE CAN BE CORRECT AND INFORMATIVE.

4.

THE STRUCTURE OF TABLE 3 IS PROPER. BUT THE P VALUES ARE REPORTED IN A NOT CONSISTENT MANNER: SOMETIMES WITH 3 DECIMALS, SOMETIMES WITH 4 DECIMALS. THE USUAL WAY WITH 3 DECIMALS SEEMS TO APPROPRIATE IN THIS TABLE.

We changed the p values from 4 decimals to 3 decimals.

PROPER MODIFICATION

FOOTNOTES HAVE TO BE ADDED TO THE TABLE ON (A) EXACT DATES OF THE FIRST AND THE SECOND PERIODS, AND ON (B) THE NAME OF TEST TO CALCULATE THE P-VALUES (HOPEFULLY, A TEST WHICH WAS ABLE TO EVALUATE THE NOT NORMALLY DISTRIBUTED DATA).

We added the footnotes for exact dates of both periods. We also added footnotes for the names of tests to calculate p-values. NB: Because the number of patients in our study was very high, the Central Limit Theorem suggests we could use these tests. This can be verified using a bootstrap estimation of the sample mean’s distribution.

REGARDING FOOTNOTES: PROPER MODIFICATION

REGARDING STATISTICAL TESTS: THIS EXPLANATION SHOULD BE WRITTEN NOT ONLY FOR REVIEWERS. IT IS MORE IMPORTANT TO BE PRESENTED FOR READERS!

We added a paragraph in the methods section describing the tests and their validity conditions.

5. THERE ARE DISEASE SPECIFIC DESCRIPTIVE STATISTICS IN THE RESULTS SECTION (PAGE 8, LINES 42-60), BUT THIS ANALYSIS IS NOT MENTIONED IN THE METHODS SECTION.

We added a mention of this analysis in the Methods section.

PROPER MODIFICATION

Regarding regression modeling:

1.

The difference-in-difference analysis by logistic regression modelling with interaction term for time and exposure is proper method to answer the study questions.

2.

Definition of the regression model (page 6; line 40) is good. Cited references are proper.

3.

TABLE 4 ON RESULTS OF MODELING HAS APPROPRIATE STRUCTURE. IT CONTAINS P-VALUES REPORTED WITH 3 DECIMALS AND 4 DECIMALS. IT IS TO BE CORRECTED.

FURTHERMORE, REPORTING P-VALUES AND 95% CONFIDENCE INTERVALS ARE REDUNDANT. CONSIDERING THAT P-VALUES ARE NOT INFORMATIVE ABOUT THE SIZE OF THE EFFECT, AND DUE TO THE BIG NUMBERS ANALYZED THE STATISTICAL SIGNIFICANT EFFECTS ARE NOT NECESSARILY CLINICALLY IMPORTANT, REPORTING ONLY THE 95% CONFIDENCE INTERVAL IS INFORMATIVE ENOUGH

We removed the p-values.

PROPER MODIFICATION

4.

RESULTS OF REGRESSION MODELS ARE REPORTED IN TABLE 4 NEEDS SOME EXPLANATION, SINCE THE TABLE DOES NOT CONTAIN THE ODDS RATIOS FOR THE PERIOD AND FOR THE LOCATION; WE CAN SEE ONLY THE CALCULATED MEASURES FOR INTERACTION TERM. WHY?

People often misinterpret the meaning of main effects in the presence of interactions. In this case, we were interested in the intervention effect in the center where the intervention was implemented. This was modelled with the interaction term, which was the quantity of interest.

We were less interested in the effect of a particular center, which is why we did not report these coefficients.

IF THE TESTED MODEL WAS APPLIED WITHOUT INSERTING LOCATION AND PERIOD AS EXPLANATORY VARIABLES THEN THE MODEL IS NOT APPROPRIATE. IF THESE VARIABLES WERE INCLUDED IN THE MODEL THEN THE RESULTS FOR THEM SHOULD BE ADDED TO THE TABLE.

These variables were included in the model simultaneously with the interaction term. We added the results in the table.

REGARDING LOCATION: THE LOCATION IS SIGNIFICANT FACTOR FOR BOTH MODEL PRESENTED IN TABLE 4 (ACCORDING TO THE ORIGINAL VERSION). IT MEANS THAT THERE WERE SIGNIFICANT DIFFERENCES BETWEEN INTERVENTION AND CONTROL HOSPITALS. IT NEEDS EXPLICIT DISCUSSION WHETHER THE CONTROL HOSPITALS WERE USEFUL FOR THE DIFFERENCE-IN-DIFFERENCE ANALYSIS.

We added these points to the discussion.

5.

THE REGRESSION MODEL CONTAINED 7-DAY READMISSION AS VARIABLE TO CONTROL FOR THE QUALITY OF CARE. THERE ARE 3 INDICATORS USED FOR QUALITY OF CARE DESCRIPTION (7-DAY READMISSION, 30-DAY READMISSION, NUMBER OF PATIENTS LEAVING WITHOUT BEING SEEN) IN THE PAPER. IT IS NOT WRITTEN EXPLICITLY WHAT WAS THE REASON FOR USING 7-DAY READMISSION ONLY IN MODELING.

Only the 7-day readmissions variable was kept in the model because early readmissions could be more relevant than late readmissions from the hospital’s point of view. The nature of 7-day readmissions could be different from 30-day readmissions. For example: Graham KL, Auerbach AD, Schnipper JL, et al. Preventability of Early Versus Late Hospital Readmissions in a National Cohort of General Medicine Patients. Ann Intern Med 2018;168:766–74. doi:10.7326/M17-1724. We added this reference to the article.

THIS ANSWER IS NOT STATISTICAL - IT CAN BE PROPER MODIFICATION

6.

IT IS WRITTEN IN PAGE 10-LINE 3 THAT EXPONENTIATED DIFFERENCE-IN-DIFFERENCE WAS ESTIMATED, AND THE REPORTED ODDS RATIO IS THE SAME SHOWN IN TABLE 4. IT IS NOT MENTIONED IN THE METHODS THAT THE PARAMETERS HAD TRANSFORMED BEFORE MODELING TO CORRECT THE LACK OF NORMALITY, AND IT IS NOT DECLARED IN THE TITLE OF TABLE 4 THAT SOME VARIABLES WERE TRANSFORMED.

We added a sentence to explain this in the Methods section : “To facilitate modelling, length of stay was transformed in a binary variable using thresholds classically found in the litterature.”

Reference: Khanna S, Boyle J, Good N, et al. New emergency department quality measure: from access block to National Emergency Access Target compliance. Emerg Med Australas EMA 2013;25:565–72. doi:10.1111/1742-6723.12139

We verified that all variables in the table were correctly identified as categorical.

PROPER MODIFICATION

7.

CRITERIA OF APPLYING DIFFERENCE-IN-DIFFERENCE ANALYSIS ARE NOT EVALUATED PROPERLY. (A) THE SIMILAR BEFORE-INTERVENTION TRENDS IN THE STUDIED HOSPITALS ARE NOT CHECKED PROPERLY. STATISTICAL EVALUATION OF BEFORE-INTERVENTION-TRENDS IN BOTH HOSPITALS NEEDED (THE GRAPHIC PRESENTATION IS NOT ENOUGH.)

Thank you for raising this point. Random fluctuations could make the common trends test indicate a significant effect, whereas from the organizational point of view, no major changes took place in the hospitals. This could be the reason why common trends are rarely tested in the litterature. As applied researchers, we are most confident when we measure statistical indicators of the effect of interventions backed by theoretical and empirical evidence. We cannot provide a better evaluation of the common trend hypothesis.

(B) THE CRITERION OF COMMON SHOCKS IS VIOLATED AS IT IS ACKNOWLEDGED BY AUTHORS (PAGE 15, LINES 36-45). WITHOUT DEMONSTRATING THAT THIS PROBLEM DOES NOT JEOPARDIZE THE VALIDITY, RESULTS ARE NOT CONVINCING AND ARE NOT ABLE TO ANSWER THE STUDY QUESTIONS.

We moved this sentence to the end of the discussion. We are not aware of any epidemic that could have affected the population served by one hospital without affecting the other during the study period. Although the design we used relies on assumptions about the intervention and control hospitals which are difficult to assess, we believe it conveys more information than a before-after study restricted to the intervention hospital.

THE DIFFERENCE-IN-DIFFERENCE ANALYSIS IS ABLE TO PRODUCE MORE CONVINCING EVIDENCES THAN THE BEFORE-AFTER ANALYSIS WHEN THE CRITERIA FOR ITS APPLICATION ARE MET. IF THESE CRITERIA ARE NOT MET OR AUTHORS HAVE NO DATA TO DEMONSTRATE THAT THOSE ARE MET, THEN THIS STATISTICAL APPROACH IS NOT ESTABLISHED. ACCORDING TO THE ANSWERS:

(A) THE SIMILAR BEFORE-INTERVENTION TREND IS NOT MET (LOCATION SPECIFIC ORs INSERTED INTO TABLE 4)

(B) THE COMMON SHOCK IS VIOLATED AS IT IS ACKNOWLEDGED BY AUTHORS.

AUTHORS COULD NOT ANSWER THE VALIDITY RELATED QUESTIONS. UNTIL THEY CAN DO IT, THE CONCLUSIONS ARE NOT CONVINCINGLY SUPPORTED BY THE PRESENTED RESULTS.

We added these points to the discussion. In the current state of the article, the reader can refer to the descriptive tables for a before-after analysis and to a complementary difference-in-differences analysis for which the reader can consider the limitations underlined in the discussion when interpreting the results. However, we can remove the difference-in-differences analysis if needed.

VERSION 3 - REVIEW

REVIEWER	János Sándor University of Debrecen, Faculty of Public Health, Department of Preventive Medicine, Hungary
REVIEW RETURNED	23-Mar-2019

GENERAL COMMENTS	My new comment/questions are in capitals-italics at the end of numbered sections. Unfortunately the main criticisms on the appropriateness of statistical methods and on the lack of validity analysis have been not addressed by authors. Some of my notions are not answered at all. If the Editor needs further peer-review for this manuscript then, please, look for an other reviewer. Regarding descriptive statistics: 1. TIME OF THE INTERVENTION IS NOT REPORTED IN THE TEXT. We added the time of the intervention in the Methods section. PROPER MODIFICATION 2. IT IS WRITTEN THAT PS CLASSIFICATION BY DATA WAS NOT POSSIBLE IN ALL RECORDS. SOMETIMES EXPERT OPINION WAS THE BASE OF CATEGORIZATION. IT IS NOT DECLARED HOW FREQUENT THIS SECONDARY APPROACH WAS. This was the case in 11.9 % of cases. We added this in the methods section of the revised manuscript. PROPER MODIFICATION, BUT I MISS THE DISCUSSION OF THE INFLUENCE OF THIS VALIDITY PROBLEM ON THE FINAL RESULTS. FURTHER, WHY WAS NOT APPLIED A STATISTICAL METHOD TO ESTIMATE MISSING DATA (E.G.: MULTIPLE IMPUTATION)? We added a paragraph regarding missing data in the Discussion. We agree that multiple imputation is a good method for treating missing data. However, it was difficult for us to carry out multiple imputation on this large dataset. IT SHOULD BE ACKNOWLEDGED AMONG LIMITATIONS. We added this to the limitations in the Discussion. THIS SENTENCE/ARGUMENTATION IS NOT PROPER: „THE OMITTION OF THE MULTIPLE COMPUTATION CAN NOT BE JUSTIFIED BY THE SIZE OF DATA (SIC!)" – RATHER DATABASE. FURTHERMORE, THE IMPACT OF THE

	CLASSIFICATION BIAS ON THE FINAL CONCLUSION IS NOT DISCUSSED YET. 3. CONTINUOUS VARIABLES (AGE, LOS) HAVE OBVIOUSLY NOT NORMAL DISTRIBUTION (PAGE 8, LINES 27-60). THE MEAN AND SD SEEMS TO BE NOT PROPER SUMMARY MEASURES TO DESCRIBE THE DISTRIBUTION OF OBSERVED DATA. We agree that the median can be a good summary statistic for asymetric data. However, it is important for us to use the mean, because by multiplying LOS by the number of patients we can estimate total bed utilization. This is a meaningful information for hospital managers. IT IS NOT A MAJOR ISSUE, BUT… ALTHOUGH, MEAN VALUE AND STANDARD DEVIATION CAN BE CALCULATED FOR VARIABLE WITH NON-NORMAL DISTRIBUTION, BUT THESE HAVE NO MEANING AT ALL. IN DESCRIPTION, THE MEDIAN WITH INTER-QUARTILE RANGE CAN BE CORRECT AND INFORMATIVE. UNFORTUNATELY, AUTHORS DID NOT ANSWER TO THE QUESTION RELATED TO THE NORMALITY OF THEIR DATA. THE OBSERVED DISTRIBUTIONS ARE NOT NORMAL OBVIOUSLY – IT IS ADMITTED BY AUTHORS. IN THIS CASE THE USE OF MEAN±SD AS SUMMARY STATISTICS IS NOT ALLOWED; AND THE USE OF STUDENT T-TEST TO COMPARE THE CALCULATED MEAN VALUES IS ALSO MISLEADING. RESULTS PRESENTED IN TABLE 2 AND THE COMPUTATION BEHIND HAVE TO BE CORRECTED. 4. THE STRUCTURE OF TABLE 3 IS PROPER. BUT THE P VALUES ARE REPORTED IN A NOT CONSISTENT MANNER: SOMETIMES WITH 3 DECIMALS, SOMETIMES WITH 4 DECIMALS. THE USUAL WAY WITH 3 DECIMALS SEEMS TO APPROPRIATE IN THIS TABLE. We changed the p values from 4 decimals to 3 decimals. PROPER MODIFICATION FOOTNOTES HAVE TO BE ADDED TO THE TABLE ON (A) EXACT DATES OF THE FIRST AND THE SECOND PERIODS, AND ON (B) THE NAME OF TEST TO CALCULATE THE P-VALUES (HOPEFULLY, A TEST WHICH WAS ABLE TO EVALUATE THE NOT NORMALLY DISTRIBUTED DATA). We added the footnotes for exact dates of both periods. We also added footnotes for the names of tests to calculate p-values. NB: Because the number of patients in our study was very high, the Central Limit Theorem suggests we could use these tests. This can be verified using a bootstrap estimation of the sample mean's distribution. REGARDING FOOTNOTES: PROPER MODIFICATION REGARDING STATISTICAL TESTS: THIS EXPLANATION SHOULD BE WRITTEN NOT ONLY FOR REVIEWERS. IT IS MORE IMPORTANT TO BE PRESENTED FOR READERS! We added a paragraph in the methods section describing the tests and their validity conditions. WHY NOT TO PRESENT RESULTS OF BOOTSTRAP ESTIMATION AS IT WAS MENTIONED IN THE FORMER ANSWER? (WHY NOT TO USE A NON-PARAMETRIC APPROACH TO TEST THE DIFFERENCE OF MEDIAN VALUES?)

5. THERE ARE DISEASE SPECIFIC DESCRIPTIVE STATISTICS IN THE RESULTS SECTION (PAGE 8, LINES 42-60), BUT THIS ANALYSIS IS NOT MENTIONED IN THE METHODS SECTION.

We added a mention of this analysis in the Methods section.

PROPER MODIFICATION

Regarding regression modeling:

1.

The difference-in-difference analysis by logistic regression modelling with interaction term for time and exposure is proper method to answer the study questions.

2.

Definition of the regression model (page 6; line 40) is good. Cited references are proper.

3.

TABLE 4 ON RESULTS OF MODELING HAS APPROPRIATE STRUCTURE. IT CONTAINS P-VALUES REPORTED WITH 3 DECIMALS AND 4 DECIMALS. IT IS TO BE CORRECTED. FURTHERMORE, REPORTING P-VALUES AND 95% CONFIDENCE INTERVALS ARE REDUNDANT. CONSIDERING THAT P-VALUES ARE NOT INFORMATIVE ABOUT THE SIZE OF THE EFFECT, AND DUE TO THE BIG NUMBERS ANALYZED THE STATISTICAL SIGNIFICANT EFFECTS ARE NOT NECESSARILY CLINICALLY IMPORTANT, REPORTING ONLY THE 95% CONFIDENCE INTERVAL IS INFORMATIVE ENOUGH

We removed the p-values.

PROPER MODIFICATION

4.

RESULTS OF REGRESSION MODELS ARE REPORTED IN TABLE 4 NEEDS SOME EXPLANATION, SINCE THE TABLE DOES NOT CONTAIN THE ODDS RATIOS FOR THE PERIOD AND FOR THE LOCATION; WE CAN SEE ONLY THE CALCULATED MEASURES FOR INTERACTION TERM. WHY?

People often misinterpret the meaning of main effects in the presence of interactions. In this case, we were interested in the intervention effect in the center where the intervention was implemented. This was modelled with the interaction term, which was the quantity of interest. We were less interested in the effect of a particular center, which is why we did not report these coefficients.

IF THE TESTED MODEL WAS APPLIED WITHOUT INSERTING LOCATION AND PERIOD AS EXPLANATORY VARIABLES THEN THE MODEL IS NOT APPROPRIATE. IF THESE VARIABLES WERE INCLUDED IN THE MODEL THEN THE RESULTS FOR THEM SHOULD BE ADDED TO THE TABLE.

These variables were included in the model simultaneously with the interaction term. We added the results in the table.

REGARDING LOCATION: THE LOCATION IS SIGNIFICANT FACTOR FOR BOTH MODELS PRESENTED IN TABLE 4 (ACCORDING TO THE ORIGINAL VERSION). IT MEANS THAT THERE WERE SIGNIFICANT DIFFERENCES BETWEEN INTERVENTION AND CONTROL HOSPITALS. IT NEEDS EXPLICIT DISCUSSION WHETHER THE CONTROL HOSPITALS WERE USEFUL FOR THE DIFFERENCE-IN-DIFFERENCE ANALYSIS.

We added these points to the discussion.

I CAN READ IN THE ANSWER THAT ("The control hospital was included because it was in the same region and was of the same size as the intervention hospital, however there were differences

	regarding the functioning of their emergency departments. Stays in the intervention hospital were more likely to last ≥ 4 hours than in the control hospital as is shown by the location-specific odds ratios.") AUTHORS ADMIT THE INAPPROPRIATENESS OF THE CONTROL. 5. THE REGRESSION MODEL CONTAINED 7-DAY READMISSION AS VARIABLE TO CONTROL FOR THE QUALITY OF CARE. THERE ARE 3 INDICATORS USED FOR QUALITY OF CARE DESCRIPTION (7-DAY READMISSION, 30-DAY READMISSION, NUMBER OF PATIENTS LEAVING WITHOUT BEING SEEN) IN THE PAPER. IT IS NOT WRITTEN EXPLICITLY WHAT WAS THE REASON FOR USING 7-DAY READMISSION ONLY IN MODELING. Only the 7-day readmissions variable was kept in the model because early readmissions could be more relevant than late readmissions from the hospital's point of view. The nature of 7-day readmissions could be different from 30-day readmissions. For example: Graham KL, Auerbach AD, Schnipper JL, et al. Preventability of Early Versus Late Hospital Readmissions in a National Cohort of General Medicine Patients. Ann Intern Med 2018;168:766–74. doi:10.7326/M17- 1724. We added this reference to the article. THIS ANSWER IS NOT STATISTICAL - IT CAN BE PROPER MODIFICATION I CANNOT UNDERSTAND (AND IT IS NOT WRITTEN IN THE MANUSCRIPT) WHY WERE USED 3 INDICATORS IN DESCRIPTION AND ONLY 1 INDICATOR IN FURTHER STATISTICAL ANALYSIS. 6. IT IS WRITTEN IN PAGE 10-LINE 3 THAT EXPONENTIATED DIFFERENCE-INDIFFERENCE WAS ESTIMATED, AND THE REPORTED ODDS RATIO IS THE SAME SHOWN IN TABLE 4. IT IS NOT MENTIONED IN THE METHODS THAT THE PARAMETERS HAD TRANSFORMED BEFORE MODELING TO CORRECT THE LACK OF NORMALITY, AND IT IS NOT DECLARED IN THE TITLE OF TABLE 4 THAT SOME VARIABLES WERE TRANSFORMED. We added a sentence to explain this in the Methods section : "To facilitate modelling, length of stay was transformed in a binary variable using thresholds classically found in the litterature." Reference: Khanna S, Boyle J, Good N, et al. New emergency department quality measure: from access block to National Emergency Access Target compliance. Emerg Med Australas EMA 2013;25:565–72. doi:10.1111/1742-6723.12139 We verified that all variables in the table were correctly identified as categorical. PROPER MODIFICATION 7. CRITERIA OF APPLYING DIFFERENCE-IN-DIFFERENCE ANALYSIS ARE NOT EVALUATED PROPERLY. (A) THE SIMILAR BEFORE-INTERVENTION TRENDS IN THE STUDIED HOSPITALS ARE NOT CHECKED PROPERLY. STATISTICAL EVALUATION OF BEFORE-INTERVENTION-TRENDS IN BOTH HOSPITALS NEEDED (THE GRAPHIC PRESENTATION IS NOT ENOUGH.) Thank you for raising this point. Random fluctuations could make the common trends test indicate a significant effect, whereas from the organizational point of view, no major changes took place in the hospitals. This could be the reason why common trends are

	rarely tested in the litterature. As applied researchers, we are most confident when we measure statistical indicators of the effect of interventions backed by theoretical and empirical evidence. We cannot provide a better evaluation of the common trend hypothesis. (B) THE CRITERION OF COMMON SHOCKS IS VIOLATED AS IT IS ACKNOWLEDGED BY AUTHORS (PAGE 15, LINES 36-45). WITHOUT DEMONSTRATING THAT THIS PROBLEM DOES NOT JEOPARDIZE THE VALIDITY, RESULTS ARE NOT CONVINCING AND ARE NOT ABLE TO ANSWER THE STUDY QUESTIONS. We moved this sentence to the end of the discussion. We are not aware of any epidemic that could have affected the population served by one hospital without affecting the other during the study period. Although the design we used relies on assumptions about the intervention and control hospitals which are difficult to assess, we believe it conveys more information than a before-after study restricted to the intervention hospital. THE DIFFERENCE-IN-DIFFERENCE ANALYSIS IS ABLE TO PRODUCE MORE CONVINCING EVIDENCES THAN THE BEFORE-AFTER ANALYSIS WHEN THE CRITERIA FOR ITS APPLICATION ARE MET. IF THESE CRITERIA ARE NOT MET OR AUTHORS HAVE NO DATA TO DEMONSTRATE THAT THOSE ARE MET, THEN THIS STATISTICAL APPROACH IS NOT ESTABLISHED. ACCORDING TO THE ANSWERS: (A) THE SIMILAR BEFORE-INTERVENTION TREND IS NOT MET (LOCATION SPECIFIC ORs INSERTED INTO TABLE 4) (B) THE COMMON SHOCK IS VIOLATED AS IT IS ACKNOWLEDGED BY AUTHORS. AUTHORS COULD NOT ANSWER THE VALIDITY RELATED QUESTIONS. UNTIL THEY CAN DO IT, THE CONCLUSIONS ARE NOT CONVINCINGLY SUPPORTED BY THE PRESENTED RESULTS. We added these points to the discussion. In the current state of the article, the reader can refer to the descriptive tables for a before-after analysis and to a complementary difference-in-differences analysis for which the reader can consider the limitations underlined in the discussion when interpreting the results. However, we can remove the difference-in-differences analysis if needed. WITHOUT EXPLICIT DISCUSSION OF THE IMPACT OF VIOLATION OF COMMON SHOCK PRE-REQUIREMENT ON TH E REPORTED STATISTICAL RESULTS, RESULTS FROM THE DIFFERENCE-IN-DIFFERENCE ANALYSIS CANNOT BE USED TO DRAW CONCLUSIONS.

VERSION 3 – AUTHOR RESPONSE

Reviewer: 3

Reviewer Name: János Sándor

Institution and Country: University of Debrecen, Faculty of Public Health, Department of Preventive Medicine, Hungary

Please state any competing interests or state ‘None declared’: None declared.

Please leave your comments for the authors below

My new comment/questions are in capitals-italics at the end of numbered sections.

Unfortunately the main criticisms on the appropriateness of statistical methods and on the lack of validity analysis have been not addressed by authors. Some of my notions are not answered at all. If the Editor needs further peer-review for this manuscript then, please, look for an other reviewer.

In this revision we have adressed all the issues discussed by the reviewer. The reviewer's concerns about underlying hypotheses led us to abandon the difference-in-differences analysis.

Regarding descriptive statistics:

1.

TIME OF THE INTERVENTION IS NOT REPORTED IN THE TEXT.

We added the time of the intervention in the Methods section.

PROPER MODIFICATION

2.

IT IS WRITTEN THAT PS CLASSIFICATION BY DATA WAS NOT POSSIBLE IN ALL RECORDS. SOMETIMES EXPERT OPINION WAS THE BASE OF CATEGORIZATION. IT IS NOT DECLARED HOW FREQUENT THIS SECONDARY APPROACH WAS.

This was the case in 11.9 % of cases. We added this in the methods section of the revised manuscript.

PROPER MODIFICATION, BUT I MISS THE DISCUSSION OF THE INFLUENCE OF THIS VALIDITY PROBLEM ON THE FINAL RESULTS. FURTHER, WHY WAS NOT APPLIED A STATISTICAL METHOD TO ESTIMATE MISSING DATA (E.G.: MULTIPLE IMPUTATION)?

We added a paragraph regarding missing data in the Discussion.

We agree that multiple imputation is a good method for treating missing data. However, it was difficult for us to carry out multiple imputation on this large dataset.

IT SHOULD BE ACKNOWLEDGED AMONG LIMITATIONS.

We added this to the limitations in the Discussion.

THIS SENTENCE/ARGUMENTATION IS NOT PROPER: „THE OMITTION OF THE MULTIPLE COMPUTATION CAN NOT BE JUSTIFIED BY THE SIZE OF DATA (SIC!)" – RATHER DATABASE. FURTHERMORE, THE IMPACT OF THE CLASSIFICATION BIAS ON THE FINAL CONCLUSION IS NOT DISCUSSED YET.

We added a new paragraph regarding the effect of missing PS classification. We performed a multiple imputation to quantify this effect as requested by the reviewer. Because the classification was missing primarily in patients who left without being seen, and these patients had a lower length of stay and were more frequent in the first period, by keeping them in the model the effect of our intervention was underestimated. After careful thought we considered that it was better to keep these patients in the model, thus accepting the risk of underestimating the effect of the intervention.

3.

CONTINUOUS VARIABLES (AGE, LOS) HAVE OBVIOUSLY NOT NORMAL DISTRIBUTION (PAGE 8, LINES 27-60). THE MEAN AND SD SEEMS TO BE NOT PROPER SUMMARY MEASURES TO DESCRIBE THE DISTRIBUTION OF OBSERVED DATA.

We agree that the median can be a good summary statistic for asymetric data. However, it is important for us to use the mean, because by multiplying LOS by the number of patients we can estimate total bed utilization. This is a meaningful information for hospital managers.

IT IS NOT A MAJOR ISSUE, BUT… ALTHOUGH, MEAN VALUE AND STANDARD DEVIATION CAN BE CALCULATED FOR VARIABLE WITH NON-NORMAL DISTRIBUTION, BUT THESE HAVE NO MEANING AT ALL. IN DESCRIPTION, THE MEDIAN WITH INTER-QUARTILE RANGE CAN BE CORRECT AND INFORMATIVE.

UNFORTUNATELY, AUTHORS DID NOT ANSWER TO THE QUESTION RELATED TO THE NORMALITY OF THEIR DATA. THE OBSERVED DISTRIBUTIONS ARE NOT NORMAL OBVIOUSLY – IT IS ADMITTED BY AUTHORS. IN THIS CASE THE USE OF MEAN±SD AS SUMMARY STATISTICS IS NOT ALLOWED; AND THE USE OF STUDENT T-TEST TO COMPARE THE CALCULATED MEAN VALUES IS ALSO MISLEADING. RESULTS PRESENTED IN TABLE 2 AND THE COMPUTATION BEHIND HAVE TO BE CORRECTED.

We replaced means by medians with the interquartile range and performed a non-parametric test (the Wilcoxon-Mann-Whitney U test) as requested by the reviewer.

4.

THE STRUCTURE OF TABLE 3 IS PROPER. BUT THE P VALUES ARE REPORTED IN A NOT CONSISTENT MANNER: SOMETIMES WITH 3 DECIMALS, SOMETIMES WITH 4 DECIMALS. THE USUAL WAY WITH 3 DECIMALS SEEMS TO APPROPRIATE IN THIS TABLE.

We changed the p values from 4 decimals to 3 decimals.

PROPER MODIFICATION FOOTNOTES HAVE TO BE ADDED TO THE TABLE ON (A) EXACT DATES OF THE FIRST AND THE SECOND PERIODS, AND ON (B) THE NAME OF TEST TO CALCULATE THE P-VALUES (HOPEFULLY, A TEST WHICH WAS ABLE TO EVALUATE THE NOT NORMALLY DISTRIBUTED DATA).

We added the footnotes for exact dates of both periods. We also added footnotes for the names of tests to calculate p-values. NB: Because the number of patients in our study was very high, the Central Limit Theorem suggests we could use these tests. This can be verified using a bootstrap estimation of the sample mean’s distribution.

REGARDING FOOTNOTES: PROPER MODIFICATION REGARDING STATISTICAL TESTS: THIS EXPLANATION SHOULD BE WRITTEN NOT ONLY FOR REVIEWERS. IT IS MORE IMPORTANT TO BE PRESENTED FOR READERS!

We added a paragraph in the methods section describing the tests and their validity conditions.

WHY NOT TO PRESENT RESULTS OF BOOTSTRAP ESTIMATION AS IT WAS MENTIONED IN THE FORMER ANSWER? (WHY NOT TO USE A NON-PARAMETRIC APPROACH TO TEST THE DIFFERENCE OF MEDIAN VALUES?)

We changed the test to a non-parametric test appropriate for asymetric distributions.

5. THERE ARE DISEASE SPECIFIC DESCRIPTIVE STATISTICS IN THE RESULTS SECTION (PAGE 8, LINES 42-60), BUT THIS ANALYSIS IS NOT MENTIONED IN THE METHODS SECTION.

We added a mention of this analysis in the Methods section.

PROPER MODIFICATION

Regarding regression modeling:

1.

The difference-in-difference analysis by logistic regression modelling with interaction term for time and exposure is proper method to answer the study questions.

2.

Definition of the regression model (page 6; line 40) is good. Cited references are proper.

3.

TABLE 4 ON RESULTS OF MODELING HAS APPROPRIATE STRUCTURE. IT CONTAINS P-VALUES REPORTED WITH 3 DECIMALS AND 4 DECIMALS. IT IS TO BE CORRECTED. FURTHERMORE, REPORTING P-VALUES AND 95% CONFIDENCE INTERVALS ARE REDUNDANT. CONSIDERING THAT P-VALUES ARE NOT INFORMATIVE ABOUT THE SIZE OF THE EFFECT, AND DUE TO THE BIG NUMBERS ANALYZED THE STATISTICAL SIGNIFICANT EFFECTS ARE NOT NECESSARILY CLINICALLY IMPORTANT, REPORTING ONLY THE 95% CONFIDENCE INTERVAL IS INFORMATIVE ENOUGH

We removed the p-values.

PROPER MODIFICATION

4.

RESULTS OF REGRESSION MODELS ARE REPORTED IN TABLE 4 NEEDS SOME EXPLANATION, SINCE THE TABLE DOES NOT CONTAIN THE ODDS RATIOS FOR THE PERIOD AND FOR THE LOCATION; WE CAN SEE ONLY THE CALCULATED MEASURES FOR INTERACTION TERM. WHY?

People often misinterpret the meaning of main effects in the presence of interactions. In this case, we were interested in the intervention effect in the center where the intervention was implemented. This was modelled with the interaction term, which was the quantity of interest. We were less interested in the effect of a particular center, which is why we did not report these coefficients.

IF THE TESTED MODEL WAS APPLIED WITHOUT INSERTING LOCATION AND PERIOD AS EXPLANATORY VARIABLES THEN THE MODEL IS NOT APPROPRIATE. IF THESE VARIABLES WERE INCLUDED IN THE MODEL THEN THE RESULTS FOR THEM SHOULD BE ADDED TO THE TABLE.

These variables were included in the model simultaneously with the interaction term. We added the results in the table.

REGARDING LOCATION: THE LOCATION IS SIGNIFICANT FACTOR FOR BOTH MODELS PRESENTED IN TABLE 4 (ACCORDING TO THE ORIGINAL VERSION). IT MEANS THAT THERE WERE SIGNIFICANT DIFFERENCES BETWEEN INTERVENTION AND CONTROL HOSPITALS. IT NEEDS EXPLICIT DISCUSSION WHETHER THE CONTROL HOSPITALS WERE USEFUL FOR THE DIFFERENCE-IN-DIFFERENCE ANALYSIS.

We added these points to the discussion.

I CAN READ IN THE ANSWER THAT ("The control hospital was included because it was in the same region and was of the same size as the intervention hospital, however there were differences regarding the functioning of their emergency departments. Stays in the intervention hospital were more likely to last ≥ 4 hours than in the control hospital as is shown by the location-specific odds ratios.") AUTHORS ADMIT THE INAPPROPRIATENESS OF THE CONTROL.

To answer the reviewer's comment, we changed the design of the article, and removed the control group. We performed a new analysis based on a before-after model fitted by logistic regression. All relevant sections, Tables and Figures were changed accordingly.

5.

THE REGRESSION MODEL CONTAINED 7-DAY READMISSION AS VARIABLE TO CONTROL FOR THE QUALITY OF CARE. THERE ARE 3 INDICATORS USED FOR QUALITY OF CARE DESCRIPTION (7-DAY READMISSION, 30-DAY READMISSION, NUMBER OF PATIENTS LEAVING WITHOUT BEING SEEN) IN THE PAPER. IT IS NOT WRITTEN EXPLICITLY WHAT WAS THE REASON FOR USING 7-DAY READMISSION ONLY IN MODELING.

Only the 7-day readmissions variable was kept in the model because early readmissions could be more relevant than late readmissions from the hospital's point of view. The nature of 7-day readmissions could be different from 30-day readmissions. For example: Graham KL, Auerbach AD, Schnipper JL, et al. Preventability of Early Versus Late Hospital Readmissions in a National Cohort of General Medicine Patients. Ann Intern Med 2018;168:766–74. doi:10.7326/M17- 1724. We added this reference to the article.

THIS ANSWER IS NOT STATISTICAL - IT CAN BE PROPER MODIFICATION

We added a new indicator variable in the model for readmissions that took place from the 8th day to the 30th day.

I CANNOT UNDERSTAND (AND IT IS NOT WRITTEN IN THE MANUSCRIPT) WHY WERE USED 3 INDICATORS IN DESCRIPTION AND ONLY 1 INDICATOR IN FURTHER STATISTICAL ANALYSIS.

In the revised version all readmissions are considered in the multivariable analysis.

We added in the manuscript: “The proportion of patients leaving without being seen was a secondary outcome. Due to the scarcity of information on these patients, it was not considered an independent variable for multivariable analysis”.

6.

IT IS WRITTEN IN PAGE 10-LINE 3 THAT EXPONENTIATED DIFFERENCE-INDIFFERENCE WAS ESTIMATED, AND THE REPORTED ODDS RATIO IS THE SAME SHOWN IN TABLE 4. IT IS NOT MENTIONED IN THE METHODS THAT THE PARAMETERS HAD TRANSFORMED BEFORE MODELING TO CORRECT THE LACK OF NORMALITY, AND IT IS NOT DECLARED IN THE TITLE OF TABLE 4 THAT SOME VARIABLES WERE TRANSFORMED.

We added a sentence to explain this in the Methods section : “To facilitate modelling, length of stay was transformed in a binary variable using thresholds classically found in the litterature.” Reference: Khanna S, Boyle J, Good N, et al. New emergency department quality measure: from access block to National Emergency Access Target compliance. Emerg Med Australas EMA 2013;25:565–72. doi:10.1111/1742-6723.12139 We verified that all variables in the table were correctly identified as categorical.

PROPER MODIFICATION

7.

CRITERIA OF APPLYING DIFFERENCE-IN-DIFFERENCE ANALYSIS ARE NOT EVALUATED PROPERLY.

We removed the difference-in-differences model as suggested by the reviewer. All analyses are now performed in the intervention hospital.

(A) THE SIMILAR BEFORE-INTERVENTION TRENDS IN THE STUDIED HOSPITALS ARE NOT CHECKED PROPERLY. STATISTICAL EVALUATION OF BEFORE-INTERVENTION-TRENDS IN BOTH HOSPITALS NEEDED (THE GRAPHIC PRESENTATION IS NOT ENOUGH.)

Thank you for raising this point. Random fluctuations could make the common trends test indicate a significant effect, whereas from the organizational point of view, no major changes took place in the hospitals. This could be the reason why common trends are rarely tested in the litterature. As applied researchers, we are most confident when we measure statistical indicators of the effect of interventions backed by theoretical and empirical evidence. We cannot provide a better evaluation of the common trend hypothesis.

(B) THE CRITERION OF COMMON SHOCKS IS VIOLATED AS IT IS ACKNOWLEDGED BY AUTHORS (PAGE 15, LINES 36-45). WITHOUT DEMONSTRATING THAT THIS PROBLEM DOES NOT JEOPARDIZE THE VALIDITY, RESULTS ARE NOT CONVINCING AND ARE NOT ABLE TO ANSWER THE STUDY QUESTIONS.

We moved this sentence to the end of the discussion. We are not aware of any epidemic that could have affected the population served by one hospital without affecting the other during the study period. Although the design we used relies on assumptions about the intervention and control hospitals which are difficult to assess, we believe it conveys more information than a before-after study restricted to the intervention hospital.

THE DIFFERENCE-IN-DIFFERENCE ANALYSIS IS ABLE TO PRODUCE MORE CONVINCING EVIDENCES THAN THE BEFORE-AFTER ANALYSIS WHEN THE CRITERIA FOR ITS APPLICATION ARE MET. IF THESE CRITERIA ARE NOT MET OR AUTHORS HAVE NO DATA TO DEMONSTRATE THAT THOSE ARE MET, THEN THIS STATISTICAL APPROACH IS NOT ESTABLISHED. ACCORDING TO THE ANSWERS: (A) THE SIMILAR BEFORE-INTERVENTION TREND IS NOT MET (LOCATION SPECIFIC ORs INSERTED INTO TABLE 4) (B) THE COMMON SHOCK IS VIOLATED AS IT IS ACKNOWLEDGED BY AUTHORS. AUTHORS COULD NOT ANSWER THE VALIDITY RELATED QUESTIONS. UNTIL THEY CAN DO IT, THE CONCLUSIONS ARE NOT CONVINCINGLY SUPPORTED BY THE PRESENTED RESULTS.

We added these points to the discussion. In the current state of the article, the reader can refer to the descriptive tables for a before-after analysis and to a complementary difference-in-differences analysis for which the reader can consider the limitations underlined in the discussion when interpreting the results. However, we can remove the difference-in-differences analysis if needed.

WITHOUT EXPLICIT DISCUSSION OF THE IMPACT OF VIOLATION OF COMMON SHOCK PRE-REQUIREMENT ON TH E REPORTED STATISTICAL RESULTS, RESULTS FROM THE DIFFERENCE-IN-DIFFERENCE ANALYSIS CANNOT BE USED TO DRAW CONCLUSIONS.