PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Assessing the association of oxytocin augmentation with obstetric anal sphincter injury in nulliparous women – a population-based, case-control study
AUTHORS	Rygh, Astrid; Skjeldestad, Finn Egil; Körner, Hartwig; Eggebø, Torbjørn

VERSION 1 - REVIEW

REVIEWER	Mika Gissler THL National Institute for Health and Welfare, Finland I am currently participating in a Nordic study on Caesarean section, led by Dr Finn Egil Skjeldestad. I have no other conflict of interest to report.
REVIEW RETURNED	13-Jan-2014

GENERAL COMMENTS	The article uses local hospital register to explore the connection between oxytocin augmentation and obstetric anal sphincter injury among nulliparous women with singleton births. The topic of this article is interesting and the study is well done. It shows that oxytocin augmentation increases the risk of obstetric anal sphincter injury, even though the causality cannot be proved in the current study design. My main comments are related to methods: - Why did the authors re-code missing variables into the reference category? Their numbers are small, so I would either remove them or consider using imputation methods for the estimates. - Since the incidence of obstetric anal sphincter injury decreased significantly during the study period, birth year should be included in the modelling. - The last sentence of results of modelling could be better placed in Methods. - In discussion, it could be mentioned that the study data are based on one hospital only. A further limitation is that there are no data on socioeconomic status, BMI or smoking. I have few minor additional comments: - Page 6: How did the authors assess ethnicity? - Table 1: Use full stop instead of comma systematically. The last category for active second stage of labour could be 60 and more, similarly as for example for maternal age. - Add confidence intervals also in unadjusted ORs in Table 3. - The ORs and CIs should be given systematically with one or two decimals.

REVIEWER	Philip Steer Imperial College London UK
REVIEW RETURNED	16-Jan-2014

GENERAL COMMENTS	1. The basic problem with this paper is that the causation of anal sphincter injury is multifactorial, and it is very difficult (as the authors state) to determine causality rather than association when one is considering only an observational study. The authors say that "a randomised controlled study would not be feasible" but the Cochrane database records no less than 14 randomised trials of augmentation with amniotomy/oxytocin, the most recent being in 2011. Altogether, more than 8000 women have been enrolled in such prospective randomised controlled trials. As the authors state, oxytocin augmentation has not been shown to decrease the need for operative deliveries, and it only shortens labour by 1-2 hours. In view of the doubtful benefits from augmentation of labour, further trials are clearly indicated and therefore the authors' statement cannot be correct. 2. The rate of third-degree tears declined substantially, from 9.6% in 1999-2000 to 2.8% in 2010-2012. The authors need to explain which variables were related to this decrease. Did the rate of oxytocin augmentation change during this period? 3. It is known that a major factor influencing the incidence of third degree tears is birthweight, large babies give rise to a higher incidence of third degree tears. Larger babies are also associated with longer labours, both the first and second stages, and the need for instrumental deliveries. Longer labours also give rise to an increased use of oxytocin augmentation. I would therefore expect any analysis of the effect of oxytocin to start with correction for birthweight, followed by correction for duration of labour, followed by correction for the presence or absence of instrumental delivery. Only once these factors had been adjusted for, would it be worth looking for any independent effect of oxytocin augmentation itself. Rather than analyse the data in this sequential way, the authors chose a stratified analysis, a technique with which I am not familiar, and I would suggest that a statistician should advise as to whether this is appropriate. The authors say on page 14 that "without testing for possible interactions, multivariable regression models, e.g. entering all variables simultaneously, would fail to reveal this information". By "this information" they mean "understanding how oxytocin augmentation interacts with other major risk factors". In fact, the hypothesis they are examining is the relationship of oxytocin augmentation with third-degree tears, and the other major risk factors are potential confounders. I would have thought that by entering the explanatory variables one by one, in the order in which it seems plausible that they would act, would be more appropriate approach. For example, the interaction of oxytocin augmentation with birthweight cannot be causal; rather, birthweight is likely to influence the use of oxytocin augmentation. Therefore, birthweight needs to be a primary confounder which is entered first in the logistic regression. Moreover, the authors have only corrected for birthweight above and below 4000 g, while I would expect the influence of birthweight be continuous across birthweights, even with birthweights below 4000 g. It is not clear to me why the authors have used a dichotomous approach rather than a technique treating birthweight as a continuum.

	4. Although the authors have accepted that they can only show associations rather than causation, they repeatedly use terminology which implies causation, for example on page 11 they say "an epidural reduced the risk of obstetric anal sphincter injury by 30%". More correctly they should say "an epidural was associated with a 30% lower incidence of anal sphincter injury". In the abstract they should say "to assess the association of oxytocin augmentation with obstetric anal sphincter injury among nulliparous women", not "to assess the effect" (which would require a prospective randomised trial). 5. Throughout the paper the authors use the word "risk" when they are actually describing events which have already occurred, and they should therefore be using the term "incidence". "Risk" is used to define the possibility that something may happen in the future, and should not be used to describe events which have already occurred. 6. On page 13, the authors say that they consider the duration of the active second stage of labour to be a proxy for oxytocin augmentation and instrumental delivery. This seems illogical to me, I consider the main associations influencing the duration of the second stage to be fetal size and the position of the fetal head; oxytocin augmentation, duration of the second stage and instrumental delivery are all likely to be a consequence of the baby being large. 7. In the discussion, the authors recommend the use of the partogram and say that this could be helpful; however the recent Cochrane review which the authors themselves quote (Lavender T, Hart A, Smyth RM. Effect of partogram use on outcomes for women in spontaneous labour at term. Cochrane Database Syst Rev 2013;7:CD005461) finds that the use of a partogram does not improve the outcome of labour – “On the basis of the findings of this review, we cannot recommend routine use of the partogram as part of standard labour management and care.”. 8. The authors suggest that fetal weight estimation by ultrasound may be considered when macrosomia is suspected; this is also against current recommendations because ultrasound estimates of the weight of large babies are not reliable, and in any case, the weight of the baby cannot be modified in any way. Minor comments: 1. P4 - Risk factors are prospective, therefore ‘primiparity’ should be ‘nulliparity’. 2. 'Caucasian' should be replaced by white European. Caucasian is poorly defined and non-specific, originating from the term for a racial group which included peoples from North India, Middle East, North Africa and Europe. It is also out of date: in customised growth charts, a classification by geographical ethnic origin (GEO) has been used for some time *www.pi.nhs.uk/manners/newsletter0306.htm* . Similarly, the National Library of Medicine replaced in 2004 racial MeSH headings on Medline citations with ethnic groups by geographical origin: *www.nlm.nih.gov/pubs/techbull/nd03/nd03_med_data_changes.html* A ‘special communication’ in JAMA 2003 (Kaplan JB and Bennett T, JAMA. 2003;289:2709-2716) commented that “At a minimum,

	journals should follow the guidance of the Council of Science Editors in prohibiting use of the term “Caucasian,” which, like “Caucasoid,” “Mongoloid,” and “Negroid,” is “based on an outmoded theory of racial distinction,” and requiring use of the term “Asian” instead of “Oriental” or “Asiatic.”

VERSION 1 – AUTHOR RESPONSE

Reviewer Name: Mika Gissler

1) *The article uses local hospital register to explore the connection between oxytocin augmentation and obstetric anal sphincter injury among nulliparous women with singleton births. The topic of this article is interesting and the study is well done. It shows that oxytocin augmentation increases the risk of obstetric anal sphincter injury, even though the causality cannot be proved in the current study design.*
   R: We appreciate the positive feedback on our study. We have changed “causality/risk” to associations/odds ratio throughout the ms.
2) *Why did the authors re-code missing variables into the reference category?*
   Their numbers are small, so I would either remove them or consider using imputation methods for the estimates.
   R: We have reviewed medical records for women with missing information and updated the database accordingly. This changed the size of the study group from 15 493 to 15 476.

3) *Since the incidence of obstetric anal sphincter injury decreased significantly during the study period, birth year should be included in the modelling.*
   R: We have not included time-period or year in the modelling. We will address this issue in a separate study. The risk factors for anal sphincter injury remained significant across all time-periods, but the strength of the estimates changed by changing medical practise.

4) *The last sentence of results of modelling could be better placed in Methods.*
   R: This sentence in Results is now changed into a descriptive one, and the issue is further elaborated on in Discussion.

5) *In discussion, it could be mentioned that the study data are based on one hospital only. A further limitation is that there are no data on socioeconomic status, BMI or smoking.*
   R: This issue is underlined in the Discussion part as a limitation of the study

I have few minor additional comments:

6) *Page 6: How did the authors assess ethnicity?*
   R: We have made changes in accordance with comments from Philip Steer (minor comments 4) and classified the population into originating from Europe or North America (Western) or not.

7) *Table 1: Use full stop instead of comma systematically. The last category for active second stage of labour could be 60 and more, similarly as for example for maternal age.*
   R: Revised.

8) *Add confidence intervals also in unadjusted ORs in Table 3.*
   R: We considered this, but chose not to include CIs for the unadjusted ORs in order to ease the reading of the table. With unadjusted ORs in the table, we think the table will be overloaded with data. However, if the Editor wishes us to do so, this can easily be provided.

9) *The ORs and CIs should be given systematically with one or two decimals.*
   R: Revised to one decimal.

Reviewer Name: Philip Steer

10) *The basic problem with this paper is that the causation of anal sphincter injury is multifactorial,*

and it is very difficult (as the authors state) to determine causality rather than association when one is considering only an observational study. The authors say that "a randomised controlled study would not be feasible" but the Cochrane database records no less than 14 randomised trials of augmentation with amniotomy/oxytocin, the most recent being in 2011. Altogether, more than 8000 women have been enrolled in such prospective randomised controlled trials. As the authors state, oxytocin augmentation has not been shown to decrease the need for operative deliveries, and it only shortens labour by 1-2 hours. In view of the doubtful benefits from augmentation of labour, further trials are clearly indicated and therefore the authors' statement cannot be correct.
R: We agree, and state that RCT are needed, and propose anal sphincter injuries as an important endpoint

11) *The rate of third-degree tears declined substantially, from 9.6% in 1999-2000 to 2.8% in 2010-2012. The authors need to explain which variables were related to this decrease.*
Did the rate of oxytocin augmentation change during this period?
R: See issue 3 – response to Mika Gissler.
The use of oxytocin augmentation was lower in the last period (2010-2012) in accordance with new guidelines; however, we observed an association between oxytocin augmentation and sphincter ruptures through all time periods. We have included this information in the manuscript.

12) *It is known that a major factor influencing the incidence of third degree tears is birthweight, large babies give rise to a higher incidence of third degree tears. Larger babies are also associated with longer labours, both the first and second stages, and the need for instrumental deliveries. Longer labours also give rise to an increased use of oxytocin augmentation. I would therefore expect any analysis of the effect of oxytocin to start with correction for birthweight, followed by correction for duration of labour, followed by correction for the presence or absence of instrumental delivery. Only once these factors had been adjusted for, would it be worth looking for any independent effect of oxytocin augmentation itself. Rather than analyse the data in this sequential way, the authors chose a stratified analysis, a technique with which I am not familiar, and I would suggest that a statistician should advise as to whether this is appropriate. The authors say on page 14 that "without testing for possible interactions, multivariable regression models, e.g. entering all variables simultaneously, would fail to reveal this information". By "this information" they mean "understanding how oxytocin augmentation interacts with other major risk factors". In fact, the hypothesis they are examining is the relationship of oxytocin augmentation with third-degree tears, and the other major risk factors are potential confounders. I would have thought that by entering the explanatory variables one by one, in the order in which it seems plausible that they would act, would be more appropriate approach. For example, the interaction of oxytocin augmentation with birthweight cannot be causal; rather, birthweight is likely to influence the use of oxytocin augmentation. Therefore, birthweight needs to be a primary confounder which is entered first in the logistic regression. Moreover, the authors have only corrected for birthweight above and below 4000 g, while I would expect the influence of birthweight be continuous across birthweights, even with birthweights below 4000 g. It is not clear to me why the authors have used a dichotomous approach rather than a technique treating birthweight as a continuum.*
R: We started with a traditional forward stepwise logistic regression analysis from the "strongest" association from the univariate analysis. However, we found strong interactions from operative vaginal delivery and birth weight. We fitted a model with interactions terms and ended nearly up with table 3.
Then we started to rethink what is taking place in the last part of active second stage. The model we had created by doing the statistical "correct" approach by testing for interactions, and subsequently included the cross products of the interaction terms, gave us a deeper understanding of what actually takes place during the expulsion of the head. This lead us to table 2, providing estimates for all the 16 possible combinations of oxytocin augmentation, episiotomy, operative vaginal delivery and birth weight. From table 2, we collapsed groups that gave meaningful combinations of order of factors taking place in the last part of the second stage. Usually oxytocin augmentation is provided when the duration of second stage lasts longer or when the contractions are considered weak (step 1), episiotomy is done when the head is at the point of passing perineum (step 2), the forces of operative vaginal delivery are exerted on the sphincter complex when the head is passing the perineum (step 3), birthweight is most often

assessed when the baby is born (step 4). By this understanding of statistics and what is taking place in nature, we have created a model that mimic the real world. However, to our knowledge this approach of "solving" interactions terms into what is taking place during bearing down is not published before.

13) *Although the authors have accepted that they can only show associations rather than causation, they repeatedly use terminology which implies causation, for example on page 11 they say "an epidural reduced the risk of obstetric anal sphincter injury by 30%". More correctly they should say "an epidural was associated with a 30% lower incidence of anal sphincter injury". In the abstract they should say "to assess the association of oxytocin augmentation with obstetric anal sphincter injury among nulliparous women", not "to assess the effect" (which would require a prospective randomised trial).*
R: We agree, and have made changes accordingly in the title and in the manuscript.

14) *Throughout the paper the authors use the word "risk" when they are actually describing events which have already occurred, and they should therefore be using the term "incidence". "Risk" is used to define the possibility that something may happen in the future, and should not be used to describe events which have already occurred.*
R: We agree, and have replaced "risk" with odds ratio/association throughout the paper.

15) *On page 13, the authors say that they consider the duration of the active second stage of labour to be a proxy for oxytocin augmentation and instrumental delivery. This seems illogical to me, I consider the main associations influencing the duration of the second stage to be fetal size and the position of the fetal head; oxytocin augmentation, duration of the second stage and instrumental delivery are all likely to be a consequence of the baby being large.*
R: During a prolonged active second stage of delivery the head is usually positioned in the true pelvis above the pelvic floor most of the time. The head may have reached the pelvic floor, but is not exerting forces on the perineum. Then, when the head is crowning, breaking the perineal border, the expulsion of the head takes place, and tears may result. This is why we don't consider the duration of the active second phase as a risk factor, but as a latency period, when lacerations are considered. In the literature this is understood very differently from our understanding. The active second stage may last for hours, however, if the head is not passing the perineal border, tears will not occur. The provision of oxytocin augmentation in this phase may cause uncontrolled, strong contractions, and the administration of oxytocin augmentation is associated with a prolonged second stage, as well as with operative vaginal delivery. We found a strong colinearity between oxytocin augmentation and instrumental delivery and the duration of the active second stage of labour. The major "players" for exerting the forces when the head is passing the perineal border are operative vaginal delivery and oxytocin augmentation, not the "latent phase" of a prolonged active second stage!

16) *In the discussion, the authors recommend the use of the partogram and say that this could be helpful; however the recent Cochrane review which the authors themselves quote (Lavender T, Hart A, Smyth RM. Effect of partogram use on outcomes for women in spontaneous labour at term. Cochrane Database Syst Rev 2013;7:CD005461) finds that the use of a partogram does not improve the outcome of labour – "On the basis of the findings of this review, we cannot recommend routine use of the partogram as part of standard labour management and care."*
R: We agree, and have modified our comment accordingly (Discussion).

17) *The authors suggest that fetal weight estimation by ultrasound may be considered when macrosomia is suspected; this is also against current recommendations because ultrasound estimates of the weight of large babies are not reliable, and in any case, the weight of the baby cannot be modified in any way.*
R: We have changed our comment on birthweight estimation accordingly, and added a recent reference addressing the uncertainty of weight estimation of large babies. (Discussion)

Minor comments:

18) P4 - Risk factors are prospective, therefore 'primiparity' should be 'nulliparity'.
R: Revised.

19) *'Caucasian' should be replaced by white European. Caucasian is poorly defined and non-specific, originating from the term for a racial group which included peoples from North India, Middle East, North Africa and Europe. It is also out of date: in customised growth charts, a classification by geographical ethnic origin (GEO) has been used for some time www.pi.nhs.uk/manners/newsletter0306.htm . Similarly, the National Library of Medicine replaced in 2004 racial MeSH headings on Medline citations with ethnic groups by geographical origin: www.nlm.nih.gov/pubs/techbull/nd03/nd03_med_data_changes.html*
A 'special communication' in JAMA 2003 (Kaplan JB and Bennett T, JAMA. 2003;289:2709-2716) commented that "At a minimum, journals should follow the guidance of the Council of Science Editors in prohibiting use of the term "Caucasian," which, like "Caucasoid," "Mongoloid," and "Negroid," is "based on an outmoded theory of racial distinction," and requiring use of the term "Asian" instead of "Oriental" or "Asiatic."
R: Revised.

VERSION 2 – REVIEW

REVIEWER	Mika Gissler THL National Institute for Health and Welfare, Finland
REVIEW RETURNED	06-Mar-2014

GENERAL COMMENTS	The article is now ready for publication after the following minor issues have been corrected. 1. Since Tage Malmström was a Swede, his name should be spelled with ö. 2. non-Western instead of not Western (Table 1 and text). 3. Table 2: Some ORs are given in two decimals (risk groups 2, 3, 4). 4. The use of oxytocin was restricted in the department in 2010 or from 2010 onwards. 5. Collinearity instead of colinearity. 6. WHO spells its name with z: Organization.

REVIEWER	Philip Steer Imperial College London UK
REVIEW RETURNED	06-Mar-2014

GENERAL COMMENTS	The authors have addressed fully the points I made in my initial review. I cannot comment further on their statistical approach as it appears novel in this context (the authors state this themselves in their accompanying letter) and I am not a qualified statistician and so

	cannot comment further. I find the 'stratified' approach rather difficult to follow, but a statistician may decide it is appropriate. The authors have changed their description of ethnic origin to Western/non-Western, which is not a standard descriptive term for ethnic group. The authors describe Western as "originating from Europe or North America, or not" so I think what they mean is white European origin. If this is correct, then I think 'white European origin' would be a preferable term. It would be interesting to know what a statistician thinks of their 'stratified' rather than sequential approach to allowance for confounders. Persoanlly, I still see labour and delivery as sequential, rather than each stage being independent, but that may just be my unfamiliarity with this statistical approach.

REVIEWER	Peter Watson Medical Research Council UK
REVIEW RETURNED	28-Mar-2014

GENERAL COMMENTS	An easily implementable more understandable analysis for the interactions can be done. Instead of using what appears to be one single 16 level factor giving odds ratios as in Table 2 on page 8 we can perform the standard sequential approach to testing for interactions by constructing product terms and entering these after a block containing the main effects which constitute them (further details are below). The results including those in the abstract can then be re-written to reflect the output form these analyses. This sequential approach allows forward stepwise procedures and chi-squares and odds ratios for each 2-way interaction obtained from the logistic regression coefficient to be reported in the paper. This is in a manner akin to the F ratio in the factorial ANOVA. It is more interpretable to enter a series of interactions rather than one 16-level factor as each interaction corresponds to specific relationships between pairs of factors and the outcome e.g. factor A x factor B on Y (all binary variables as in this study) may be interpreted as the comparison between the odds ratio of factor B with Y for factor A=0 with the odds ratio of factor B with Y at factor A=1. The pooling in Table 3 on page 10 of the sixteen groups of Table 2 into seven pooled groups looks arbitrary with overlapping confidence intervals for odds ratios of groups from Table 2 (e.g. groups 2 and 3) which are placed in separate pooled groups (A and B) in Table 3. I don't see why this secondary pooling is necessary when one can simply test for two-way interactions using product terms entered in a sequential manner in a logistic regression. The strata presented in Tables 2 (page 8) and 3 (page 10) do not clearly interpret the odds ratios associated with interactions involving the four predictors and can be improved upon (see below). I wonder if the reason for this unusual approach to interpreting interactions is because the logistic regression procedure used in this paper (in SPSS) does not fit interactions by default. Interactions, however, can be forced into the model by creating product terms and adding these into the model as predictors in the 'covariate' box in a second block. The main effects (ie the terms used to construct any interactions) are entered in the first block. So, for example, if we had three

	factors, factor A and factor B and factor C, these can each be entered as predictors in block one and then two-way interactions involving multiplying together pairs of combinations of these three factors added in block two. The three-way interaction could then be entered in a third block, if desired, providing all lower order main effects and the three pairwise two-way interactions are entered in earlier blocks. If interactions are involved logistic regression models involving stepwise or backward elimination can still be used. Care must be taken, though, in using these stepwise procedures. The highest order interactions are tested first. So for example any two-way interaction which adds the highest statistically significant improvement using the likelihood ratio chi-square is added to the model containing all main effect if a forward stepwise procedure is used. Further two-way Interaction terms are then considered for addition in a model containing the recently added two-way interaction and all main effects. If there is an interaction between two factors we don't look at the main effects of the factors constituting this interaction since logically an interaction is implying that one factor effect depends on the level of another. Instead we only explain the two-way interaction by comparing 2x2 tables of frequencies and their odds ratios which can be obtained either using the 'risk' option in CROSSTABS or by the exp(B) column in the logistic regression output. The effect of the interaction can be presented as a chi-square representing the difference in log-likelihoods with and without the interaction term in the model. The regression coefficient representing the comparison between odds ratios from 2x2 tables involving the two factors in the interaction can then be used to give the effect size and interpret the interaction. In other words a sequential approach can be used to the testing of interactions adding main effects and interactions of the same order in separate blocks as is done with the General Linear Model. There are, incidentally, Generalized Linear Mixed Models (GENLIN in SPSS) which can be used for correlated responses but I don't think these are required here. If such an approach to looking at interactions was used in this study the results (particularly involving odds ratios) and interpretation of the two-way interactions would be a lot clearer than using what appears to be a single factor of 16 levels (giving the odds ratios in Table 2 on page 8). On what basis were the groups in Table 2 (on page 8) merged to form the groups in Table 3 on pager 10? For example, the confidence interval for group 3 in Table 2 (1.3-1.9) overlaps with the confidence interval for group 2 (0.8-1.5) so I am not sure that the odds ratios for groups 2 and 3 are statistically different yet these are merged into separate groups (A and B) in Table 3 on page 10. Other comments Page 7. Not clear from Table 1 to what the p-value in the rightmost column refers. Page 10, last sentence at end of first paragraph. I am not clear what is the difference between adjusted and unadjusted odds ratios. Are other different covariates in the model for one and not the other ie saying something along the lines that the adjusted odds ratios are

	assuming the same level of Factors A, B and C?

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Mika Gissler

(R) We have corrected the following orthographic mistakes:

1. Malmstrøm to Malmström
2. Not Western to non-Western
3. ORs in Table 3 with two decimals to one decimal
4. The use of oxytocin was restricted in the department in 2010 to … from 2010 onwards.
5. Colinearity to collinearity
6. WHO to Organization with 'z'

Reviewer: 2
Reviewer NamePhilip Steer

The authors have changed their description of ethnic origin to Western/non-Western, which is not a standard descriptive term for ethnic group. The authors describe Western as "originating from Europe or North America, or not" so I think what they mean is white European origin. If this is correct, then I think 'white European origin' would be a preferable term.

(R) We have used the term Western to classify women originating from Europe and North America. However, as there might be some non-white women living in a country within Western Europe or in North America classified into this category, we find it most prudent not to state explicitly that everyone was of white European origin. We have made a minor change according to Gissler's suggestion (point 2).

It would be interesting to know what a statistician thinks of their 'stratified' rather than sequential approach to allowance for confounders. Personally, I still see labour and delivery as sequential, rather than each stage being independent, but that may just be my unfamiliarity with this statistical approach.

(R) See below

Reviewer: 3
Reviewer Name Peter Watson

An easily implementable more understandable analysis for the interactions can be done. Instead of using what appears to be one single 16 level factor giving odds ratios as in Table 2 on page 8 we can perform the standard sequential approach to testing for interactions by constructing product terms and entering these after a block containing the main effects which constitute them (further details are below). The results including those in the abstract can then be re-written to reflect the output form these analyses.

(R) We highly appreciate the comments by the statistical reviewer with valuable suggestions for statistical analysis of our data.

Prior to our study we carefully reviewed the literature dealing with factors of importance for the occurrence of obstetric anal sphincter injury (OASIS). To the best of our knowledge, the interaction

between major clinical factors had so far not been discussed in a way that takes into account the various events of delivery in a clinically appropriate way.

Operative vaginal delivery (vacuum extraction and forceps) and large newborns (≥4000 g) are established as significant risk factors for OASIS. To our knowledge, analysis of the possible additive or multiplicative effects of the presence of both factors had not been explored earlier. Episiotomy is used in order to prevent OASIS, often when an operative vaginal delivery is undertaken, or when a large baby is expected. However, we realized that the possibility of a two- or three-ways interaction of the involved factors seemed to be unaddressed in the literature. Our clinical experience with oxytocin augmentation, causing too strong and uncontrolled contractions in the expulsive phase of labour and subsequent anal sphincter tears suggested that the possible association of oxytocin with OASIS should be explored, and this is the aim of our study. This association was hardly studied before, and the widespread use of stimulation lacked evidence-based support.

We wanted to design our analysis, based on a biological, dynamic understanding of what takes place at the end of the second stage of labour. In our guidelines, episiotomy is recommended for instrumental deliveries and when a large baby is expected. However, for spontaneous deliveries of normal-sized babies, routine episiotomy is not recommended. The prevalence of OASIS in this scenario is much lower, and episiotomies are expected to be few because the number needed to treat is assumed to be high.

On this basis, we started to analyze the data in a forward stepwise approach, starting with the most important variable from the univariate analysis, and then adding the second strongest etc. We entered instrumental delivery at step 1, birth-weight ≥4000 g at step 2, found an interaction, and continued in a new analysis with the cross product of instrumental delivery/large birth weight, before entering episiotomy. When a second interaction term was significant, we resolved this by cross products of instrumental delivery (C), high birth-weight (D) and episiotomy (A). Finally, we entered oxytocin augmentation (B). In this way we created a model of our statistical approach.

We assumed that this model could be difficult to understand for clinicians. Therefore we chose the opposite way, and started with 16 strata from 2x2x2x2 tables (Table 2). We considered this approach as appropriate as our dataset includes 15 000 subjects. From Table 2 we collapsed strata that were plausibly associated with the causal pathway of OASIS – for example, episiotomy (+/-) in the absence of B, C and D, and oxytocin augmentation in the absence of C and D, regardless of A (episiotomy). Furthermore, we studied the effect of episiotomy (A) on prevention of OASIS in instrumental deliveries (C) in view of large infants and oxytocin augmentation. Then we collapsed strata in a consistent way and within a “causal” model of OASIS in presence/absence of other factors (from Table 2 to table 3 in the paper).

The pooling in Table 3 on page 10 of the sixteen groups of Table 2 into seven pooled groups looks arbitrary with overlapping confidence intervals for odds ratios of groups from Table 2 (e.g. groups 2 and 3), which are placed in separate pooled groups (A and B) in Table 3. I don't see why this secondary pooling is necessary when one can simply test for two-way interactions using product terms entered in a sequential manner in a logistic regression.

The strata presented in Tables 2 (page 8) and 3 (page 10) do not clearly interpret the odds ratios associated with interactions involving the four predictors and can be improved upon (see below). I wonder if the reason for this unusual approach to interpreting interactions is because the logistic regression procedure used in this paper (in SPSS) does not fit interactions by default. Interactions, however, can be forced into the model by creating product terms and adding these into the model as predictors in the 'covariate' box in a second block. The main effects (ie the terms used to construct any interactions) are entered in the first block. So, for example, if we had three factors, factor A and factor B and factor C, these can each be entered as predictors in block one and then two-way interactions involving multiplying together pairs of combinations of these three factors added in block

two. The three-way interaction could then be entered in a third block, if desired, providing all lower order main effects and the three pairwise two-way interactions are entered in earlier blocks. If interactions are involved logistic regression models involving stepwise or backward elimination can still be used. Care must be taken, though, in using these stepwise procedures. The highest order interactions are tested first. So for example any two-way interaction which adds the highest statistically significant improvement using the likelihood ratio chi-square is added to the model containing all main effect if a forward stepwise procedure is used. Further two-way Interaction terms are then considered for addition in a model containing the recently added two-way interaction and all main effects. If there is an interaction between two factors we don't look at the main effects of the factors constituting this interaction since logically an interaction is implying that one factor effect depends on the level of another.

Instead we only explain the two-way interaction by comparing 2x2 tables of frequencies and their odds ratios which can be obtained either using the 'risk' option in CROSSTABS or by the exp(B) column in the logistic regression output. The effect of the interaction can be presented as a chi-square representing the difference in log-likelihoods with and without the interaction term in the model. The regression coefficient representing the comparison between odds ratios from 2x2 tables involving the two factors in the interaction can then be used to give the effect size and interpret the interaction. In other words a sequential approach can be used to the testing of interactions adding main effects and interactions of the same order in separate blocks as is done with the General Linear Model. There are, incidentally, Generalized Linear Mixed Models (GENLIN in SPSS) which can be used for correlated responses but I don't think these are required here.

If such an approach to looking at interactions was used in this study the results (particularly involving odds ratios) and interpretation of the two-way interactions would be a lot clearer than using what appears to be a single factor of 16 levels (giving the odds ratios in Table 2 on page 8).

On what basis were the groups in Table 2 (on page 8) merged to form the groups in Table 3 on pager 10? For example, the confidence interval for group 3 in Table 2 (1.3-1.9) overlaps with the confidence interval for group 2 (0.8-1.5) so I am not sure that the odds ratios for groups 2 and 3 are statistically different yet these are merged into separate groups (A and B) in Table 3 on page 10.

(R) We have done new analyses including all possible interaction terms into a forward stepwise logistic regression analysis.

compute A=episiotomy
compute B=oxytocin augmentation
compute C=operative vaginal delivery (vacuum extraction/forceps)
compute D=Birth weight >= 4000 g

LOGISTIC REGRESSION VARIABLES rupture
/METHOD=fstep A B C D A*B A*C A*D B*C B*D C*D A*B*C A*B*D A*C*D B*C*D A*B*C*D
/PRINT=CI(95)
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	**C by D**	1,273	,108	138,619	1	,000	3,573	2,890	4,416
	Constant	-2,746	,034	6346,354	1	,000	,064		
Step 2[b]	**C**	,643	,075	73,924	1	,000	1,903	1,643	2,203
	C by D	,794	,120	43,788	1	,000	2,211	1,748	2,797
	Constant	-2,909	,042	4903,477	1	,000	,055		
Step 3[c]	**C**	,794	,078	103,032	1	,000	2,212	1,897	2,578
	D	,898	,100	80,688	1	,000	2,454	2,017	2,985
	C by D	-,104	,156	,445	1	,505	,901	,664	1,224
	Constant	-3,060	,047	4175,498	1	,000	,047		
Step 4[c]	**C**	,768	,068	128,223	1	,000	2,154	1,886	2,460
	D	,855	,077	123,367	1	,000	2,350	2,021	2,733
	Constant	-3,050	,045	4606,553	1	,000	,047		
Step 5[d]	**C**	1,113	,086	168,102	1	,000	3,044	2,573	3,602
	D	,866	,077	125,984	1	,000	2,377	2,044	2,765
	A by C	-,630	,107	34,598	1	,000	,533	,432	,657
	Constant	-3,052	,045	4600,071	1	,000	,047		
Step 6[e]	**B**	,155	,070	4,821	1	,028	1,167	1,017	1,340
	C	1,055	,090	138,467	1	,000	2,873	2,410	3,425
	D	,852	,077	121,131	1	,000	2,344	2,014	2,728
	A by C	-,623	,107	33,809	1	,000	,536	,435	,662
	Constant	-3,110	,053	3476,906	1	,000	,045		
Step 7[f]	**B**	,311	,084	13,612	1	,000	1,364	1,157	1,609
	C	1,368	,128	114,175	1	,000	3,926	3,055	5,046
	D	,849	,077	120,425	1	,000	2,338	2,009	2,721

A by C	-,637	,107	35,286	1	,000	,529	,429	,653
B by C	-,480	,145	10,928	1	,001	,619	,465	,822
Constant	-3,176	,058	3025,996	1	,000	,042		

a. Variable(s) entered on step 1: C * D .

b. Variable(s) entered on step 2: C.

c. Variable(s) entered on step 3: D.

d. Variable(s) entered on step 5: A * C .

e. Variable(s) entered on step 6: B.

f. Variable(s) entered on step 7: B * C .

By doing this analysis, we ended up with B, C, D and two-ways interactions A*C and B*C, whereas D became an independent risk factor. We solved the interaction terms into a three-way interaction between A*B*C, and made 8 strata out of A, B and C, and let D=1 as the 9th strata.

```
Compute Strat_mod_a=10.
if  ((A=0) and (B=0) and (C=0)) strat_mod_a=11.
if  ((A=1) and (B=0) and (C=0)) strat_mod_a=2.
if  ((A=0) and (B=0) and (C=1)) strat_mod_a=3.
if  ((A=1) and (B=0) and (C=1)) strat_mod_a=4.
if  ((A=0) and (B=1) and (C=0)) strat_mod_a=5.
if  ((A=1) and (B=1) and (C=0)) strat_mod_a=6.
if  ((A=0) and (B=1) and (C=1)) strat_mod_a=7.
if  ((A=1) and (B=1) and (C=1)) strat_mod_a=8.
if (D=1) strat_mod_a=9.
```

(by this construction D is absent in category 2 thru 8, and 11) (category 10 is empty – just a control that all cases where captured).

We continued with another logistic regression analysis:

```
LOGISTIC REGRESSION VARIABLES rupture
/METHOD=ENTER strat_mod_a
/contrast (strat_mod_a)=indicator
/PRINT=CI (95)
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)

							Lower	Upper
Step 1[a]	Strat_mod_a			270,135	8	,000		
	Strat_mod_a(1)	,123	,150	,673	1	,412	1,131	,842	1,519
	Strat_mod_a(2)	1,510	,174	75,404	1	,000	4,527	3,219	6,365
	Strat_mod_a(3)	,813	,175	21,667	1	,000	2,255	1,601	3,176
	Strat_mod_a(4)	,439	,111	15,507	1	,000	1,551	1,246	1,929
	Strat_mod_a(5)	,480	,151	10,165	1	,001	1,616	1,203	2,171
	Strat_mod_a(6)	1,213	,127	90,815	1	,000	3,365	2,622	4,319
	Strat_mod_a(7)	,694	,130	28,394	1	,000	2,001	1,551	2,583
	Strat_mod_a(8)	1,345	,098	187,700	1	,000	3,840	3,168	4,655
	Constant	-3,255	,072	2019,352	1	,000	,039		

a. Variable(s) entered on step 1: Strat_mod_a.

Which reads as follows:

	Episiotomy (A)	Oxytocin Augment.	Instrumental delivery (C)	OR	95% CI
Episiotomy in absence of B, C, D	-	-	-	1,0	
	+	-	-	1,13	0,84-1,52
Effect episiotomy in absence of B, D	**-**	**-**	**+**	**4,5**	**2,3-6,4**
	+	**-**	**+**	**2,3**	**1,6-3,2**
Effect oxytocin in absence of C, D, regardless of A	**-**	**+**	**-**	**1,55**	**1,3-1,9**
	+	**+**	**-**	**1,62**	**1,3-2,2**
Effect episiotomy (+B, -D)	**-**	**+**	**+**	**3,4**	**2,6-4,3**
	+	**+**	**+**	**2,0**	**1,6-2,6**
BW >= 4000 g (+D)				3,8	3,2-4-7

We were able to reproduce our stratified approach in a logistic regression analysis of all possible combinations of interactions, except for the interaction term of high birth- weight*instrumental delivery and episiotomy. We find it important to stress that there might be an additive effect of instrumental delivery*high birth-weight in presence/absence of episiotomy (not overlapping confidence intervals in table 3; strata F and G). In our opinion, this interaction is of greatest clinical relevance, and we think it is important to report it, and thus encourage other researcher in the future to continue building models from a dynamic understanding of what is going on during the last part of the active second stage of labour.

In our opinion the literature dealing with the risk of OASIS would benefit from applying statistical analyses that take into account dynamics of the physiological events during delivery. We believe that our study adds novel knowledge in the efforts to prevent OASIS by assessing the associative role of oxytocin augmentation in the large group of women having a normal delivery. Further by depicting the associations of episiotomy with OASIS in various clinically recognizable scenarios, represented by the strata in our model.

Based on our considerations as outlined above, we think it is justified to remain the text of the manuscript unchanged. However, if the Editor advises us to include more of the above reasoning in the manuscript, we will be happy to follow such advice and add the appropriate information.

Other comments

Page 7. Not clear from Table 1 to what the p-value in the rightmost column refers.

(R) P-values from x^2-test, added to text in Table 1.

Page 10, last sentence at end of first paragraph. I am not clear what is the difference between adjusted and unadjusted odds ratios. Are other different covariates in the model for one and not the other ie saying something along the lines that the adjusted odds ratios are assuming the same level of Factors A, B and C?

(R) Some of the ORs changed more than 10% - indicating confounding after including epidural anesthesia in the model.

There were no confounding effects of origin, occiput posterior presentation or age. Only epidural analgesia had a confounding effect as stated in text (not included in table 3).

VERSION 3 - REVIEW

REVIEWER	Peter Watson Medical Research Council UK
REVIEW RETURNED	23-Apr-2014

GENERAL COMMENTS	To help interpret and compare the odds ratios in Table 2 (page 8) we need to redo the stepwise logistic regression on page 6 and page 39 and present any statistically significant results explaining them in the text. This should not take too long to do in SPSS but requires a manual backward elimination approach which I describe below in the comments to authors but which is also illustrated in Agresti's 1996 Introduction to Categorical Data Analysis book. The

	logistic regression on page 39 does not seem to me to take into account the need to look at higher order terms with other terms included in the model. For example you can't look at a main effect if it is involved with a moderator factor in an interaction because that variable's effect is influenced by the factor with which it interacts. We can also only interpret interaction terms if the lower order terms comprising them are also in the regression e.g. looking at an A by B interaction needs the main effects of A and B to be also in the model. The authors have made a valiant effort to test interaction terms using SPSS logistic regression. Looking at the SPSS syntax on page 39, however, they do not test these interactions in a sequential fashion and do not construct product terms representing the interactions. A backwards manual stepwise approach needs to be adopted since one looks at higher order interaction terms first only looking at lower order ones if the higher order ones are not statistically significant. For example we do not usually interpret a main effect of a factor A if it interacts with B since the interaction is telling us that the effect of factor A is moderated by factor B so can only be interpreted if looked at in conjunction with the factor B. Agresti (1996, pages 127-129) describes and illustrates a stepwise approach to the testing of interactions in logistic regression using backward elimination. I provide a summary of this approach below. A similar approach can be used for any regression model involving interactions. The four factors are coded as 0/1 binary variables. The interactions are then formed by multiplying these factors together so for example the A by B by C by D interaction term is composed of factor A multiplied by factor B multiplied by factor C multiplied by factor D. In this way we end up with six products representing each of the six two-way interactions, four products for the three-way interactions and one product representing the four-way interaction. This gives four factors + six two-way interactions + four three-way interactions + one four-way interaction as (15) predictors of the rupture variable used as outcome in the logistic regressions on pages 39 and 41. We can now start model fitting. Firstly we fit all 15 predictors together as on page 39 of the paper ie upto the four way interaction A by B by C by D. If the A by B by C by D interaction is significant explore this interaction by for example looking at the three-way interaction at a level of the fourth factor. If the A by B by C by D interaction is not significant delete the A by B by C by D interaction and refit the model without it looking at the four three-way interactions and then delete the least non-significant three-way interaction. Keep refitting the logistic models deleting the least non-significant three-way interaction each time until either the remaining three-way interactions are all significant ($p<0.05$) or they have all being deleted. You can then look at two-way interactions in the same way deleting the least non-significant interaction from the logistic model, refitting without the deleted interaction and examining the remaining two-way interactions. Remember though you can only assess two-way interactions if the factors that they consist of are NOT BOTH involved in any statistically significant three-way interaction. So for example if A by B by C is statistically significant you cannot look at A by B (or A by C or B by C) since A by B is moderated by C but you could assess the A by D interaction since no two-way interaction

	involving D is moderated by the other factors. The results from such an analysis can then be described in the body of the text, perhaps, quoting a chi-square value (either the Wald chi-square or the likelihood ratio chi-square representing the difference in fit from a model with and without a particular term) and then interpreted rather than as on page 6 (sixth line from the bottom) mentioning vaguely that the 'interaction terms were significant'. The clean interpretation above relates each of the factors to a particular response such as the 'rupture' variable on page 39 rather than using all 16 possible combinations as in Table 2 on pages 8-9 where it is difficult to see how the factors relate to the injury. Reference Agresti A (1996) An introduction to categorical data analysis. Wiley:New York.

VERSION 3 – AUTHOR RESPONSE

To help interpret and compare the odds ratios in Table 2 (page 8) we need to redo the stepwise logistic regression on page 6 and page 39 and present any statistically significant results explaining them in the text. This should not take too long to do in SPSS but requires a manual backward elimination approach which I describe below in the comments to authors but which is also illustrated in Agresti's 1996 Introduction to Categorical Data Analysis book.

The logistic regression on page 39 does not seem to me to take into account the need to look at higher order terms with other terms included in the model. For example you can't look at a main effect if it is involved with a moderator factor in an interaction because that variable's effect is influenced by the factor with which it interacts. We can also only interpret interaction terms if the lower order terms comprising them are also in the regression e.g. looking at an A by B interaction needs the main effects of A and B to be also in the model.

Assessing the association of oxytocin augmentation with obstetric anal sphincter injury in nulliparious women – a population-based, case-control study. bmjopen-2013-004592.R2

The authors have made a valiant effort to test interaction terms using SPSS logistic regression. Looking at the SPSS syntax on page 39, however, they do not test these interactions in a sequential fashion and do not construct product terms representing the interactions. A backwards manual stepwise approach needs to be adopted since one looks at higher order interaction terms first only looking at lower order ones if the higher order ones are not statistically significant. For example we do not usually interpret a main effect of a factor A if it interacts with B since the interaction is telling us that the effect of factor A is moderated by factor B so can only be interpreted if looked at in conjunction with the factor B. Agresti (1996, pages 127-129) describes and illustrates a stepwise approach to the testing of interactions in logistic regression using backward elimination. I provide a summary of this approach below. A similar approach can be used for any regression model involving interactions.

The four factors are coded as 0/1 binary variables. The interactions are then formed by multiplying these factors together so for example the A by B by C by D interaction term is composed of factor A multiplied by factor B multiplied by factor C multiplied by factor D. In this way we end up with six

products representing each of the six two-way interactions, four products for the three-way interactions and one product representing the four-way interaction. This gives four factors + six two-way interactions + four three-way interactions + one four-way interaction as (15) predictors of the rupture variable used as outcome in the logistic regressions on pages 39 and 41. We can now start model fitting.

Firstly we fit all 15 predictors together as on page 39 of the paper ie upto the four way interaction A by B by C by D.

If the A by B by C by D interaction is significant explore this interaction by for example looking at the three-way interaction at a level of the fourth factor. If the A by B by C by D interaction is not significant delete the A by B by C by D interaction and refit the model without it looking at the four three-way interactions and then delete the least non-significant three-way interaction. Keep refitting the logistic models deleting the least non-significant three-way interaction each time until either the remaining three-way interactions are all significant ($p<0.05$) or they have all being deleted. You can then look at two-way interactions in the same way deleting the least non-significant interaction from the logistic model, refitting without the deleted interaction and examining the remaining two-way interactions. Remember though you can only assess two-way interactions if the factors that they consist of are NOT BOTH involved in any statistically significant three-way interaction. So for example if A by B by C is statistically significant you cannot look at A by B (or A by C or B by C) since A by B is moderated by C but you could assess the A by D interaction since no two-way interaction involving D is moderated by the other factors.

The results from such an analysis can then be described in the body of the text, perhaps, quoting a chi-square value (either the Wald chi-square or the likelihood ratio chi-square representing the difference in fit from a model with and without a particular term) and then interpreted rather than as on page 6 (sixth line from the bottom) mentioning vaguely that the 'interaction terms were significant'. The clean interpretation above relates each of the factors to a particular response such as the 'rupture' variable on page 39 rather than using all 16 possible combinations as in Table 2 on pages 8-9 where it is difficult to see how the factors relate to the injury.

Reference
Agresti A (1996) An introduction to categorical data analysis. Wiley:New York.

The following text is added to the M&M chapter:

"The intention of this study was to explore the effect of three obstetric practices (oxytocin augmentation (O), episiotomy (E) and vacuum/forceps (VF)) and birth weight (BW) on obstetric anal sphincter injuries before other risk factors were considered. These main risk factors correlate as episiotomy is often used for instrumental deliveries and when large babies are expected. Furthermore, oxytocin augmentation is provided for failure to progress because of dystocia. Women with dystocia are more often delivered instrumentally than women without dystocia. This basic understanding of the birth dynamics of the first and second stage of labour indicates that the main risk factors may have a direct or indirect effect on obstetric anal sphincter injuries, and that the effects of categories across different explanatory variables are not constant on the outcome.

We analysed our dataset using the Chi-squared test and backward manual stepwise logistic regression analyses with $p<0.05$ as significance level. We built and checked the fit of our regression model as proposed by Agresti [21]. At step one we compared a model of the highest order interaction

term (four-way product term; E*O*VF*BW) and the main risk factors (E+O+VF+BW) with a model comprising only the main risk factors. If the highest order product term is not significant, Agresti propose to continue with second highest order terms by removing the term with the highest p-value until the model of best fit is reached. Confounders, possible risk factors in addition to the main factors of interest, were tested one by one and set to at least 10% change in any estimate in the model of best fit. Interaction terms were significant at $p<0.05$. Statistical analyses were performed with IBM SPSS Statistics for Windows, v. 19.0 Armonk, NY: IBM Corp."

In the Results section we have rewritten the text as follows:

"The log likelihood-ratio score from the highest order model (O*E*VF*FW+O+E+VF+BW; -2 LR: 7213.8) did not differ from the model comprising the main effects (O+E+VF+BW, -2 LR: 7215.9). After removing insignificant three-way interaction product terms, and playing with the remaining two-ways interaction terms, the model that gave the best fit comprised the interaction of oxytocin augmentation, episiotomy and instrumental delivery (O*E*VF), in addition to episiotomy/birth weight (E*BW) and instrumental delivery/birth weight (VF*BW) (-2 LR: 7371.2) (Model A). We could resolve interaction terms into stratified analysis of 8 strata of combinations of oxytocin augmentation, episiotomy and instrumental delivery for birth weights <4000 g, and 4 strata of combinations of episiotomy, instrumental delivery and birth weight ≥4000 g, independent of oxytocin augmentation. The results are displayed in Table 2."

Table 2 follows

From a clinical perspective we can simplify model A into model B by collapsing groups that comprise similar risks for sphincter injury by obstetric interventions despite overlapping confidence intervals. Spontaneous delivery of an infant weighing <4000 g without oxytocin augmentation and episiotomy was chosen as the reference group (group 1). We collapsed group 1 and 2 as the odds for sphincter injury was similar with and without episiotomy in unstimulated, spontaneous births of normal-sized infants. Groups 3 to 6 display the odds for sphincter injury in instrumental deliveries of normal-sized infants with and without oxytocin augmentation and episiotomy. A marked difference in the odds for sphincter injury was observed between women delivered instrumentally with (group 3 and 5) and without (group 4 and 6) episiotomy, despite the fact that those stimulated with oxytocin had a non-significant lower odds for sphincter injury. It was therefore reasonable to collapse groups 3 and 5, and

4 and 6. Furthermore, we collapsed groups 7 and 8 as the odds for sphincter injury was similar with and without episiotomy during spontaneous deliveries of infants <4000 g, regardless of oxytocin augmentation. Finally, the use of episiotomy appeared to be strongly associated with lower odds for sphincter injury in instrumental deliveries of infants ≥4000 g (groups 11 and 12). The modified model B (Table 3) comprises a clinically relevant risk estimation of anal sphincter injury among the main modified risk factors for sphincter injury."

Table 3 follows.

By this we hope that we have addressed the issues from the reviewer satisfactorily.

VERSION 4 - REVIEW

REVIEWER	Peter Watson Medical Research Council Cambridge UK
REVIEW RETURNED	11-Jun-2014

GENERAL COMMENTS	I give a suggestion for interpreting further the three (O x E x VF) and two-way interactions (E x BW) and (VF x BW) by constructing two 2x2 tables for each delivery method of the eight proportions of injured women given in Table 2 and comparing the differences in using episiotomy in the two tables which clearly show episiotomy reduced injuries by a larger amount than oxytocin for VF='+'. (Neither are effective in reducing proportions injured for VF = '-'). The two-way interactions can be similarly explained (see below). This seems a more objective way to explain the interactions than using blocks of groups where groups from different blocks still have overlapping confidence intervals for their odds ratios. I also was not clear how (page 6 line 38) you can fit a four-way interaction when the oxytocin is unspecified for higher birth weights as apparently show by the '+/-' for groups 9-12 in Table 2 on page 9. I think you, therefore, have to fit two logistic regressions one with O, E and VF as predictors of proportion of injuries and the other with E, VF and birth weight as predictors of proportion of injuries and admit one cannot fit interactions involving O and birth weight given the lack of a specific oxytocin value for higher birth weights (>=4000g = '+'). The authors do now adopt the backward elimination strategy of Agresti and describe it in lines 35-41 on page 6 with the results in lines 37-42 on page 8. Just a bit of fine tuning is needed to clarify the method and particularly the results. On lines 37-38 in the methods section on page 6 I would say step one compares the model including the highest order four-way interaction with a model without the four-way interaction. If the highest order product term is not significant, Agresti proposes continuing removing the highest order term with the highest non-significant p-value until all remaining terms have statistically significant p-values. Four predictors (O= Oxytocin, E=Episiotomy, VF= Vaginal Delivery and birth weight) are used to predict the proportion of women with sphincter injuries.

	I would remove lines 36-39 on page 8. There is confusion in what is written on these lines. The sentence on lines 37-38, for example, implies that there are no interactions since a model containing the interactions has no better fit than one involving only the main effects. Yet the next sentence states that the best fit comprises a model with interactions. We can say in the paragraph from lines 37-42 on page 8 that after adopting the strategy of Agresti (as described in the methods section earlier) by deleting the highest statistically non-significant terms in the models until all remaining terms are statistically significant we end up with a best fitting model involving the three-way interaction of O x E x VF and two two-way interactions E x BW and VF x BW as described on lines 40-42 on page 8. To explain the O x E x VF interaction we can present a two by two table of proportions injured with the four proportions injured for the four oxytocin and episotomy combinations for vacuum and the four proportions injured for the four oxtycin and episotomy combinations for forceps. By using the proportions in the injury N (%) column in Table 2 (third column from the right) which is simply using the frequencies in the columns Women N and injury N (%) in Table 2 on page 9 we get proportions injured of 0.037 = 198/5328 (O=not used, E=not used), 0.042 for (O=not used, E=used), 0.056 for (O=used, E=not used) and 0.059 (O=used, E=used) all for VF= vacuum and 0.149 (O=not used, E=not used) 0.080 (O=not used, E=used), 0.115 for (O=used, E=not used) and 0.072 for (O=used, E=used) for VF= forceps. Looking at these proportions one could conclude that the O x E x VF interaction is telling us that the use of episotomy is more successful when forceps are used compared to oxtycin augmentation. This method of interpreting the interaction is I think a less ambiguous and clearer way of presenting the results than that used on page 10 because the rationale for pooling the groups to give Table 3 on page 10 is suspect since groups in different pooled blocks have overlapping confidence intervals e.g. group 9 has an= 95% CI for its odds ratio using the right hand column in Table 2 on page 9 which overlaps with groups 3,4,5 and 6 which are, however, in separate blocks in Table 3 on page 10. We could also fit the O x E interaction (with O and E in the model) with obstetric and sphincter injury as response in a logistic regression and see if this is statistically significant when VD = ‘vacuum‘ and repeat when VD is ‘forceps’ if wished. Similarly for E x BW we can compare the two proportions injured for episotomy used and not used for birth weights < 4000g with the two proportions injured for episotomy used and not used when birth weights are >= 4000g to assess how successful episotomy is for low and high birth weights. The VF x BW interaction would compare proportions injured using vacuum and forceps in high birth weights with those with low birth weights and explain how the difference between the proportion injured using forceps as opposed to vacuum differs in babies with a high birth weight from the difference between the proportion injured using forceps as opposed to vacuum in babies having a low birth weight. We would simply pool the frequencies in Table 2 on page 9 to obtain these proportions. On pages 9-10 the conclusions of the results would therefore involve three statements explaining each of the three interactions found and therefore there would be no need for Table 3 and the associated references to the pooling of the pairs of the groups in Table 2 (e.g. line 37 on page 11 in the discussion) which could be deleted. The three interactions obtained from the logistic regressions would be

	the results that would be discussed. I also think one should state in the results or methods section that, due to groups 9-12 in Table 2 on page 9 having a non-specific oxytocin augmentation which is neither used or unused for birth weights of 4000g or more, two logistic regression models must be fitted. The first looking at the Oxytocin augmentation, episiotomy and delivery on proportion of injuries and the second looking at episiotomy, delivery and birth weight on proportion of injuries There is a suggestion in line 38 on page 6 that a four-way interaction (E x O x VF x BW) is fitted involving oxytocin augmentation and birth weight but this is not possible since the lack of a single value for oxytocin augmentation for higher birth weights, as given in groups 9-12 with a '+/-' value in Table 2, would appear to preclude the testing of an interaction involving both oxytocin augmentation and birth weight. Please clarify that you used two logistic regressions or how you fitted a four-way interaction when there is no specific value of oxytocin for higher birth weights in Table 2 on page 9. Page 8, line 39. What is meant by 'playing' with the two way interactions? In stating the statistical significance of the interaction terms analysed I would quote the chi-squares obtained from comparing the differences in log likelihoods for models with and without terms of interest instead of the log likelihood in lines 37-42 on page 8 with the p-value. This is easily obtained from SPSS. One could simply take -2(difference in log likelihoods) or equivalently the difference in model deviances both of which are outputted. See Agresti's book which is referenced on line 37 of page 6 for further details. Page 9, Table 2. It would be more informative to use 'unused/used', 'vacuum/forceps' and '<4000g/>=4000g' instead of the '-' and '+' signs in the body of the table to explain the variable combinations comprising the data. I was not clear from the table or text to what levels the '-' and '+' referred but assumed the '-' was corresponding in the four respective predictor variables to unused (O and E), vacuum (VF) and <4000g (birth weight).

VERSION 4 – AUTHOR RESPONSE

Thank you for a most helpful review of our third revision. We have changed our manuscript accordingly and believe that this has made our article easier to understand.

The aim of our study was to explore if there is an association of oxytocin augmentation with anal sphincter injuries (OASIS). Our hypothesis was that uncontrolled, augmented contractions could hamper the slow stretching and manual protection of the perineum, and thus be an independent risk factor. This association has to our knowledge not yet been the primary outcome in previous studies. Our intention was to investigate this association in clinically recognizable scenarios; taking into account three other established risk factors for obstetric anal sphincter injury. These factors are "working" at different points of time during labour – oxytocin augmentation first to strengthen uterine contractions; secondly episiotomy that widens the vaginal outlet in the last minutes of delivery; third

operative vaginal delivery (vacuum or forceps) where the pulling forces of the birth attendant are added to the expulsive forces produced by the birthing woman. The size of the baby, measured through its birth weight, excerts a direct force on the sphincter complex as the baby passes the vaginal outlet (but not during the early part of the pushing). We will comment in more detail below why this is important for the model displayed in Table 3.

1. The authors do now adopt the backward elimination strategy of Agresti and describe it in lines 35-41 on page 6 with the results in lines 37-42 on page 8. Just a bit of fine tuning is needed to clarify the method and particularly the results.
On lines 37-38 in the methods section on page 6 I would say step one compares the model including the highest order four-way interaction with a model without the four-way interaction. If the highest order product term is not significant, Agresti proposes continuing removing the highest order term with the highest non-significant p-value until all remaining terms have statistically significant p-values. Four predictors (O= Oxytocin, E=Episiotomy, VF= Vaginal Delivery and birth weight) are used to predict the proportion of women with sphincter injuries.

R: We have replaced lines 37-41 p 6 by the text suggested.
2. I would remove lines 36-39 on page 8. There is confusion in what is written on these lines. The sentence on lines 37-38, for example, implies that there are no interactions since a model containing the interactions has no better fit than one involving only the main effects. Yet the next sentence states that the best fit comprises a model with interactions. We can say in the paragraph from lines 37-42 on page 8 that after adopting the strategy of Agresti (as described in the methods section earlier) by deleting the highest statistically non-significant terms in the models until all remaining terms are statistically significant we end up with a best fitting model involving the three-way interaction of O x E x VF and two two-way interactions E x BW and VF x BW as described on lines 40-42 on page 8.

R: We have replaced these lines by the text suggested.

3. To explain the O x E x VF interaction we can present a two by two table of proportions injured with the four proportions injured for the four oxytocin and episiotomy combinations for vacuum and the four proportions injured for the four oxytocin and episiotomy combinations for forceps. By using the proportions in the injury N (%) column in Table 2 (third column from the right) which is simply using the frequencies in the columns Women N and injury N (%) in Table 2 on page 9 we get proportions injured of 0.037 = 198/5328 (O=not used, E=not used), 0.042 for (O=not used, E=used), 0.056 for (O=used, E=not used) and 0.059 (O=used, E=used) all for VF= vacuum and 0.149 (O=not used, E=not used) 0.080 (O=not used, E=used), 0.115 for (O=used, E=not used) and 0.072 for (O=used, E=used) for VF= forceps. Looking at these proportions one could conclude that the O x E x VF interaction is telling us that the use of episiotomy is more successful when forceps are used compared to oxytocin augmentation. This method of interpreting the interaction is I think a less ambiguous and clearer way of presenting the results than that used on page 10 because the rationale

for pooling the groups to give Table 3 on page 10 is suspect since groups in different pooled blocks have overlapping confidence intervals e.g. group 9 has an= 95% CI for its odds ratio using the right hand column in Table 2 on page 9 which overlaps with groups 3,4,5 and 6 which are, however, in separate blocks in Table 3 on page 10. We could also fit the O x E interaction (with O and E in the model) with obstetric and sphincter injury as response in a logistic regression and see if this is statistically significant when VD = 'vacuum' and repeat when VD is 'forceps' if wished.

R: With reference to existing knowledge in the field our study is exploratory. We designed our study to answer whether oxytocin is a risk factor for OASIS in a dynamic model of the four driving forces during the last minutes of 2nd active phase of birth. As explained in the first paragraph of our response, the order of the factors relative the occurrence of OASIS is important.

The tree-way-interaction "O * E * VF", in the absence of BW (normal birth weight infants), is telling us that "*the use of episiotomy is more successful when vacuum/forceps are used compared to oxytocin augmentation*". In our words, oxytocin augmentation, which is applied prior to episiotomy and vacuum/forceps, does not exert an independent effect on OASIS in the presences of the two others (E/VF). This is an important clinical observation. This finding fits perfect with our dynamic understanding of what is taking place. However, in the absence of vacuum/forceps, among normal sized infants, episiotomy does not have an effect on OASIS when oxytocin augmentation is applied. **This is a new finding**. Finally, we are of the opinion that the difference in the effect of episiotomy on OASIS in the absence of "O" and "VF" is of a less important magnitude. This is the clinical rationale for collapsing the groups in Table 2 to Table 3. The transition of Table 2 to Table 3 may be questionable from a statistical point of view; however, interactions may also be treated in a meaningful clinical context. We prefere to share this interpretation with our readers. In this way we take care of the complexity of our model, and choose to display the model in such a way that it is easier to interpret/understand.

4. Similarly for E x BW we can compare the two proportions injured for episiotomy used and not used for birth weights < 4000g with the two proportions injured for episiotomy used and not used when birth weights are >= 4000g to assess how successful episiotomy is for low and high birth weights. The VF x BW interaction would compare proportions injured using vacuum and forceps in high birth weights with those with low birth weights and explain how the difference between the proportion injured using forceps as opposed to vacuum differs in babies with a high birth weight from the difference between the proportion injured using forceps as opposed to vacuum in babies having a low birth weight. We would simply pool the frequencies in Table 2 on page 9 to obtain these proportions.

R: In the present study 90% of operative vaginal deliveries (VF) are vacuum deliveries. We don't have power to analyze forceps deliveries separately.

Similarly for E x BW we can compare the two proportions injured for episiotomy used and not used for birth weights < 4000g with the two proportions injured for episiotomy used and not used when birth weights are >= 4000g to assess how successful episiotomy is for low and high birth weights.

These results are displayed in Table 2, group 9 and 10, in the absence of VF, regardless of "O". The estimates (with confidence intervals) on OASIS are very close. The effect of BW on OASIS is mediated through VF and E. Within a complete model, having O * E * VF as the stronger interaction term compared to the others, we created 4 strata among BW (+) across episiotomy and VF. We didn't see any other solution. We can see a clear effect of episiotomy in normal sized infants delivered by VF (group C/D in Table 3), and in infants 4000 g or more (group F/G in Table 3). The effect of BW+ relative BW- across VF and episiotomy is doubled! We have not stressed this in the discussion, as the main research issue is the effect of oxytocin augmentation on OASIS.

5. On pages 9-10 the conclusions of the results would therefore involve three statements explaining each of the three interactions found and therefore there would be no need for Table 3 and the associated references to the pooling of the pairs of the groups in Table 2 (e.g. line 37 on page 11 in the discussion), which could be deleted. The three interactions obtained from the logistic regressions would be the results that would be discussed.

R: We have studied these views carefully, however, referring to our replies above, we still think that the groups we have collapsed (Table 3) corresponds to the clinically important scenarios. The timing and dynamic effect of the combinations (strata) represents clinical situations that are independent, and are clinical entities displayed independently in Table 3. As stated previously, we are of the opinion that Table 3 eases the interpretation of the results for the reader.

6. I also think one should state in the results or methods section that, due to groups 9-12 in Table 2 on page 9 having a non-specific oxytocin augmentation which is neither used or unused for birth weights of 4000 g or more, two logistic regression models must be fitted. The first looking at the Oxytocin augmentation, episiotomy and delivery on proportion of injuries and the second looking at episiotomy, delivery and birth weight on proportion of injuries There is a suggestion in line 38 on page 6 that a four-way interaction (E x O x VF x BW) is fitted involving oxytocin augmentation and birth weight but this is not possible since the lack of a single value for oxytocin augmentation for higher birth weights, as given in groups 9-12 with a '+/-' value in Table 2, would appear to preclude the testing of an interaction involving both oxytocin augmentation and birth weight. Please clarify that you used two logistic regressions or how you fitted a four-way interaction when there is no specific value of oxytocin for higher birth weights in Table 2 on page 9.
There is no interaction of oxytocin augmentation and birthweight on OASIS. We have changed the text on page 6.

7. Page 8, line 39. What is meant by 'playing' with the two-way interactions? In stating the statistical significance of the interaction terms analysed I would quote the chi-squares obtained from comparing the differences in log likelihoods for models with and without terms of interest instead of the log likelihood in lines 37-42 on page 8 with the p-value. This is easily obtained from SPSS. One could simply take -2 (difference in log likelihoods) or equivalently the difference in model deviances both of

which are outputted. See Agresti's book, which is referenced on line 37 of page 6 for further details.

R: This is probably a language/semantic problem when not writing English as a native language. We dissolved the interactions O*E*VF, E*BW and VT*BW into 8 strata of O*E*VF among BW negative (normal birth weight) and 4 strata of E*VF among BW positive (≥ 4000 g). The action of dissolving E*BW and *VF*BW into 4 strata of E*VF among BW positive was the meaning of "playing with". This was the solution we were able to go for.

8. Page 9, Table 2. It would be more informative to use 'unused/used', 'vacuum/forceps' and '<4000g/>=4000g' instead of the '-' and '+' signs in the body of the table to explain the variable combinations comprising the data. I was not clear from the table or text to what levels the '-' and '+' referred but assumed the '-' was corresponding in the four respective predictor variables to unused (O and E), vacuum (VF) and <4000g (birth weight).

R: We have included these options as a footnote in Table 2 and 3.