PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Association between perinatal interventional activity and 2-year outcome of Swiss extremely preterm born infants: a population-based cohort study
AUTHORS	Adams, Mark; Berger, Thomas; Borradori-Tolsa, Cristina; Bickle-Graz, Myriam; Grunt, Sebastian; Gerull, Roland; Bassler, Dirk; Natalucci, Giancarlo

VERSION 1 – REVIEW

REVIEWER	Amber Reichert University of Alberta, Canada
REVIEW RETURNED	27-Jun-2018

GENERAL COMMENTS	Please include discussion about the limitations of including a composite outcome (specifically death or NDI - where they are competing outcomes). This is a commonly reported outcome but there are limitations associated with its use. I wonder also about the adjustment for defining moderate cognitive disability by shifting the cut off by an entire 15 point standard deviation. Please refer to North American publications regarding comparison of the Bayley III vs BSID II in the preterm population (rather than babies with HIE). Also, possible to explore the potential bias resulting from increased proportion of outborn in those lost to follow-up. Would imputational analysis truly correct for this?

REVIEWER	Noelle Younge Duke University, USA
REVIEW RETURNED	03-Jul-2018

GENERAL COMMENTS	This is an interesting and nicely presented study of 2 year outcomes of extremely preterm infants born in Swiss centers classified as high or low perinatal interventional activity centers. The investigators found that centers with high perinatal interventional activity scores had lower mortality among infants born <26 weeks' gestation, but there was no significant difference in morbidity or neurodevelopmental impairment among surviving infants. These findings are similar to those reported in other international studies. The strengths of the study include the large sample size, inclusion of the vast majority of all liveborn infants in the country, and high-quality data on neonatal and early childhood outcomes. The paper is very well-written and provides a thorough discussion of the findings in relation to other recent studies. Comments/Questions:

	Three different neurodevelopmental assessments were used during the study. Could the authors please include additional information on how the GMDS compares to the Bayley examinations? Could the authors provide more information about the outborn infants? Including the outborn infants seems somewhat problematic as perinatal interventions would not reflect the center they were admitted to, and selects for infants who survive for transfer to the perinatal centers. Also, including the proportion of inborn infants in each center in the activity score does not necessarily reflect perinatal interventional activity as in the Serenius study, where the activity score were region-based rather than center-based. Last, since the study focuses on the effect of perinatal interventions on outcomes, the rationale for including inborn status in both the interventional score and as a variable in the regression model was unclear. The authors provide a brief justification in the Discussion section for the cut-off gestational age points used to describe Cohorts A and B. However, inclusion of the 25 weeks infants in Cohort A may obscure the effects of active perinatal intervention at lower gestational ages. I would consider adding additional information regarding how the receipt of perinatal interventions varied by gestational week in Cohort A. I would also add the actual proportion of infants in each Cohort who received each perinatal intervention that corresponds to the top center (i.e. the center given the 100 score for each of the components). This would give the reader a greater understanding of how the approach to perinatal intervention compares with other international studies. Were the centers similar in volume (i.e. number of infants)? A brief footnote in the Tables 1 and 3 regarding the SES variable (i.e. definition and scale) would be helpful. This statement was not clear: “Data per item was missing in 1.2% of datasets for the outcome “major morbidity” or less than 0.1% in all other cases.” Please clarify why are GA and GA2 terms both included in the final regression model? I would consider noting limitations of the perinatal activity score in the Discussion. While it seems a reasonable measure of perinatal interventions, the equal weighing of each component, interrelatedness of some of the components, and arbitrary cut-off score between high and low centers may not fully capture the variations in perinatal treatment approach. In Table 1 and 3, consider noting which variables are significantly different between groups.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Please include discussion about the limitations of including a composite outcome (specifically death or NDI - where they are competing outcomes). This is a commonly reported outcome but there are limitations associated with its use.

OUR REPONSE:
-> As requested by bmj open, we listed our strengths and limitations under the required heading "Strengths and Limitations". Among them we list the limitations of using composite measures in general. We however believe to have covered the limitation of competing outcomes in the composite outcomes "death or NDI" and "death or major morbidity" by providing the results for the composite and the individual outcomes in the same chart (i.e. separately for mortality, NDI and morbidities). We did add "the major morbidities" to the list of composite outcomes with limitations under "Strengths and

Limitations"
I wonder also about the adjustment for defining moderate cognitive disability by shifting the cut off by an entire 15 point standard deviation. Please refer to North American publications regarding comparison of the Bayley III vs BSID II in the preterm population (rather than babies with HIE).

OUR REPONSE:
-> We thank Reviewer 1 for this important comment. We agree on the inadequateness of Reference 15 (Jary et al, 2013), where data of term-born infants with hypoxic ischemic encephalopathy are described. Therefore, we deleted it and we added two other references from North America (Vohr et al., Pediatrics 2012; Sharp and DeMauro, J Dev Behav Pediatr 2017) where the focus is on the development of preterm infants.
->Concering the adjustment for defining moderate cognitive disability:
->The comparison of BSID-II and Bayley-III data is a matter of debate in many countries. Previous studies suggest that average scores on the Bayley-III may be significantly higher than scores on the BSID-II, but the magnitude of differences between the assessments is not precisely known yet. While the main question is whether the BSID-II underestimates or the BSID-III overestimates infant development, the best way to handle the comparison of the “old” and “new” developmental scores (i.e. of the two version of the Bayley Scales) would probably be to define local (national or regional) normative data rather than applying the test-normative data. In Switzerland, we do not have local-normative data for BSID-II or Bayley-III and we did not investigate a collective of preterm infants with both the BSID-II and the Bayley-III. We however observe that the mental and motor developmental indices of preterm infants tested with the BSID-II are 13 and 12 points, respectively, lower than cognitive and motor composite scores of successive preterm infants tested with the Bayley-III. Based on this observation and on the results (and recommendations) published in the literature (including USA groups), we opted to define moderate disability by shifting the cut-point by 15 points.
->UK
- Johnson et al. (for the Epicure Study, UK) in Pediatric Research 2014, propose that a Bayley-III cognitive score <85 provides the best definition of moderate to severe neurodevelopmental delay for equivalence with MDI <70.
->Victoria, Australia
- In Pediatrics (135; 5:e1258-65), Spencer et al. described how the use of local Bayley-III reference data with a cut-point of 85 in the cognitive composite score for the identification of moderate delay.
->USA
- In a small sample of 77 very preterm infants (<32 weeks) at age 18 to 22 months, Sharp and De Mauro (Philadelphia, USA) found that Bayley-III scores were significantly higher than BSID-II scores in all domains: cognitive and motor composite scores (Bayley-III) 1 and 0.6 SD higher (14.1 ± 12.9 and 9.0 ± 11.9 points, $p < .001$) than MDI and PDI scores (BSID-II), respectively. [J Dev Behav Pediatr, 2017]
- Finally, the findings of Vohr et al. (for the Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network) support a cut-point of 80-85 of the cognitive composite score (Bayley-III) for the definition of moderate NDI. [J Pediatr, 2012]

Also, possible to explore the potential bias resulting from increased proportion of outborn in those lost to follow-up. Would imputational analysis truly correct for this?

OUR RESPONSE
-> As this issue is raised in greater detail with reviewer 2, we will respond below.

Reviewer: 2
This is an interesting and nicely presented study of 2 year outcomes of extremely preterm infants born in Swiss centers classified as high or low perinatal interventional activity centers. The investigators found that centers with high perinatal interventional activity scores had lower mortality among infants

born <26 weeks' gestation, but there was no significant difference in morbidity or neurodevelopmental impairment among surviving infants. These findings are similar to those reported in other international studies. The strengths of the study include the large sample size, inclusion of the vast majority of all liveborn infants in the country, and high-quality data on neonatal and early childhood outcomes. The paper is very well-written and provides a thorough discussion of the findings in relation to other recent studies.
Comments/Questions:
Three different neurodevelopmental assessments were used during the study. Could the authors please include additional information on how the GMDS compares to the Bayley examinations?

OUR RESPONSE
-> Only 4% of survivors with 2 year follow-up were tested using GMDS. We considered the GMDS developmental quotient as equivalent to BSID-II MDI. We have adjusted the methods section to provide more detail and added the reference by Cirelli, Infant Behavior and Development 2015.

Could the authors provide more information about the outborn infants? Including the outborn infants seems somewhat problematic as perinatal interventions would not reflect the center they were admitted to, and selects for infants who survive for transfer to the perinatal centers. Also, including the proportion of inborn infants in each center in the activity score does not necessarily reflect perinatal interventional activity as in the Serenius study, where the activity score were region-based rather than center-based. Last, since the study focuses on the effect of perinatal interventions on outcomes, the rationale for including inborn status in both the interventional score and as a variable in the regression model was unclear.

OUR RESPONSE
-> We thank Reviewer 2 for this important comment. We agree that including outborn infants is somewhat problematical for the reason listed by the reviewer. We chose not to exclude outborn infants as our aim was to remain as close to the observed reality of everyday practice and as complete as possible with regards to the observed population. Also, being delivered in a perinatal center (level III) without restrictions in respiratory care, rather than a periphery center with subsequent transfer to a perinatal center, is known to benefit the newborn (Marlow, Fetal Neonatal Ed. 2014), particularly in the observed gestational age cohorts. Excluding outborn infants would thus benefit those centers with higher incidence of outborns. A prenatal transfer of a child to a perinatal center however can be interpreted as a perinatal decision to initiate life-supporting intensive care, which is why we included 'inborn' into the activity score. It may help to know that Switzerland has a web-based NICU bed availability coordination system. Including outborn in the regression model however was a mistake which we corrected: we replaced Fig. 2 in the manuscript accordingly and corrected the adjusted ORs in the results – the difference was marginal, probably due to the overall small incidence of outborn infants in Switzerland (4%-5%). The uneven distribution of outborn infants between patients with and without follow-up may however lead to bias in the crude evaluation as well as the evaluation after imputation as they are not truly missing at random. We thank both reviewers for making us aware of this. In order to address the respective issues of both reviewer 1 and 2, we performed a sensitivity analysis with prior exclusion of all outborns both for the regression and the activity score. The result was equivalent to the main result and was added to the supplemental section. We also added a sentence to the the methods section (under Data completeness) and to the result section to this effect.

The authors provide a brief justification in the Discussion section for the cut-off gestational age points used to describe Cohorts A and B. However, inclusion of the 25 weeks infants in Cohort A may obscure the effects of active perinatal intervention at lower gestational ages. I would consider adding additional information regarding how the receipt of perinatal interventions varied by gestational week in Cohort A. I would also add the actual proportion of infants in each Cohort who received each

perinatal intervention that corresponds to the top center (i.e. the center given the 100 score for each of the components). This would give the reader a greater understanding of how the approach to perinatal intervention compares with other international studies.

OUR RESPONSE
-> Switzerland is known for its restrictive approach to initiation of intensive care 'at the limit of viability'. Prior to the publication of the 2nd edition of the guidelines to care for infants at the limit of viability in Switzerland in 2011 (Berger, Swiss Medical Weekly, 2011), providing intensive care to infants below 25 weeks gestation was considered optional and often not used, particularly below 24 weeks. As of 2011, intensive care was provided mostly as of 24 weeks but only rarely below (Berger, BMJ open, 2017). Selecting a cut-off value below 25 weeks would not have yielded useful results in this study. We address this towards the end of the 2nd paragraph in the discussion. An overview over receipt of perinatal interventions by gestational age was provided in previous publications (Berger, BMJ open, 2017 and Berger, Arch.Dis.Child. 2012) which is why refer to these studies in our manuscript rather than repeating the results in this study, where they are not in the focus. We appreciate the idea of adding the actual proportion corresponding to the top center and included them in the activity score table 2 (cohort A) and supplemental table 1 (cohort B).

Were the centers similar in volume (i.e. number of infants)?

OUR RESPONSE
-> No. Swiss centers have different volumes. The number of infants are however evenly distributed between low and high activity centers which is relevant to this study. We have added the number of infants per activity group to the results section in par. 2. We are not at liberty to list each center's number of infants as this could reveal their identity.

A brief footnote in the Tables 1 and 3 regarding the SES variable (i.e. definition and scale) would be helpful.

OUR RESPONSE
-> We added: Socioeconomic status was calculated by means of a score reflecting both maternal education and paternal occupation, with a maximum and minimum scores of 12 and 2, indicating lower and higher status, respectively.

This statement was not clear: "Data per item was missing in 1.2% of datasets for the outcome "major morbidity" or less than 0.1% in all other cases."

OUR RESPONSE
-> We reformulated the passage to read: "1.2% of the datasets were missing information on "major morbidity" and were therefore eliminated in the non-imputed analyses including this outcome. Other data was missing in less than 0.1% of cases."

Please clarify why are GA and GA2 terms both included in the final regression model?

OUR RESPONSE
-> Our aim was to achieve the highest possible predictive validity (measured by c-statistics) in order to enhance the validity of risk adjustment (Adams et al, BMC Pediatr. 2017). We routinely achieve best results by combining GA and GA2. We assume this best models the non-linear dependency of most outcome incidences in very preterm infants on gestational age. We reformulated the passage in the methods section under the heading "Statistical analysis" accordingly.

I would consider noting limitations of the perinatal activity score in the Discussion. While it seems a

reasonable measure of perinatal interventions, the equal weighing of each component, interrelatedness of some of the components, and arbitrary cut-off score between high and low centers may not fully capture the variations in perinatal treatment approach.

OUR RESPONSE
-> We recognize the limitation raised by Reviewer 2 with respect of the interrelatedness of some of the components of the perinatal activity score that we used (e.g. surfactant and respiratory support). As BMJ open requests limitations in a separate list, we added another bullet point to the subheading "strengths and limitation" it in the revised manuscript version.
-> The score was however aimed to quantitatively describe the level of proactive interventions (high vs low proactive care, i.e. more vs less interventions) and not their quality or effectiveness in perinatal support. Therefore, we think that the lack of equal weighting of each score's component should not be mentioned as a limitation.
-> As far as the arbitrary cut-off point is concerned, as the study protocol is complex, we considered an easy to understand dichotomic analysis separating the units into two similar sized groups superior to a linear analysis or to determining a cut-off point via ROC (Youden index or similar principle). We rephrased the paragraph "Perinatal interventional activity score" in the methods section accordingly.

In Table 1 and 3, consider noting which variables are significantly different between groups.

OUR RESPONSE
-> Thank you for this suggestion. We are aware that this is often customary. But we prefer not listing p-values in patient characteristics overviews as they are often rendered meaningless due to multiple testing and thus easily convey a false message of difference where there is none or no difference where there actually is. We prefer qualitatively analyzing the differences and addressing necessary issues by additional analysis, such as by using multiple imputation (table 1), adjusted odds or risk ratio analysis with details on predictive validity (table 3), or restrictive analysis excluding outborn infants.

VERSION 2 – REVIEW

REVIEWER	Noelle Younge, MD, Assistant Professor of Pediatrics Duke University School of Medicine USA
REVIEW RETURNED	26-Oct-2018

GENERAL COMMENTS	I appreciate the thorough responses from the authors and the changes made in the revised manuscript. I have no further concerns.