PEER REVIEW HISTORY

BMJ Paediatrics Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Paediatrics Open. The paper was subsequently accepted for publication at BMJ Paediatrics Open.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Bronchopulmonary dysplasia: impact of severity and timing of diagnosis of on neurodevelopment of preterm infants. A retrospective study
AUTHORS	Natalucci, Giancarlo; Malavolti, Anna; Bassler, Dirk; Arlettaz Mieth, Romaine; Faldella, Giacomo; Latal, Beatrice

VERSION 1 - REVIEW

REVIEWER	Ballot, Daynia University of the Witwatersrand Johannesburg, South Africa Competing interests: I declare no competing interests
REVIEW RETURNED	24-Jul-2017

GENERAL COMMENTS	This is a well written manuscript. The research methodology is sound and the results are relevant. The authors have a balanced discussion of their findings- including strengths and limitations.

REVIEWER	Koenig, Kai Kinderarztpraxis am Bahnhof Lucerne & Children's Hospital Lucerne, Switzerland Competing interests: None
REVIEW RETURNED	31-Jul-2017

GENERAL COMMENTS	The authors investigated the effect of bronchopulmonary dysplasia severity and time-point of diagnosis on neurodevelopmental outcome in a large cohort of very preterm infants born <30 weeks gestation. This a very relevant topic as BPD is the most common morbidity in very preterm infants, however, the true impact on longterm outcome, particularly for infants with only mild or moderate BPD, remains unclear. Overall, the manuscript is well written. The study conduct is clear, the analysis is sound, and the discussion balanced. I only have a couple comments: - Page 7: the authors state that of 1270 live-born infants during the study period, 458 infants died, this translates into a mortality rate of

	36% which appears a bit on the higher side. Please clarify. - It would be interesting to know how many babies of those excluded died because of severe BPD (according to the study definitions and time-points). Is this information available for the study cohort? - The study cohort includes infants born over a 14-year period. During this period, a number of high quality studies have been published addressing new strategies and modifications of respiratory support in very preterm infants (increased use of CPAP, introduction of HFNC, new modalities of surfactant administration etc etc). The readers may benefit of a short description of the NICUs respirstory support management preferences throughout the study period.

REVIEWER	Datta, Vikram Department of Neonatology, Lady Hardinge Medical College, New Delhi, India Competing interests: None
REVIEW RETURNED	05-Aug-2017

GENERAL COMMENTS	The authors need to be congratulated for addressing an important cause of neurodevelopment impairment in the preterm neonates. The title should be rephrased to include the study design. Why preterm neonates <30 weeks were included. What efforts were made to address potential sources of bias. A lost to follow up rate of 19% is on the higher side ,does that become one of the limitations as well. In the study flow diagram the number of deaths have been depicted as 458 whereas they have been cited on page 7 as 516 in text , can authors explain this discrepancy. In table 3 authors have mentioned Sensibility , i presume it is sensitivity ,can this be kindly addressed. Can the authors mention in detail about generalisability and recommendations for future research in concluding paragraph.

REVIEWER	Carrara, Greta IRCCS - "Mario Negri" Institute for Pharmacological Research, Italy Competing interests: None declared
REVIEW RETURNED	12-Sep-2017

GENERAL COMMENTS	The manuscript entitled “Bronchopulmonary dysplasia: impact of severity and timing of diagnosis of on neurodevelopment of preterm infants.” presents an interesting study of the impact of bronchopulmonary dysplasia (BPD), its severity, and timing of diagnosis on neurodevelopment impairment (NDI). Patients considered in the study are preterm (<30 weeks) infants, undergoing follow-up at the corrected age of 24 months. Results suggest an association between severe BPD and NDI. I find the manuscript to be generally well written and easy to follow. Nevertheless, I would like to submit the authors some questions and suggestions: 1. Some concerns about BPD definition:

	a) the authors use only the 2000-NICHD definition by Jobe AH and Bancalari E., but they also state that BPD is defined heterogeneously in the literature. Did the authors perform some sensitivity analyses using other BPD definitions? It might be interesting to add the results of these analyses to the online supplement in order to corroborate the results. b) Where is information on BPD diagnosis derived from? Is it explicitly written in the neonatal charts or was it deduced from some parameters written in neonatal charts (e.g. FiO2)? In the latter situation, which is the role played by clinician's subjectivity? It might be useful that at least two different clinicians deduce diagnosis, assessing then their agreement. 2. Some concerns about NDI definition: a) Why do the authors use three different tests in order to assess NDI? How do they allocate children to a test in place of another? b) To define NDI, authors use a cut-off of -2SD: if a child has a score < -2SD then he has the NDI. Is this cut-off validated? Might this cut-off be variated, using for example -1SD, for sensitivity analyses? How the results would change by varying this cut-off? 3. Regarding the estimate of 'gestational age', does the exact date of initiation of pregnancy appear in the neonatal charts? Alternatively, is this date deduced from other data? If so, BPD might be scored in a wrong week? 4. Authors state that children were invited to a follow-up at the corrected age of 18 to 24. According to me, this gap is very large because neurodevelopment of 18 months old child may be very different from the neurodevelopment of 2 years old child. Are the scores (Bayley 2, Bayley 3 and GMDS) weighted on the basis of the age? In the results the range of the corrected age at follow-up is even wider: from a minimum of 16.5 months to a maximum of 37.6. I would ask the authors to add some more statistics regarding the distribution of the corrected age at follow-up: e.g., a boxplot and value of interquartile range. Which is the mean of corrected age at follow-up in BPD group and no BPD one (with p-value of difference)? 5. Authors suggest that the higher prevalence of BPD in their study, with the respect to the literature, may be due to a possible selection bias. What could be the reason of this bias? 6. Some concerns about statistical methods: a) Table 1: it may be useful to add column with the statistics of all patients (N=610) b) Table 3: sensibility/specificity/positive predictive value/negative predictive value can be estimated only for dichotomous variables. Thus, for first column 'BPD' it is easy to understand that authors test patients with BPD versus patients without BPD. For the second column 'Mild BPD', did the authors test Mild versus Not Mild (i.e. No BPD, but also moderate and severe BPD)? I think that this way is not meaningful, because patients with moderate or severe BPD are grouped together with patients without BPD. There is the same problem, eventually, also for Moderate BPD. c) Table 3: authors report a high negative predictive value

	(87%); I think that it is better to state in the manuscript that, by definition, a low prevalence of disease (NPI) implies high value of negative predictive value. d) I think that it is necessary to present the full results of the multivariable models (all variables with their coefficients/OR, standard errors, …), at least in the online supplement. e) Authors do not show any information about the goodness of fit of the built models, in term of both calibration and discrimination. Regarding discrimination, authors could show the Area Under the ROC Curve. Regarding calibration, authors could use the Hosmer-Lemeshow test and the 'calibration belt' method (Stat Med. 2014;33(14):2390-407 and Stat Med. 2016;35(5):709-20). f) In the models there are some numeric and continuous variables: how did the authors check the linearity assumption? What are the results? g) In table 2 there are some very high ORs (e.g. 5.6 and 16.6) with very wide 95%CI (2.0-15.9 and 4.6-59.9, respectively). What is the explanation of that? The same problem arises also in Table A, B and C. h) In order to state that there is a difference in the prediction capability between severe BPD at 40 weeks' PMA and 36 weeks' PMA, the respective confidence intervals must not overlap after correcting CI with the method proposed by Payton (J Insect Sci. 2003; 3:34). Otherwise, statistically it is not correct to claim that severe BPD at 40 weeks' PMA allows a better prediction of NDI. Specifically, correcting CI with Payton method, one obtains 5.6 (95%CI: 2.7-11.8) and 16.6 (6.6-41.5), for 36 and 40 weeks respectively. Although this result is suggestive, it does not prove that severe BPD at 40 weeks' PMA allows a better prediction of NDI.

VERSION 1 – AUTHOR RESPONSE

Reviewer1
Reply 1: We thank Reviewer 1 for the encouraging comments.

Reviewer2
Comment 1: Page 7: the authors state that of 1270 live-born infants during the study period, 458 infants died, this translates into a mortality rate of 36% which appears a bit on the higher side. Please clarify.
Reply 1: We thank Reviewer 2 for this very important comment that helped us to notice an unclear aspect of the Methods section (*Study subjects*), that need to be corrected. In fact, the information concerning the deceased (and so excluded) infants is incomplete. Within the group of 458 deceased infants, there are 179 infants who were a priori treated in a palliative way in the delivery room and died there [gestational age 22 weeks (n=12), 23 weeks (n=69), 24 weeks (n=70, primary palliative care up to 2010-2011) weeks, and major congenital (including chromosomal) anomaly (n=28, i.e. trisomy 13 and 18, anencephaly)]. Hoping that the editor and the reviewers will accept this change, we opted to modify the Methods section regarding the *Study subjects* adding to the exclusion criteria following criterion: "a priori palliative care". Accordingly, we added information about palliative care management in Figure 1. (*Please see Methods, page 4, lines 5-6; Results, page 7, lines 2-4; and Figure 1*).

Comment 2: It would be interesting to know how many babies of those excluded died because of severe BPD (according to the study definitions and time-points). Is this information available for the study cohort?
Reply 2: Among the excluded infants, 12 infants with a diagnosis of BPD died. Among them:

- 7 infants died before 36 weeks' PMA (complications of NEC and Sepsis in 4 and 3 cases, respectively;

- 1 infant died between 36 and 40 weeks' PMA, because of complications of CMV infection, including pneumonia (moderate BPD before development of pneumonia);

- 4 infants died at a PMA range of 41-to-47 weeks' PMA, because of complications of NEC and Sepsis in 3 and 1 cases, respectively (among them, 3 infants had moderate BPD and 1 infant had severe BPD before development of the neonatal complications).

Comment 3: The study cohort includes infants born over a 14-year period. During this period, a number of high quality studies have been published addressing new strategies and modifications of respiratory support in very preterm infants (increased use of CPAP, introduction of HFNC, new modalities of surfactant administration, etc.). The readers may benefit of a short description of the NICUs respiratory support management preferences throughout the study period.

Reply 3: This comment of Reviewer 2 is important. A description of the NICUs respiratory support management practices was initially planned by the authors. For many reasons we opted to omit this information for following main reasons:

a: The aim of the study is to describe the neurodevelopmental outcome of infants with a diagnosis of BPD according to different definitions.

b: Respiratory support strategy did not relevantly change during the study period (primary surfactant administration by means of an endotracheal tube in delivery room for infants < 26 weeks of gestation; N-CPAP primarily for infants ≥ 26 weeks of gestation; no use of HFNC on a regular basis until 2016; HFO ventilation as rescue intervention except for infants with congenital diaphragmatic hernia or lung emphysema.

c: As we do not provide any long-term respiratory outcome data of the study infants, we think that an unspecific description of the respiratory support strategy of the study infants goes behind the scope of the study.

Reviewer3
Comment 1: The title should be rephrased to include the study design.

Reply 1: We rephrased the title of the manuscript accordingly ("*… A retrospective study.*"). Please see Title and Editor's Comment 1.

Comment 2: Why preterm neonates <30 weeks were included.

Reply 2: We thank reviewer 3 for this pertinent question. We included preterm neonates <30 weeks of gestation because for this population we were able to provide clinical follow-up data at 2 years of corrected age.

Comment 3: What efforts were made to address potential sources of bias.

Reply 3: We tried to reduce selection bias by means of a clear definition of the study population of very preterm infants with almost (>95%) complete neonatal dataset and representative follow-up data at 2 years (>80%).

- Neonatal (including days on FiO2>21%, days on positive pressure support and differentiated information regarding mechanical ventilation and CPAP/NIPPV) and follow-up data analysed within the present study was prospectively collected according to the recommendation of the Swiss Neonatal Network and Follow-up Group. Only the specific information about the respiratory support of

the infant at 40 weeks' PMA, which is not included in the Swiss Minimal Neonatal Data Set (yet), was retrospectively collected.

- We used outcome measures that are validated (developmental tests, GMFCS), and outcome composite definitions that used and already published from the major research groups focussing on outcome after prematurity.

- Models applied for the assessment of the relation between BPD severity and outcome were adjusted for variables that were significantly unequally distributed among groups, and post hoc analyses were performed by adjusting the models to include known neonatal predictors of poor neurodevelopment.

- Finally, a sensitivity analysis that focused purely on infants tested with the most often used developmental test in the cohort (Bayley-II) was performed in order to verify the study observations and support its conclusions.

Comment 4: A lost to follow up rate of 19% is on the higher side, does that become one of the limitations as well.
Reply 4: This comment is pertinent. We added and commented this information in the Discussion section accordingly. (*Please see page 11, lines 3-7*)

Comment 5: In the study flow diagram the number of deaths have been depicted as 458 whereas they have been cited on page 7 as 516 in text, can authors explain this discrepancy.
Reply 5: We thank Reviewer 3 for this comment but, after an attentive control of the text, we notice that there is no discrepancy between the two pieces of information: 516 is the total number of excluded infants, while 458 is the number of deceased infants (the other 58 infants have been excluded for other reasons).

Comment 6: In table 3 authors have mentioned Sensibility, I presume it is sensitivity, can this be kindly addressed.
Reply 6: We apologize for the typing error. We corrected the text accordingly. (*Please see page 20, Table 3*)

Comment 7: Can the authors mention in detail about generalisability and recommendations for future research in concluding paragraph.
Reply 7: We add a comment in the conclusive paragraph of the Discussion section accordingly. (*Please see page 11, second to last line*)

Reviewer:4
Comment 1: Some concerns about BPD definition:
a) the authors use only the 2000-NICHD definition by Jobe AH and Bancalari E., but they also state that BPD is defined heterogeneously in the literature. Did the authors perform some sensitivity analyses using other BPD definitions? It might be interesting to add the results of these analyses to the online supplement in order to corroborate the results.
Reply 1.a: We thank Reviewer 4 for this comment. We did not perform other analyses of the association between BPD severity according to other definitions because in some cases we did not provide enough information (further specific clinical or lung imaging data) to do it.

b) Where is information on BPD diagnosis derived from? Is it explicitly written in the neonatal charts or was it deduced from some parameters written in neonatal charts (e.g. FiO2)? In the latter situation, which is the role played by clinician's subjectivity? It might be useful that at least two different clinicians deduce diagnosis, assessing then their agreement.
Reply 1.b: This is a very interesting point. The BPD diagnosis and its severity up to 36 weeks' PMA

were derived from a prospectively collected dataset (Swiss Minimal Neonatal Dataset). This information was verified (for each study infant) by means of an attentive study of the neonatal charts.

The information about BPD severity at 40 weeks' PMA was derived by the retrospective review of the clinical charts (FiO2 and/or respiratory support).

BPD and its severity were objectively identified at each observation time-points (28 days of life, 36 and 40 weeks' PMA) according to the definitions used in the study (100% agreement of all authors).

The problem of the clinician's subjectivity might have played a role in the clinical setting during the treatment of some study infants. While we (the authors) were able to objectively identify (and not deduce) BPD definition criteria based on the explicitly documented level of respiratory support at a certain time-point, we were not always able to identify whether the indication of the started respiratory support was based on BPD or on other health problems.

Comment 2: Some concerns about NDI definition:
2.a) Why do the authors use three different tests in order to assess NDI? How do they allocate children to a test in place of another?
Reply 2.a: During the study period (14 years) the follow-up of preterm infants in Switzerland evolved and developmental tests changed. The three tests have been used because of following reason. According to the recommendation of the Swiss Neonatal Network, the follow-up examinations were based from 2000 to 2013 on the Bayley Scales of Infant Development (2. Ed.) and thereafter on the Bayley scales of infant and toddler development (3. Ed.). Early on ("transition period" up to 2002), the Griffiths mental development scales-Revised (GMDS) was still used in a few children and starting from 2011 (year of introduction of the Bayley-III in many Swiss Follow-up centres) an increasing number of infants has been assessed with the Bayley-III.

2.b) To define NDI, authors use a cut-off of -2SD: if a child has a score < -2SD then he has the NDI. Is this cut-off validated? Might this cut-off be variated, using for example -1SD, for sensitivity analyses? How the results would change by varying this cut-off?
Reply 2.b: There is no international consensus on the definition of NDI. With respect to the neurodevelopmental outcome at age 2 years, many study groups (reporting multicentre or population based data, including the Swiss Follow-up Group) apply the outcome definitions according to the guidelines of the working group of the British Association of Perinatal Medicine and the National Neonatal Audit Project on the Classification of Health Status (2008), where severe and moderate neurodevelopmental disability is (among others) defined as a (mental and/or psychomotor) development index <-3SD and <-2SD, respectively.

In a previous study (Schlapbach et al, BMC Pediatrics 2012) we introduced the concept of favourable outcome at 2 years, defined as (among others) a developmental index ≥ -2SD.

The cut-off that we used has not been validated in the literature, as it is the case for other cut-offs. It seems (based on the literature) that a neurodevelopmental index or score at age 2 years (derived from a form of standardized developmental test) <-2SD has a good predictive value for neurodevelopmental impairment in later life, while an index or score ≥-2SD;<-1SD at age 2 years is not predictive for later developmental performance. From a clinical point of view (information for clinicians, for parents, for therapists) we did not considered the cut-off -1SD as useful because poorly informative or the cut-off -3SD because of the low number of affected infants.

Comment 3: Regarding the estimate of 'gestational age', does the exact date of initiation of pregnancy appear in the neonatal charts? Alternatively, is this date deduced from other data? If so, BPD might be scored in a wrong week?
Reply 3: We thank reviewer 4 for this pertinent comment. In the neonatal charts, the gestational age of the infants was based on the best estimated initiation of pregnancy which is also documented in the (maternal) charts. On overall, the estimation of the gestational age of a foetus could vary (week/day + error range) according to the method used. In our study sample the estimation made by the obstetrician in charge was based on early (I trimester) prenatal ultrasound findings or obstetric measurements based on the last menstrual period. That implies an estimate error range of 3 to (max) 7 days. The problem of a range of error in the estimation of the gestational age of a newborn infant is a major challenge that does not concern only BPD definition but many other aspects of neonatal care.

To date, the definition of BPD based (among others) on periods of time under respiratory support is matter of debate.

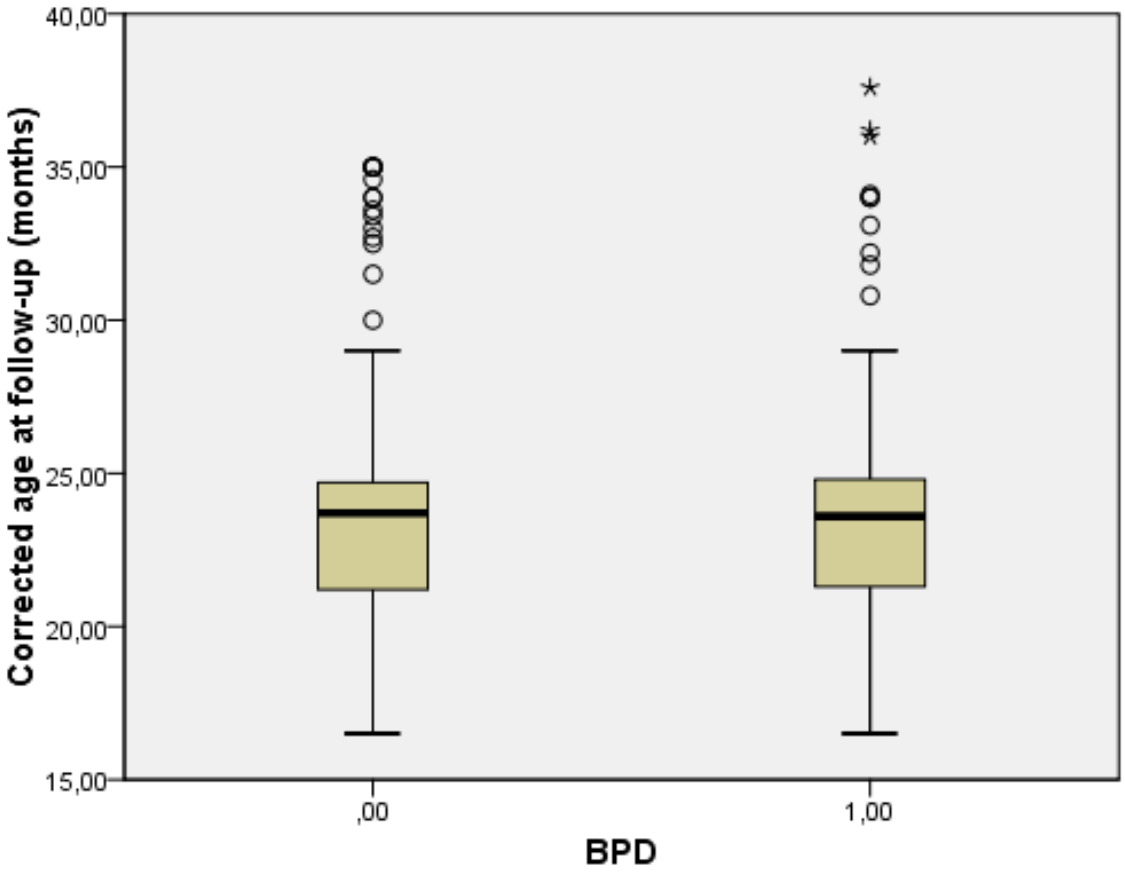
Comment 4: Authors state that children were invited to a follow-up at the corrected age of 18 to 24. According to me, this gap is very large because neurodevelopment of 18 months old child may be very different from the neurodevelopment of 2 years old child. **4.a)** Are the scores (Bayley 2, Bayley 3 and GMDS) weighted on the basis of the age?
4.b) In the results the range of the corrected age at follow-up is even wider: from a minimum of 16.5 months to a maximum of 37.6. I would ask the authors to add some more statistics regarding the distribution of the corrected age at follow-up: e.g., a boxplot and value of interquartile range. Which is the mean of corrected age at follow-up in BPD group and no BPD one (with p-value of difference)?
Reply 4.a: Yes, the scores of the standardised tests used in this study are weighted on the basis of the age (in months) of the infants.

Reply 4.b: The Swiss Neonatal Network and Follow-up Group collect systematically neurodevelopmental follow-up data of very preterm infants at 18-to-24 months of corrected age and at age 4-to-6 years. These ranges are wide defined because the Network is aware of the difficulties to organize coordinated visits (at the very same age period) of very preterm infants in a clinical setting that has no financial support from third parties.

In the present study, the range of the infants' corrected age at follow-up is wide primarily because of following reasons: the clinical follow-up examinations were planned for a target period of life but some families were not able to come at the planed date because of (among others) private reasons, transport problems, holydays, health problems of the infant.

The median age at follow-up of the study infants was 23.1 months, SD 3.6, median 23.6 months, IQR 21.2 – 24.8 months. There was no difference in the mean age at follow-up between infants with [23.1 (3.4)] and without [23.2 (3.8)] BPD (p=0.770). We added this information in the revised version of the manuscript. (*Please see page 7, second to last sentence*).

Please find attached the boxplot representing the comparison of corrected age of infants with and without BPD at follow-up period.



Boxplot with median value (dark horizontal line in the middle of the boxes), upper (75th) and lower (25th) percentile (top and bottom of the box, respectively). The T bars represent maximum (up) and minimum (down) values excluding outliers and extremes. Outliers and extremes values are represented by the points and the asterisks.

Comment 5: Authors suggest that the higher prevalence of BPD in their study, with the respect to the literature, may be due to a possible selection bias. What could be the reason of this bias? **Reply 5:** A comparison with data from the literature is difficult because the differences in the BPD definitions. We suppose that the reason of the high prevalence of BPD might be the caused by a selection of study subjects (those visited for follow-up had a higher prevalence than those lost to follow-up). Additionally we suppose that the clinicians' subjective judgments might have influenced the respiratory support of the study infants and so the classification of BPD (especially in the case of infants with mild and moderate forms of BPD). While we can partially prove the first hypothesis, the second hypothesis remains of speculative nature. We added this aspect in the Discussion section. (Please see page 10, second paragraph).

Comment 6: Some concerns about statistical methods:

6.a) Table 1: it may be useful to add column with the statistics of all patients (N=610)

Reply 6.a: We add this information in Table 1 accordingly. Additionally, we noticed an error regarding the information about the socioeconomic score of the infants' families. We corrected this information accordingly. (*Please see Page 17, Table 1*)

6.b) Table 3: sensibility/specificity/positive predictive value/negative predictive value can be estimated only for dichotomous variables. Thus, for first column 'BPD' it is easy to understand that authors test patients with BPD versus patients without BPD. For the second column 'Mild BPD', did the authors test Mild versus Not Mild (i.e. No BPD, but also moderate and severe BPD)? I think that this way is not meaningful, because patients with moderate or severe BPD are grouped together with patients without BPD. There is the same problem, eventually, also for Moderate BPD.

Reply 6.b: We thank Reviewer 4 for this comment and we are sorry, that we were not able to better inform readers about this point. We tested all forms of BPD (yes, mild, moderate, severe) with the reference No BPD (examples: Mild BPD at 36 weeks' PMA vs No BPD; Severe BPD at 40 weeks' PMA vs No BPD). For this reason we mentioned in tables 2a-c and all supplementary (online only) tables that No BPD is the reference for the tests. We reworded the legend of each concerned table in the revised manuscript trying to enhance this important information. We did not highlight changes in the supplemental material files by using the track changes because of the format of the documents (PDF). We apologize for that. (*Please see page 19 for tables 2a, 2b, 2c, and suppl. Material A, B,C, and D*)

6.c) Table 3: authors report a high negative predictive value (87%); I think that it is better to state in the manuscript that, by definition, a low prevalence of disease (NPI) implies high value of negative predictive value.

Reply 6.c: We thank Reviewer 4 for this pertinent comment. We added this information in the Discussion section of the revised manuscript, commenting the strengths and limitations of the study. (*Please see page 10, last 2 lines; page 11 first line*)

6.d) I think that it is necessary to present the full results of the multivariable models (all variables with their coefficients/OR, standard errors, …), at least in the online supplement.

Reply 6.d: We agree with Reviewer 4 with respect to the relevance of presenting the results of the models used in the study. Consequently, we added an online supplement file (supplemental Table D) with full information regarding the models analysing the relationship between BPD/BPD severity (at 36 and 40 weeks' PMA) and the outcome measure NDI. Table D is cited in the main text on page 8, last to 8th and 7th lines. However we opted not to add the information regarding the complete set of models a) due to file size concerns (7 defined BPD forms, 8 outcome measures, 2 models adjusted for 8 variables that were significantly unequally distributed among infants with and without BPD, and for 7 known neonatal predictors of poor neurodevelopment, that is 112 tables); and b) because of the limited impact of the information.

6.e) Authors do not show any information about the goodness of fit of the built models, in term of both calibration and discrimination. Regarding discrimination, authors could show the Area Under the ROC Curve. Regarding calibration, authors could use the Hosmer-Lemeshow test and the 'calibration belt' method (Stat Med. 2014;33(14):2390-407 and Stat Med. 2016;35(5):709-20).

Reply 6.e: We agree with Reviewer 4' with respect to the importance of the quality of the models that we used. Concerning the measure of goodness-of-fit of the logistic regression models, we would like to focus on the four models of major interests, that is the models where the association between severe BPD at 36 (a-b) and at 40 (c-d) weeks' PMA is assessed (both models adjusted for variables that are significantly unequally distributed among groups and for known neonatal predictors of poor neurodevelopment).

- Regarding the calibration of the models, we report following results of the Hosmer & Lemeshow test, that suggest that the four models are a good fit to the data. (Please see below)

- Regarding discrimination of the models, we report following area under the ROC curve values of the models, that suggest that the logistic regression models classify the group significantly better than by chance (AUC relevantly and significantly different form 0.5). (Please see below)

a) Multivariate association between severe BPD at 36 weeks' PMA and NDI adjusted for variables that were significantly unequally distributed among groups

Hosmer & Lemeshow test: Chi-square 6.513, df 8, p=0.590

Area under the ROC curve (95%-CI): 0.875 (0.824-0.926), SE 0.026, p<0.001

b) Multivariate association between severe BPD at 36 weeks' PMA and NDI adjusted for known neonatal predictors of poor neurodevelopment

Hosmer & Lemeshow test: Chi-square 10.032, df 8, p=0.263

Area under the ROC curve (95%-CI): 0.802 (0.725-0.880), SE 0.040, p<0.001

c) Multivariate association between severe BPD at 40 weeks' PMA and NDI adjusted for variables that were significantly unequally distributed among groups

Hosmer & Lemeshow test: Chi-square 7.058, df 8, p=0.530

Area under the ROC curve (95%-CI): 0.992 (0.978-1.000), SE 0.002, p<0.001

d) Multivariate association between severe BPD at 40 weeks' PMA and NDI adjusted for known neonatal predictors of poor neurodevelopment

Hosmer & Lemeshow test: Chi-square 4.853, df 8, p=0.773

Area under the ROC curve (95%-CI): 0.976 (0.957-0.996), SE 0.010, p<0.001

6.f) In the models there are some numeric and continuous variables: how did the authors check the linearity assumption? What are the results?

Reply 6.f: This point raised form Reviewer 4 is important.

Concerning the logistic regression models (association between BPD and NDI):

We assumed that the log odds were related to the predictors in a linear fashion and we used the Box-Tidwell test to evaluate this assumption. We included in each model interactions between each continuous predictor and its natural logarithm and we observed that no interaction was significant.

Concerning the linear regression models (association between BPD and MDI or PDI):

We checked the Pearson's bivariate correlation and we found that the continuous variables were significantly correlated (r range 0.140-0.691). Thereafter, we checked whether there was a linear relationship in the continuous (dependent and independent) data by mean of scatter plots. The scatter plots indicated a good linear relationship between the dependent (outcome) variables MDI and PDI and the independent variables, except for the weak association between MDI and gestational age. Based on our previous observations and those of other groups, this was surprising but indicates that long-term outcome in preterm infants is influenced by both biological and non-biological (psychosocial, familial) factors that we were not able to document in this study.

Finally, we checked for multivariate normality (in the models) by means of normal Q-Q plots of each variable and we found that multivariate normality was present in the data.

We would like to add supplemental information concerning the insertion of the continuous co-variables in the regression models:

We added these co-variables in the regression models because they were either significantly unequally distributed among groups or because they represented known neonatal predictors of poor neurodevelopment.

Other regression models, where those co-variables were excluded, and regression models where the co-variables were included but in a dichotomized form (using standard cut-offs used in clinical setting, i.e. gestational age < 28 weeks or birth weight z-score < -1.27) provided similar results to those presented in the manuscript.

6.g) In table 2 there are some very high ORs (e.g. 5.6 and 16.6) with very wide 95%CI (2.0-15.9 and 4.6-59.9, respectively). What is the explanation of that? The same problem arises also in Table A, B and C.
Reply 6.g: We agree with Reviewer 4's Comment regarding the high OR and the wide confidence intervals reported in this manuscript. We think that these observations reflect the low volume of severe BPD and NDI cases observed and the variability of outcome measures within the studied group. While think, that the measured ORs are of clinical relevance, we commented this aspect in the Discussion section of the revised manuscript regarding the strengths and limitations of the study, pointing out the width of the confidence intervals. (*Please see page 10, last sentence*)

Similarly, the conclusions of the study have been slightly reworded, accordingly. (*Please see page 2 Abstract's Conclusion, page 9 1st Discussion's paragraph, and pages 11-12 conclusive Discussion's paragraph*)

6.h) In order to state that there is a difference in the prediction capability between severe BPD at 40 weeks' PMA and 36 weeks' PMA, the respective confidence intervals must not overlap after correcting CI with the method proposed by Payton (J Insect Sci. 2003; 3:34). Otherwise, statistically it is not correct to claim that severe BPD at 40 weeks' PMA allows a better prediction of NDI. Specifically, correcting CI with Payton method, one obtains 5.6 (95%CI: 2.7-11.8) and 16.6 (6.6-41.5), for 36 and 40 weeks respectively. Although this result is suggestive, it does not prove that severe BPD at 40 weeks' PMA allows a better prediction of NDI.
Reply 6.h: We thank Reviewer 4 for this specific and very useful comment. The authors tried to cautiously suggest that the study results are evocative for a better prediction of NDI at 2 years of age by severe BPD at 40 that at 36 weeks' PMA. We are sorry, that this point has not been expressed more clearly. We reworded the abstract's conclusion and key parts of the Discussion section regarding the prediction of NDI of severe BDI at 36 versus at 40 weeks' PMA, trying to give more accents to the fact that the present findings do not prove that severe BPD at 40 weeks' PMA allows a better prediction of NDI. (*Please see page 2, last Abstract's sentence; page 9, Discussion, lines 9-11; pages 10-11, second sentence of the conclusive paragraph*)

VERSION 2 – REVIEW

REVIEWER	Koenig, Kai Kinderarztpraxis am Bahnhof & Children's Hospital Lucerne, Switzerland Competing interests: None
REVIEW RETURNED	06-Oct-2017

GENERAL COMMENTS	Thank you very much for your detailed response to the reviewers' comments.

REVIEWER	Ballot, Daynia University of the Witwatersrand South Africa Competing interests: I have no competing interests to declare
REVIEW RETURNED	11-Oct-2017

GENERAL COMMENTS	This is an interesting article that reviews the relationship between the severity of BPD and neurodevelopmental outcome at 2 years of age in a group of preterm infants. The conclusion drawn is that severe BPD at 40 weeks PMA is associated with neurodevelopmental impairment. The study is well done and the research methodology is clearly explained. My main concern is that the babies were assessed using different tools - Baylyey 2, Bayley 3 and GMDS. The authors have done quite a lot of statistics to try and remedy this, but point out that there is no local normative data for the Bayley 3 vs Bayley2. They have chosen to use a cut-off of 85 as NDI on the Bayley 3, based on other reports. The authors admit that this may overestimate NDI. My concern is that most babies were assessed with BSID 2 and the numbers are really small in BSID 3- particularly in the "Severe BPD group at 40 weeks. In addition, there are very small numbers of infants with deafness and blindness (table b - multivariable regression). I do not think that including the BSID 3 assessments adds anything to the paper and I would recommend excluding this group. There are a few typos - Page 3 line 36 despite (not espite) Page 11 - Line 7 - the follow up rate is not 19% - it was 89% (which is actually very good) Page 12 Line 14 Acknowledgements spelt incorrectly

REVIEWER	Carrara, Greta IRCCS - "Mario Negri" Institute for Pharmacological Research, Italy Competing interests: None declared
REVIEW RETURNED	19-Oct-2017

GENERAL COMMENTS	Thank you for properly addressing most of the comments. Remain only few suggestions and some minor concerns: a. The authors could add among the study limitations the problem of clinician's subjectivity cited in the Reply 1.b b. Regarding the Reply 6.b, I think that there was a misunderstanding: I was referring to the table 3 and not to the tables 2(a-c). So, it is still unclear the way of estimation of Sensitivity/Specificity/PPV/NPV in Table 3. c. Regarding the Reply 6.e, authors reported excellent results about goodness of fit, with good calibration and high values of AUC. I would suggest the authors to report these in the manuscript (or at least in the online supplement). d. Regarding the Reply 6.g and 6.h, thank you for rewording your conclusions. Could you reword also the second point of the paragraph 'What the study adds'? e. Some typos: - Page 1 of supplementary D: 95%CI of BPD (vs no BPD) is 0.8-1.2.5 instead of 0.8-2.5 - Page 11 of the manuscript: there is, I think, an error in the phrase 'follow-up rate of 19%' - In general, authors can rewrite p-values 0.000 as <0.001

REVIEWER	Datta, Vikram Department of Neonatology, Lady Hardinge Medical College, New Delhi, India

	Competing interests: None
REVIEW RETURNED	24-Oct-2017

GENERAL COMMENTS	Kindly mention the lost to follow up rate of 19% and not follow up rate in the limitations.

VERSION 2 – AUTHOR RESPONSE

REVIEWER 2

Comment 1: My main concern is that the babies were assessed using different tools - Bayley 2, Bayley 3 and GMDS. The authors have done quite a lot of statistics to try and remedy this, but point out that there is no local normative data for the Bayley 3 vs Bayley2. They have chosen to use a cut-off of 85 as NDI on the Bayley 3, based on other reports. The authors admit that this may overestimate NDI. My concern is that most babies were assessed with BSID 2 and the numbers are really small in BSID 3- particularly in the "Severe BPD group at 40 weeks. In addition, there are very small numbers of infants with deafness and blindness (table b - multivariable regression). I do not think that including the BSID 3 assessments adds anything to the paper and I would recommend excluding this group.

Reply 1: We thank Reviewer 2 for this important comment. To solve the problem of the use of different developmental tests, we performed a sensitivity analysis excluding children tested with the Bayley-III and GMDS (online Table B). The results of this analysis confirm the results of the main analysis concerning the whole cohort of infants. Additionally, we analysed mean differences between developmental scores (indices) of infants with and without BPD, that have been tested only with Bayley-II and we mentioned the results in the Results' section.

Concerning the part of the Discussion, where we comment on the use of Bayley-III, we are sorry for the misunderstanding. We only aimed to comment on the fact that in our study we were not able to show, that Bayley-III tend to overestimate the developmental level (not the NDI rate) of infants as it has been reported in many studies.

Comment 2: There are a few typos - Page 3 line 36 despite (not espite) Page 11 - Line 7 - the follow up rate is not 19% - it was 89% (which is actually very good) Page 12 Line 14 Acknowledgements spelt incorrectly
Reply 2: We thank Reviewer 2 for this correction and we apologize for the typos. We corrected all of them accordingly. Concerning the rate of 19% mentioned in the Discussion section we apologize for the typing error. In fact, we aimed to mention the rate of infants lost to follow-up.

REVIEWER 3

Comment 1: The authors could add among the study limitations the problem of clinician's subjectivity cited in the Reply 1.b
Reply 1: We added this information in the Limitations' paragraph of the Discussion section. (*Please see page 11, lines 15-17)*.

Comment 2: Regarding the Reply 6.b, I think that there was a misunderstanding: I was referring to the table 3 and not to the tables 2(a-c). So, it is still unclear the way of estimation of Sensitivity/Specificity/PPV/NPV in Table 3.

Reply 2: We are sorry for having misunderstood comment 6b of Reviewer 3 (first revision) and we recognize the pertinence of the point raised by the Reviewer. We calculated predictive values of any form of BPD (BPD / severity form /at 36 and 40 weeks' PMA) in relation to 'No BPD' (as Reference). We added this information in the revised text of the Statistics section and the legend of Table 3 accordingly. (*Please see page 6, line 18 and Table 3 on page 21*)

Comment 3: Regarding the Reply 6.e, authors reported excellent results about goodness of fit, with good calibration and high values of AUC. I would suggest the authors to report these in the manuscript (or at least in the online supplement).
Reply 3: We thank Reviewer 3 for this suggestion. For easier reading of the text and tables we mentioned the two methods (AUC and Hosmer & Lemeshow test) in the statistics section (Please see page , lines) and we opted to add the information concerning the results in the legend of each table (*please see page 6, lines 13-15, Table 2 (pages 19-20) and Supplemental table B*).

Comment 4: Regarding the Reply 6.g and 6.h, thank you for rewording your conclusions. Could you reword also the second point of the paragraph 'What the study adds'?
Reply 4: We reworded the second point of the paragraph 'What the study adds' accordingly.

Comment 5: Some typos: **a)** Page 1 of supplementary D: 95%CI of BPD (vs no BPD) is 0.8-1.2.5 instead of 0.8-2.5. **b)** Page 11 of the manuscript: there is, I think, an error in the phrase 'follow-up rate of 19%'. **c)** In general, authors can rewrite p-values 0.000 as <0.001
Reply 5: We thank Reviewer 3 for these comments. We are sorry for these typos. We corrected all of them accordingly: a) 0.8-2.5 is the correct form; b) Lost to follow-up rate is the correct form; c) all 0.000 have been changed in < 0.001.

REVIEWER 4
Comment 1: Kindly mention the lost to follow up rate of 19% and not follow up rate in the limitations.

Reply 1: We thank Reviewer 4 for this comment. The limitation concerning the rate of infants lost to follow-up has been corrected accordingly. (*Please see page 11, line 8*).