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.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Influence of timing of initiation of therapeutic hypothermia on brain MRI and neurodevelopment at 18 months in infants with HIE: a retrospective cohort study
AUTHORS	Guillot, Mireille; Philippe, Marissa; Miller, Elka; Davila, Jorge; Barrowman, Nicholas; Harrison, Mary-Ann; Ben Fadel, Nadya; Redpath, Stephanie; Lemyre, Brigitte

VERSION 1 – REVIEW

REVIEWER	Reviewer name: Malcolm Battin Institution and Country: Newborn Services, Auckland City Hospital, Auckland, New Zealand Competing interests: None
REVIEW RETURNED	25-Jan-2019

GENERAL COMMENTS	This paper reports a single centre experience with therapeutic hypothermia between 2009 and 2016. The study objective was to examine the effect of timing of therapeutic hypothermia (TH) onset (before and after 3 hrs of age) on the pattern and severity of brain injury on MRI and neurodevelopmental outcomes. The conclusion was that “early TH started before 3 h of life was neither associated with less brain lesions on MRI nor better neurodevelopmental outcomes”. Although this wording is quite cautious it is presented as “challenging” the preclinical data demonstrating that TH is more effective when started early (see section on “What this study adds). My major concern with the paper, as presented, is that it does not provide the reader with enough detail to judge the robustness of the conclusions based on analysis of a modest retrospective cohort. First, there is the question of power to detect a difference between the two (<3 h Vs >3h) groups. Obviously, a retrospective cohort is limited to the number of cases during the given time frame. In the manuscript the size of the cohort is listed as a strength, “the largest to report on influence of timing of TH of the study” (p12). However, on the previous page it is stated that “establishing whether early TH has additional benefits may require similarly larger cohorts”. The size of previously reported cohorts is not the question and this paper needs to be very clear whether the sample size was sufficient to adequately detect differences between the two groups and fulfil the study objective or not. Second, in a retrospective cohort where onset of cooling is not randomized it would be important to consider and account for other clinical factors that might affect outcome. For example the arrival time of the transport team and thus initiation of hypothermia could be influenced by condition at presentation and/or referring team concerns.

	Figure 1 indicates that the early TH group included more babies who died before MRI (7 vs 1) and more who died before discharge (6 Vs 2) suggesting that the groups may be different in clinical status. Table 2 summarizes the clinical variables and shows minor differences in resuscitation only. However, in this table all of the non survivors have been removed. A table that included both non survivors and survivors divided by early and late TH would be a better way to establish whether the two groups were clinically comparable with regard to clinical status before the onset of hypothermia. Third, there is no data given on changes in practice over the time period 2009 to 2016 and the impact that would have on time of transport team arrival or on the use of passive cooling prior to arrival. Fourth, the ASQ has been assessed as a tool to detect severe neurodevelopmental disability at 12 months. The results state that 75 patients completed neurodevelopmental assessment at 18 months and that 12 were reported to have moderate to severe impairment. Use of a different tool would have given more granularity to the developmental outcomes, so strengthened the ability to detect differences between the two groups. Finally, Table 2 includes Neurologic exam on admission - specifically Sarnat stage 1,2,3. I note that there is a higher number of Stage 3 in the early TH group. However, no details are given on the time of admission in relation to the time of birth. If the infants in the early TH group were admitted earlier than the late TH group and encephalopathy severity progresses over time there could be further variation between the groups that is not accounted for in a regression using stage on admission. In summary, I would suggest use of caution in analysis of a small retrospective cohort to address the question of timing of TH with respect to beneficial effect. In the clinical paradigm we do not have the accurate information on timing of insult that is available in preclinical models and so use time from birth as a proxy. Furthermore factors such as the clinical presentation, severity of insult and degree of encephalopathy may have influence on timing of TH. In a retrospective cohort study it may not be simple to account for everything and accordingly it is important to be guarded in interpreting findings.

REVIEWER	Reviewer name: Khorshid Mohammad Institution and Country: University of Calgary , Canada Competing interests: none
REVIEW RETURNED	28-Jan-2019

GENERAL COMMENTS	This is an important study to bring the focus back to resuscitation and stabilization in the first hours of life which may impact the outcome more the initiation of TH that early. I have few comments and suggestions for the authors : 1. Why the primary outcome wasn't death/brain injury and death/neurodevelopmental outcome as death is a competing outcome and my understanding that infants had higher rate of mortality in the early cooling initiation 2. Nothing was mentioned about the method of cooling at the referring centers and on transport and how that played a role in all this 3. it will be nice to have a table comparing infants who died between the early and late group rather than mortality vs no mortality 4. Where these infants all outborns? if not it will be an important confounding factor to include

REVIEWER	Reviewer name: Floris Groenendaal Institution and Country: Assoiate professor. Department of Neonatology. University Medical Center Utrecht. Utrecht. The Netherlands Competing interests: None
REVIEW RETURNED	01-Feb-2019

GENERAL COMMENTS	In their manuscript Guillot et al describe the timing of initiation of therapeutic hypothermia on brain MRI and neurodevelopment at 18 months in infants with HIE. The paper is interesting, and the study addressed a relevant clinical issue. The following questions, however, arise on reading the manuscript. First, and most importantly, why is time to initiate hypothermia dichotomized, since it is a continuous variable? The time point of 3 hours appears to be an arbitrary time point. In addition: was hypothermia used during transport (I think it has been), and if so, how? And which target temperatures were used during transport? Secondly, the question arises why the time to reach target temperature after start of TH has been so long (almost 4 hours in Early TH, and 3.2 hours in Late TH)? By using the Blanketrol target temperature could have been reached well within one hour. In fact, the study population therefore consists of infants cooled <6 hours versus >6 hours as far as target temperature is concerned. Thirdly, it is not sure that both groups (Early vs Late) are comparable. More Sarnat 3 cases were included in the Early TH group, and more infants in this group received mechanical ventilation following intubation. To summarize, the dichotomization in timing is not supported regarding a longitudinal clinical variable, and time to reach target temperature is far beyond 3 hours. It is hardly within the aim of 6 hours in the early group. Furthermore it cannot be excluded that the Early group was clinically more affected by the asphyxia (more intubation than the Late group). If the outcome (MRI and neurodevelopment) would indeed be similar, whereas asphyxia has been more severe, then earlier treatment of those infants would be beneficial compared to later treatment. This would be in contrast with the authors' conclusions. Minor points In table 1 the timing of aEEG has not been mentioned. In the Apgar score at 10 min mv (mechanical ventilation?) has not been reported for the Early TH group, and may differ from the Late TH group. In table 3 several patients are missing in the MR analysis. Could the authors mention cranial ultrasound findings or provide other information on these missing MRI cases?

	In table 5 of the 8 infants who died 5 had severe Basal Ganglia (BG) abnormalities, whereas 5 had severe Watershed (WS) lesions, 1 had total brain injury. Seven infants of the 8 who died 7 had moderate–severe brain injury. Then more than 1 of the infants had both BG and WS injury, i.e. total brain involvement?

REVIEWER	Reviewer name: Emmanouil Bagkeris Institution and Country: University College London Competing interests: No competing interest
REVIEW RETURNED	08-Feb-2019

GENERAL COMMENTS	1. The statistical analysis section refers to the logistic regression as multivariate. Please correct to multivariable. Multivariate analysis is performed when more than one outcomes are involved, which is not the case in this manuscript. Moreover, there are no results of this logistic regression anywhere in the main text or tables apart from its mention at the statistical analysis. Has the analysis been performed? If so, please edit the relevant section accordingly. 2. The abstract of the manuscript does not report death as an outcome of interest. I think that the analysis should be focused on survivors only, considering that neurodevelopmental outcomes cannot be deemed from non-survivors. The study population and results should include only those with data on both the main exposure (TH) and the main outcome(s). Table 1 can be moved to the supplementary materials. 3. The second sentence of the second paragraph of the statistical analysis is a result and should be moved to the appropriate section. 4. Be consistent with the decimal places used to report the p-values of table 1. Please advise and reference T J Cole, “Too many digits: the presentation of numerical data” https://adc.bmj.com/content/100/7/608 5. In table 1, for all continuous factors specify what the parenthesis presents. It is not clear that it is the (IQR) for all continuous factors. Moreover, use parenthesis instead of using the ± symbol when you report standard deviations. 6. Perhaps edit the title of table 3. A more appropriate title could be: MRI findings stratified by timing of therapeutic hypothermia. 7. The table 4 can be moved to the supplementary material of the manuscript considering that there is no reference of the specific neurodevelopmental abnormalities in the main text. 8. The title of the figure 1 poorly describes the context of the figure.

REVIEWER	Reviewer name: Marianne Thoresen Institution and Country: University of Bristol , UK, and University of Oslo, Norway Competing interests: None
REVIEW RETURNED	20-Feb-2019

GENERAL COMMENTS	To the editor. This is an important and well conducted retrospective cohort study on therapeutic hypothermia after perinatal asphyxia.

	The major limitation is the very mild cohort, 33% had normal or grade 1 Sarnat at entry. It is of course not possible to find an effect of early or late cooling on children who are already normal.I hope they are able to present the data I suggest, that they calculate a motor score as well as a separate cognitive score from their ASQ-3 which is a parental questionaire and not a clinical examination. Also thay they compare these outcome data by rank ordering the results before comparison. The numbers with moderate or severe NE at entry are only 50, few are injured hence the power is were low to be able to detect a difference.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: Malcolm Battin, Auckland, NZ

1 - C1: My major concern with the paper, as presented, is that it does not provide the reader with enough detail to judge the robustness of the conclusions based on analysis of a modest retrospective cohort.

First, there is the question of power to detect a difference between the two (<3 h Vs >3h) groups. Obviously, a retrospective cohort is limited to the number of cases during the given time frame. In the manuscript the size of the cohort is listed as a strength, “the largest to report on influence of timing of TH of the study” (p12). However, on the previous page it is stated that “establishing whether early TH has additional benefits may require similarly larger cohorts”. The size of previously reported cohorts is not the question and this paper needs to be very clear whether the sample size was sufficient to adequately detect differences between the two groups and fulfil the study objective or not.

1 - R1: Thank you very much for your comments and suggestions. This report is one of the first ones exploring timing of initiation of TH as a predictor of brain injury and/or neurodevelopmental impairment in neonates with HIE. Not having a good grasp on a possible effect size, it was difficult to estimate power or sample size. While this question is clinically extremely important and could potentially change our practice, larger population cohorts are likely needed to better understand this relationship. This report is a first step in investigating this question.

We adjusted our discussion and conclusions consequently. As suggested by the reviewer, we are now more cautious regarding our conclusions on timing of initiation of TH and outcomes.

The following sentence has been removed from the manuscript:

Our study is the largest reporting on early (≤3h of life) vs. late (>3h of life) initiation of TH and comparing findings on MRI and 18 months outcomes.

We now underline the limitation of our sample size in several sections of the manuscript:

1) Introduction

- Our objective was to examine the effects of early TH (started ≤ 3h of life) on the pattern and severity of brain injury on MRI and neurodevelopmental outcomes in a large regional cohort of infants with HIE.

2) Discussion

- The strengths of our study are the size of the cohort, the largest to report on influence of timing of TH,

- Our data In this retrospective observational cohort, showed that early TH was neither associated with a difference in brain injury on MRI nor better outcomes at 18 months in comparison with TH initiated later.

- Also, the size of the cohort is relatively small and might have been insufficient to detect a difference between the early and late TH groups.

3) Conclusion

- In our this retrospective observational cohort, early TH started before 3h of life was neither associated with less brain lesions on MRI nor better neurodevelopmental outcomes.

- Large population studies are needed in the future to better establish the effect of timing of TH.

1 - C2: Second, in a retrospective cohort where onset of cooling is not randomized it would be important to consider and account for other clinical factors that might affect outcome. For example the arrival time of the transport team and thus initiation of hypothermia could be influenced by condition at presentation and/or referring team concerns. Figure 1 indicates that the early TH group included more babies who died before MRI (7 vs 1) and more who died before discharge (6 Vs 2) suggesting that the groups may be different in clinical status. Table 2 summarizes the clinical variables and shows minor differences in resuscitation only. However, in this table all of the non survivors have been removed. A table that included both non survivors and survivors divided by early and late TH would be a better way to establish whether the two groups were clinically comparable with regard to clinical status before the onset of hypothermia.

1 - R2: We agree with the reviewer's suggestion. Table 1 now describes all the eligible patients included in the cohort. We also highlighted the important differences between the 2 groups in the "Results" section:

In the early TH group, Caesarian section delivery was more common, resuscitation was more extensive, more neonates suffered from severe encephalopathy and more neonates died (Table 1). Also, as described in Table 1, for neonates in early group, TH was more often initiated by the birthing centre before advice and referral to CHEO was done earlier.

We acknowledge that these differences are clinically important and despite our efforts to control for confounding, there are likely residual confounding factors not included in our analysis. We underline this limitation in our discussion:

Second, the early and late TH groups possibly have some clinical and pathophysiological differences, which can influence their outcomes. Although we minimized the known confounding factors by controlling for encephalopathy severity (regression models and sub-group analysis) and using instrumental variable analysis, there are likely residual confounding factors not included in our analysis.

Confounding by indication was a significant challenge in analyzing data from this study since various factors that influence the timing of TH also affect outcomes. This was addressed using adjustment for potential confounding factors such as HIE severity, as well as an alternative instrumental variables analysis.

Importantly, following reviewers' recommendations, the outcomes (brain MRI and neurodevelopmental outcomes at 18 months) are now reported for patients assessed at 18 months of age. The characteristics of these patients are reported in Table 4. Again, the baseline characteristics of the 2 groups (early and late) are different. We report the differences in the result section:

The descriptive data for these patients, divided in early vs. late TH, are presented in Table 4. Again, the 2 groups were different with more extensive resuscitation and a trend towards more severe encephalopathy in the early group.

1 - C3: Third, there is no data given on changes in practice over the time period 2009 to 2016 and the impact that would have on time of transport team arrival or on the use of passive cooling prior to arrival.

1 - R3: Thank you for the comment. Our practice did change over time, as it did in general in centers offering TH, with ongoing quality improvement effort and increasing experience and comfort with the therapy. We began using aEEG in 2011, after 15 patients were enrolled in our cohort. In 2015 our review of TH practice on transport revealed delays in commencing passive cooling awaiting arrival of the Transport team and on occasion central line insertion. Subsequently, we created a community support package for patient management, completed a series of outreach education workshops and the Neonatology group agreed to recommend passive cooling as soon as the patient was referred and transport team dispatched. We also revised our Transport Team TH practice guideline and algorithm with enhanced documentation. These collective measures contributed to a reduction in time to initiate TH and time to attain target temperature (unpublished data). Unfortunately, the time to patient referral has not changed over time.

We underlined the limitation regarding the aEEG data and changes in practice over time in the discussion:

Additionally, some changes in practice were noted over the time period of the study. Particularly, aEEG was not used before 2011, explaining the several missing data for this variable.

1 - C4: Fourth, the ASQ has been assessed as a tool to detect severe neurodevelopmental disability at 12 months. The results state that 75 patients completed neurodevelopmental assessment at 18 months and that 12 were reported to have moderate to severe impairment. Use of a different tool would have given more granularity to the developmental outcomes, so strengthened the ability to detect differences between the two groups.

1 - R4: In our center, the ASQ-3 is used by the neonatal follow up team to guide the interview with parents and observe specific tasks performed by the child at 18 months of age. The ASQ tool is used as a screening tool for developmental delay and guide the decision to refer to developmental resources. Previous studies support that ASQ is a valid neurodevelopmental tool at 18 to 24 months.(Gollenberg, Lynch, Jackson, McGuinness, & Msall, 2010; Mackin et al., 2017; Noeder et al., 2017)

We agree with the reviewer that a different tool, such as Bayley Scales, would have given more granularity to detect specific area of developmental impairment. We recognize this limitation of the ASQ in the discussion:

In this study we used the ASQ-3, as our neurodevelopmental assessment tool while Thoresen's study and large RCTs used Bayley scales. Despite reports of concurrent validation of ASQ and Bayley scales (Gollenberg et al., 2010; Lindsay, Healy, Colditz, & Lingwood, 2008), comparing our results is more complex when using 2 different scales. Moreover, the ASQ-3 provides an overall assessment of development, based on 5 domains – one of which is fine motor skills and one is gross motor skills, without a precise normative value like the PDI or the Motor Composite Score. Also, based on the number of applicable questions, PDI is influenced more by gross motor skills than fine motor skills. Consequently, the ASQ can difficultly be compared to the PDI or Motor Composite score and might is not precise enough to detect an improvement in one specific area of development, such as motor outcomes. can't assess more precisely the impact of TH on mental and psychomotor development alone.

1 - C5: Finally, Table 2 includes Neurologic exam on admission - specifically Sarnat stage 1,2,3. I note that there is a higher number of Stage 3 in the early TH group. However, no details are given on the time of admission in relation to the time of birth. If the infants in the early TH group were admitted earlier than the late TH group and encephalopathy severity progresses over time there could be further variation between the groups that is not accounted for in a regression using stage on admission.

1 - R5: Thank you very much for noting that. In fact, the severity of encephalopathy referred to the clinical assessment, in the first 6h of life, before initiating TH. The following has been modified accordingly:

Postnatal variables included severity of encephalopathy in the first 6h of life, before initiating TH on admission (Sarnat score)

In tables: Neurologic exam on admission Degree of encephalopathy before TH

We controlled for encephalopathy severity in our analysis, both in our regression and alternative instrumental models.

1 - C6: In summary, I would suggest use of caution in analysis of a small retrospective cohort to address the question of timing of TH with respect to beneficial effect. In the clinical paradigm we do not have the accurate information on timing of insult that is available in preclinical models and so use time from birth as a proxy. Furthermore factors such as the clinical presentation, severity of insult and degree of encephalopathy may have influence on timing of TH. In a retrospective cohort study it may not be simple to account for everything and accordingly it is important to be guarded in interpreting findings.

1 - R6: We would like to thank the reviewer for this summary. We report the experience of TH in one outborn centre with a limited sample size. We agree that timing of initiation of TH depends on multiple factors and despite our best efforts, it's probably impossible to account for all of them. We believe that we now clearly state the important limitations of our study and we are more cautious in the interpretation of our analysis.

We added the following sentence to our discussion:

First, unlike what's observed in preclinical models, the timing of injury in neonates with HIE is uncertain and the time of birth might not accurately reflect this timing.

Reviewer 2: Khorshid Mohammad, Calgary, Canada

2 - C1: Why the primary outcome wasn't death/brain injury and death/neurodevelopmental outcome as death is a competing outcome and my understanding that infants had higher rate of mortality in the early cooling initiation

2 - R1: Thank you for your comments.

Based on preclinical studies, our initial hypothesis was that an earlier initiation of TH in neonates with HIE would lead to less severe brain injury and therefore, improved neurodevelopmental outcomes. The decision not to use a composite outcome including death is that we don't believe that mortality is on the same causality pathway linking timing of initiation of TH and brain injury and neurodevelopmental outcomes.

We now compare the characteristics of non survivors in early TH vs. late TH group (Table 2). Both groups had similar characteristics.

We believe that the differences observed between survivors and non survivors (lower Apgar at 10 minutes, more extensive resuscitation, more severe encephalopathy and more abnormal cerebral function monitoring, as shown in Table 3) are closely related to mortality, and not the timing of initiation of TH.

2 - C2: Nothing was mentioned about the method of cooling at the referring centers and on transport and how that played a role in all this

2 - R2: Thank you for this suggestion. We now added details in the descriptive tables about 1) who initiated TH, 2) How the cooling was initiated and 3) timing of referral to CHEO.

As we can now see in our updated Table 1, in the early TH group, TH was initiated by the birthing centre before advice from our NICU more often and referral to CHEO was done earlier. There was no difference between the groups on the method of cooling at referring centre. We underlined this information in the results section:

Also, as described in Table 1, for neonates in early group, TH was more often initiated by the birthing centre before advice and referral to CHEO was done earlier.

2 - C3: It will be nice to have a table comparing infants who died between the early and late group rather than mortality vs. no mortality.

2 - R3: Thank you. As per your suggestion, we added a table (Table 2, supplementary material) comparing the characteristics of non survivors in early vs. late TH group.

2 - C4: Where these infants all outborns? if not it will be an important confounding factor to include

2 - R4: CHEO is a level 3 outborn NICU. All the neonates described in the cohort were outborn. We described the outborn status of our NICU in the methods and discussion section:

Methods: CHEO is a university-affiliated level 3 outborn unit, with 400 admissions per year.

Discussion: Our outborn NICU serves a very large geographical region of almost 440,000 km2.

Reviewer 3: Floris Groenendaal, Utrech, Netherlands

3 - C1: First, and most importantly, why is time to initiate hypothermia dichotomized, since it is a continuous variable? The time point of 3 hours appears to be an arbitrary time point.

3 - R1: Thank you for your comments.

We explored the timing effect both as dichotomized variable and continuous variable. We haven't observed any differences in the outcomes using both strategies. The ultimate decision to present our results with time to initiate TH as early vs. late was to be congruent with the previously published cohort study (Thoresen, 2013).

We underlined our analysis of time as a continuous variable in our manuscript in the following sections:

1) Statistical Analysis

The relationship between timing of initiation of TH and outcomes was assessed using a multivariate logistic regression adjusting for severity of encephalopathy at baseline. This relationship was also analyzed using timing of initiation of TH as a continuous predictor.

2) Results

Analogous analyses using time as a continuous predictor were not significant for moderate to severe impairment.

3) Discussion

In this retrospective observational cohort, timing of initiation of TH, assessed both as a dichotomous and continuous variable, was neither associated with a difference in brain injury on MRI nor better outcomes at 18 months.

3 - C2: Was hypothermia used during transport (I think it has been), and if so, how? And which target temperatures were used during transport?

3 - R2: Yes, our transport team provide passive cooling followed by active cooling (cool packs) following an algorithm, to target a rectal temperature of 33.0 to 34.0 degrees Celsius.

3 - C3: Secondly, the question arises why the time to reach target temperature after start of TH has been so long (almost 4 hours in Early TH, and 3.2 hours in Late TH)? By using the Blanketrol target temperature could have been reached well within one hour.

In fact, the study population therefore consists of infants cooled <6 hours versus >6 hours as far as target temperature is concerned.

3 - R3: Thank you for your observation. In our cohort of outborn HIE infants, TH was initiated outside of the NICU for 85 patients out of the 91 included in our study. While the Blanketrol was used in our NICU, TH was initially passive for the majority of our cohort (n=57, 63%). This information is now available in Table 1.

From our updated Table 1, the median delay between initiation of TH and target core temperature is 2.9h for the early TH group and 3h for the late TH group. Although this delay in reaching target core temperature of 33.0 to 34.0ºC is relatively long, it's comparable to other cohorts of neonates admitted to a tertiary NICU and transported by a neonatal transport team and is also in keeping with our large geographical area. We have listed a few examples of neonatal studies using TH for HIE with delay to reach target above 1h:

1. (Tsuda et al., 2017) : Baby Cooling Registry of Japan, mean 94 min

2. (Thoresen et al., 2013): median delay of 142.5 min in the early TH group and 72.5 min in the late TH group

3. (Lemyre et al., 2017): 2 Canadian cohorts (2009 – 2013) are described, CHEO (3.8h) and Sick Kids (3h)

3 - C4: Thirdly, it is not sure that both groups (Early vs Late) are comparable. More Sarnat 3 cases were included in the Early TH group, and more infants in this group received mechanical ventilation following intubation.

3 - R4: We acknowledge that both groups were different. We clarified these differences in our updated Table 1 and in the results section:

In the early TH group, Caesarian section delivery was more common, resuscitation was more extensive, more neonates suffered from severe encephalopathy and more neonates died (Table 1).

Furthermore, we now better describe:

1. The full cohort of eligible patients (Table 1, n=91 (54 Early, 37 Late)

2. The non survivors, according to TH timing of initiation (Table 2, n=16 (13 Early, 3 Late))

3. The survivors vs. non survivors across the full cohort (Table 3, n=91)

4. Patients with neurodevelopmental assessment at 18 months (Table 4, n=64 (36 Early, 28 Late))

We acknowledge that the differences between the groups are clinically important and despite our efforts to control for confounding factors by controlling for severity of encephalopathy in our regression models and using an instrumental variable approach, there are likely residual confounding factors not included in our analysis. We underline this limitation in our discussion. (please refer to our response R2 to the first reviewer for examples)

3 - C5: To summarize, the dichotomization in timing is not supported regarding a longitudinal clinical variable, and time to reach target temperature is far beyond 3 hours. It is hardly within the aim of 6 hours in the early group. Furthermore it cannot be excluded that the Early group was clinically more affected by the asphyxia (more intubation than the Late group).

If the outcome (MRI and neurodevelopment) would indeed be similar, whereas asphyxia has been more severe, then earlier treatment of those infants would be beneficial compared to later treatment. This would be in contrast with the authors' conclusions.

3 - R5: Thank you for your comments.

Time was explored as both a dichotomous and a continuous exposure and both analysis showed no association with outcomes (brain injury on MRI and/or moderate to severe neurodevelopmental impairment at 18 months).

In regards to the differences between the characteristics of early and late TH groups, we agree that the early group appears sicker. While we tried to correct for the differences between the groups (as described in our previous response), given our relatively small sample size, we preferred to be cautious in our conclusions. We included your suggestion in the discussion:

Given that infants who received TH earlier were sicker at birth and more severely encephalopathic, perhaps this is positive, as one might have expected more brain injury on MRI and/or worse outcomes at 18 months in that group.

Minor comments:

3 - C6: In table 1 the timing of aEEG has not been mentioned.

3 - R6: We clarified the timing of aEEG in the methods section:

Neonates were monitored with cerebral function monitoring (BrainZ Instruments, New Zealand) from their admission to NICU and for the duration of the TH,

3 - C7: In the Apgar score at 10 min mv (mechanical ventilation?) has not been reported for the Early TH group, and may differ from the Late TH group.

3 - R7: We clarified the amount of resuscitation at birth by using a resuscitation score. We described the resuscitation in the data collection section and reported the score in the descriptive tables. Indeed, the resuscitation score is higher in the early TH group vs. late group.

The amount of resuscitation at birth was summarized by a previously described resuscitation score grade from 1 to 6:

1 = no intervention, 2 = blow-by oxygen, 3= endotracheal suctioning, 4 = bag-mask positive pressure ventilation, 5= endotracheal intubation with positive pressure ventilation, and 6 = endotracheal intubation with ventilation and medication.(Miller et al., 2005)

3 - C8: In table 3 several patients are missing in the MR analysis. Could the authors mention cranial ultrasound findings or provide other information on these missing MRI cases?

3 - R8: Among the 8 deceased patients, 6 patients had at least one cranial ultrasound. They all showed signs in keeping with brain edema such as increased echogenicity, loss of gray-white matter differentiation, small ventricles and low cerebral resistance index. One patient had early cystic transformation.

However, as suggested by other reviewers, we now present the outcomes (both brain MRI findings and neurodevelopmental outcomes) exclusively in survivors with neurodevelopmental assessment completed at 18 months (n=64). Table 4 reports the characteristics of the 64 patients, divided according to timing of initiation of TH and Table 5 describes their MRI findings. In this context, we did not add the cranial ultrasound findings of the deceased patients.

3 - C9: In table 5 of the 8 infants who died 5 had severe Basal Ganglia (BG) abnormalities, whereas 5 had severe Watershed (WS) lesions, 1 had total brain injury. Seven infants of the 8 who died 7 had moderate–severe brain injury. Then more than 1 of the infants had both BG and WS injury, i.e. total brain involvement?

3 - R9: Thank you for this observation. The 2 patients identified as having a total brain injury pattern had a Watershed (WS) score of 5 and Basal Ganglia (BG) score of 4.

As mentioned by the reviewer, some patients were in the moderate-severe brain injury category for both patterns (WS and BG) but, one specific pattern was predominant. For example, a patient with BG score of 2 and WS score of 4 would fall in the WS predominant pattern of injury.

Reviewer 4: Emmanouil Bagkeris, University College London

4 - C1: The statistical analysis section refers to the logistic regression as multivariate. Please correct to multivariable. Multivariate analysis is performed when more than one outcomes are involved, which is not the case in this manuscript. Moreover, there are no results of this logistic regression anywhere in the main text or tables apart from its mention at the statistical analysis. Has the analysis been performed? If so, please edit the relevant section accordingly.

4 - R1: Thank you. We corrected the word multivariate for multivariable, as suggested. We also edited the results section:

Logistic regression analyses using TH initiation time as a dichotomous predictor (≤3h vs. >3h), and controlling for severity of encephalopathy, revealed no significant differences between groups for moderate to severe impairment and/or death.

4 - C2: The abstract of the manuscript does not report death as an outcome of interest. I think that the analysis should be focused on survivors only, considering that neurodevelopmental outcomes cannot be deemed from non-survivors. The study population and results should include only those with data on both the main exposure (TH) and the main outcome(s). Table 1 can be moved to the supplementary materials.

4 - R2: We now report outcomes (brain injury on MRI and neurodevelopmental outcomes at 18 months) exclusively in survivors with neurodevelopmental assessment completed at 18 months (n=64). We now describe this cohort of patients in our updated Table 4.

As suggested, the table describing survivors vs. non survivors (now, updated Table 3) will be moved to the supplementary materials.

4 - C3: The second sentence of the second paragraph of the statistical analysis is a result and should be moved to the appropriate section.

4 - R3: This sentence has been moved to the results section, thank you.

4 - C4: Be consistent with the decimal places used to report the p-values of table 1. Please advise and reference T J Cole, “Too many digits: the presentation of numerical data” https://adc.bmj.com/content/100/7/608

4 - R4: We adjusted our numerical data as per the reference mentioned. Thank you.

4 - C5: In table 1, for all continuous factors specify what the parenthesis presents. It is not clear that it is the (IQR) for all continuous factors. Moreover, use parenthesis instead of using the ± symbol when you report standard deviations.

4 - R5: Thank you, we now clarified whether we present mean (SD) or median (IQR) in the tables.

4 - C6: Perhaps edit the title of table 3. A more appropriate title could be: MRI findings stratified by timing of therapeutic hypothermia.

4 - R6: The updated Table 5 now present the MRI findings, the title now reads: MRI findings in patients with neurodevelopmental assessment at 18 months of age, stratified by timing of therapeutic hypothermia

4 - C7: The table 4 can be moved to the supplementary material of the manuscript considering that there is no reference of the specific neurodevelopmental abnormalities in the main text.

4 - R7: Thank you for your suggestion. The tables 2, 3 and 6 have been moved to the supplementary material.

4 - C8: The title of the figure 1 poorly describes the context of the figure.

4- R8: We changed the title for: Flow chart for the study population

Reviewer 5: Marianne Thoresen, Bristol UK and Oslo, Norway

This is an important and well conducted retrospective cohort study on therapeutic hypothermia after perinatal asphyxia from Ottawa, Canada. Additional data is most likely available and should be included as suggested.

5 - C1: Power calculation stating which difference in ASQ-3 score they aim to detect, how many patients do they need in each group to detect such a difference.

5 - R1: We acknowledge that our sample size is limited and establishing if earlier TH is beneficial may require larger cohorts. As mentioned in our first response to reviewer 1 (1 - R1), not having a good grasp on a possible effect size, it was difficult to estimate power or sample size. We have adjusted our discussion and conclusion to underline this limitation.

We believe that this retrospective study conducted in our centre can contribute to better understand the multiple factors involved in timing of initiation of TH and outcomes.

5 - C2: Below in this review I suggest that you calculate a new motor and cognitive score from the raw scores of your results. If the scores are rank ordered, by comparing the ranks in the two groups, (nonparametric) you are more likely to find a difference than with a binary test with a cut off.

I suggest that the authors use the raw scores from their two motor domains and make a new composite score of (gross motorx2 + fine motor x1) which is weighted similar to Bayley II PDI and test whether this ASQ-3motor score is different between the early and late cooled groups. Repeat this analysis also after removing the mild 33% of each group (who are normal anyway). There were only moderate and severe HIE in the 65 survivors the Thoresen study. They would be comparable to the 50 survivors with Sarnat 2 or 3 in your study. You can also try a regression between time of start cooling and the ASQ-3motor score in the group where the mild ones have been excluded.

5 – R2: Thank you for this great suggestion. While it's not possible to calculate a new score for this cohort (our site changed their records to an electronic record in 2017 and access to previous paper charts is limited), we will keep this suggestion for our future cohorts. Still, we now underline in the discussion this important concept that you raised about the different influence of the gross motor skills in the PDI compared to the ASQ.

Moreover, the ASQ-3 provides an overall assessment of development, based on 5 domains – one of which is fine motor skills and one is gross motor skills, without a precise normative value like the PDI or the Motor Composite Score. Also, based on the number of applicable questions, PDI is influenced more by gross motor skills than fine motor skills. Consequently, the ASQ can difficultly be compared to the PDI or Motor Composite score and might not be precise enough to detect an improvement in one specific area of development, such as motor outcomes.

5 - C3: aEEG was not used as entry criteria but background pattern as defined by Lena Westas (pls give the reference) as used. When was the aEEG monitoring started?

5 - R3: The aEEG monitoring was initiated at time of admission to the NICU. The aEEG background was not considered a criteria to initiate TH. However, the aEEG background contributes to the neurological assessment and can sometimes be used as an additional argument to initiate or not TH in ambiguous cases. We clarified the timing of aEEG in the methods section:

Neonates were monitored with cerebral function monitoring (BrainZ Instruments, New Zealand) from their admission to NICU and for the duration of the TH

We also updated our references (Hellström-Westas L, 1995), as suggested by the reviewer.

5 - C4: This paper does not find difference in outcome between the early and late cooling groups. The only other paper asking the same question (Thoresen) found that PDI but not MDI showed a difference so that early cooling had better outcome than late cooling. ASQ-3 has five domains, one domain is gross motor and one domain is fine motor. Using Bayley II, the relative weight from raw scores on gross motor tests versus fine motor is 2:1. PDI is somewhat different from Motor score in Bayley III. The authors state they could only use the full ASQ-3 as the outcome variable and since 3 domains are non-motor it is difficult to find a difference in motor scores after early or late cooling using this test.

5-R4: We want to thank the reviewer for this important observation. As mentioned in our response to comment 2 of the reviewer (5 – R2), we now recognize this limitation of the ASQ assessment in our limitation section.

5 - C5: Did the 33% with normal or Sarnat 1 on admission fulfill the NICHD entry criteria for cooling?

5 – R5: As per our chart review, two patients were classified as having no encephalopathy before initiation of TH. One of these 2 patients developed seizures in the first 6 hours of life. The other patient had extensive resuscitation, requiring intubation and chest compression, severe asphyxia (cord pH <6.8) and became encephalopathic in the first hours of TH. Among the 23 patients with Sarnat 1, 4 presented with seizures in the first 6 hours of life. As far as we can extract from the medical records, these “mild” infants all met biochemical criteria.

It’s one important limitation of retrospective studies to rely on chart review for encephalopathy assessment. There is a risk that the information provided in the chart is incomplete or missing. We also underlined this limitation in the discussion:

The limitations of our study include its retrospective nature, particularly given the topic of encephalopathy which can be challenging to diagnose and can evolve over time.

C6: I note that you group Burst Suppression, Low Voltage and Flat Trace as abnormal aEEG trace. There are 7 out of 41 with early TH and 7 out of 34 with late TH who has these three patterns. That means 34 early HT and 27 late HT had early aEEG patterns of CNV or DNV. This shows that you have a very mild cohort. It would be useful with a table showing Sarnat me, two and three for each group and the corresponding aEEG patterns CNV, DNV, BS, LA and FT. If you prefer fewer groups, they should be CNV, DNV+BS , LA+FT. This would correspond to the classification used in the TOBY and CoolCap trial who classified according to voltage pattern so that groups were normal (CNV), moderate (DNV and BS) or severe (LA or FT) background aEEG.

5 – R6: Thank you. Our tables 1 to 4 now include the description of the abnormal aEEG background (discontinuous, burst suppression, low amplitude or flat trace).

It’s important to note that the practice has evolved during the study period (2009 – 2016). We began using aEEG in 2011, after 15 patients were enrolled in our cohort. Therefore, our aEEG data are available for 72 out of the 91 eligible patients and 35% (n=7) of our patients with Sarnat 3 did not have aEEG monitoring.

C7: References: Please use the original references for a new technique so that ref 30 can be replaced with Lena Westas, Linda de Vries from the mid 1990-ties. Thoresen, Westas, Liu, de Vries, Pediatrics 2010 describe aEEG in infants undergoing TH as well as normothermia.

5 – R7: As suggested, we updated our discussion and included the recommended references. It now reads:

It’s been previously demonstrated that the type of background pattern in the first 6h of life is a strong predictor of neurodevelopmental outcome in normothermic HIE infants.(Hellström-Westas, Rosén, & Svenningsen, 1995) Importantly in infant treated with TH, the time to normalization of background activity is a better predictor of outcomes.(Thoresen, Hellstrom-Westas, Liu, & de Vries, 2010)

30. Hellström-Westas L, Rosén I, Svenningsen NW. Predictive value of early continuous amplitude integrated EEG recordings on outcome after severe birth asphyxia in full term infants. Arch Dis Child Fetal Neonatal Ed [Internet]. 1995 Jan [cited 2019 Feb 25];72(1):F34-8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7743282

31. Thoresen M, Hellstrom-Westas L, Liu X, de Vries LS. Effect of hypothermia on amplitude-integrated electroencephalogram in infants with asphyxia. Pediatrics [Internet]. 2010;126(1):e131-9. Available from: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med7&NEWS=N&AN=20566612