Detection of ‘EEG bursts’ in the early preterm EEG: Visual vs. automated detection

Kirsu Palmu a,b,⁎, Sverre Wikström c,d,e, Eero Hippeläinen a,f, Geraldine Boylan g, Lena Hellström-Westas c,d, Sampsa Vanhataloa

aDepartment of Clinical Neurophysiology, University Hospital of Helsinki, Finland
bDepartment of Biomedical Engineering and Computational Science, Aalto University, Finland
cDepartment of Women’s and Children’s Health, Uppsala University, Sweden
dDepartment of Clinical Sciences, Lund University, Sweden
eDepartment of Clinical Research, County Council of Värmland, Sweden
fDepartment of Physical Sciences, University of Helsinki, Finland
gNeonatal Brain Research Group, Department of Paediatrics & Child Health, University College Cork, Ireland

ARTICLE INFO

Article history:
Accepted 8 February 2010
Available online xxxx

Keywords:
Neonatal intensive care unit
SAT
Brain monitoring
Neonatal EEG
Premature EEG

ABSTRACT

Objective: To describe the characteristics of activity bursts in the early preterm EEG, to assess inter-rater agreement of burst detection by visual inspection, and to determine the performance of an automated burst detector that uses non-linear energy operator (NLEO).

Methods: EEG recordings from extremely preterm (n = 12) and very preterm (n = 6) infants were analysed. Three neurophysiologists independently marked bursts in the EEG, the characteristics of bursts were analyzed and inter-rater agreement determined. Unanimous detections were used as the gold standard in estimating the performance of an automated burst detector. In addition, some details of this automated detector were revised in an attempt to improve performance.

Results: Overall, inter-rater agreement was 86% for extremely preterm infants and 81% for very preterm infants. In visual markings, bursts had variable lengths (≈1–10 s) and increased amplitudes (and power) throughout the frequency spectrum. Accuracy of the original detection algorithm was 87% and 79% and accuracy of the revised algorithm 93% and 87% for extremely preterm and very preterm babies, respectively.

Conclusion: Visual detection of bursts from the early preterm EEG is comparable albeit not identical between raters. The original automated detector underestimates the amount of burst occurrence, but can be readily improved to yield results comparable to visual detection. Further clinical studies are warranted to assess the optimal descriptors of burst detection for monitoring and prognostication.

Significance: Validation of a burst detector offers an evidence-based platform for further development of brain monitors in very preterm babies.

© 2010 International Federation of Clinical Neurophysiology. Published by Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Due to remarkable developments in neonatal intensive care in the past few decades, an increasing number of extremely low birthweight preterm infants (ELBW) are now surviving. Increasing attention is now focused on the high incidence of significant neurocognitive morbidities that still remain in ex-preterm populations (Marlow et al., 2005; Tommiska et al., 2007). Hence, there are increasing demands to develop methods to monitor brain function in this population so that early neuroprotective interventions can be implemented. A few studies have shown that some EEG features, e.g. burst density and averaged interburst intervals, derived from EEG activity during the first days of life in very preterm infants are predictive of later neurological outcome (Wikström et al., 2008; Hellström-Westas et al., 2001; see also Greisen et al., 1987).

The time period of highest vulnerability in extremely preterm infants coincides with a particular developmental phase in EEG activity. During this period, the EEG is mainly discontinuous (‘tracé discontinue’; Lamblin et al., 1999), with periods of relative inactivity interrupted with bursts of higher amplitude activity. As the infant approaches full term, this dichotomous pattern of activity is gradually replaced by more continuous activity, including increasing activity (i.e. ongoing EEG) between the activity bursts (then called ‘tracé alternant’; Lamblin et al., 1999).
In this context, it is highly relevant, that recent studies in basic developmental neurobiology have established that these ‘bursts’, also called SATs (spontaneous activity transients; see Vanhatalo et al., 2005b; Vanhatalo and Kaila, 2006) comprise a crucial type of endogenous brain activity that is needed for brain wiring at the time when sensory systems are not yet fully functional. They may provide an opportunity to evaluate how electrophysiological brain functions develop and are organized during the weeks of rapid growth of connections, i.e. a time period when the most preterm infants are treated in the NICU (Kostović and Jovanov-Milosević, 2006).

In the present paper, we use the terms ‘burst’ and ‘SAT’ interchangeably, and in many places we use the term ‘burst’ only for simplicity. In the early preterm infant, it is important to note that these two concepts refer to the same physiological event (Vanhatalo and Kaila, 2006, 2009), i.e. a short transient of cerebral activity that arises from a state of relative EEG silence. This ‘relatively silent’ period is typically referred to as the ‘interburst’ period (or interburst interval, IBI). The resulting alternating pattern of EEG with successions of bursts/SATs and interburst periods is termed trace discontinuity (Lamblin et al., 1999). This pattern is completely distinct from the pathological EEG pattern of burst-suppression, which is seen in older subjects after major cerebral insults (e.g. post-ischaemic state).

Based on these considerations, detection of ‘bursts’ in the preterm EEG are a feature that could prove very useful for the development of an automated brain monitoring system for neonates. Technically, it should be less challenging than detection algorithms for adult EEG, due to the dichotomous nature of the very preterm EEG. There are, however, two closely related bottlenecks in the clinical implementation of such a ‘preterm burst monitor’: (i) No universal description of bursts exist in the literature such that it would have precluded a valid comparison between visual and automated (i.e. post-ischaemic state).

In the present paper, we use the terms ‘burst’ and ‘SAT’ interchangeably, and in many places we use the term ‘burst’ only for simplicity. In the early preterm infant, it is important to note that these two concepts refer to the same physiological event (Vanhatalo and Kaila, 2006, 2009), i.e. a short transient of cerebral activity that arises from a state of relative EEG silence. This ‘relatively silent’ period is typically referred to as the ‘interburst’ period (or interburst interval, IBI). The resulting alternating pattern of EEG with successions of bursts/SATs and interburst periods is termed trace discontinuity (Lamblin et al., 1999). This pattern is completely distinct from the pathological EEG pattern of burst-suppression, which is seen in older subjects after major cerebral insults (e.g. post-ischaemic state).

Based on these considerations, detection of ‘bursts’ in the preterm EEG are a feature that could prove very useful for the development of an automated brain monitoring system for neonates. Technically, it should be less challenging than detection algorithms for adult EEG, due to the dichotomous nature of the very preterm EEG. There are, however, two closely related bottlenecks in the clinical implementation of such a ‘preterm burst monitor’: (i) No universal description of bursts exist in the literature such that it could be used for validation of a detection algorithm, and (ii) to date, no algorithms have been validated for the detection of bursts in preterm EEG.

In the earlier literature, bursts were often technically defined as activity with a certain duration and an amplitude above a given threshold, whereas ‘clinical detection’ (i.e. visual reading of EEG) of bursts relies on pattern recognition. Strict amplitude criteria have significant limitations mainly related to recording settings (e.g. interelectrode spacing or filter settings). Visual detection by an individual EEG interpreter may compensate for such, but it is bound to be subjective.

Many algorithms have been used to detect EEG bursts (Särkelä et al., 2002; Pfurtscheller et al., 2005; Löfhede et al., 2008), and the algorithm of Särkelä et al. (2002), utilizing non-linear energy operator (NLEO), is already available in commercial devices (NicoOne, Cardinal Healthcare, Nicolet Biomedical, Madison, WI). The NLEO combines information from both the amplitude and frequency content of the signal, which makes it an attractive candidate for detection of preterm bursts (Vanhatalo et al., 2005b).

The aim of this study was to test the inter-rater reliability of burst detection in the early preterm EEG, to investigate the characteristics of these bursts and their occurrence, and to assess the performance of a commercially available NLEO-based burst detector.

2. Methods

2.1. Material

The dataset consisted of single channel EEG recordings from 12 extremely preterm neonates with gestational ages (GA) 23–27 weeks (named hereafter EP1–EP12 or EP-neonates), and six very preterm neonates (GA 28–30 weeks, VP1–VP6, VP-neonates). All recordings were made in the Neonatal Intensive Care Unit at Lund University Hospital, Sweden (Nervus/NicOne 3.3 EEG system with U16 amplifier, Cardinal Healthcare, Nicolet Biomedical, Madison, WI) from P3 to P4 derivation, with a sampling rate of 256 Hz. Fz was used as the recording reference. All recordings were made during the first 3 days of life. Total duration of the original recordings was 6–24 h. For this analysis, we selected 11 min epochs from each recording from places where the traces contained minimal artefacts and consisted of clear epochs of trace discontinuity. Workflow of the study is visualized in Fig. 1.

Written informed consent was obtained from all parents. The Regional Ethics Committees in Lund approved the research protocol.

2.2. Visual detections

2.2.1. Manual marking

Three experienced EEG readers (LHW, SW, SV, hereafter rater A–C) chose common visual criteria (see below) for burst identification by using a “training EEG epoch”, which was not included in the actual analysis dataset. Thereafter, each reader marked all bursts in all 11 min EEG epochs using the event marking feature in NicoOne software. All raters used the same display settings: filtering 0.16–30 Hz, sensitivity 200 μV/cm, timebase 15 mm/s. Artefacts were marked by one of the raters (C) based on visual analysis.

A burst was defined as a distinct occurrence of cerebral activity with a slow (<2 Hz) component and associated faster activity. Burst onset was defined as a clear deviation from the relatively inactive background activity (i.e. interburst period of trace discontinuity (cf Lamblin et al., 1999)). Due to the filter-generated rebounds of the slow components in the preterm EEG (Vanhatalo and Kaila, 2008; Vanhatalo et al., 2005a, b), the end of the burst was defined by the visual appearance of the trace returning to ‘baseline’ (i.e. ‘inactive, interburst mode’) rather than a strict return to zero line. We deliberately avoided using amplitude and/or frequency criteria for two reasons: First, despite their numerical accuracy, they are always arbitrary, and they cannot take into account the inter-individual variations in EEG amplitudes. A visual criterion is readily ‘normalized’ in this respect. Second, use of numerical criteria would have precluded a valid comparison between visual and automated (i.e. by definition numerical) detection.

Inspection of manual markings in EP-neonates revealed that the initial “visual detection threshold” of one of the raters differed compared to the other two raters, and consequently this rater completed a new set of markings. All subsequent analysis was done from this second set of markings.

Epochs not marked as bursts were called interbursts and the time between two consecutive bursts was defined as the interburst interval (IBI).

![Workflow of the study.](https://example.com/workflow.png)
2.2.2. Preprocessing

After visual marking in the NicOne Reader software, the files were converted to CNT-format (with ASA software, Version 4.6.2.0, ANT, Enschede, Holland). All subsequent data analysis was performed using Matlab (Version 7.6.0, MathWorks, Natick, MA, USA).

In order to preclude “edge effects” by filters in the analysed data, 30 s of data at the beginning and end of each EEG epoch were discarded, leaving a 10 min epoch from each recording. All manually marked artefacts were discarded (<6% of EEG in all babies, except for baby EP5 with 16% of data with artefacts).

2.2.3. Inter-rater agreement

Visual detections of the raters were compared at each sampled time point (256 points per second) and the results were given as confusion matrices. A confusion matrix describes the observed (relative) frequencies for each possible combination of markings (e.g. A: burst, B: interburst, C: burst). In addition, the proportion of overall agreement was evaluated in order to obtain a single value describing the inter-rater agreement in each recording. For a binary classification the three raters’ proportion of overall agreement is defined as:

\[
P_0 = \frac{\text{number of agreements}}{\text{number of possible agreements}}
\]

\[
= \frac{1}{3N} \left( \sum_{i=1}^{N} (n_{\text{burst},i}^2 - 3n_{\text{burst},i}) \right) + 1
\]

where \( i \) is the case classification (a single sample), \( N \) number of all cases and \( n_{\text{burst},i} \) the number of times case \( i \) is classified as burst \( n_{\text{burst},i} \in \{0, 1, 2, 3\} \), calculation based on (Fleiss, 1971).

We deliberately avoided the use of kappa statistics; because it is not suited for dependent samples like EEG time series (see also Norman and Scott, 2007). Notably, visual markings in this study are processed as continuous time series, and each time point has a classification as either burst or interburst in the visual detection of each rater.

2.2.4. Derived parameters

Three parameters to describe bursts were derived for each recording from the markings of each rater: number of bursts, average burst duration and proportion of time covered by bursts (burst\%), for an example see Fig. 2. The number of bursts was taken as the number of epochs (i.e. series of consecutive sample points) with a manual rating as burst. Hence the “burst number” is distinct from the number of points in the EEG time series classified as bursts (see also Fig. 2). Correlations were calculated pairwise between the raters for each parameter and group, and considered significant, if the null hypotheses of no correlation could be rejected at \( p = 0.05 \) level.

2.2.5. Burst characteristics

In order to further characterise burst activity in EP and VP neonates, the distributions of burst durations in the manual markings of each rater were also evaluated. Additionally, the frequency spectrum of epochs that were unanimously detected as either burst or interburst were calculated. Spectra were calculated using fast Fourier transformation (FFT) from epochs longer than one second. Each epoch was divided in one second segments from which the FFT (Hamming windowing, zeropadding by factor 4) was calculated separately. Finally, spectra from all epochs in each recording were averaged and a grand average calculated from the individual averages.

2.3. Automated detections

Two different versions of an algorithm for automated detection were implemented. The first version was constructed in Matlab to mimic as closely as possible the algorithm used in the NicOne software for burst-suppression-detection (“original” algorithm). The implementation was done in close cooperation with the vendor. In the second version, modifications were done in order to further improve the accuracy of detection (“revised” algorithm). Exact definitions of original and revised algorithms used in this article are given in Supplementary data.

2.3.1. Original algorithm

The NicOne burst-suppression-detection is modified from an algorithm developed by Särkelä et al. (2002) for adult EEG. The basic idea is to use the output of a non-linear energy operator (NLEO), which was presented in a generalized form by Plotkin and Swamy (1992). Särkelä et al. (2002) use the following definition of the NLEO output:

\[
\text{NLEO}(x(i)) = x(i)x(i-3) - x(i-1)x(i-2)
\]

where \( i \) is the current sample and \( x(i) \) the value of band filtered EEG at that sample.

NLEO values reflect both amplitude and frequency content of the analyzed signal. In the detection algorithm, EEG is filtered into two frequency bands (EEG band and artefact band) and NLEO values are calculated from each frequency band (Fig. 3). Absolute values of NLEO values are then smoothed by calculating the sum in a sliding window. If the smoothed NLEO value in EEG band (“bs”) exceeds a burst threshold for a longer period of time (1–2 s, see below), EEG is classified as burst. If the same value stays below another threshold (“suppression threshold”) for a certain period of time, then EEG is classified as suppression. If neither of these conditions is true or if the NLEO value of the artefact band (“artefact”) exceeds an artefact threshold for a longer period of time (1–2 s), the EEG is classified as artefact/continuous EEG.

The EEG band was defined to be 0.1–8 Hz (high pass: 1st order Butterworth, low pass: 6th order elliptic filter) and artefact band 47–49 Hz (8th order elliptic band pass filter). Burst threshold was set to 300, suppression threshold to 40 and minimum length for bursts to 1 s (if previous state was suppression) or 2 s (if previous state was artefact/continuous EEG). Minimum suppression length was 2 s. Artefact threshold was 10,000. The sliding window used for smoothing the absolute NLEO values was defined to be the past 1 s for each sample point. These values are the same as those used as default in NicOne software.


Fig. 2. Visual detections of bursts in a very preterm (VP2) baby. Red bars indicate burst detections by three independent raters (A–C) from the P3 to P4 derivation filtered at 0.16–30 Hz. Periods of unanimous detections are shown in red (SATs/bursts) and in blue (no detection; interburst) background colour. Grey colour indicates periods where rating was not unanimous within three raters. Note that the variation in detections comes mainly from the differences in defining the onset and end of each detection. The gold standard dataset used in this study is taken from the unanimous detections only (i.e. red and blue background). Parameters of each rater were also evaluated. Additionally, the frequency spectrum of epochs that were unanimously detected as either burst or interburst were calculated. Spectra were calculated using fast Fourier transformation (FFT) from epochs longer than one second. Each epoch was divided in one second segments from which the FFT (Hamming windowing, zeropadding by factor 4) was calculated separately. Finally, spectra from all epochs in each recording were averaged and a grand average calculated from the individual averages.

2.3. Automated detections

Two different versions of an algorithm for automated detection were implemented. The first version was constructed in Matlab to mimic as closely as possible the algorithm used in the NicOne software for burst-suppression-detection (“original” algorithm). The implementation was done in close cooperation with the vendor. In the second version, modifications were done in order to further improve the accuracy of detection (“revised” algorithm). Exact definitions of original and revised algorithms used in this article are given in Supplementary data.

2.3.1. Original algorithm

The NicOne burst-suppression-detection is modified from an algorithm developed by Särkelä et al. (2002) for adult EEG. The basic idea is to use the output of a non-linear energy operator (NLEO), which was presented in a generalized form by Plotkin and Swamy (1992). Särkelä et al. (2002) use the following definition of the NLEO output:

\[
\text{NLEO}(x(i)) = x(i)x(i-3) - x(i-1)x(i-2)
\]

where \( i \) is the current sample and \( x(i) \) the value of band filtered EEG at that sample.

NLEO values reflect both amplitude and frequency content of the analyzed signal. In the detection algorithm, EEG is filtered into two frequency bands (EEG band and artefact band) and NLEO values are calculated from each frequency band (Fig. 3). Absolute values of NLEO values are then smoothed by calculating the sum in a sliding window. If the smoothed NLEO value in EEG band (“bs”) exceeds a burst threshold for a longer period of time (1–2 s, see below), EEG is classified as burst. If the same value stays below another threshold (“suppression threshold”) for a certain period of time, then EEG is classified as suppression. If neither of these conditions is true or if the NLEO value of the artefact band (“artefact”) exceeds an artefact threshold for a longer period of time (1–2 s), the EEG is classified as artefact/continuous EEG.

The EEG band was defined to be 0.1–8 Hz (high pass: 1st order Butterworth, low pass: 6th order elliptic filter) and artefact band 47–49 Hz (8th order elliptic band pass filter). Burst threshold was set to 300, suppression threshold to 40 and minimum length for bursts to 1 s (if previous state was suppression) or 2 s (if previous state was artefact/continuous EEG). Minimum suppression length was 2 s. Artefact threshold was 10,000. The sliding window used for smoothing the absolute NLEO values was defined to be the past 1 s for each sample point. These values are the same as those used as default in NicOne software.
The algorithm defines the EEG class ("burst", "suppression" or "artifact/ongoing EEG") for every sample but only one value per second is kept (so the detection signal is down sampled to 1 Hz). Additionally, the output of the algorithm is changed only after the criterion for minimum duration for bursts is fulfilled. Therefore, the algorithm keeps the previous classification for the first 1–2 s of each burst and in this way underestimates their duration.

2.3.2. Revised algorithm

The revised algorithm is identical to the original algorithm except for two modifications: (1) The classification results are kept for every sample (256 per second in this dataset), i.e. no down sampling is performed, and (2) the beginning of each burst gets the correct classification (see Appendix A).

2.4. Comparison between visual and automated detections

Results of the automated detections were compared with a "gold standard", built on the basis of the visual detections. In the gold standard dataset, only epochs where all three raters agreed on the classification of the EEG (burst/interburst) were considered (see Fig. 1). In comparisons, the classification results of the algorithms were summarized into two classes, bursts and interbursts (see also Fig. 3). Automated detection was shifted 0.5 s backwards to compensate for the delay caused by the smoothing window.

Confusion matrices were calculated at "point-by-point" level for original algorithm vs. gold standard as well as revised algorithm vs. gold standard. Based on these, accuracy, sensitivity and specificity of the algorithms was evaluated.

The difference in the performance of the algorithms was evaluated by Mann–Whitney test and was considered significant if \( p < 0.05 \).

3. Results

3.1. Visual detections

3.1.1. Inter-rater agreement

Analysis of manual detections at "point-by-point" level showed that there was a high level of overall agreement between the raters (86% and 81% for EP- and VP-neonates, respectively). Confusion matrices of visual detections are given in Table 1.

3.1.2. Derived parameters

Analysis of manual detections was also done from the derived, descriptive parameters (burst number, average burst duration and burst%). Results given in Fig. 4 show differences in these parameters between and within groups (EP vs. VP). There were, however, also notable inter-rater differences.

Burst% was the only derived parameter that showed a statistically significant pair wise correlation between all raters and in both groups (EP and VP), and the coefficients of determination \( R^2 \) were highest for burst%. All values of correlation analysis are given in Table 2 (see also Supplementary Figures S1 and S2).

3.1.3. Burst characteristics

Fig. 5 shows the distribution of the burst durations in markings of each rater. Number of bursts were 456, 523 and 705 for raters A–C, respectively in EP1–EP12 babies. For VP1–VP6 babies, the numbers were 308, 336 and 326, respectively. As a general observation, very few bursts were shorter than 1 s or longer than 10 s. As expected, the bursts also became longer with increasing maturational median burst length in EP babies was 4.3, 3.1 and 2.9 s (raters A–C), and 4.4, 4.8 and 3.3 s in VP babies.

Frequency content of the EEG during bursts and interburst periods (Fig. 5) shows that bursts have a remarkably higher amplitude (and power) than interbursts through all frequencies.

3.2. Comparison of visual and automated detections

Confusion matrices for the original and revised algorithm compared to the gold standard are given in Table 3. All analysis was done on a "point-by-point" level.

The accuracy (proportion of correctly classified time points) of both detection algorithms was relatively high: for the original algorithm on average 87% and 79% (extremely preterm and very preterm babies), and for the revised algorithm 93% and 87%, respectively, see Fig. 6.

The sensitivity of the original algorithm was low, on average 64% and 69% (EP and VP babies). This could, however, be readily explained by the misclassification of the beginning of each burst (see 2.3 above as well as Fig. 6). The sensitivity of the revised algorithm was on average 95% and 95% (EP and VP babies).

The specificity of the algorithms varied very much between the babies, this was especially notable for the revised algorithm. The average specificity was 96% and 88% in original algorithm and
92% and 80% in the revised algorithm for EP and VP babies, respectively.

Comparison of the detection performance of the two algorithms showed a significant difference in sensitivity in both EP and VP babies (Mann–Whitney $U = 0$, $n_1 = n_2 = 12$ for EP babies, $U = 2$, $n_1 = n_2 = 6$ for VP babies, $p < 0.01$ two-tailed in both cases). In the group of EP babies, the change in accuracy was significant (Mann–Whitney $U = 0$, $n_1 = n_2 = 12$, $p < 0.01$ two-tailed). From Fig. 6 it is obvious that the burst detection performed poorly in the case of VP6. This was likely due to the higher amplitude in the ongoing EEG between bursts in this baby at the upper end (GA 30 weeks) of the age range of our study population (cf. Vanhatalo and Kaila, 2006). If only babies VP1–VP5 are considered, there was a significant improvement in accuracy between the original and revised algorithm (Mann–Whitney $U = 0$, $n_1 = n_2 = 5$, $p < 0.01$ two-tailed in both) in VP babies too. Changes in specificity were not significant.

4. Discussion

4.1. Towards an ideal detector

Development of an ‘ideal burst detector’ is a challenge that searches for a balance between clinical and biological aims.

4.1.1. Clinical ideal

Clinicians need brain monitoring options that are able to detect or predict pathophysiology with high sensitivity and specificity. There are always trade-offs between sensitivity and specificity in clinical tools and the optimum level depends on the intended use of the information (e.g. guidance of treatment options). Our data show that already the present NLEO-based ‘burst/SAT detector’ performs relatively well in the youngest age groups, but future work is still needed to make the detector ‘evidence based’ by systematically optimizing its in-built parameters as well as the derived indexes. Moreover, clinicians may wish to have detectors that have both high diagnostic and prognostic value. Current literature suggests that quantification of bursts in the youngest preterm population may provide information about both of these (Hellström-Westas et al., 1991, 2001; Wikström et al., 2008). It has been shown, however, that early brain activity is modified by many environmental conditions, including pathologies as well as respiratory or drug treatments (Hellström-Westas, 2004; Victor et al., 2005; Hellström-Westas and de Vries, 2009). Hence a detector may be more readily used as a diagnostic or monitoring tool than as a prognostic test device.

4.1.2. Biological ideal

Current demands of evidence-based medicine have set the ideal that monitoring paradigms should be based on known biological mechanisms. In the context of early preterm EEG, this ideal may have a relatively clear target: There is overwhelming neurobiological evidence now available that emphasizes the critical developmental importance of the intermittent, spontaneous activity transients (SAT) of cerebral activity in preterm infants (cf. Vanhatalo and Kaila, 2006, 2009). These developmental events are likely to be the most relevant index of the brain’s wellbeing in the youngest preterm infants. They may have varying spatial configurations, leading to an unpredictable way of spatial summation and integration in recordings with only a few electrodes (one or two channels) as used in most current neonatal units. This means that there are no perfect, i.e. fully unambiguous approaches to quantifying burst/SAT occurrence. Our current empirical data, together with the known data from basic science, suggests that the proportional occurrence (here called “burst%”) may give a less ambiguous and

![Fig. 4](image-url) Results from the visual detections from each baby and every rater. Comparison of the numbers of bursts, the average durations of bursts, as well as the burst% in EP- and VP-babies shows that the differences between raters varies across babies. For results of correlation analysis of these values, see also Table 2 and Supplementary Figs. S1 and S2.

Table 2

Results of correlation analysis between pairs of raters. The values given are the coefficients of determination ($R^2$), stars indicate the significance of the correlation when null hypothesis is no correlation ( for $p < 0.05$, ** for $p < 0.01$). Detailed figures available in supplementary Figs. S1 and S2.

<table>
<thead>
<tr>
<th></th>
<th>Number of bursts</th>
<th>Burst average duration</th>
<th>Burst%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EP babies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A vs. B</td>
<td>0.24</td>
<td>0.33*</td>
<td>0.67**</td>
</tr>
<tr>
<td>A vs. C</td>
<td>0.42**</td>
<td>0.51</td>
<td>0.77**</td>
</tr>
<tr>
<td>B vs. C</td>
<td>0.59</td>
<td>0.70*</td>
<td>0.77</td>
</tr>
<tr>
<td><strong>VP babies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A vs. B</td>
<td>0.56</td>
<td>0.70</td>
<td>0.91**</td>
</tr>
<tr>
<td>A vs. C</td>
<td>0.70</td>
<td>0.06</td>
<td>0.85**</td>
</tr>
<tr>
<td>B vs. C</td>
<td>0.10</td>
<td>0.05</td>
<td>0.98</td>
</tr>
</tbody>
</table>

more reproducible measure of early preterm brain activity (see also below). Prior literature has shown that the neonatal EEG matures to a less dichotomous form (cf Vanhatalo and Kaila, 2006; Dreyfus-Brisac, 1962; Lamblin et al. 1999) and ‘bursts’ become less prominent. Hence the distinction between ‘bursts’ and ‘interbursts’ is progressively less detectable by NLEO-based algorithms, and our results may be applicable to earliest preterm ages only. Prior studies from our group suggest that EEG findings in this particular age range may have a particular relevance to longer term neurological outcome (Wikström et al., 2008; Hellström-Westas et al., 2001).

4.2. Visual detections

Our present results show that bursts (SATs; see Table 1 in Vanhatalo and Kaila, 2009) in the preterm EEG can be identified reliably using a visual classification scheme and that inter-rater reliability is high. We performed a visual rating of the EEG on a point-by-point level, implying that no a priori limits were imposed on temporal or amplitude dimensions of bursts. As expected, most inter-rater variability came from the varying perceptions of the exact onset and end of bursts. We felt that a reliable ‘definition of burst’ should be based on a common perception (here, unanimous marking) among several raters.

4.2.1. Descriptive parameters

Detection of EEG features, such as bursts, provides the basis for the subsequent calculation of parameters that are routinely used in clinical assessment of the preterm neonatal EEG. In the present study, we calculated burst number, average burst duration, as well as burst%.

The present data suggest that burst number and average duration of bursts are both sensitive to variations in detection thresholds. This is largely due to the fact that epochs with multiple closely succeeding bursts can easily be defined as one longer or as a series of many shorter bursts (see also Fig. 2). Burst% is less sensitive to this confounder. Indeed, inter-rater correlations showed that burst% is the only parameter with strong, significant correlation between all raters in both groups of babies. This implies that burst% reflects the occurrence of bursts in the EEG in a more consistent way.

4.2.2. Burst characteristics

The dataset allowed us to assess burst characteristics in more detail, which is needed for later optimization of automated detectors. Our findings demonstrate clearly that bursts have variable

Table 3

<table>
<thead>
<tr>
<th></th>
<th>Gold standard:</th>
<th>Gold standard:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>burst</td>
<td>interburst</td>
</tr>
<tr>
<td>EP babies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>original/revised:</td>
<td>18.6 / 27.6</td>
<td>2.7 / 5.4</td>
</tr>
<tr>
<td>burst</td>
<td>10.3 / 1.4</td>
<td>68.3 / 65.6</td>
</tr>
<tr>
<td>original/revised:</td>
<td>14.1 / 2.4</td>
<td>47.5 / 43.5</td>
</tr>
<tr>
<td>interburst</td>
<td>31.8 / 43.4</td>
<td>6.7 / 10.7</td>
</tr>
<tr>
<td>VP babies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>original/revised:</td>
<td>31.8 / 43.4</td>
<td>6.7 / 10.7</td>
</tr>
<tr>
<td>burst</td>
<td>14.1 / 2.4</td>
<td>47.5 / 43.5</td>
</tr>
<tr>
<td>original/revised:</td>
<td>10.3 / 1.4</td>
<td>68.3 / 65.6</td>
</tr>
<tr>
<td>interburst</td>
<td>14.1 / 2.4</td>
<td>47.5 / 43.5</td>
</tr>
</tbody>
</table>

Fig. 5. Distributions of burst/SAT lengths (left column) and amplitude spectra (right column) in visual detections by the raters. The burst duration values are shown as proportions (%) of bursts with given length. The proportion of bursts shorter than 1 s was below 3% in all distributions. Averaged amplitude spectra show that bursts/SATs have much higher amplitudes than interbursts throughout the frequency range of interest, and that there is a prominent “hump” in the spectra of bursts/SATs at around 4–8 Hz.
lengths with a non-Gaussian distribution between 1 and 10 s. Amplitude spectra of bursts and interburst periods show that (i) the two EEG forms have distinctly different amplitudes throughout the frequency spectrum, and that (ii) there is a ‘peak’ in the middle frequency range (4–8 Hz) in bursts only. Both observations may have clear implications for further development of such automated detection paradigms that are based on EEG frequency content.

4.3. Constraints of the algorithms

4.3.1. Minimum burst length

Our results have shown that a large number of bursts in neonates are short (1–2 s). In the original algorithm the first 1–2 s of each burst is, by definition, dismissed (i.e. not classified as burst), leading to an underestimation of bursts. The revised algorithm still requires the bursts to have a given minimum length (as also implied by the analysis of manual ratings; see Fig. 5), but the whole burst is subsequently included in the detection.

4.3.2. Smoothing of NLEO transformation

Smoothing of NLEO transformation is necessary to compensate for its rapid fluctuations, but it poses two compromises on detection: First, the automated detection has a temporal delay in comparison with the actual EEG (see Fig. 6). The delay does not bias the descriptive parameters (i.e. no harm to potential clinical use), but it skews the comparisons between manual and automated detections. In our study, this effect was compensated for by time shifting the automated detections (see also Fig. 6).

Second, in both the original and the revised algorithm, smoothed NLEO values are calculated from a 1 s window without taking into account of the sampling frequency. Notably, sampling frequency affects the output of NLEO transformation in a non-linear way and therefore our observations apply to the currently used sampling frequency (256 Hz) only. Use of any other sampling frequency would require its compensation within the algorithm.

4.3.3. Definition of classes

In both algorithms the classification is performed between three classes: “burst”, “suppression” and “ongoing EEG/artefact”. This is a plausible division for the detection of burst-suppression patterns during adult anaesthesia which is what this algorithm was originally designed for (see Särkelä et al., 2002). However, it may not be that well suited for the characterisation of very preterm EEG which has negligible ‘ongoing EEG activity’ (Vanhatalo and Kaila, 2006; Khazipov and Luhmann, 2006), and mainly displays a dichotomous pattern of burst and interburst periods (Lamblin et al., 1999). Hence, we propose that automated classification of the early preterm EEG should include classes “burst”, “interburst” and “artefact”. Exclusion of epochs with artefacts would make it much easier and straightforward to build clinically suitable indexes and trends (‘descriptive parameters’) from the detection. Moreover, seizure activity might interfere with burst detection, because both patterns have multifrequency, complex waveform characters. In future, neonatal brain monitors, coincident and independent automated detection of both seizures and bursts will be required.

It is possible that our results apply only to the detection of physiological bursts in the early preterm EEGs. This is because of the developmental increase in the ongoing oscillations (Vanhatalo and Kaila 2006), as well as an absolute and relative decrease in the EEG activity nested within the SAT/bursts (cf. Vanhatalo and Kaila 2006) will likely rapidly compromise the distinction of EEG classes in older preterm babies.
4.3.4. Definition of frequency bands for calculation of NLEO values
The algorithms used in our study filter the EEG signal into two frequency bands: 0.1–8 Hz (burst band) and 47–49 Hz (artefact band). Based on frequency spectra, the choice of the burst band seems to be reasonable, although not empirically defined (i.e. not optimized yet). The artefact band at 47–49 Hz, instead, is clearly suboptimal for excluding artefacts in the preterm EEGs, because their EMG is more prominent at much lower frequencies (cf. Norman et al., 2008). This band might help in the detection of prominent 50/60 Hz artefacts, which, however, are unlikely to obscure the detection of bursts that are characterized by much lower frequencies. Automated artefact detection is challenging, and attempts to address this are ongoing in our laboratory. It is fortunate, however, that the EEGs from the centroparietal area of the earliest preterm babies have usually relatively few phasic (i.e. short duration) movement artefacts that would interfere with burst detections.
Despite these constraints, the revised version of the algorithm developed in this study is already capable of detecting bursts in early preterm EEG with high accuracy.

Acknowledgements
This work was supported by Grants from University Hospital of Helsinki (KP and SV), Finnish Academy (SV), Sigrid Juselius Foundation (KP and SV), County Council of Värmland (SV), Lund and Uppsala Universities (LHW), and Medical Research Council (Grant 0037, LHW). The authors would like to thank Mr. Lars Johan Ahnlide and Mr. Heiðar Einarrsson for their help in the implementation of the original algorithm.

Appendix A. Supplementary data

References