Line Length as a Robust Method to Detect High-Activity Events

Automated Burst Detection in Premature EEG Recordings

Ninah Koolen[1,2], Katrien Jansen[3], Jan Vervisch[3], Vladimir Matic[1,2],
Maarten De Vos[4,1], Gunnar Naulaers[5] and Sabine Van Huffel[1,2]

[1]*Department of Electrical Engineering (ESAT), division SCD, KU Leuven, Leuven, Belgium*
[2]*iMinds-KU Leuven Future Health Department, Leuven, Belgium*
[3]*Department of Pediatrics, University Hospital Gasthuisberg, Leuven, Belgium*
[4]*Cluster of Excellence "Hearing4all" & Methods in neurocognitive Psychology, University of Oldenburg, Oldenburg, Germany*
[5]*Neonatal Intensive Care Unit, University Hospital Gasthuisberg, Leuven, Belgium*

Keywords

Preterm; Brain monitoring; Background EEG; Brain maturation; Automated detection; Bursts; Interburst intervals; Line length

Highlights

- We present a novel method for automated burst detection in premature EEG recordings, using a single feature *Line Length* based on both the amplitude and frequency content of the signal.
- The accuracy of the burst detection method is within the range of 79.54%-89.82%, resulting from both high sensitivity and high specificity and a comparable inter-rater agreement.
- The implemented algorithm combines the knowledge of multiple neurologists, thereby objectively assessing premature brain function and reducing the costs for the labor intensive visual analysis.

Abstract

Objective: EEG is a valuable tool for evaluation of brain maturation in preterm babies. Preterm EEG constitutes of high voltage burst activities and more suppressed episodes, called interburst intervals (IBIs). Evolution of background characteristics provides information on brain maturation and helps in prediction of neurological outcome. The aim is to develop a method for automated burst detection.
Methods: Thirteen polysomnography recordings were used, collected at preterm postmenstrual age of 31.4 (26.1-34.4) weeks. We developed a burst detection algorithm based on the feature line length and compared it with manual scorings of clinical experts and other published methods.
Results: The line length–based algorithm is robust (84.27% accuracy, 84.00% sensitivity, 85.70% specificity). It is not critically dependent on the number of measurement channels, because two channels still provide 82% accuracy. Furthermore, it approximates well clinically relevant features, such as median IBI duration 5.45 (4.00-7.11) seconds, maximum IBI duration 14.02 (8.73-18.80) seconds and burst percentage 48.89 (35.45-60.12) %, with a median deviation of respectively 0.65s, 1.96s and 6.55%.
Conclusion: Automated assessment of long-term preterm EEG is possible and its use will optimize EEG interpretation in the NICU.
Significance: This study takes a first step towards fully automatic analysis of the preterm brain.

Corresponding author:
Ninah Koolen
From the Department of Electrical Engineering (ESAT) – STADIUS , Centre for Dynamical Systems, Signal Processing and Data Analysis – BIOMED
University of Leuven (KU Leuven)
Kasteelpark Arenberg 10 - bus 2446, 3001 Leuven, Belgium
+3216329621
ninah.koolen@esat.kuleuven.be

1. Introduction: towards improving preterm prognosis

Each year, more than 15 million babies are born preterm worldwide. Over 1 million babies die each year due to complications of early birth (The Global Action Report on Preterm Birth, 2012). Two principal brain lesions that result in neurological problems for premature infants are periventricular hemorrhagic infarction and periventricular leukomalacia (Volpe et al., 1998). As a consequence, premature newborns risk to develop disabilities, including learning impairments, motor and cognitive problems (Institute of Medicine, 2007). These disabilities excert pressure and raise cost on society and on families. Therefore, it is important that preterm born babies are diagnosed, prognosed and treated as quickly and as accurately as possible.

In large hospitals, the diagnosis and special care after birth is provided by the Neonatal Intensive Care Unit (NICU). Recording of long term EEG can help in evaluating brain function in a non-invasive way. EEG recordings are interpreted by experts through visual analysis. However, interpretation of the EEG has several important limitations. For instance, the analysis of relevant features varies across reviewers and quantification is very restricted due to visual analysis. Moreover, it is very labor intensive and thus expensive. In small hospitals, the expertise is often not available. The analysis of the evolution of the EEG over time requires long term monitoring and thus significant effort for clinicians. Hence, the automatic analysis of the preterm EEG would be a welcome addition in the tool set of hospitals.

Currently, algorithms developed for automatic EEG analysis in newborns are limited. Neonatal seizures are an important clinical phenomenon (Legido et al., 1991), and as many seizures have only very subtle or no clinical manifestations, there is a need for monitoring of seizures with long term EEG. Solutions have been proposed for seizure detection (Deburchgraeve et al., 2009; Cherian et al., 2011; Temko et al., 2011), but have not yet been extended to preterm babies. However, we believe much more information can be found within the background EEG on brain functionality and maturation (Menache et al., 2002; André et al., 2010; Cherian et al., 2010; Niemarkt et al., 2010). Quantification of parameters within the background EEG will make it possible to objectively assess brain function and detect abnormalities. For preterm babies, the EEG of the first week provides a good basis for prediction, even though more information about the evolution can be obtained from sequential EEG measurements. In this way, brain maturation can be evaluated and patients at risk for abnormal neurological development can be identified.

Long term prognosis is based on background EEG analysis. That means, clinically relevant information is visually extracted from the EEG background pattern, to discover acute and chronic abnormalities. The background parameters change in function of postmenstrual age and parallel brain maturation. Abnormal background characteristics can predict adverse neurological outcome (Le Bihannic et al., 2011) Critical factors, describing this pattern, are degree of (dis-)continuities, amplitude and frequency content (Vanhatalo and Kaila, 2006; Löfhede et al., 2010). The best studied parameter in preterm EEG is continuity. Discontinuous activity is the characteristic EEG pattern below 30 weeks of gestational age. A discontinuous EEG pattern, or the so-called tracé discontinue, consists of high amplitude and frequency periods (bursts) interrupted by periods of low brain activity with low-voltage EEG, defined as the interburst intervals (IBIs). A good neurological outcome can be expected, if the duration of the interburst intervals decrease gradually and the EEG evolves to a continuous pattern. At the same time, the burst length and the activity within the burst increases with age (Vecchierini et al., 2007). Furthermore, the time continuous tracings are observed over the EEG measurement, increases with age, as Niemarkt demonstrated using the interburst/burst ratio in preterm babies with normal neurological outcome (Niemarkt et al., 2010). However, amplitude depression and increased discontinuity give rise to an increased risk of brain dysfunctions (Hellström-Westas and Rosén, 2005; Le Bihannic et al., 2011).

Many people have been correlating the EEG pattern to a prognosis or outcome (Menache et al., 2002; Vanhatalo and Kaila, 2006; Niemarkt et al., 2010; Le Bihannic et al., 2011; Hayashi-Kurahashi et al.,

2012), but so far the research in automatically detecting these patterns has been restricted. For this purpose, we want to extract the bursts and IBIs from the preterm EEG in an automated and novel way. Commercial algorithms and algorithms published earlier in literature only make use of the amplitude content of the bursts to detect them (Hayakawa et al., 2001; Niemarkt et al., 2010). However, accuracy can be enhanced by taking frequency content into account. In this way, less false positives will be introduced by moving artefacts, while medication and filter settings will have less influence on the detection system. Some recent studies already make use of both amplitude and frequency content to detect the bursts in the background EEG, such as the non-linear energy operator (NLEO) (Särkelä et al., 2002; Palmu et al., 2010a) and threshold detection on the envelope of the EEG (Jennekens et al., 2011). However, the clinical aim is to maximize both detection sensitivity and specificity above 80% (Mani et al., 2012), since respectively both burst and IBI derived parameters are of prognostic importance.

In this paper, we present a novel method by using a single feature, *Line Length*, to accurately detect bursts and to find suppressed parts (IBIs) in long EEG recordings. In addition, this feature has already been proven to detect the onset of high frequency activity in the EEG, such as seizures (Esteller, 2001; Logesparan et al., 2012; Paz et al., 2013). Furthermore, we applied a variable detection threshold, which automatically adapts to changing amplitude levels, e.g. when medication is administrated. We estimated prognostic maturational features and their manifestation. We also investigated whether there is an influence of the number of input EEG channels on burst detection. As such, we decreased the number of channels from 9 to 2 and 4 channels. Using less channels has the advantage that it is applicable in a broader field, easier to apply and to interpret and hence helpful in secondary and tertiary care, but obviously less information will become available. Our goal was to get accuracy, sensitivity and specificity above 80% or as close as possible to inter-rater agreement. High values are not easily obtained for the inter-rater agreement, because there is no golden standard for the description of bursts in literature. However, it is believed to be the best possible estimate combining the experiences and capacities of all assessors/neurologists involved.

2. Methods: automatically reading EEG recordings

2.1 Data acquisition

EEG data were recorded between June 2011 and November 2012 at the Neonatal Intensive Care Unit of the University Hospitals of Leuven, Belgium. In this large hospital, 2400 babies are delivered each year with approximately 350 born prematurely. The study recruited only preterm infants, of which a polysomnographic dataset is available including long-term video-EEG recordings. No exclusion criteria were used. This resulted in a dataset of 10 subjects born with a postmenstrual age (PMA) of 24-32 weeks. Two subjects had a suspicion of convulsions, two suffered from mild peripartal asphyxia and three suffered from intracranial hemorrhage. Measurements were started on the preterm infants from a few hours to a few weeks after birth (26-34 weeks PMA) and were typically 3-12 hours long. If necessary, a second and a third measurement were taken, respectively at day 14 and on the day the patient was allowed to leave the hospital. For two out of these ten patients, an evaluation of these three EEG recordings was carried out to see an evolution of the brain. OSG (Rumst, Belgium) equipment was used, with full head EEG recordings at 9 electrode locations (Fp1, Fp2, T3, T4, C3, C4, Cz, O1, O2) and sampling frequency of 250 Hz. The reference electrode was Cz, set up was in unipolar mode. Afterwards, a pre-processing step was performed consisting of a 50 and 100 Hz Notch filter followed by a 1-20 Hz band pass filter. The latter one was applied because most information about neurological processes is within this frequency band. Two experienced clinicians have scored twenty minutes of each EEG measurement into bursts and IBIs for further research. The study had full approval from the Clinical Ethics Committee of the University Hospitals of Leuven and written informed parental consent was obtained for all infants studied.

2.2 Burst Detection

a) Detection based on Line Length

Fractal dimension (FD) is a method for transient detection, requiring no prior knowledge of the transient characteristics (Accardo et al., 1996). Since a line and a plane have respectively the

dimension of 1 and 2, the FD of an EEG signal will always lie within the interval [1, 2]. The more the line fluctuates, the more the plane is 'covered', thus increasing the FD. Line length, a simplified version of the FD, is reported as successful at detecting EEG transients, such as seizures (Esteller et al., 2001). Line length is defined as the running sum of the absolute differences between all consecutive samples within a predefined window. The value of this feature will grow as the data sequence magnitude or signal variance increases. Hence, line length can be seen as an amplitude and frequency demodulator (Esteller et al., 2001). We propose an efficient burst detection algorithm based on this feature, consisting of the following steps:

1. The firing of neurons is random, resulting in non-stationary EEG signals, i.e. statistical properties of the frequency and amplitude content change. Moreover, we can expect these differences in the content of epochs in the EEG, such as bursts and IBIs (Selton et al., 2000; Palmu et al., 2010a, 2010b). Therefore, the signal is divided into smaller 'stationary' parts of fixed length with similar characteristics, to extract useful neural information (Wong and Abdulla, 2006; Azami et al., 2011). Here, we consecutively segmented each EEG channel for the duration of 1 second, with an overlap of 0.12 second (Accardo et al., 1996). Short duration segments are essential for reliable detection of transient events, such as bursts. For every segment, we want to decide whether it constitutes of bursts or suppressed EEG (IBIs).
2. Line length is calculated for each segment i as the sum of all absolute distances between the consecutive 250 sample points x_j. The total length of each segment is represented by $L(i)$ and calculated as in formula 1.

$$L(i) = \sum_{j=1}^{250-1} \left| x_{j+1} - x_j \right| \tag{1}$$

After computing line length $L(i)$ for each segment i in each EEG channel n, normalization is achieved through dividing the line length by the total sum of all $L(i)$ values of the same channel (Accardo, 1996). This is defined in formula 2.

$$L_n(i) = \frac{L(i)}{\sum_i L(i)} \tag{2}$$

Clinical experts define bursts as high activity on more than one or more than half of the EEG channels. Consequently, burst segments will have high line length values for several EEG channels in which a burst appears. In order to detect the burst segments, we determine the median value for each segment i over all channels n:

$$Me_L_n(i) = median(L_n(i)) \tag{3}$$

We take the median since this variable is more robust than the mean in the presence of outliers. For example, a moving electrode can cause an artefact on one channel resulting in high amplitude content, which would have a dominant unwanted influence on the mean. The median line length values are represented by the curve in Figure 1 c for an example of 65 seconds EEG recording.

3. In the last step, burst detection is performed. Bursts are detected when the amplitude of the curve Me_L_n is above a certain threshold, which we denote by *Thr_Det*. In order to find an optimal threshold, we divided the dataset into two sets: training and validation. First set includes 5 patients, of which 2 patients have had multiple EEG recordings, resulting in 8 recordings. The validation set consists of 5 patients. For each patient, two human EEG readers have scored 20 minutes data into bursts/IBIs. The threshold *Thr_Det* is adapted in the following way:

$$Thr_Det = F * mean(Me_L_n) \tag{4}$$

Thereby, factor F runs from 0.5 to 1.1, in steps of 0.05. Thus, *Thr_Det* has 13 different values during the training phase. As a result, the ROC curve in Figure 2 is obtained by comparing burst detected samples with samples clinically classified as bursts, as further explained in section 2.3a. Because we want to maximize both the sensitivity and the specificity of the algorithm, we deduct from this figure that F=0.8 or F=0.85 results in good detection rates for the comparison with both clinicians. Namely, for F=0.8, it gives an averaged sensitivity of 85.49% and a specificity of

83.03%. Respectively, it provides for F=0.85 a 82.85% sensitivity and a 85.75% specificity. Considering both are suitable, we decided to further work with a threshold *Thr_Det=0.85*mean(Me_L_n).* This threshold is applied in the validation set.

An additional condition for detection is that the difference in amplitude between a successive non-detected and detected point and vice versa should be large enough. That is to say, larger than *0.4*std(Me_L_n)*, so only pronounced peaks are detected. Furthermore, all IBIs shorter than 2 seconds are removed since they are not considered by clinical experts. An example of 65 seconds EEG partitioned into bursts and IBIs is shown in Figure 1. Each circle or diamond in Figure 1 c represent a segment EEG of 150seconds. Segments detected as bursts are indicated by diamonds. In Figure 1 b, we compare the detected bursts with clinical labeling of high activity periods by two experienced clinicians.

b) Suppression curve

To make an objective validation, the original aim was to segment and detect the bursts in the last 10 minutes of every hour in a 4 hour EEG recording. However, neurologists are not looking into these same parts of the recording when they are interested in features to make a prognosis about the neurological outcome. Instead, they are searching for the most discontinuous parts within the EEG, because this is best related to the outcome of the patient (Menache et al., 2002). Consequently, we developed another part of the algorithm which indicates the 'suppression of the EEG', defined as the suppression curve.

This curve is deducted from the median curve, determined every 150 seconds, as defined in the previous section. We calculate the median and the mean of the curve *Me_L_n*, respectively shown by a grey and dashed line in Figure 3. Then, the difference between the mean and the median is taken as follows:

$$supp = 1 - \frac{median(Me_L_n)}{mean(Me_L_n)} \tag{5}$$

This difference, called *supp*, is depicted in the suppression curve in Figure 4 as a single point. Thus, every point of this curve represents the grade of suppression for an EEG segment of 150 seconds. Therefore, EEG parts with long IBIs resulted in high *supp* values, whereas continuous EEG with short or even no IBIs resulted in lower *supp* values. This is also shown in Figure 3, where the difference between mean and median is larger for a discontinuous pattern (a), compared to a more continuous pattern (b). An example of the suppression curve of one patient is shown in Figure 4, in which we can see that the EEG recording is more or less continuous (very low *supp* values), except two suppressed periods around the 6000th and the 16000th second of the measurement. In general, neurologists are scrolling through a measurement to find these suppressed periods to make a diagnosis. A suppression curve is created for every premature patient and visually verified by an expert human EEG reader. Isolated extreme peaks in the suppression curve are often due to artefacts, so discarded for the further analysis. For each patient, we have selected two periods of 10 minutes with consecutive high *supp* values, to be sure that we incorporate the most informative and practically artifact-free periods in the burst detection examination.

c) Comparison to other detection algorithms

Whereas most developed methods for burst detection only make use of the amplitude content of the EEG background pattern, a few methods exploit both the amplitude and the frequency content. One of these methods applies threshold detection on the envelope of the EEG signals (Jennekens et al., 2011). This envelope *EV(i)* is derived from the average signal power *P(i)*:

$$EV(i) = \sqrt{2P(i)} = \sqrt{\frac{2}{Nw} \sum_{i=1}^{Nw} x(i)^2} \tag{7}$$

Thereby, *x(i)* represents the amplitude of the EEG signal and *Nw* the number of samples within 1 second of data (250 samples). If two or more of the nine EEG channels have an envelope value *EV(i)*

higher than the predefined amplitude threshold, then sample *x(i)* will be classified as burst. Otherwise it is an IBI sample. Similarly to the line length-based method, IBIs shorter than two seconds are removed and changed into burst samples.
Another non-linear method makes use of the non-linear energy operator (NLEO) (Palmu et al., 2010a), where *x(i)* is the value at sample *i* (Palmu et al., 2010a):

$$\mathrm{NLEO}\left(\mathrm{x}\left(\mathrm{i}\right)\right) = \left|\mathrm{x}\left(\mathrm{i}\right)\mathrm{x}\left(\mathrm{i}-3\right) - \mathrm{x}\left(\mathrm{i}-1\right)\mathrm{x}\left(\mathrm{i}-2\right)\right| \quad (8)$$

This signal is smoothed by the average value of a 1.5 second sliding window centred around the time sample NLEO(x(i)). Next, a baseline correction is achieved by subtracting the minimum signal value within the second before the current sample to remove continuous artefacts. Similar to the envelope-based method, a sample is classified as a burst if the processed signal is higher than a predefined amplitude on two or more channels.
Those two methods are compared to the developed line length-based method by statistical analysis, analyzing the discrepancies between automated and clinical detection of bursts. However, we should keep in mind that the line length algorithm has an adaptive threshold whereas the other two methods have a fixed threshold. Furthermore, it is not only important to have high sensitivity and specificity, it is even more crucial to get similar values for prognostic maturational features. In this paper, we present the median length of IBIs, the maximum length of IBIs (longest period of EEG inactivity) and the percentage of bursts (proportion of time covered by bursts). Although more features have been explored (Koolen et al., 2013), these three features are highly informative. For example, with a good development of the premature brain, the median and maximum IBI length will decrease when the PMA is increasing (André et al., 2010; Niemarkt et al., 2010). This means, there will be more EEG activity in the form of bursts due to more connectivity between neurons in the premature brain. In conclusion, we want these prognostic values to be well approximated to assess the evolution of the premature EEG pattern.

d) Comparison of a variable number of EEG channels

Small hospitals do not have the possibility and the equipment to measure a full head EEG. Therefore, we reduced the number of input channels of the line-length detection algorithm from 9 to only 2 and 4 EEG channels. We compared the detection accuracy to the full head clinical scoring of bursts and IBIs based on 9 channels. EEG channels recorded at C3 and C4 are taken as 2 input channels and C3,C4, O1 and O2 are used for 4 channels detection. In case that the accuracy would be similar, many hospitals could advantage considering that most hospitals in Europe only measure two channel EEG (C3 and C4), resulting in the so-called amplitude-integrated EEG (aEEG) (Hellström-Westas et al., 2006, Tao and Mathur, 2010).

2.3 Performance evaluation

a) Defining accuracy, sensitivity, specificity

Bursts are detected by the algorithm and are visually marked by two experienced clinicians (KJ and JV). Since there is no 'golden standard' available for burst detection, visual scoring of more than one rater is taken into account. Moreover, raters often did not agree on the start and ending sample of a burst, resulting into a different number of bursts/IBIs. In addition, bursts should sometimes be merged together to get the same burst as the other rater marked. Therefore, statistical analysis of the detection accuracy is performed sample-by-sample (Palmu et al., 2010a, 2010b). Confusion matrices are obtained by comparing the burst detection of the algorithm with the markings of the clinician. This matrix is composed for each clinician separately and consists of:

- Number of true positives (TP): samples detected as burst by the expert AND by the algorithm
- Number of false positives (FP): samples detected as burst by the algorithm, but not by the expert
- Number of true negatives (TN): samples classified as IBI by the algorithm and by the expert
- Number of false negatives (FN): samples marked as a burst by the expert, but missed by the algorithm

Based on these values, we can determine the robustness of the developed algorithm by calculating the sensitivity, the specificity and the accuracy:

$$sensitivity = \frac{TP}{TP + FN}$$
$$specificity = \frac{TN}{TN + FP} \quad (6)$$
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Some false positives are introduced by movement artefacts. We did not filter out these movement artefacts, because we have selected EEG segments practically free of artefacts based on the suppression curve. Movement artefacts are detected by thresholding the line length signal obtained from the tremor or respiratory signal. A false positive sample, caused by movements, is defined by the following three conditions: the data sample is detected as movement on the tremor or respiratory signal *and* it is detected by the algorithm as a burst *and* both clinicians did not label the sample as a burst. The amount of false positives due to movements over the total amount of false positives is calculated.

b) Relationship among statistical parameters

A high accuracy will be achieved through a high number of correctly classified samples: true positives (bursts) and true negatives (IBIs). However, it can be the case that a high number of only TP or TN can still result in a high accuracy. In other words, we can obtain high accuracy for the combination of low sensitivity and high specificity or vice versa. Nevertheless, the clinical aim is to maximize both the sensitivity and specificity, above 80% (Mani et al., 2012). Maximizing both values can be achieved for the applied thresholds, although increasing or decreasing the threshold will have an influence on both sensitivity and specificity. Therefore, it is of key importance to set patient specific and adapting thresholds. Maximization of both statistical parameters is crucial for clinical EEG examination to make an accurate prognosis about the brain's state of activity. That is, false negatives or false positives will affect prognostic values, such as the median IBI length, the maximum IBI length or the burst percentage.

c) Statistical analysis

Numerical values calculated for the statistical measurements – accuracy, sensitivity and specificity – and for the prognostic values – maximum IBI length, median IBI length and burst percentage – are compared between different detection methods to see whether there is one performing significantly better. Thereto, we first applied Anova test, to see whether there ia a difference between the three detection algorithms. Next, we used Student's t-test. Each time two detection methods are compared. The null hypothesis states that the statistical measurement is similar for both detection methods. If the null hypothesis is rejected (h=1), then the detection methods are performing significantly differently on this measurement with a p-value of $p<0.05$, $p<0.01$ or $p<0.001$.

3. Results: detection characteristics applying real-world data

3.1 Comparing burst detection methods

In a first step of the validation part, the suppression curves of all patients were manually verified by an experienced neurologist (KJ). It is confirmed that the peaks of all suppression curves represent the most discontinuous parts of the EEG measurements. Next, for both training and validation set, automatically detected bursts are compared with the manual scoring of the bursts by two clinical experts (KJ and JV). In Table 1, we present the sensitivity, specificity and accuracy for both sets of the three detection algorithms: based on line length, envelope calculation and NLEO. These methods show respectively a mean accuracy of 84.27%, 80.58% and 82.04%, whereas the inter-rater agreement is 86.20%. There is no remarkable difference for the values of these statistical performance measurements between the training and validation set. The mean percentages of false positives introduced by movement artefacts over the total amount of false positives are respectively for the three

methods 4.84%, 8.36% and 10.91%. Nevertheless, movement artefacts are rare in the selected and analyzed EEG segements.

All three algorithms can be used to reliably and accurately detect bursts. The boxplots, depicted in Figure 5, show the range of sensitivity, specificity and accuracy, including values of all 13 measurements (training and validation set). The accuracy of the line length method is within the range of 79.54%-89.82%, resulting from both high sensitivity and high specificity. For this data set, the accuracy values are good and comparable to clinician's inter-rater agreement. In addition, both sensitivity and specificity are above the aim of 80% predefined by clinical experts (Mani et al., 2012). Furthermore, we can see that the envelope based algorithm gives lower accuracy values within the range of 76.05-83.99% with two extremes of 70.3% and 91.53%. Overall, this method has lower accuracy due to false positives introduced by movement artefacts. The reason therefore is that only an amplitude threshold is used upon the envelope of the EEG signal (around 30 μV). Therefore, bursts are detected by the algorithm, but not by the expert. In addition, a training phase is needed to tune the different parameter values for the algorithm settings, which requires a broader dataset. The lower extreme value is caused by a low sensitivity of 61.84%, resulting in low accuracy. The other extreme has resulted in a very high accuracy for all three detection methods (89.63%-91.97%), because the transitions between bursts and IBIs were clearly pronounced (Figure 6 a), as the EEG was in the state of the so-called tracé discontinue (André et al., 2010).

The NLEO method has a high accuracy (above 80%), although this is due to the averaging of high sensitivity of 93.90% and low specificity of 70.38%. This lower specificity is not desirable, as the aim is to maximize both the sensitivity and specificity of the detection algorithm (>80%). For one in thirteen cases , the NLEO method clearly has a lower accuracy of 68.6%, due to a low specificity of 51.26% and high sensitivity of 96.85%. This patient's EEG contains more activity periods of higher voltage, which are not marked by clinicians as bursts (Figure 6 b). However, these periods are detected by the algorithm, because this high activity is smoothed out. This smoothing step results into more detected burst samples due to a fixed amplitude threshold in the algorithm.

3.2 Prognostic features

It is not only important to have the highest detection accuracy, in addition prognostic features should be well approximated after automated burst detection. Especially, the evolution of these maturational EEG features is of great importance for diagnosis and prognosis. A good prognosis can be expected, in case that the median IBI length decreases (Le Bihannic et al., 2011). An example is presented in Figure 7, where the median IBI length decreased over 14 days. That is, the grey line shifted to the left comparing the histograms in Figure 7 a and b. Moreover, the whole histogram shifted to the left, thus including shorter detected IBIs. Hence, more EEG discontinuities are interrupted by bursts, resulting in shorter IBIs. This is also shown by the increase of the burst percentage in these EEGs: from 42.55% to 57.82% two weeks later (after automated detection of the bursts). These results are in accordance with a normal evolution of the brain maturation.

In Figure 8, boxplots are provided which represent the error between automated and manual detection for three prognostic values: median IBI length, maximum IBI length and the percentage of bursts in the EEG. Mean values over all 13 measurements are respectively 5.45 (4.00-7.11) seconds, 14.02 (8.73-18.80) seconds and 48.89 (35.45-60.12) %. In Figure 8, the errors on clinical values are shown, instead of the real values, since we try to approximate the clinical features as good as possible. In this way, we want the error to be minimized. For median IBI, the error for automated detection is small for all three detection methods; the median error ranges between 0.65-1.75 seconds. The error on the maximum IBI length is larger (1.96 - 3.25 seconds), because this variable varies more within the EEG data. Namely, over all measurements, clinicians found a maximum IBI length between 3.93 and 66.61 seconds. For all 3 methods, the burst percentage deviates respectively 6.55 %, 5.50 % and 10.22% from the clinical values. For the NLEO detection method, it deviates the most (2.63%-31.49%). The statistical significance is described in the next section.

3.3 Statistical significance

We tested whether one of the detection methods was performing significantly better for burst detection. Therefore, we first applied ANOVA tests, because we have values obtained by three groups (detection methods). The ANOVA tests have revealed the variability between the means of the three groups, for values of all statistical measurements and all prognostic features ($p<0.05$). Then, we compared the detection methods in pairs using two-group t-tests. We applied t-tests with an input of two vectors, both sized 26x1. For example, one vector includes the accuracy values of method 1 compared to clinician 1 (13 values for 13 EEG measurements), and the accuracy values of method 1 compared to clinician 2 (also 13 values). The second vector is analogously constructed, but contains the accuracy of method 2. In a similar way, we have built all other input vectors containing the statistical measurements and prognostic features. The results can be found in Table 2 and Table 3 for t-tests on respectively statistical measurements and most prognostic features. For the accuracy, the null hypothesis is rejected by comparing the line length method with the envelope method. Based on numerical values, we can conclude that the former one is better. In addition, this method's specificity is better too, since the null hypothesis is rejected twice for both t-tests where the line length was compared with the other two methods ($p<0.01$). However, the sensitivity of the NLEO based method was the best (null hypothesis rejected twice for the analysis with the two other methods). Nevertheless, the line length method performs good as well for the statistical analysis of the prognostic values (Table 3). Namely, the error of these IBI lengths is small for the line length method, as well as for the envelope method. There is even a significant difference with the NLEO method. In fact, the null hypothesis is rejected twice by comparison of the line length method with the NLEO method ($p<0.05$).

3.4 Detection on a variable number of channels

The investigation of the developed line-length detection algorithm with less input EEG channels showed only a minor difference in accuracy. As expected, the accuracy will increase when information of more channels is included (Figure 9). In fact, the accuracy only increases a few percent (3-5%), respectively for comparison with scoring of clinician 1 and clinician 2. More concretely, when using the the two most informative EEG channels (C3 and C4), the accuracy is still around 82%.

A higher accuracy can be achieved for two channels when the neurologists manually put an amplitude threshold (14-18 μV). Notwithstanding, this higher accuracy can never be achieved in an automatic way by setting a fixed amplitude threshold, since the threshold is patient dependent and should be adapted visually to the patient's EEG pattern.

4. Discussion

EEG in preterms shows dynamic changes parallel to brain maturation (André et al., 2010). The aim of this study was to automatically detect the bursts in the background EEG, which will be used to calculate these maturational features. In this way, normal brain maturation can be assessed and abnormal EEG patterns can be detected identifying patients at risk for abnormal neurological outcome (Hellström-Westas et al., 2005). Moreover, influences of different treatment options can be studied and an accurate prognosis and early counseling can be achieved. In the future this can help in identifying preterm babies that could benefit from neuroprotective measures.

Mimicking the way clinicians observe phenomena in the EEG signals, we have introduced the suppression curve, so that the most discontinuous parts can be examined and used as the input of the burst detection algorithm. The overall results show that the feature line length in the detection algorithm catches the high-frequency phenomena, the bursts, in an accurate way. The accuracy of this line length detection method is in general good and comparable to clinician's inter-rater agreement. In addition, both sensitivity and specificity are above the aim of 80% predefined by clinical experts (Mani et al., 2012). Moreover, the detection threshold is adapted every 150 seconds for each patient, because the EEG could be attenuated when medication is provided and the pattern is corresponding to the age. Only using the two most informative channels, as in aEEG, the detection accuracy resembles the high accuracy already obtained for 'full head' EEG. These findings result in usability for the secondary care as well, where only two channels are measured instead of a 'full head' EEG. Therefore,

not only hospitals doing research, but every hospital can benefit from the use of this method for automatic detection and analysis of the background EEG.

Assessment of the evolution over time of the unique characteristics of the EEG is very valuable and beneficial for better understanding of the preterm brain development. Consequently, good approximation of clinical features is of high importance. Discontinuity is a normal pattern in the very preterm infant. It is an alternating pattern of bursts of high amplitude slow waves and sharp transients, with intervals of low amplitude activity (interburst interval). Increased discontinuity increases the risk of brain dysfunction and abnormal neurological outcome. Dysmaturity of the EEG pattern, with prolonged IBI as the most important feature, is the predominant EEG pattern found in infants who later develop moderate or severe neurological sequelae (Le Bihannic et al., 2011). Conde et al. showed that in preterm babies with major ultrasound brain lesions, the interburst interval was prolonged and the burst duration was shorter when compared to preterm babies without major brain lesions at 2 and 4 weeks postnatal life (Conde et al., 2005). Predictive EEG characteristics, describing the severity of the discontinuity, are the length of the consecutive interburst intervals, but also the amplitude and duration of bursts (Biagone et al., 1994). IBI and burst lengths, as well as burst interburst ratio, are dependent on postconceptional age (Niemarkt et al., 2010). Burst percentage was considered as the statistically most significant parameter for correlations between all raters (Palmu et al., 2010b). The line length just give small deviations from clinical values for this parameter. Not only the burst percentage is informative, even more the IBI lengths are indicative for normal maturation of the brains. By using the developed technique, the median and maximum IBI length are well approximated too. Using these parameters, normal maturation can be assessed and abnormal background EEG patterns can be easily identified. The line length and NLEO methods have a significantly higher accuracy than the envelope method, for which many false positives were introduced. Whereas the NLEO method reached the highest sensitivity in detecting bursts, our goal was to maximize both sensitivity and specificity. Therefore, the line length method was performing significantly better at the approach of the three selected prognostic features (median IBI length, maximum IBI length and burst percentage). For these prognostic features, the errors between algorithm and clinical values were minimized. Nevertheless, the threshold of the other two detection methods can be made adaptive as well.

Some limitations should be included. The validation is only carried out on a small patient database (13 measurements), further research would need to include more patients. In addition, no artefact removal is applied on these signals to remove artefacts introduced by ECG, respiration or movements (De Vos et al., 2011). In fact, automated artefact removal cannot be easily applied, since the ECG and respiration signal should be clear and continuous signals, with a high SNR, which is often not the case for these preterm measurements. Furthermore, a first step towards finding the most discontinuous parts in the EEG has been made, by introducing the suppression curve. Nevertheless, automation of selecting the largest peaks in the suppression curve could be incorporated, since they are yet manually chosen. Future work will focus on fine-tuning the algorithm, based on a larger dataset of validated EEG segments. In addition, more features and their clinical relevance require exploration.

5. Conclusion

The EEG provides vital information about the state of the brains of premature babies and their evolution prognosis. However, it is generally difficult and time-consuming work to analyse the recordings and requires the trained eye of one or more specialized neonatologists. This paper shows that it is possible to read the EEG and accurately analyse the pattern by a computer. The implemented algorithm combines the knowledge and expertise of several neurologists and automatically adapts to specific conditions of the individual patient. The accuracy of the analysis is generally equivalent to that of clinical experts in the NICU. It shows that access to all patient data coupled with an automated analysis is a very useful tool to assess the evolution of the EEG pattern over longer periods of time and to plan immediate treatment. Furthermore, the detection method works accurately with only 2 or 4 measurement channels allowing it to be used in small hospitals, significantly enhancing their capabilities.

Acknowledgements

Research supported by
Research Council KUL: GOA MaNet, PFV/10/002 (OPTEC), IDO 08/013 Autism, several PhD/postdoc & fellow grants;
Flemish Government:
FWO: PhD/postdoc grants, projects: G.0427.10N (Integrated EEG-fMRI), G.0108.11 (Compressed Sensing) G.0869.12N (Tumor imaging) G.0A5513N (Deep brain stimulation);
IWT: TBM070713-Accelero, TBM080658-MRI (EEG-fMRI), TBM110697-NeoGuard, PhD Grants;
iMinds 2013;
Flanders Care: Demonstratieproject Tele-Rehab III (2012-2014);
Belgian Federal Science Policy Office: IUAP P719/ (DYSCO, `Dynamical systems, control and optimization', 2012-2017);
ESA AO-PGPF-01, PRODEX (CardioControl) C4000103224;
EU: RECAP 209G within INTERREG IVB NWE programme, EU HIP Trial FP7-HEALTH/ 2007-2013 (n° 260777), EU MC ITN Transact 2012 # 316679

References

Accardo A, Affinito M, Carrozzi M, Bouquet F. Use of the fractal dimension for the analysis of electroencephalographic time series. Biol Cybern 1996;77:339-350.

André M, Lamblin MD, d'Allest AM, Curzi-Dascalova L, Moussalli-Salefranque F, Nguyen S, et al. Electroencephalography in premature and full-term infants. Developmental features and glossary. Clin Neurophysiol 2010;40: 59-124.

Biagioni E, Bartalena L, Boldrini A, Cioni G, Giancola S, Ipata AE. Background EEG activity in preterm infants: correlation of outcome with selected maturational features. Electroencephalogr Clin Neurophysiol 1994;91:154–162.

Cherian JP, Swarte RM, Visser GH. Technical standards for recording and interpretation of neonatal electroencephalogram in clinical practice. Ann Indian Acad Neurol 2009;12:58-70.

Cherian JP, Deburchgraeve W, Swarte RM, De Vos M, Govaert P, Van Huffel S, et al. Validation of a new automated neonatal seizure detection system: A clinician's perspective. Clin Neurophysiol 2011;122:1490–99.

Conde JR, de Hoyos AL, Martinez ED, Campo CG, Perez AM, Borges AA. Extrauterine life duration and ontogenic EEG parameters in preterm newborns with and without major ultrasound brain lesions. Clin Neurophysiol 2005;116:2796–809.

Deburchgraeve W, Cherian JP, De Vos M, Swarte RM, Blok JH, Visser GH, et al. Neonatal seizure localization using PARAFAC decomposition. Clin Neurophysiol 2009;120:1787 – 96.

De Vos M, Deburchgraeve W, Cherian PJ, Matic V, Swarte RM, Govaert P, et al. Automated artifact removal as preprocessing refines neonatal seizure detection. Clin Neurophysiol 2011;122:2345-2354.

Esteller R, Echauz J, Tcheng T, Litt B, Pless B. Line length: An efficient feature for seizure onset detection. Proc of the 23rd IEEE EMBS Ann Intern Conf 2001, Istanbul, Turkey;1707-10.

Hayakawa M, Okumura A, Hayakawa F, Watanabe K, Ohshiro M, Kato Y, et al. Background electroencephalographic (EEG) activities of very preterm infants born at less than 27 weeks gestation: a study on the degree of continuity. Arch Dis Child – Fetal and Neonatal Edition 2001;84:163–167.

Hayashi-Kurahashi N, Kidokoro H, Kubota T, Maruyama K, Kato Y, Kato T, et al. EEG for predicting early neurodevelopment in preterm infants: an observational cohort study. Pediatrics 2012;130:891-7.

Hellström-Westas L and Rosén I. Electroencephalography and brain injury in preterm infants. Early Hum Dev 2005;81:255-61.

Hellström-Westas L, Rosén I, de Vries LS, Greisen G. Amplitude-integrated EEG classification and interpretation in preterm and term infants. Neoreviews 2006;7:76-87.

Institute of Medicine. Preterm Birth: Causes, Consequences,and Prevention. National Academy Press 2007. Washington, D.C., USA.

Jennekens W, Ruijs LS, Lommen CML, Niemarkt HJ, Pasman JW, van Kranen-Mastenbroek VH, et al. Automatic burst detection for the EEG of the preterm infant. Physiol Meas 2011;32:1623–37.

Koolen N, Jansen K, Vervisch J, Matic V, De Vos M, Naulaers G, et al. Automatic burst detection based on line length in the premature EEG. Proc of the Intern Conf on bio-inspired systems and signal processing 2013. Barcelona, Spain;105-11.

Le Bihannic A, Beauvais K, Busnel A, de Barace C, Furby A. Prognostic value of EEG in very premature newborns. Arch Dis Child - Fetal and Neonatal Edition 2011;97:106-9.

Legido A, Clancy RR, Berman PH. Neurologic outcome after electroencephalographically proven neonatal seizures. Pediatrics 1991;88:583-96.

Löfhede J, Thordstein M, Löfgren N, Flisberg A, Rosa-Zurera M, Kjellmar I, et al. Automatic classification of background EEG activity in healthy and sick neonates. J Neural Eng 2010;7:016007.

Logesparan L, Casson AJ, Rodriguez-Villegas E. Optimal features for online seizure detection. Med Biol Eng Comput 2012;50:659-69.

Mani R, Arif H, Hirsch LJ, Gerard EE, LaRoche SM. Interrater reliability of ICU EEG research terminology. Clin Neurophysiol 2012;29:203-12.

Menache CC, Bourgeois BFD, Volpe JJ. Prognostic value of neonatal discontinuous EEG. Pediatr Neurol 2002;27:93-101.

Niemarkt HJ, Andriessen P, Peters CHL, Pasman JW, Zimmermann LJ, Bambang Oetomo S. Quantitative analysis of maturational changes in EEG background activity in very preterm infants with a normal neurodevelopment at 1 year of age. Early Hum Dev 2010;86:219-24.

Palmu K, Stevenson N, Wikström S, Hellström-Westas L, Vanhatalo S, Palva JM. Optimization of an NLEO-based algorithm for automated detection of spontaneous activity transients in early preterm EEG. Physiol Meas 2010a;31:85-93.

Palmu K, Wikström S, Hippeläinen E, Boylan G, Hellström-Westas L, Vanhatalo S. Detection of 'EEG bursts' in the early preterm EEG: visual vs. automated detection. Clin Neurophysiol 2010b;121:1015-22.

Paz JT, Davidson TJ, Frechette ES, Delord B, Parada I, Peng K, et al. Closed-loop optogenetic control of thalamus as a tool for interrupting seizures after cortical injury. Nat Neurosci 2013;16:64-70.
Särkelä M, Mustola M, Seppänen T, Koskinen M, Lepola P, Suominen K, et al. Automatic analysis and monitoring of burst suppression in anesthesia. J Clin Monit Comput 2002;17:125-34.
Tao JD, Mathur AM, Using amplitude-integrated EEG in neonatal intensive care. J Perinatol 2010;30:73-81.
Temko A, Thomas E, Marnane W, Lightbody G, Boylan G. EEG-based neonatal seizure detection with Support Vector Machines. Clin Neurophysiol 2011;122:464-73.
Vecchierini MF, Andre M, d'Allest AM: Normal EEG of premature infants born between 24 and 30 weeks gestational age: terminology, definitions and maturation aspects. Neurophysiol Clin 2007;37:311–323.
Vanhatalo S, Kaila K, Development of neonatal EEG activity: from phenomenology to physiology. Semin Fetal Neonatal Med 2006;11:471-78.
Volpe JJ. Neurologic Outcome of Prematurity. Arch Neurol 1998;55:297-300.
Wikström S, Pupp IH, Rosén I, Norman E, Fellman V, Ley D, et al. Early single-channel aEEG/EEG predicts outcome in very preterm infants. Acta Paediatr 2012;101:719-26.
Wong L and Abdulla W. Time-frequency evaluation of segmentation methods for neonatal EEG signals. Proc of the 28th IEEE EMBS Ann Intern Conf 2006, New York City, USA;1303-06.

Legends Figures

Figure 1: a) Example of 65 seconds 9-channel EEG recording with reference electrode Cz, b) Burst detection: by 2 clinicians and by algorithm, c) Black curve: median $L_n(i)$ as calculated in step 2, grey line: threshold for detection of bursts (*Thr_Det*), diamonds: detected burst segments after step 3.

Figure 2: ROC curve of the burst detection algorithm, *F* is adapted to get different *Thr_Det*.

Figure 3: Comparison of the line length curve, for a discontinuous pattern (a) and a more continuous pattern (b). Median of the curve is shown by a grey line; mean is shown by a dashed line.

Figure 4: Suppression curve of one patient. Discontinuities are revealed around the 6000th second and the 16000th second of the 5 hour long measurement.

Figure 5: Box plots of sensitivity, specificity and accuracy for the three detection algorithms, all 13 measurements are included.

Figure 6: a) clean EEG with clear burst transitions, which results into high accuracy values for all detection methods; b) Unclear transitions between bursts and IBIs (lot of activity), leading to a low accuracy for the NLEO method.

Figure 7: The evolution of the premature brain leads to a change in the IBI duration histogram: histogram in Figure 7 b are derived from a measurement taken two weeks later than the measurement resulting in the histograms in Figure 7 a. The histogram has shifted to the left; including shorter IBIs. The median IBI length (grey line) decreased too.

Figure 8: Error in automatic detection by comparing to clinicians' scoring for three prognostic values (median IBI, maximum IBI and burst percentage). The box plots represent the differences for all patients calculated for three detection methods (1: line length based method, 2: envelope based method, 3: NLEO based method).

Figure 9: Detection accuracy by using a different number of channels as input for the line length detection method: 2 channels (C3-C4), 4 channels (C3-C4-O1-O2), 9 channels (C3-C4-O1-O2-T3-T4-Fp1-Fp2-Cz).

Figure 1

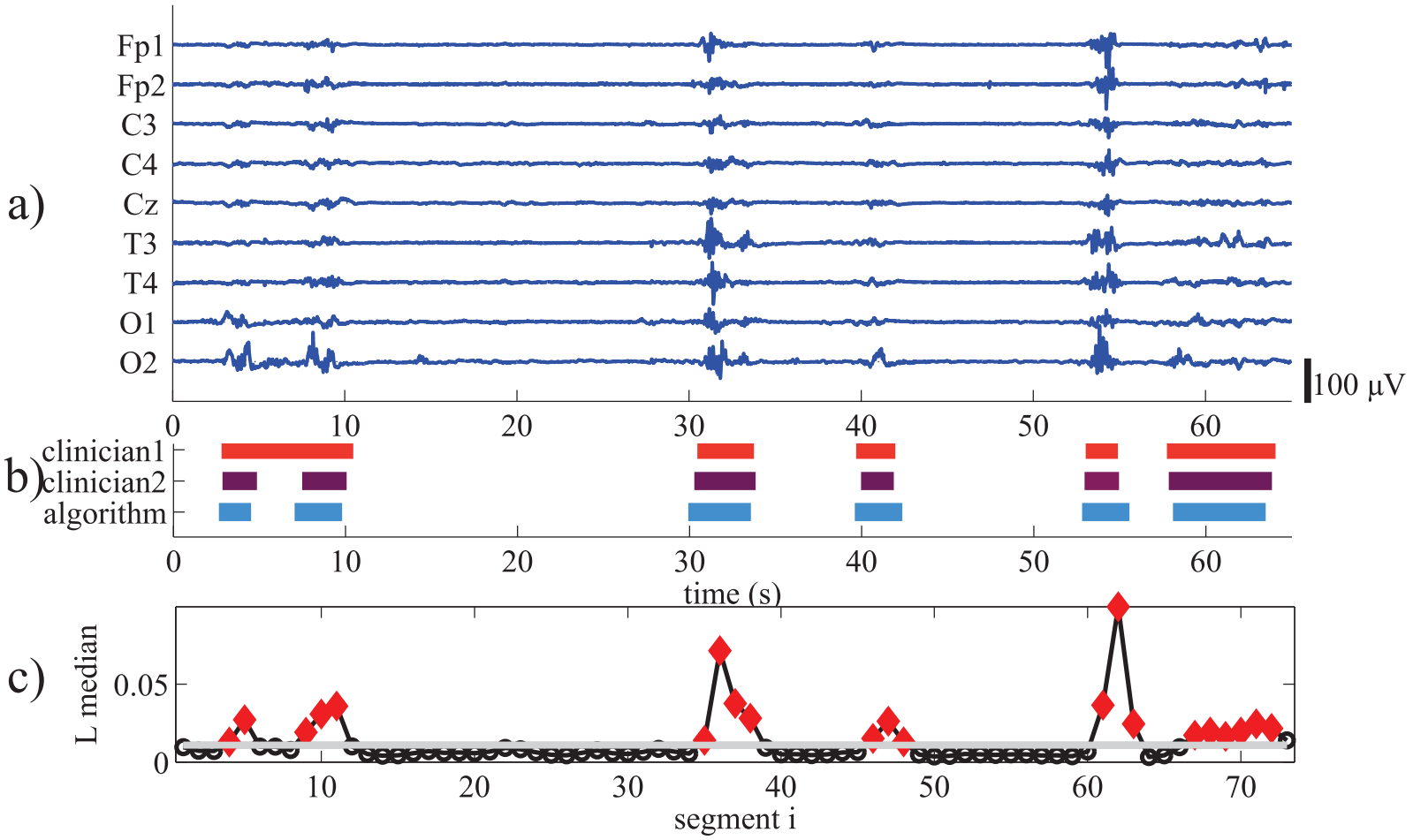
a)
Fp1
Fp2
C3
C4
Cz
T3
T4
O1
O2
100 µV
0
10
20
30
40
50
60
b)
clinician1
clinician2
algorithm
time (s)
c)
L median
0.05
0
70
segment i

Figure 2

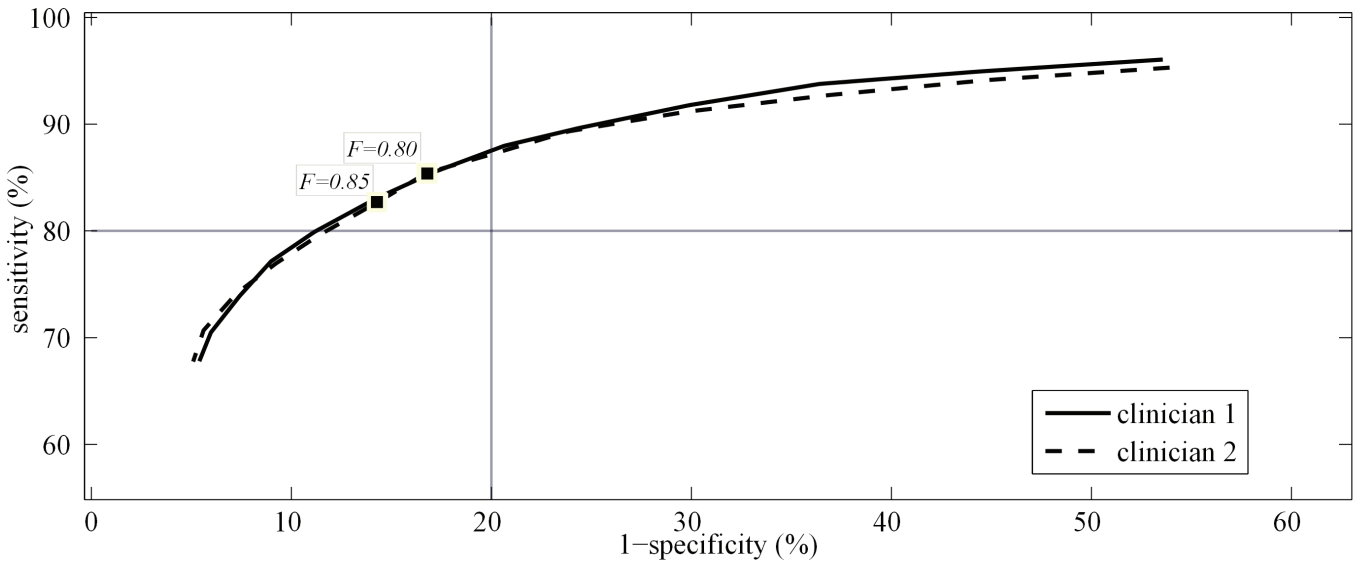
100
90
80
70
60
sensitivity (%)
0
10
20
30
40
50
60
1−specificity (%)
F=0.80
F=0.85
clinician 1
clinician 2

Figure 3 a

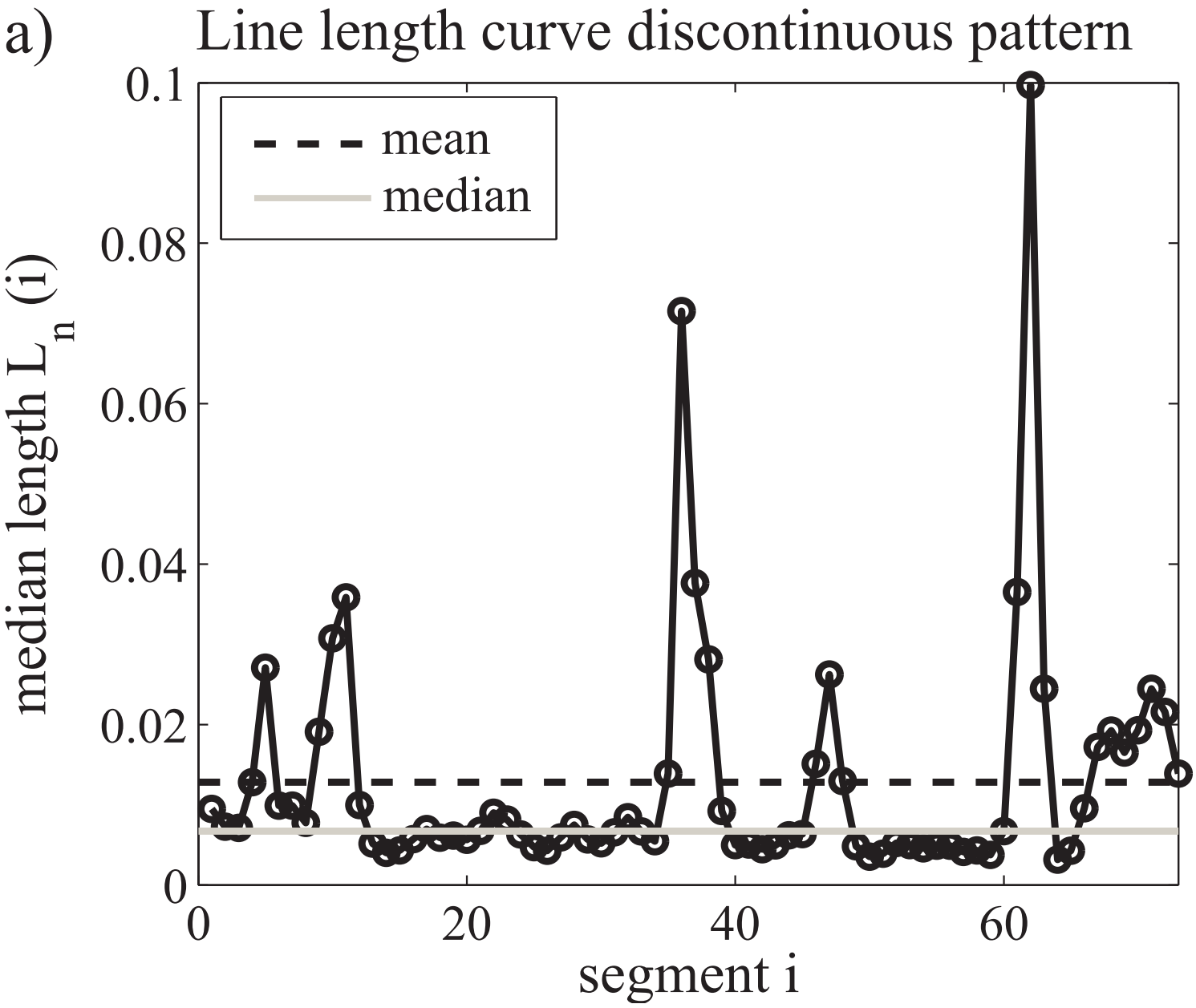
a)
Line length curve discontinuous pattern
mean
median
median length L_n (i)
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
segment i

Figure 3 b

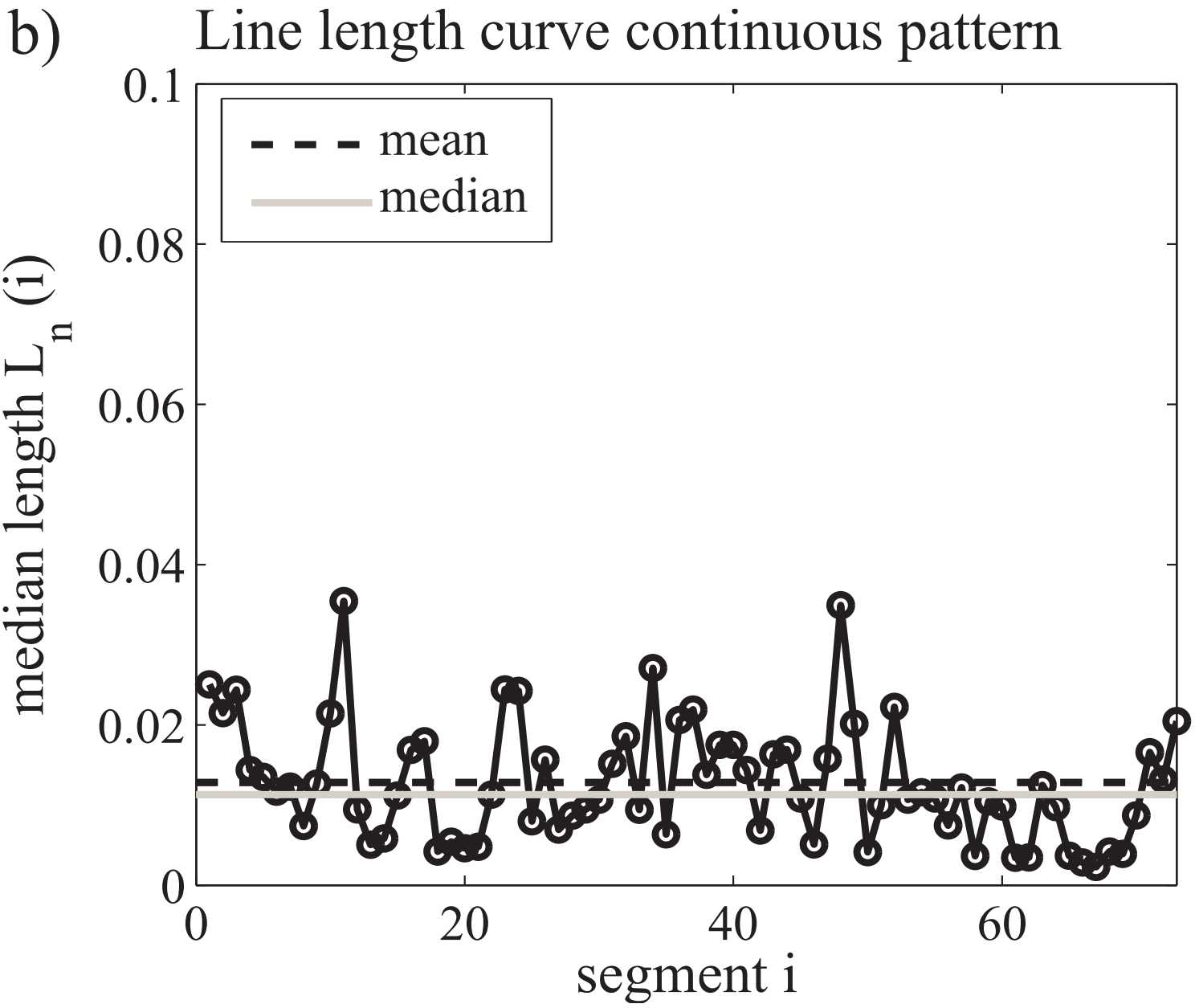
b)
Line length curve continuous pattern
mean
median
median length L_n (i)
segment i
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60

Figure 4

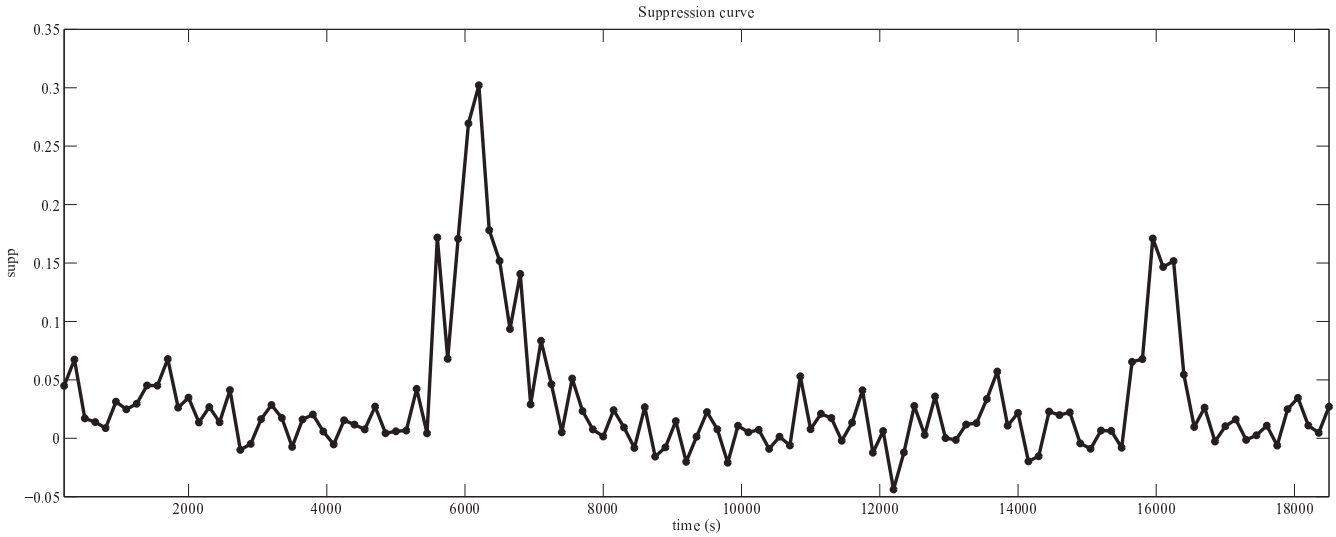
Suppression curve
supp
time (s)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
2000
4000
6000
8000
10000
12000
14000
16000
18000

Figure 5

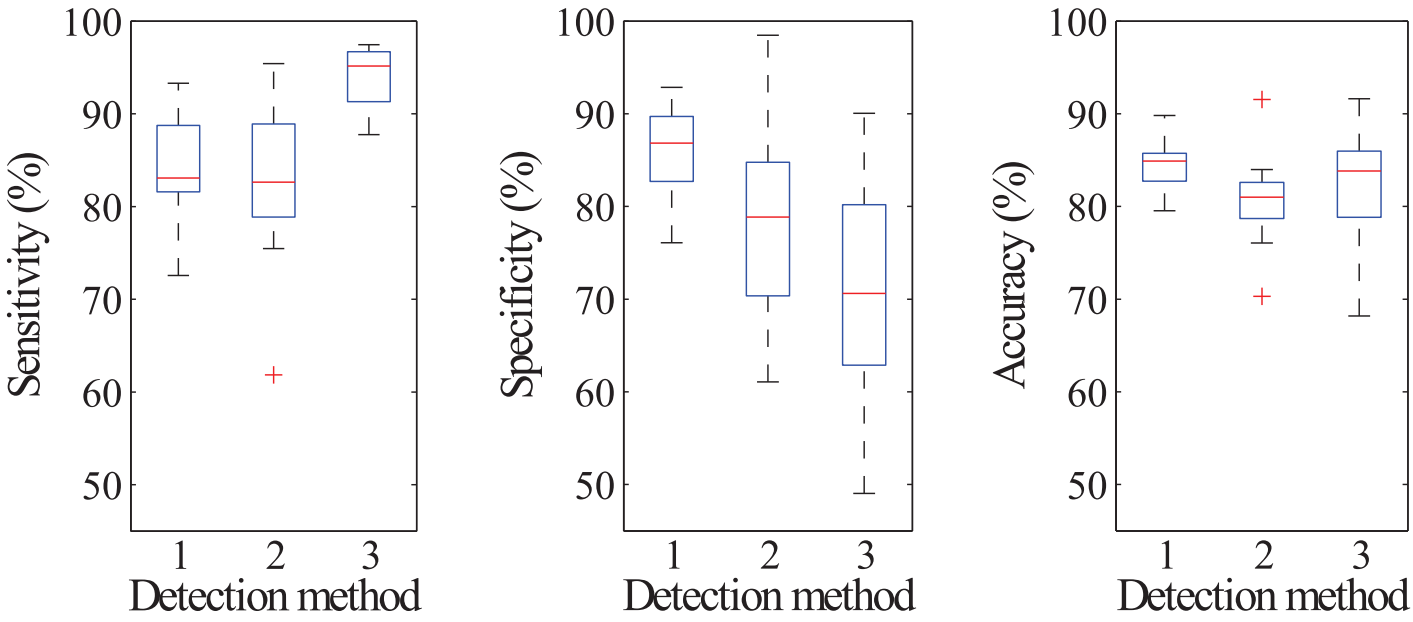
Sensitivity (%)
100
90
80
70
60
50
1
2
3
Detection method
Specificity (%)
100
90
80
70
60
50
1
2
3
Detection method
Accuracy (%)
100
90
80
70
60
50
1
2
3
Detection method

Figure 6 a

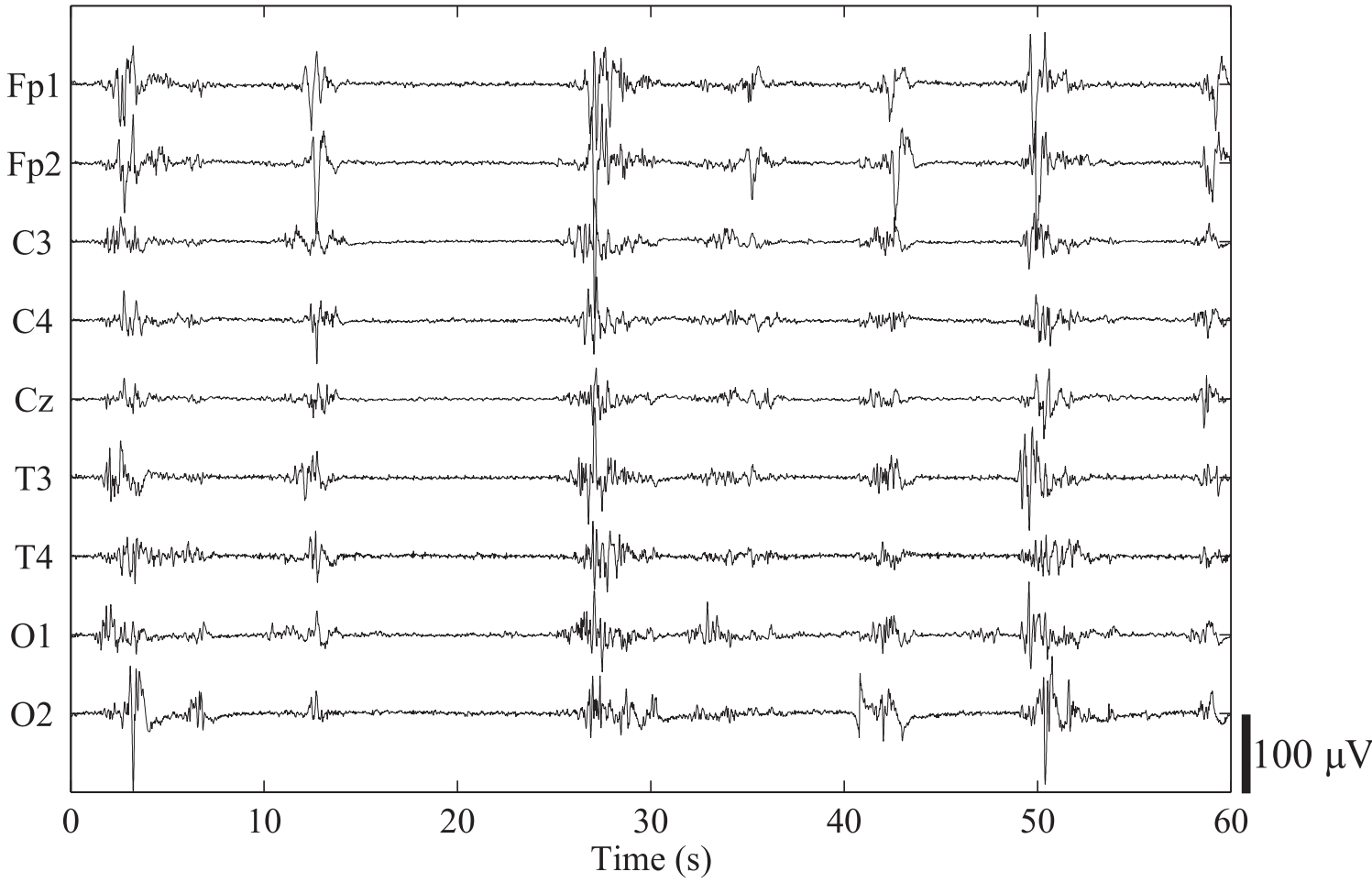

Figure 6 b

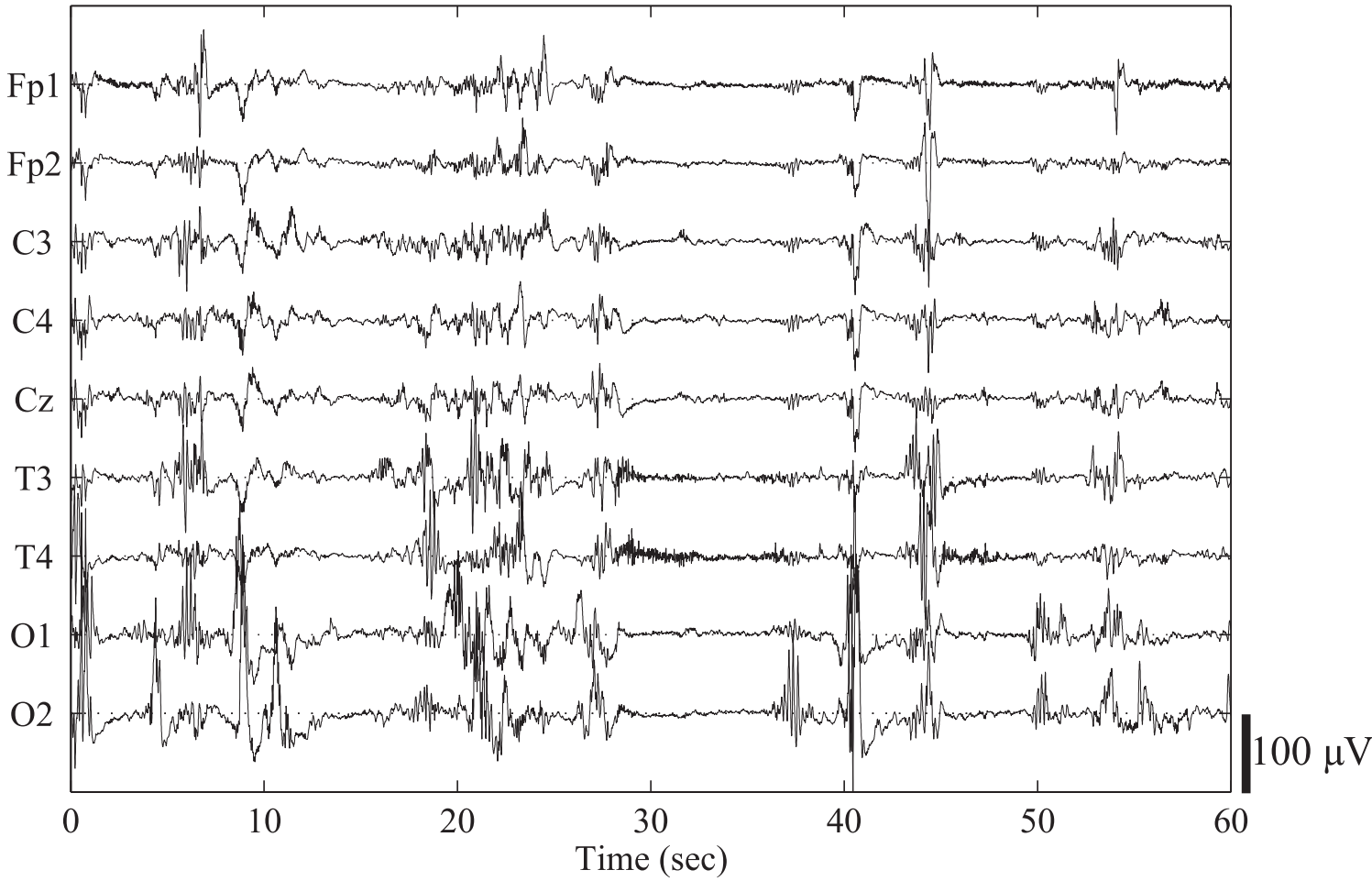

Figure 7 a

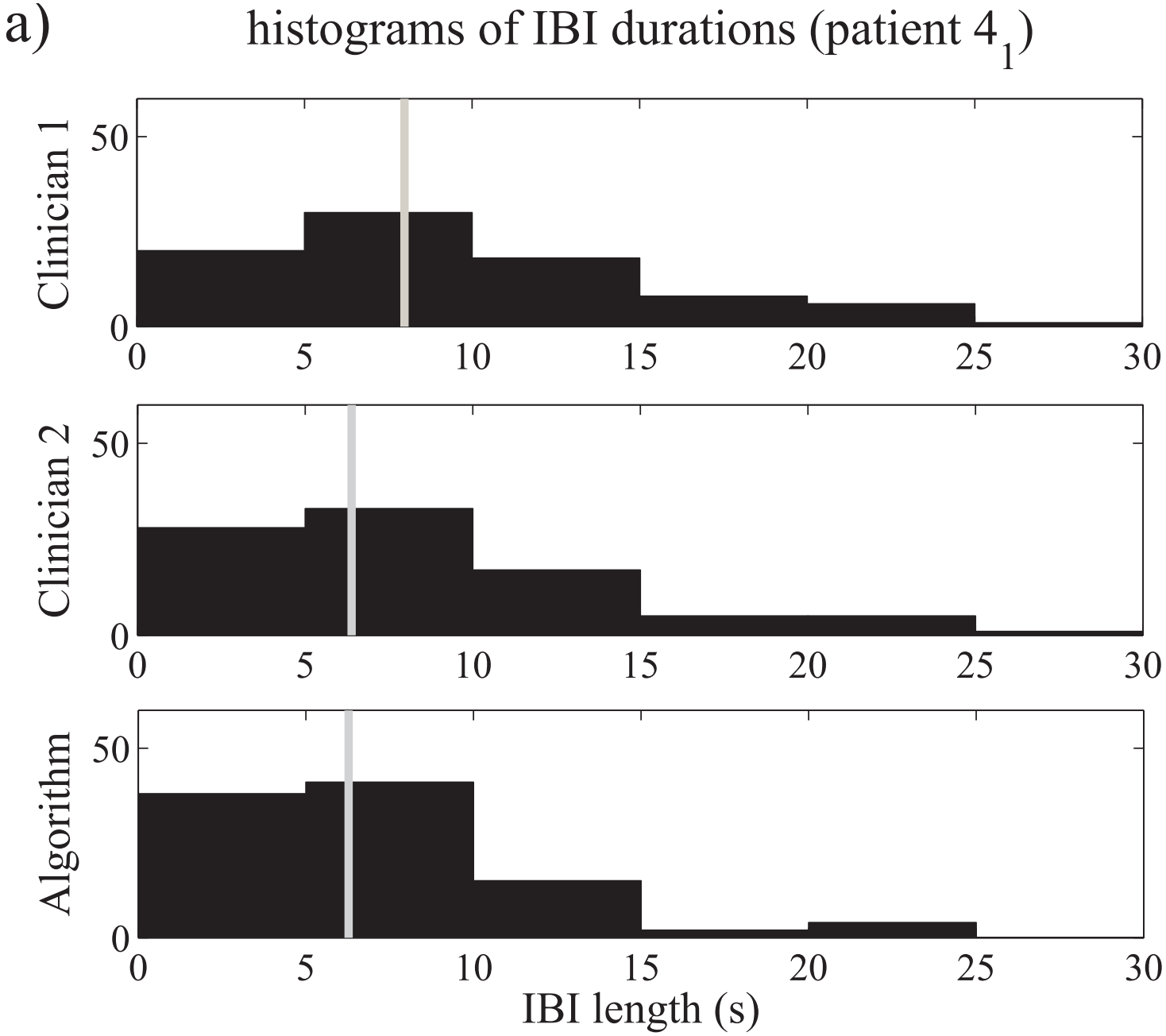
a)
histograms of IBI durations (patient 4 1)
Clinician 1
Clinician 2
Algorithm
50
0
0
5
10
15
20
25
30
IBI length (s)

Figure 7 b

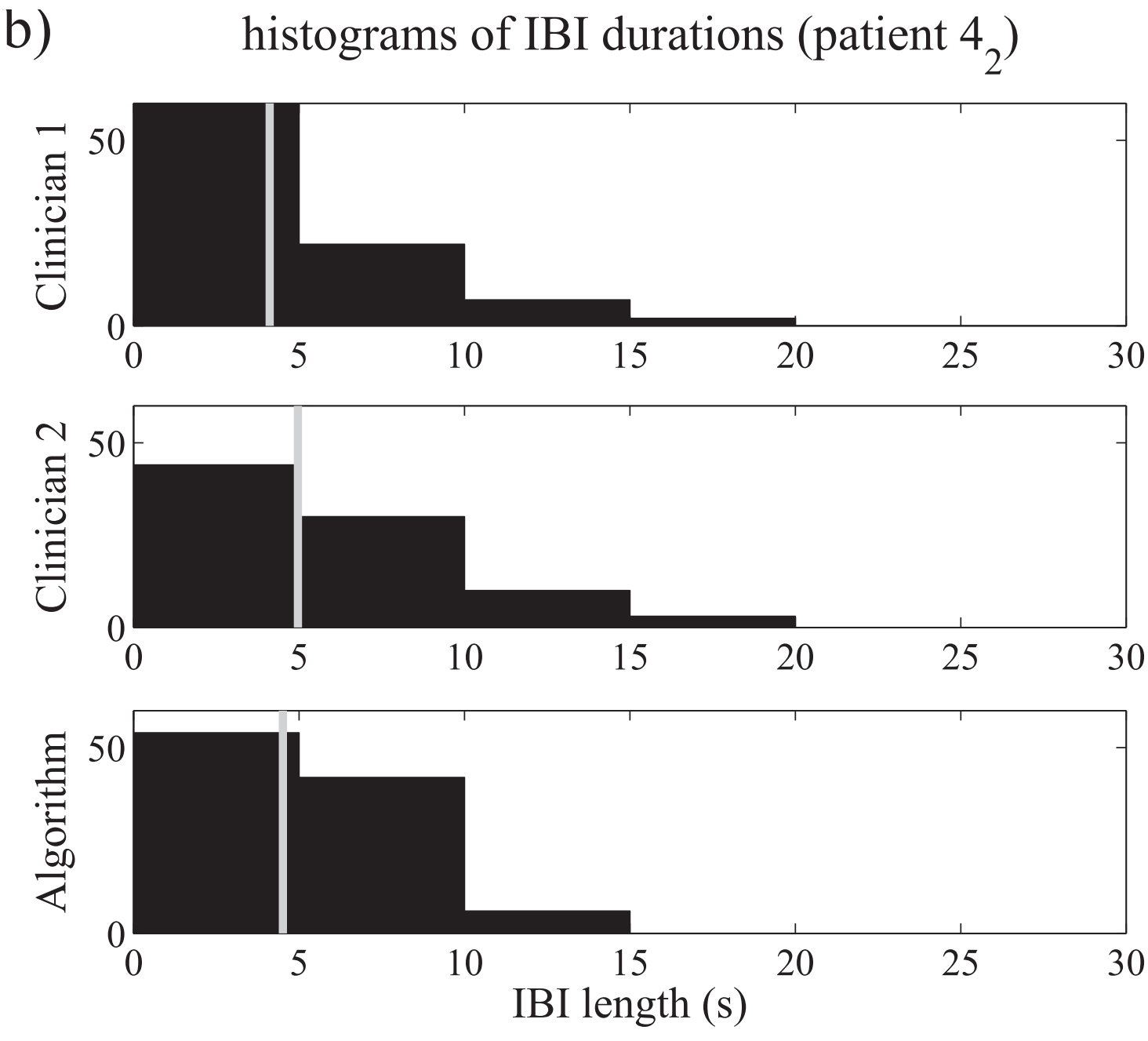
b)
histograms of IBI durations (patient 4$_2$)
Clinician 1
Clinician 2
Algorithm
50
0
0
5
10
15
20
25
30
IBI length (s)

Figure 8

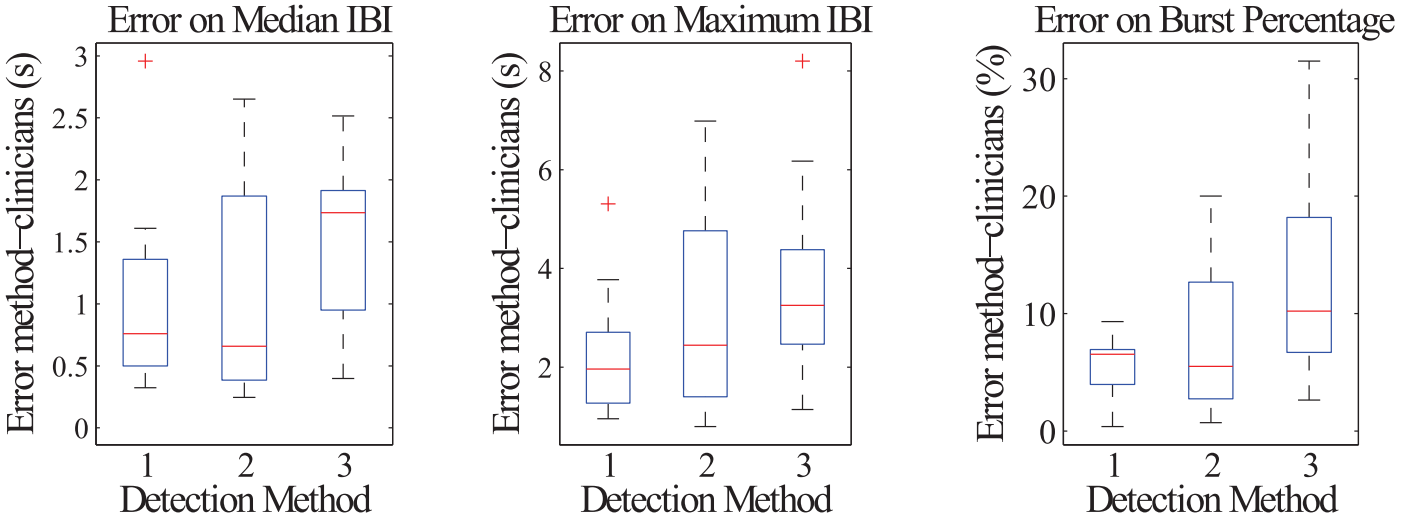
Error on Median IBI
Error method–clinicians (s)
3
2.5
2
1.5
1
0.5
0
1
2
3
Detection Method
Error on Maximum IBI
Error method–clinicians (s)
8
6
4
2
1
2
3
Detection Method
Error on Burst Percentage
Error method–clinicians (%)
30
20
10
0
1
2
3
Detection Method

Figure 9

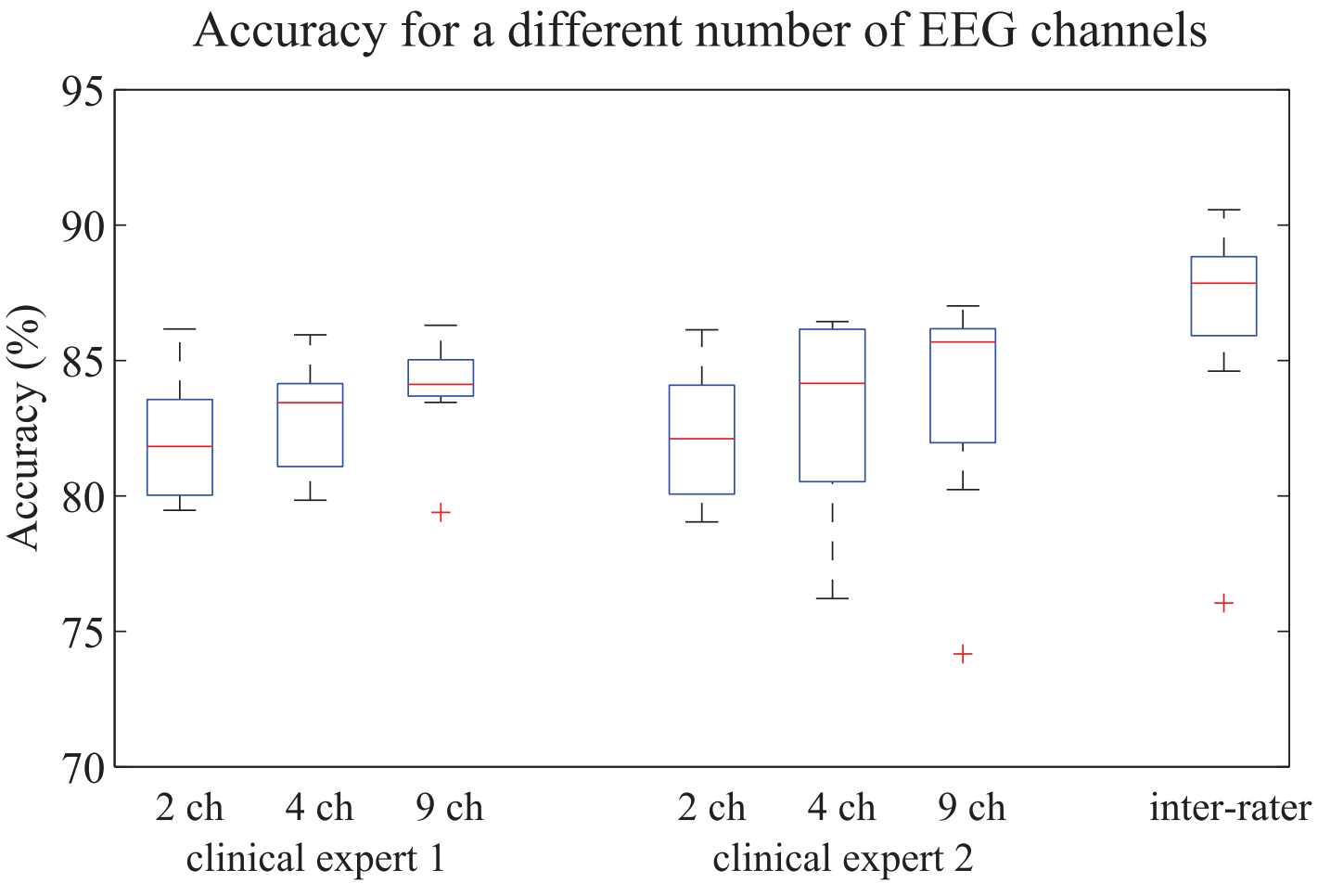
Accuracy for a different number of EEG channels
Accuracy (%)
95
90
85
80
75
70
2 ch
4 ch
9 ch
2 ch
4 ch
9 ch
inter-rater
clinical expert 1
clinical expert 2

Table 1

Table 1: Sensitivity, specificity and accuracy of three detection methods for both the training and validation set. Detection methods are based on line length (meth1), based on envelope calculation (meth2) and based on NLEO (meth3).

	Training set (8 measurements)				Validation set (5 measurements)			
	meth1	meth2	meth3	inter-rater	meth1	meth2	meth3	inter-rater
Sensitivity (%)	85.49	81.33	93.38		81.63	84.62	94.73	
Specificity (%)	83.49	77.79	68.16		89.23	79.97	73.96	
Accuracy (%)	84.02	79.07	80.22	86.56	84.70	83.00	84.95	85.63

Table 2: Statistical performance t-test on statistical measurements - accuracy, sensitivity and specificity - (see boxplots in Figure 5). Two detection methods are compared each line; m1: line length based method, m2: envelope based method, m3: NLEO based method. In the table, the differences between the mean values of these statistical measurements are given. Thereby, a significant difference is presented as * ($p<0.05$), ** ($p<0.01$), *** ($p<0.001$).

	Difference in mean accuracy (%)	Difference in mean sensitivity (%)	Difference in mean specificity (%)
m1-m2	3.7 ***	1.41	7.07 **
m1-m3	2.24	-9.9 ***	15.31 ***
m2-m3	-1.46	-11.31 ***	8.24 ***

Table 3: Statistical performance t-test on three prognostic features: median IBI length, maximum IBI length and the burst percentage. Two detection methods are compared each line; m1: line length based method, m2: envelope based method, m3: NLEO based method. In the table, the differences between errors on the prognostic features are given. Thereby, significant difference is presented as * ($p<0.05$), ** ($p<0.01$), *** ($p<0.001$).

	Difference in error on Median IBI (s)	Difference in error on Max IBI (s)	Difference in error on Burst Percentage (%)
m1-m2	0.1	-0.48	1.04
m1-m3	-0.98	-1.29 *	-3.67 **
m2-m3	-1.08 *	-0.81	-4.72 *