Interrater agreement in visual scoring of neonatal seizures based on majority voting on a web-based system: the Neoguard EEG database.

Anneleen Dereymaeker[1*], Amir H. Ansari[2,3*], Katrien Jansen[1,4], Perumpillichira J Cherian[5,6], Jan Vervisch[1,4], Paul Govaert[7,8], Leen De Wispelaere[7], Charlotte Dielman[9], Vladimir Matic[2,10], Alexander Caicedo Dorado[2,3], Maarten De Vos[11], Sabine Van Huffel[2,3], Gunnar Naulaers[1]

These authors are joint first authors

[1]Department of Development and Regeneration, University Hospitals Leuven, Neonatal Intensive Care Unit, KU Leuven (University of Leuven), Leuven, Belgium

[2]Division STADIUS, Department of Electrical Engineering (ESAT), KU Leuven (University of Leuven), Leuven, Belgium

[3]iMinds-KU Leuven Medical IT Department, Leuven, Belgium

[4]Department of Development and Regeneration, University Hospitals Leuven, Child Neurology, KU Leuven (University of Leuven), Belgium

[5]Section of Clinical Neurophysiology, Department of Neurology, Erasmus MC, University Medical Center Rotterdam, The Netherlands

[6]Division of Neurology, Department of Medicine, McMaster University, Hamilton, Canada

[7]Section of Neonatology, Department of Pediatrics, Sophia Children's Hospital, Erasmus MC, University Medical Center Rotterdam, The Netherlands

[8]Section of Neonatology, Middelheim Ziekenhuis-ZNA, Antwerp, Belgium

[9]Section of Child Neurology, Paola Ziekenhuis- ZNA, Antwerp, Belgium

[10]Faculty of Technical Science, Singidunum University, Belgrade, Serbia

[11]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK

Corresponding author:

Anneleen Dereymaeker, Department of Development and Regeneration, University Hospitals Leuven, Neonatal Intensive Care Unit, KU Leuven (University of Leuven), Herestraat 49, 3000 Leuven, Belgium E-mail address: anneleen.dereymaeker@uzleuven.be tel. +32 16 34 26 52

KEY WORDS: electrographic seizures, interrater agreement, neonatal multichannel EEG, consensus agreement, seizure detection algorithms, Neoguard.

number of text pages: 19, number of words: 4300, number of references: 40, number of figures: 6, number of tables: 0

SUMMARY

Objective: To assess interrater agreement based on majority voting in visual scoring of neonatal seizures.

Methods: Multichannel EEGs of term neonates with seizures were uploaded to an online database and independent review of suspected events was done by 4 raters. First, consensus decision was defined based on 'majority voting' and interrater agreement was estimated using Fleiss' Kappa. Second, the influences of different factors on interrater agreement were determined.

Results: A total of 1919 events in 280h EEG from 71 neonates with seizures caused by different aetiologies, were reviewed by 4 observers. Majority voting was applied to assign a binary classification (seizure/non-seizure) to all the events. From a total of 1919, 44% events were classified with high, 36% with moderate, and 20% with poor agreement, resulting in an interrater agreement for all events in a Kappa value of 0.39. Sixty-eight percent of events were labelled as seizures, and in 46%, all raters were convinced about electrographic seizures. The most common seizure duration was below 30s. Raters agreed best for seizures lasting 60-120s. There was a significant difference in the electrographic characteristics of seizures versus dubious events, with seizures having longer duration, higher power and amplitude. Upon introducing a less experienced rater, total Kappa decreased to 0.29.

Significance: This study provides an extensive assessment of interrater agreement of neonatal seizures based on majority voting. There is a wide variability in identifying rhythmic ictal and non-ictal EEG events, and only the most robust ictal patterns are consistently agreed upon. Database composition and electrographic characteristics are important factors that influence seizure recognition. Quantitative analysis helps to establish unequivocal measures which can be useful for evidence-based studies of seizure recognition and management. The use of well-described databases and input of different experts will help to improve neonatal EEG interpretation guidelines and develop uniform seizure definitions.

KEY WORDS: electrographic seizures, interrater agreement, neonatal multichannel EEG, consensus agreement, seizure detection algorithms, Neoguard.

1. INTRODUCTION

In recent years, there is an increasing focus on neuromonitoring of neonates at high risk for cerebral dysfunction. Neonatal encephalopathy and seizures are the most common and distinctive signs of underlying brain injury and warrant neuromonitoring. There is emerging evidence from experimental research that prolonged or recurrent seizures add injury to the developing brain[1-4]. Clinical studies supporting these findings are still scarce. However, few studies have reported that seizures worsen brain damage in children with hypoxic-ischemic encephalopathy

(HIE)[5; 6], a higher seizure burden is correlated with adverse outcome[4; 7-9] and that neonatal seizures are independently associated with long-term neurological impairment[10]. These observations suggest that early identification and treatment might diminish long-term morbidity, however there is a pressing need for well-designed prospective studies to support these observations.

Diagnosing neonatal seizures remains a challenge due to their unique characteristics[11]. First, bedside clinical recognition of seizures is not accurate since the clinical manifestations of neonatal seizures are often subtle and highly variable[12-14]. Second, critically ill neonates often manifest only electrographic seizures[14], and anti-epileptic drugs (AED) may suppress clinical seizures while ictal EEG discharges persist (electroclinical dissociation)[15]. Continuous video-electroencephalography (cEEG) is the gold standard for accurate diagnosis of both electroclinical and electrographic seizures[16; 17] and prolonged bedside EEG monitoring is mandatory to evaluate the efficacy of their treatment with AED[18].

The implementation of continuous video-EEG monitoring is resource intensive and requires expertise not available to all neonatal units. Automated quantitative analyses of EEG, including background activity and seizure detection algorithms (SDAs), will improve clinical applicability. In the last decade, SDAs have been developed by our group[19; 20] and others[21-23] with the ultimate goal to aid the clinician in the field, however the results vary widely, with different performance outputs reported. The challenge of reproducing published results might be due to the nature of visual scoring criteria used by the clinician or due to the nature of the data of a particular database[24-28]. First, visual detection of seizures in multichannel EEG can only be accepted as gold standard when interrater variability is low. The algorithms used for automated detection of seizures are designed based on the current EEG definition of a neonatal seizure[29]. The algorithms at their best will match the decision made by the clinician. If there is not a good agreement on the labels of a 'true' seizure, then the performance of the algorithm will be affected by it. Interrater agreement for assessing electrographic seizures varies widely with a Kappa value of 0.4 to 0.85 in different studies[26; 27; 30-33]. High interrater agreement is achieved when scoring typical seizure patterns with well-defined morphology[15], whereas the interrater agreement may differ significantly in more dubious seizure patterns[26]. Moreover, the inherent ambiguity of visual classification might be influenced by the composition of the databases used in the different studies, and cause differences in output performances. Both electrographic characteristics as well as technical factors can result in the construction of databases with highly different levels of arduousness. Features such as total seizure number, duration, burden, morphology and rhythmicity of the electrographic seizure pattern in relation to the overall background activity and the use and effect of AED[33], might affect the feasibility of seizure pattern recognition, both visually and technically[26; 27]. Finally, assessment of interrater agreement is problematic in itself and at present, there is a lack of comparable measures to assess interrater agreement. Different approaches presented in literature are Kappa coefficients or Intraclass Correlation Coefficients (ICC), which can be used on 'event-based'(comparisons of annotation of an event by event basis/any-overlap)[26; 31; 33], and 'time-based' or 'epoch-based' metrics (comparisons of annotation on a second by second basis)[32]. Time-based metrics will reflect the amount of seizures (seizure burden) but are labour intensive to assess on long duration EEG monitoring, whereas the event-based metrics provide the percentage of seizures that will be detected[24]. Both measures are significantly influenced by the length of recordings, the difference in number of patients and the prevalence of seizures. This underscores the necessity to define more accurately interrater agreement estimates based on well-classified datasets which contain clinically relevant events. Only this will allow assessment of reliability of visual EEG interpretation and comparison of automated seizure detection algorithms with the same performance measure as used for visual annotation.

The aim of the present study was twofold: the first aim was to provide an objective and comprehensive assessment of interrater agreement of neonatal seizures based on a novel approach: majority voting. For this, we designed an online platform based on a large, multicentre seizure database of multichannel EEG recordings (Neoguard database) recorded at two university hospitals. A previously published SDA[19; 20] was applied to detect suspected seizure events and 4 raters reviewed the events. The second aim was to determine the influences of different factors on interrater agreement. In-depth analysis of relevant seizure characteristics, level of experience and technical factors that affect the majority voting and therefore the recognition of seizures, was performed.

2. METHODS

2.1 Database

The database is partitioned into 4 datasets according to the centres and durations of the used EEG recordings. Neonates in DB1-DB3 were enrolled from the neonatal intensive care unit (NICU) of Erasmus MC, University Medical Center Rotterdam (EMCR) from 2003-2014. These datasets, DB1-DB3, contain recordings from 48 neonates with gestational age ≥36 weeks and presumed birth asphyxia, and were selected for cEEG monitoring for at least 24h if they had clinical features of encephalopathy and at least one of the following features: (a) arterial pH of umbilical cord blood ≤ 7.1, (b) Apgar score ≤ 5 at 5 min and (c) high clinical suspicion (fœtal distress, umbilical cord prolapse, difficult labour or a history of convulsions). The 23 neonates in DB4 were enrolled from the NICU of University Hospital Leuven who presented with neonatal seizures in the first week of life and had cEEG monitoring during the time period 01/2013-07/2015 as part of standard clinical care. The majority of these patients had acute symptomatic seizures (e.g. HIE, stroke, haemorrhage, brain infection), however some neonates with suspected metabolic and genetic causes were also included.

DB1-3 were recorded at 256 Hz, (Nervus TM monitor, Taugagreining hf, Reykjavik, Iceland) using 17 scalp electrodes, according to the full 10-20 International System. The following electrodes were applied: F1, F2, F3, F4, F7, F8, T3, T4, T5, T6, P3, P4, C4, C3, Cz, O2, O1. DB4 was recorded at 250 Hz, using Brain RT (OSG BVBA Rumst, Belgium). Nine scalp electrodes were applied according to the restricted 10-20 system (F1, F2, T3, T4, C4, C3, Cz, O2, O1). Twenty bipolar channels were used in DB1-DB3 and 12 bipolar channels for DB4. The polygraphic signals included electrocardiogram (ECG), electro-oculogram (EOG), chin or limb surface electromyogram (EMG), and a respiratory movement signal (RESP). Band-pass filter was 0.3-70Hz. DB4 was resampled to 256Hz and all databases were filtered in the software between 1-20Hz before analysis. EEGs were interpreted by a clinical neurophysiologist as part of clinical care and AED treatment was initiated and followed according to the local hospital protocol. This study (Neoguard: the development of a neonatal EEG monitor with real-time, bedside data visualization and automated decision support) was conducted with approval from the Ethical Review Board in both hospitals. Recordings were fully anonymized for this retrospective data analysis.

2.2 Neoguard Website design: creating the online database

2.2.1 Seizure annotation

The EEG recordings of DB1-DB3 and DB4 were first annotated by two independent, experienced clinical neurophysiologists (primary raters, respectively PJC for DB1-DB3 and KJ for DB4) with extensive experience in neonatal cEEG who have been trained and practiced at different institutions. An electrographic seizure was defined as a paroxysmal and evolving repetitive stereotyped EEG event with a definite start, middle and end lasting for at least 10s on at least one EEG channel[29]. Status epilepticus was defined as recurrent seizure activity lasting ≥ 50% of the time for a minimal EEG record duration of 1 hour[34]. Seizures were annotated by the primary raters with the beginning and end of each seizure collectively over all channels. The raters had no access to video or clinical information.

2.2.2 Data transfer

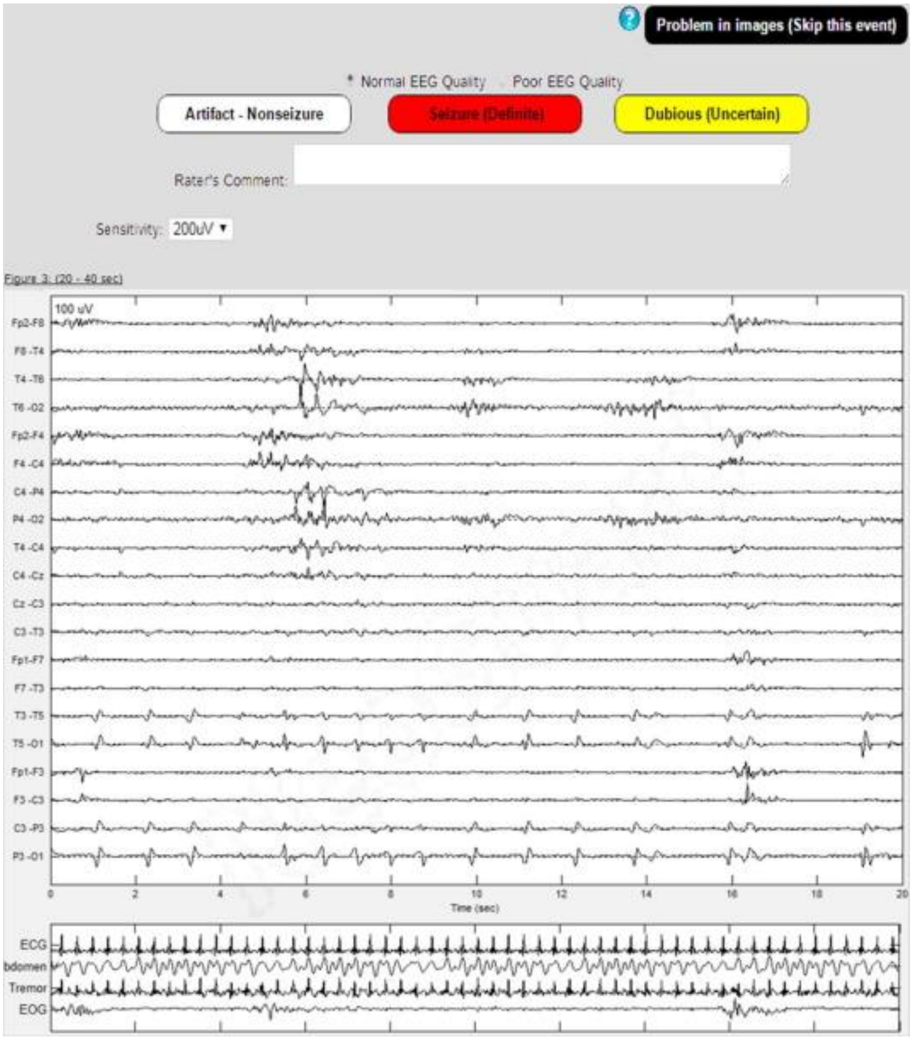
All annotated EEGs, were converted to European Data Format (EDF) files and anonymously loaded in Matlab. Montage setting was the same as in the original EEG file. From this long duration monitoring records, epochs of 8h, 4h, 2h and 2h, respectively of DB1, DB2, DB3 and DB4, containing at least one seizure observed by the primary rater, were randomly selected from the full EEG registration of each neonate. There was no preselection of EEG data based on the quality of registration or presence of artefacts.

2.2.3 Seizure detection algorithm (SDA)

A SDA was applied on the clipped 2h-8h EEG epochs of DB1-DB4. The algorithm used in this study has been described in detail in a previous article[19]. The algorithm consists of two detectors running in parallel: (a) a spike-train detector that detects high-energy segments of the EEG and analyses the correlation between them and (b) oscillatory seizure detector that detects increase in low-frequency activity (1–8 Hz) with high autocorrelation. The algorithm provides an output of event-based detection. Both events of the SDA detections and the primary raters (missed seizures= annotated by the primary raters but *not-detected* by the SDA) formed the data event pool (Neoguard database).

2.2.4 Neoguard Website design

Online, secured access was provided for all raters to review the events on the Neoguard Website. Events of the database were displayed with 5 images as follows: the first image presented the whole detected event in seconds, image 2 gave a detailed view of the first 20s of the event, images 3 and 4 gave a detailed view of the subsequent 20-40s and 40-60s, and image 5 gave a prospective view (from start of the event, to 200 sec later) (Fig. 1). Polygraphic signals were displayed simultaneously. There was no access to video or clinical information. The temporal view was fixed and there was no extra filtering option or ability to change montage. The sensitivity was adaptable (20/50/100/150µV). Raters were allowed to score without any time limitation and the whole scoring process was conducted over one year. An *experienced* rater was defined for the purpose of this study as: 'a rater who had reviewed at least 200 neonatal EEGs'.



● Normal EEG Quality ○ Poor EEG Quality

Rater's Comment:

Sensitivity: 200uV ▾

Fig. 1: Events displayed on the Neoguard Website for scoring. Screenshot of image 2, a detailed view of the 20-40s of the event.

2.3 Reviewing SDA detections

2.3.1 Scoring of the events

Three secondary raters (SR 1-3), blinded to the primary scores and each other's annotations, reviewed independently all events (all SDA detections + missed seizures= annotated by the primary raters but missed by the SDA) as 1. Definite seizure, 2. Dubious seizure, 3. Non-seizure/artefact. EEG reviewing was restricted to the selected events. Discharges were classified as 'dubious seizures' if the rater was unsure about ictal appearing electrographic discharges, such as: (a) runs of > 10s sharp waves/oscillations or a mixture of both, occurring arrhythmically (with marked variability in the interval and morphology between individual complexes in a seizure discharge for the major part of its duration) or (b) rhythmic discharges of shorter duration ("brief rhythmic discharges"), or periodically occurring complexes consisting predominantly of sharp waves with no clear-cut evolution[26]. Non-seizure/artefact included typical artefacts (such as pulsation, respiration, ECG, tremor) which mimicked the stereotyped rhythmic seizure patterns that are often seen in neonates.

2.3.2 Measures of agreement

Majority vote

A binary classification with definite label was assigned to the events in the database based on the label of the 4 raters. Events were categorised as seizures and non-seizures. Dubious events were categorised in the non-seizure group, since those events, although paroxysmal, do not fit in the operational definition of seizures and their clinical significance is controversial. For each event, majority voting was applied to assign the final label and the percentage of the majority voting was calculated. All detected events were classified into 'poorly agreed' (2 raters agreed), 'moderately agreed' (3 raters agreed) and 'highly agreed' (all 4 raters agreed).

Interrater agreement

Interrater agreement was assessed using Fleiss' Kappa coefficients for categorical variables. The concept of our study only allowed assessment of event-based measures. Fleiss' Kappa is applicable for any number of raters[35]. Kappa coefficient is a measure of agreement for categorical data while controlling for agreement by chance. Kappa values often range from 0 (interrater agreement does not differ from chance) to +1 (total agreement). However, if the observed agreement is less than chance agreement (disagreement), Kappa could be negative. The level of agreement measured by Kappa was classified as follows: 0-0.20 slight agreement; 0.21-0.40 fair agreement; 0.41-0.60 moderate agreement: 0.61-0.80 substantial agreement; 0.81-1.00 almost perfect agreement.

Measures of agreement and correlation with electrographic characteristics

Factors that may affect seizure recognition and interrater agreement were examined. First, the effect of experience was assessed by introducing a 5th rater, with experience in neonatal EEG but still in the training-phase. Second, in-depth analysis of database composition and electrographic characteristics such as duration, amplitude (peak to peak amplitude) and power ($\mu V^2/s$) were assessed to evaluate the effect of these characteristics on ictal pattern recognition. Electrographic characteristics were compared between seizures and dubious classified events. Additionally, the interrater agreement for the classification of seizure intensity was also assessed using Fleiss' Kappa. Seizure intensity was defined using the maximum accumulated duration (burden) of seizures within any hour of the clipped recording, based on event-based measures. Three groups with different amount of seizure burden per hour (<10 min, 10-20 min and >20 min) were compared for interrater agreement. This classification was based on local clinical practice, treatment protocol and literature review[36; 37]. Less than 10 min: the cumulative seizure burden in clinical practice before initiation of treatment and time to check easily correctable causes (glucose, electrolytes, start antibiotics), 10-20 min: if seizures continue, additional dose of first-line AED treatment (phenobarbitone), >20 min of cumulative seizure activity, indication for second-line AED, as there is high risk of evolution to status epilepticus[36; 37].

Observer interrater agreement was also assessed in the context of the presumed seizure aetiology in DB4. Data of seizure characteristics were log-transformed (due to the very large range of values noted) to guarantee the symmetry of the data and allow further statistical testing[38]. Wilcoxon rank-sum test was used to investigate the effect of seizure features on seizure recognition by comparing the differences in electrographic characteristics between seizures and dubious event. Data were analysed using Matlab™ (The Mathworks, Natick, MA, USA). All tests were 2-sided and a *p*-value <0.05 was considered significant.

3. RESULTS

3.1 Description of the database

A total of 1919 events in 280h of multichannel EEG from 71 neonates with seizures, were rated by 4 independent observers. The majority of the events in these databases were less than 30s (Fig. 2). The most common diagnosis was HIE (DB1-DB3: 48 patients with HIE, 3/48 treated with therapeutic hypothermia). The presumed aetiology of seizures for the 23 neonates in DB4 was: HIE (*n*=6), stroke (*n*=5), intracranial haemorrhage (*n*=2), metabolic disease (*n*=5) genetic causes (2), meningitis (*n*=1), encephalitis (*n*=1) and malformation of cortical development (*n*=1). The majority of these neonates received treatment with at least one AED during the recordings.

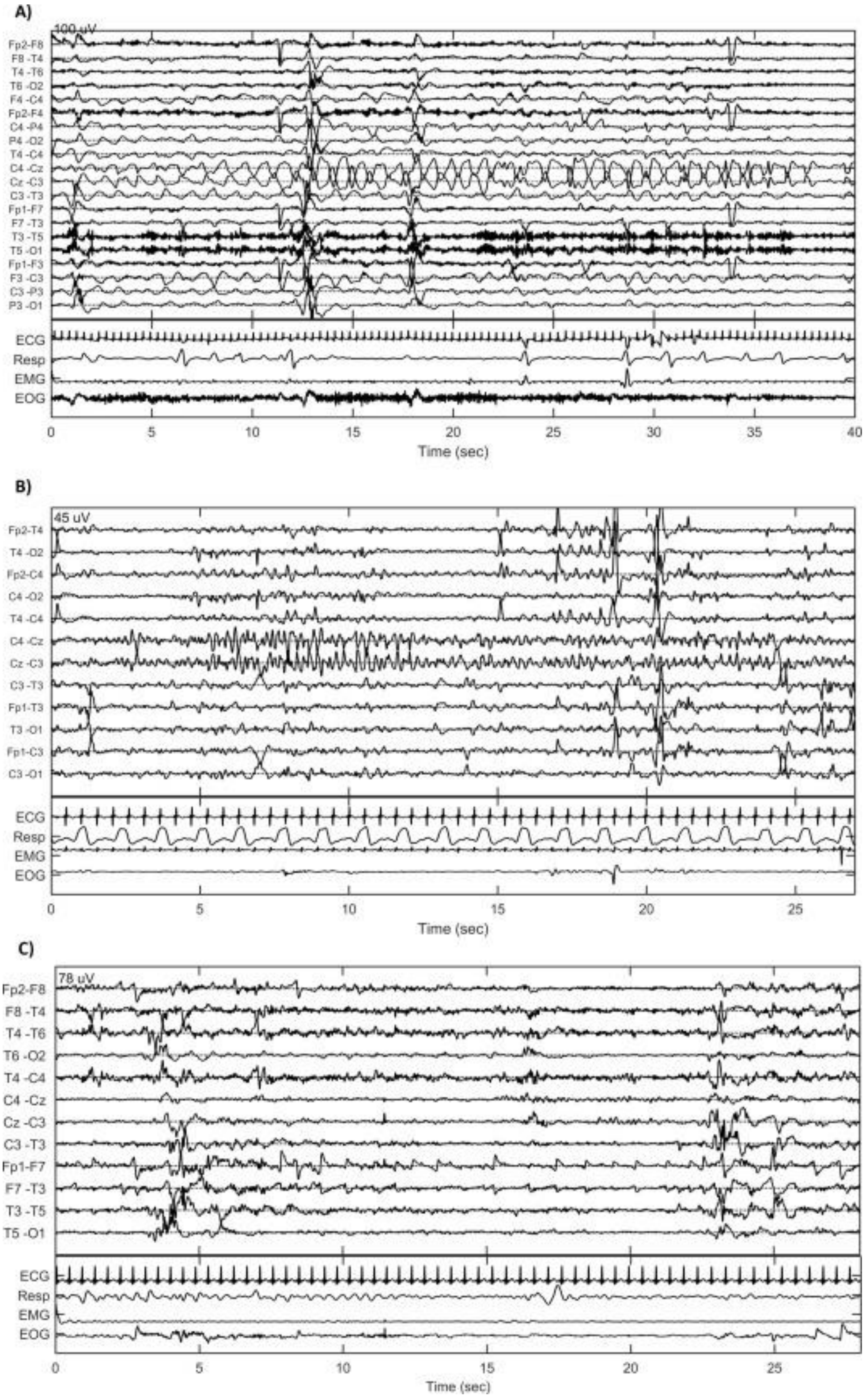


Fig. 2: Database overview: the number of all the events, classified as seizures and non-seizures with a given median duration based on majority voting.

3.2 Majority vote and interrater agreement

A majority voting system was applied to assign a binary classification (seizure/non-seizure) to the detected events. From a total of 1919 events, 44% events were classified with high agreement, 36% with moderate agreement, and 20% with poor agreement (Fig. 3).

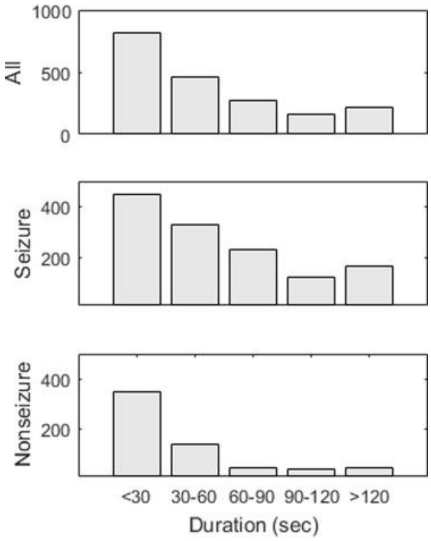


Fig. 3: Majority voting for all the events, seizures and non-seizures. Poor agreement (Poor): 2 raters agreed, moderate agreement (Mod.): 3 raters agreed, high agreement (High): 4 raters agreed.

Interrater agreement among all raters and for all the events (seizures/non-seizures), resulted in a total Fleiss' Kappa value of 0.39 (±0.007). Kappa values were comparable for of DB1, DB3 and DB4 whereas Kappa was slightly lower for DB2: DB1: 0.39 (±0.012), DB2 0.29 (±0.013), DB3 0.42 (±0.018), DB4 0.42 (±0.020). Kappa between individual raters varied among 0.30-0.47 (±0.020) (Fig. 4).

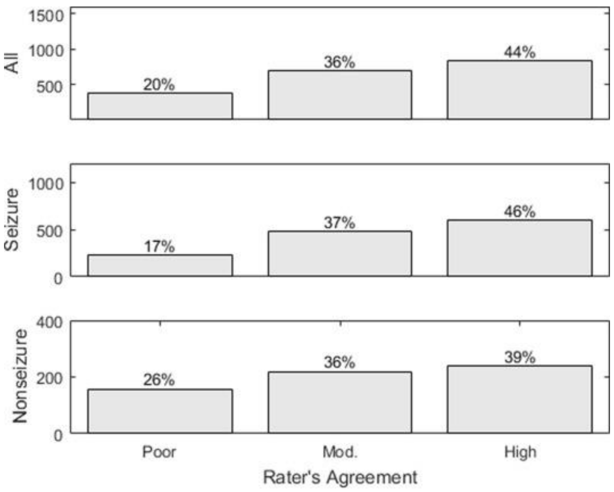


Fig. 4: Interrater agreement among all raters and for all the events, presented with Fleiss' Kappa (k). P: primary rater. SR 1-3: secondary raters.

3.3 Measures of agreement and correlation with electrographic characteristics

68% of the events were labelled as seizures, and in 83% of them, at least 3 raters were convinced about electrographic seizure activity and in 46%, all of them agreed (Fig. 3). The most common seizure duration in this dataset was below 30s (43% of all seizure events). Raters agreed the best for events of 60 to120s. The seizure duration with the highest level of agreement was 90-120s (63%) (Fig. 5). Higher levels of disagreement were noted in brief seizures (<30s) and long-duration seizures (>120s) (poor interrater agreement: respectively 21% and 23% versus 7% in seizures with duration of 60-90s). In non-seizure events, events of 60-90s and 90-120s were recognised with the highest certainty (respectively 45% and 41% with agreement for all raters).

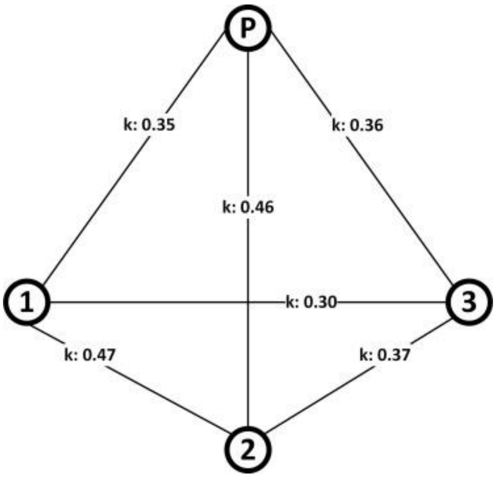


Fig. 5: Effect of seizure duration on majority voting. Poor agreement (Poor): 2 raters agreed, moderate agreement (Mod.): 3 raters agreed, high agreement (High): 4 raters agreed.

There was a significant difference in the electrographic characteristics of the seizures with moderate and high agreement versus dubious events, with seizures having higher median duration (*p*<0.001), higher median power (*p*<0.001) and higher median peak to peak amplitudes (*p*< 0.01) (Fig. 6).

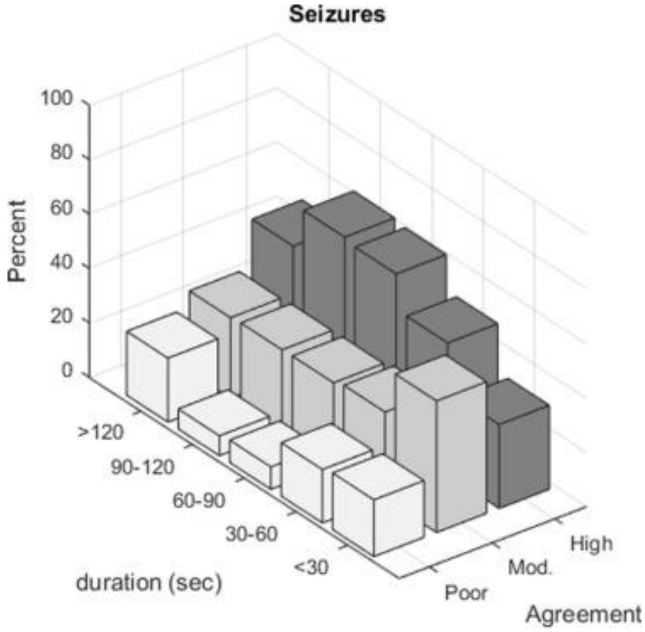


Fig. 6: Significant electrographic differences between seizures (S) and dubious events (D) for duration (*p*<0.001), power (*p*<0.001), amplitude (*p*<0.01). Data are log-transformed and analysed with Wilcoxon rank-sum test.

Upon introducing a 5th, less experienced rater, total Kappa decreased to 0.29 (±0.006). Iteratively leaving out 1 experienced rater, resulted in total Kappa values of 0.24-0.29 (±0.007) versus a total Kappa of 0.39 (±0.007) for the 4 experienced raters.

Assessment of maximum seizure burden resulted in Kappa values of 0.42 (±0.013) for <10 min maximum seizure burden per hour, 0.39 (±0.012) for 10-20 min maximum seizure burden per hour and dropped to 0.26 (±0.013) for >20 min maximum seizure burden per hour.

Interrater agreement was assessed according the different aetiology of seizures in DB4. Only a limited number of patients (n=23) were available for this analysis, however a total of 251 events were included. 3 groups were compared: group 1 (HIE: 99 events), group 2 (stroke + intracranial haemorrhage: 66 events)), group 3 (metabolic, genetic, meningitis, encephalitis and malformations of cortical development: 89 events). Fleiss' Kappa for group 1 was 0.42 (±0.03), group 2 0.48 (±0.04), but Kappa decreased notably in group 3 to 0.34 (±0.03). Subgroup analyses revealed that events in group 3 were of shorter duration (median 24s, IQR: 17-49) whereas the median duration of all events was 35s (±0.017). More events were classified as dubious in this group compared to group 1 and 2.

Fleiss' Kappa for group 1 with HIE (0.42 ±0.020), was comparable to Kappa of DB1-DB3, respectively 0.39 ((±0.012) and 0.42 (±0.018), with a median duration of events of 45-47.5s. Total Kappa of DB2 with HIE was noticeable lower: 0.29 (±0.013) as well as event duration (median 30s, IQR: 18-60).

4 DISCUSSION

This is the first study providing a comprehensive assessment of interrater agreement of neonatal seizures based on majority voting with 4 independent raters, using multichannel cEEG. This study also assessed the effects of the characteristics of the database, in-depth analysis of electrographic features of suspicious events and their effect on interrater agreement. The current results of our study suggest that there is a wide variability in identifying some rhythmic ictal and non-ictal EEG features, and that only the most robust ictal patterns are consistently agreed upon by EEG readers. The length of the events influences seizure recognition and interrater reliability. Both visual and automated recognition of very short seizures lasting less than 30s is very challenging. Stevenson et al.[32] have shown that the reliability of definite seizure recognition was low for seizures with a duration less than 30s. The results of our majority voting support these findings and confirms that greater certainty could be achieved when the ictal pattern was long enough and displayed a clear-cut evolution. Moreover, we and others have shown that automated detection of brief seizures is challenging[22; 27; 28]. Quantitative analysis of electrographic features of the events in our database indicates that dubious events can be discerned from definite seizures, based on their shorter duration, lower amplitude, and power. Correlation of human raters for seizure recognition has also identified that longer duration seizures with ambiguous offset and associated with high seizure burden, mixed up seizure annotation and resulted in lower agreement[26; 39], which is further depicted in our results of interrater agreement and seizure burden. Interestingly, subgroup analysis based on aetiologies, showed a striking decrease in Kappa in the group with metabolic and genetic causes compared to HIE and seizures due to a focal lesion. Those seizures were of shorter duration, however we observed also more variation in seizure morphology and spatial distribution in individual patients, which might have complicated seizure annotation. The drop in Kappa value for DB2 (HIE) can be explained by the lower duration of seizures (median duration of 30s) and very suppressed, low amplitude EEG background, as described by Cherian et al[26]. Finally, as expected, Kappa decreased upon introducing a less experienced rater.

Interrater agreement is fair at best, with Fleiss' Kappa values around 0.4. Untangling the exact value of the assessment of interrater agreement in previous studies is very challenging due to the different approaches for the assessment of interrater agreement and databases that have been used. First, as mentioned before, the length of the events does influence seizure recognition. The database used in Stevenson et al.[32] included around 55% of seizures between 30-120s and only 16% of seizures <30s. In a study of Mathieson et al.[27], SDA performance was also reduced when detecting short seizures and the database used in that study, was mainly constructed of seizure duration of 60-120s, whereas our database contains 45% events of shorter duration <30s, which increases the degree of arduousness. Second, a comparison of our results with other studies assessing interrater agreement, is complicated by the different metrics used, and different output of measures of agreement. We have used event-based measures to calculate Fleiss' Kappa to assess seizure versus non-seizure events, which is how a clinician would score and treat seizures detected on cEEG. Time-based measures for seizure occurrence may yield different, possible higher estimates since it takes into account also the longer epochs of negative agreement[33], in particular when using long duration monitoring, which could be more straightforward to interpret, however this measure will also be influenced by the prevalence of seizures. Studies comparing Fleiss' Kappa for seizure occurrence in neonates, based on event-based measure for this amount of patients are rare, but reported Kappa values in the same range for two raters[26]. Abend et al.[31] found a Kappa of 0.4 in older infants, assessed by three experienced raters. In contrast, Shah et al.[33] reported a much higher interrater agreement κ (Cohens Kappa) of 0.84 for two raters, however only 7 neonates with seizures were included. Stevenson et al.[32] reported high levels of interrater agreement for three experienced raters in 70 neonates (35 neonates with seizures, 35 neonates without seizures) with long duration multichannel EEG monitoring, using time-based measures for Fleiss' Kappa estimates. While we aimed to create a realistic, multicentre database, this study may have underestimated the agreement among EEG raters. First, this database includes critically ill neonates and a substantial part of them have been treated with AED and sedative medications at the time of recording. These drugs might have suppressed EEG background (amplitude and continuity)[40], influenced seizure morphology and rhythmicity and affected seizure recognition[26], which was not taken into consideration in the visual assessment, nor was this accounted for in the quantitative analyses. Undoubtedly, including 'less recognizable seizures or more dubious events' in this database has influenced interrater agreement[26; 28]. Second, the option to display the complete EEG selection in detail, extra drug information (sedative and AED) and access to video-EEG segments may improve characterisation of these events in future applications of the website. This information can assist EEG readers to make a seizure diagnosis, interpret dubious patterns or recognise the cause of seizure-like artefacts. However, only this kind of scoring methods mimics the actual working mechanism of a SDA and allows direct comparison with SDA performance. Furthermore, although long duration cEEG monitoring was done in all these neonates, due to the design of the website, only 60s EEG snapshots of the event were displayed in detail, whereas others have used the full recording for seizure annotation. This may have introduced bias, being unclear to the rater whether periodic epileptiform discharges represented an inter-ictal pattern or the start or end of a longer ictal pattern. However, since events of 60-120s are classified with the highest agreement, we assume that this effect is probably small, since this only affects events with longer duration (>120s). Improvements to the website based scoring can be easily made by displaying the whole event in detail, with successive 20s epochs. We are willing to share this web-based system with other researchers as well as with interested clinicians.

Further improvement of automated analysis will eventually help the clinicians as well as researchers by decreasing the effects of subjective interpretation of many events in prolonged neonatal EEG monitoring. Future studies need to provide unambiguous definitions and classifications of neonatal seizures which can be reproduced by EEG readers as well as used in automated algorithms. Based on the objective assessment of the 'core characteristics' of visual patterns in the neonatal EEG, research into the pathophysiological basis at cellular, synaptic and network levels, can be conducted. This will help to elucidate the effect of modulators such as severity and nature of underlying brain injury and effect of treatment on neurodevelopmental outcome.

In conclusion, visual interpretation of neonatal seizures remains challenging, however it is imperative to know the accuracy of visual classification to further improve our knowledge. This study provides an extensive assessment of interrater agreement of neonatal seizures based on majority voting. In-depth analysis of database composition and electrographic characteristics, revealed important factors that influenced seizure recognition. The results suggest that there is a wide variability in identifying rhythmic ictal and non-ictal EEG features, and that only the most robust ictal patterns are consistently agreed upon. Quantitative analysis helps to establish objective and unequivocal measures which can be useful in evidence-based studies of seizure recognition and management. The use of well-described databases and input of different EEG experts will help to improve neonatal EEG interpretation and define uniform, general accepted seizure definitions.

KEY POINTS BOX:

- Visual interpretation of electrographic neonatal seizures is very challenging.
- Objective assessment of interrater agreement of neonatal seizures and factors influencing it is crucial to improve our understanding of these phenomena.
- Majority voting can be used for clinical validation and to develop improved and uniform seizure definitions.
- The duration of events and the composition of databases influences seizure recognition and interrater agreement.
- The use of well-described datasets is useful to validate clinical EEG interpretation and serve as learning platform.

ACKNOWLEDGEMENTS:

AD, AA, VM and SV are supported by IWT grant- TBM 110697-NeoGuard. AA, VM, and SV are supported by: iMinds Medical Information Technologies: Dotatie-Strategisch basis onderzoek (SBO-2015); Belgian Federal Science Policy Office: IUAP P7/19/ (DYSCO, 'Dynamical systems, control and optimization', 2012-2017); EU: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Advanced Grant: BIOTENSORS (n° 339804). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. AC is a postdoctoral fellow of the Research Foundation Flanders (FWO). MDV is funded by the Wellcome Trust Centre [grant numbers 098461/Z/12/Z] (Sleep, Circadian Rhythms & Neuroscience Institute), the

RCUK Digital Economy Programme [grant number EP/G036861/1] (Oxford Centre for Doctoral Training in Healthcare Innovation.

DISCLOSURE:

None of the authors has any conflict of interest to disclose.

We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

REFERENCES :

1. Ben-Ari Y, Holmes GL. Effects of seizures on developmental processes in the immature brain. *Lancet Neurol* 2006;5:1055-1063.
2. Jensen FE. Developmental factors in the pathogenesis of neonatal seizures. *J Pediatr Neurol* 2009;7:5-12.
3. Holmes GL. The long-term effects of neonatal seizures. *Clin Perinatol* 2009;36:901-914, vii-viii.
4. Ramantani G. Neonatal epilepsy and underlying aetiology: to what extent do seizures and EEG abnormalities influence outcome? *Epileptic Disord* 2013;15:365-375.
5. Miller SP, Weiss J, Barnwell A, et al. Seizure-associated brain injury in term newborns with perinatal asphyxia. *Neurology* 2002;58:542-548.
6. Shah DK, Wusthoff CJ, Clarke P, et al. Electrographic seizures are associated with brain injury in newborns undergoing therapeutic hypothermia. *Arch Dis Child Fetal Neonatal Ed* 2014;99:F219-224.
7. van Rooij LG, de Vries LS, Handryastuti S, et al. Neurodevelopmental outcome in term infants with status epilepticus detected with amplitude-integrated electroencephalography. *Pediatrics* 2007;120:e354-363.
8. Pisani F, Cerminara C, Fusco C, et al. Neonatal status epilepticus vs recurrent neonatal seizures: clinical findings and outcome. *Neurology* 2007;69:2177-2185.
9. Pisani F, Facini C, Pavlidis E, et al. Epilepsy after neonatal seizures: literature review. *Eur J Paediatr Neurol* 2015;19:6-14.
10. Pisani F, Spagnoli C. Neonatal Seizures: A Review of Outcomes and Outcome Predictors. *Neuropediatrics* 2016;47:12-19.
11. Jensen FE. Neonatal seizures: an update on mechanisms and management. *Clin Perinatol* 2009;36:881-900, vii.
12. Murray DM, Boylan GB, Ali I, et al. Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures. *Arch Dis Child Fetal Neonatal Ed* 2008;93:F187-191.
13. Malone A, Ryan CA, Fitzgerald A, et al. Interobserver agreement in neonatal seizure identification. *Epilepsia* 2009;50:2097-2101.
14. Wietstock SO, Bonifacio SL, Sullivan JE, et al. Continuous Video Electroencephalographic (EEG) Monitoring for Electrographic Seizure Diagnosis in Neonates: A Single-Center Study. *J Child Neurol* 2015;31:328-332.
15. Boylan GB, Stevenson NJ, Vanhatalo S. Monitoring neonatal seizures. *Semin Fetal Neonatal Med* 2013;18:202-208.

16. Wusthoff CJ. Diagnosing neonatal seizures and status epilepticus. *J Clin Neurophysiol* 2013;30:115-121.
17. Shellhaas RA. Continuous long-term electroencephalography: the gold standard for neonatal seizure diagnosis. *Semin Fetal Neonatal Med* 2015;20:149-153.
18. Abend NS, Wusthoff CJ, Goldberg EM, et al. Electrographic seizures and status epilepticus in critically ill children and neonates with encephalopathy. *Lancet Neurol* 2013;12:1170-1179.
19. Deburchgraeve W, Cherian PJ, De Vos M, et al. Automated neonatal seizure detection mimicking a human observer reading EEG. *Clin Neurophysiol* 2008;119:2447-2454.
20. De Vos M, Deburchgraeve W, Cherian PJ, et al. Automated artifact removal as preprocessing refines neonatal seizure detection. *Clin Neurophysiol* 2011;122:2345-2354.
21. Navakatikyan MA, Colditz PB, Burke CJ, et al. Seizure detection algorithm for neonates based on wave-sequence analysis. *Clin Neurophysiol* 2006;117:1190-1203.
22. Mitra J, Glover JR, Ktonas PY, et al. A multistage system for the automated detection of epileptic seizures in neonatal electroencephalography. *J Clin Neurophysiol* 2009;26:218-226.
23. Temko A, Thomas E, Marnane W, et al. EEG-based neonatal seizure detection with Support Vector Machines. *Clin Neurophysiol* 2011;122:464-473.
24. Vanhatalo S. Development of neonatal seizure detectors: an elusive target and stretching measuring tapes. *Clin Neurophysiol* 2011;122:435-437.
25. Temko A, Thomas E, Marnane W, et al. Performance assessment for EEG-based neonatal seizure detectors. *Clin Neurophysiol* 2011;122:474-482.
26. Cherian PJ, Deburchgraeve W, Swarte RM, et al. Validation of a new automated neonatal seizure detection system: a clinician's perspective. *Clin Neurophysiol* 2011;122:1490-1499.
27. Mathieson SR, Stevenson NJ, Low E, et al. Validation of an automated seizure detection algorithm for term neonates. *Clin Neurophysiol* 2015.
28. Ansari AH, Cherian PJ, Dereymaeker et al. Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor. *Clin Neurophysiol* (in press 2016).
29. Clancy RR, Legido A. The exact ictal and interictal duration of electroencephalographic neonatal seizures. *Epilepsia* 1987;28:537-541.
30. Smit LS, Vermeulen RJ, Fetter WP, et al. Neonatal seizure monitoring using non-linear EEG analysis. *Neuropediatrics* 2004;35:329-335.
31. Abend NS, Gutierrez-Colina A, Zhao H, et al. Interobserver reproducibility of electroencephalogram interpretation in critically ill children. *J Clin Neurophysiol* 2011;28:15-19.
32. Stevenson NJ, Clancy RR, Vanhatalo S, et al. Interobserver agreement for neonatal seizure detection using multichannel EEG. *Ann Clin Transl Neurol* 2015;2:1002-1011.
33. Shah DK, Mackay MT, Lavery S, et al. Accuracy of bedside electroencephalographic monitoring in comparison with simultaneous continuous conventional electroencephalography for seizure detection in term infants. *Pediatrics* 2008;121:1146-1154.
34. Scher MS, Hamid MY, Steppe DA, et al. Ictal and interictal electrographic seizure durations in preterm and term neonates. *Epilepsia* 1993;34:284-288.
35. J.L. F. The measurement of interrater agreement Statistical Methods for Rates and Proportions, Third Edition; 2004:29.
36. Vento M, de Vries LS, Alberola A, et al. Approach to seizures in the neonatal period: a European perspective. *Acta Paediatr* 2010;99:497-501.

37. Slaughter LA, Patel AD, Slaughter JL. Pharmacological treatment of neonatal seizures: a systematic review. *J Child Neurol* 2013;28:351-364.

38. Piryatinska A, Terdik G, Woyczynski WA, et al. Automated detection of neonate EEG sleep stages. *Comput Methods Programs Biomed* 2009;95:31-46.

39. Wilson SB, Scheuer ML, Plummer C, et al. Seizure detection: correlation of human experts. *Clin Neurophysiol* 2003;114:2156-2164.

40. Olischar M, Davidson AJ, Lee KJ, et al. Effects of morphine and midazolam on sleep-wake cycling in amplitude-integrated electroencephalography in post-surgical neonates >/= 32 weeks of gestational age. *Neonatology* 2012;101:293-300.