Development of neonatal seizure detectors: An elusive target and stretching measuring tapes

Neonatal seizures of epileptic origin (NS) are serious events associated with high mortality and morbidity, and their prompt, effective treatment is currently an issue of high priority (Silverstein et al., 2008). Due to the absence or shortage of neurophysiological services in the neonatal intensive care units (NICU), treatment of neonatal seizures has been often based on clinical observations. Recent studies have, however, convincingly demonstrated that a diagnosis or treatment of NS without EEG confirmation will be erroneous in most cases (Murray et al., 2008; Malone et al., 2009), and would thus lead to over/under diagnosis and over/under treatment. This established knowledge, coupled with the known high incidence of NS (up to 20–40%) in risk babies, has dramatically increased the need for long term EEG monitoring in NICUs. Wide scale EEG monitoring, in turn, has led to a need for the development of technical solutions that enable continuous, reliable evaluation of NS in the absence of live EEG expert consultation (Boylan et al., in press). Recent studies have clearly indicated that the widely used amplitude integrated EEG (aEEG, or CFM; Hellström-Westas et al., 2006) paradigm can detect many clinically unrecognized seizures, but it does miss many kinds of seizures, and may lead to a false NS diagnosis in the hands of inexperienced users (cf. Rennie et al., 2004). Hence, automated seizure detector paradigms have been intensively developed for standard neonatal EEG recordings over the past two decades, and a range of challenges have emerged as a result of this work.

In this issue, Temko and colleagues present a novel technical solution for a well-functioning, automated NS detector (Temko et al., 2011a). In addition, they develop the challenges and potential solutions (Temko et al., 2011b) for how automated NS detectors should be evaluated to improve comparisons between different detectors. As it turns out, the fundamental bottlenecks in this work are now bouncing back from engineers to clinicians who are challenged with the quest of defining what is relevant, both physiologically and clinically.

1. First, the target must be known before shooting

It has been known for decades now that the ictal EEG of NS may exhibit waveforms of markedly different morphology. Current EEG classification of seizures is based on subjective (‘Gestalt perception’) assessment of waveforms using a combination of unspecific and technically unclear descriptors: a clear change in the signal appearance relative to background, repetitive/rhythmic, perceived as epileptiform, any amplitude, and duration $\geq 10$ s. In the absence of any bullet-proof, unequivocal measure, we are left with the typical clinical diagnostic dilemma: dichotomous decisions (yes vs. no) need to be based on sets of vague perceptions. The future development and testing of detector machines should therefore be built on the definition of seizures by a consensus agreement among observers; a group of EEG experts.

It is striking in this context, that to the best of my knowledge (including an international survey among colleagues), there are no studies on interrater agreements in assessing EEG during NSs. Similar studies on other EEG phenomena, such as scoring sleep arousals (Crowell et al., 2002) or EEG in older children (Stronik et al., 2006), have shown that only the most robust features are consistently agreed upon by larger groups of EEGers. Knowing the accuracy of the ground truth (the visual classification; see below) is a prerequisite to truly assess how well automated detectors really work. Current literature is always comparing a given algorithm to a subjective marking of seizures of the given EEGer from the given EEG set. Hence, later assessments of the same algorithm against a different EEG dataset classified by a different EEGer (with unknown agreement with the first EEGer) will often lead to worse performance estimates. The study of Temko et al. (2011b) in this issue therefore makes an important contribution to this area by proposing comparable, and more comprehensive ways to compare detectors against each other.

2. Second, seizure features are many but not equal

Over the past two decades, it has become apparent, that NS detection must be based on an approach that relies on multiple features of the EEG signal (Murray et al., 2008a; Malone et al., 2009; Shellhaas and Clancy, 2007; Vanhatalo, 2007). Other attempts in NS detection have utilized non-EEG measures, such as ECG signal (Greene et al., 2007; Doyle 2010), or ictal video analysis (Karayiannis et al., 2006). It has become clear, however, that the variability in ictal physiological signs makes them tricky to use in NS detection, and that the variability or mere absence of clinical signs in any NS (Murray et al., 2008; Malone et al., 2009) makes video recognition alone fallible (cf. Boylan and Rennie, 2006).

From the perspective of design strategy, NS detectors have been based either (i) on a set of heuristic ideas stemming from how the EEG waveforms are perceived (e.g. Deburghgraev et al., 2008; Navakatikyan et al., 2006; Gotman et al., 1997) or (ii) on ap-
proaches with classifiers that combine sets of non-specific features from the EEG time series (Temko et al., 2011a; Aarabi et al., 2006, 2007; Greene et al., 2008). Both approaches have pros and cons: heuristic design leaves the user with an intuitive understanding of what is measured, but it may develop into a computationally quite intensive solution when it really attempts to mimic human perception (cf. Debruchgraeve et al., 2008). Classifier design, such as the one used in the study of Temko et al. (2011a), is by design generic, and leaves the user/developer little idea about how individual EEG features contribute to the ultimate seizure detection. At the same time, it can be computationally lighter in design, and importantly, development can easily evolve over time by incorporating more "EEG experience".

While single features (e.g. amplitude, autocorrelation, frequency content or synchrony between channels) may nicely characterize many seizure events (Greene et al., 2008; Stevenson et al., 2007), they have all turned out to have a rather low discriminative power between epileptic seizures and other EEG-recorded signals in real life situations, i.e. longer and noisy recordings. Increase in computing power has made it possible to combine larger sets of signal features, but it has led to a new challenge: the number of possible combinations among the tens of independently set parameter values is far beyond that which is feasible to thoroughly optimize with a limited amount of data. Thus, current detector algorithms are bound to have several in-built "guestimates", and the optimality of the detector is assessed by running field tests, i.e. testing the performance on real, clinical EEG datasets.

3. Third, performance figures can describe but also distract

Performance of any detector is measured against a 'ground truth', which in the EEG context is considered to be the visual detection by EEG experts. This is technically unequivocal with a "unanimously marked" trace only (i.e. consider as seizure/non-seizure only those periods where every expert agrees; cf. Palmu et al., 2010). This is based on marking every moment in the trace. It may lead to a substantial amount of unsure trace, which may be detected technically as "false alarms" by the algorithm. The other options would be a "majority vote", or a "consensus decision" giving a dichotomous (yes vs. no seizure) report, but the latter could be influenced by social group dynamics among the EEGers, and it no longer reflects the genuine view of any one expert. Taken together, any solution in generating 'ground truth' datasets will have inherent limitations, and its natural ambiguity (i.e. interrater agreement) should be known before assessing the 'clinically relevant performance/accuracy' of the detector paradigms. Meanwhile, detector performance has been, and will be evaluated by using a locally collected dataset whose credibility is based on the international recognition of the given EEG experts.

Once the comparison data (ground truth) is known, the next challenge is to decide what to measure and how. Here the paper by Temko et al. (2011b) makes a very welcome point by introducing the concepts 'event-based' and 'epoch-basced' metrics. The former is based on the idea that the number of seizures matters, while the latter is based on assessing the duration of seizures. As depicted in the schematic drawing in Fig. 1, these two approaches may give dramatically different figures with the same NS load or may also produce similar figures from dramatically different NS loads. Clinicians at the bedside do implicitly believe that several very short seizures are less harmful than fewer but very long seizures, and the adverse impact of seizure burden (proportion of time the brain exhibits seizure activity) is emphasized in many recent studies (e.g. Murray et al., 2008a,b; Pisani et al., 2007; Shellhaas and Clancy, 2007).

In addition, Temko et al. (2011b) also presents receiver–operator curves (ROC) which are good for showing the trade-off between sensitivity and specificity. This is a particularly useful approach for developers of detectors, and should become a standard of practice in this area. However, at the ultimate user (clinician) level, they may be less useful, because the parameters (i.e. position in the ROC curve) must be set before entering real life praxis.

The traditional way of assessing diagnostic/detection accuracy has been based on sensitivity and specificity values, supplemented by the rate of false alarms per hour (FA/h). Sensitivity and specificity are good measures in diagnostic use, but they are difficult to interpret in the context of seizure detectors.

Clinical decisions made from the information given by automated NS detectors would be to initiate, adjust or stop treatment for NS. Therewith, the actual measures of interest for most clinicians would be to know:

(i) how often are my individual babies misdiagnosed? and
(ii) how reliably does the given detector estimate the 'seizure burden' in a given baby?

![Fig. 1. Illustrates the challenge in detecting and quantifying detections of neonatal seizures. The uppermost time line (1 h) shows the actual seizure events (red bars; ground truth), and five lines below it demonstrate detections by five fictional detectors (D#1–5) of the same seizures. The right hand column shows the performance of each detector: the number of seizures (SZ#), false alarms per hour (FA/h), seizure burden (SZ%), as well as the percentual errors in FA and SZ%. Note that detector performance in these three measures (SZ#, FA/h, SZ%) may be strongly discordant. Hence, a distinct a priori contention for the priority of these measures is needed before a prudent assessment of clinical utility is feasible.](image-url)
In this context, the number of false alarms may not be so critical. There are frequent alarms from the various vital monitors attached to critically ill babies in NICUs anyhow, and staff is accustomed to interpreting false alarms of various kinds. From a clinical perspective, it is difficult to see how an additional one or two false alarms per hour from a seizure detector would cause much trouble. If, and when the NS treatments are based on cumulative seizure burden, the only false alarms of clinical concern would be those of long duration, but they are also easier to recognize technically and visually.

In conclusion, NS detection presents an unusually complex challenge. The clinical need for NS detectors is pressing, and clinical utility is more important than perfection. All producers of commercial EEG monitors are keen to implement such a paradigm into their system, and a validated detector (human or machine) will be imperative for all prospective studies of NS treatments. In this situation, it is of utmost importance that work in this area continues in a transparent manner. The studies of Temko et al. (2011a,b) have taken a number of important steps by implicitly reminding us that credible development and validation of prospective clinical tools can only be done in an open manner: there should be neither commercial protections, nor conflicts of interest in the paradigms, ground truth dataset or performance values.

Such efforts will need both excellent collaborative networks (such as e.g. www.nemo-europe.com) and considerable amounts of goodwill from resourcing bodies to sustain longterm technical development and clinical validation.

References