A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial

Andrea M Pavel, Janet M Rennie, Linda S de Vries, Mats Blennow, Adrienne Foran, Divyen K Shah, Ronit M Pressler, Olga Kapellou, Eugene M Dempsey, Sean R Mathieson, Elena Pavlidis, Alexander C van Huffelen, Vicki Livingstone, Mona C Toet, Lauren C Weeke, Mikael Finder, Subhabrata Mitra, Deirdre M Murray, William P Marnane, Geraldine B Boylan

Summary

Background Despite the availability of continuous conventional electroencephalography (cEEG), accurate diagnosis of neonatal seizures is challenging in clinical practice. Algorithms for decision support in the recognition of neonatal seizures could improve detection. We aimed to assess the diagnostic accuracy of an automated seizure detection algorithm called Algorithm for Neonatal Seizure Recognition (ANSeR).

Methods This multicentre, randomised, two-arm, parallel, controlled trial was done in eight neonatal centres across Ireland, the Netherlands, Sweden, and the UK. Neonates with a corrected gestational age between 36 and 44 weeks and, with or at significant risk of, seizures requiring EEG monitoring, received cEEG plus ANSeR linked to the EEG monitor displaying a seizure probability trend in real time (algorithm group) or cEEG monitoring alone (non-algorithm group). The primary outcome was diagnostic accuracy (sensitivity, specificity, and false detection rate) of health-care professionals to identify neonates with electrographic seizures and seizure hours with and without the support of the ANSeR algorithm. Neonates with data on the outcome of interest were included in the analysis. This study is registered with ClinicalTrials.gov, NCT02431780.

Findings Between Feb 13, 2015, and Feb 7, 2017, 132 neonates were randomly assigned to the algorithm group and 132 to the non-algorithm group. Six neonates were excluded (four from the algorithm group and two from the non-algorithm group). Electrographic seizures were present in 32 (25·0%) of 128 neonates in the algorithm group and 38 (29·2%) of 130 neonates in the non-algorithm group. For recognition of neonates with electrographic seizures, sensitivity was 81·3% (95% CI 66·7–93·3) in the algorithm group and 89·5% (78·4–97·5) in the non-algorithm group; specificity was 84·4% (95% CI 76·9–91·0) in the algorithm group and 89·1% (82·5–94·7) in the non-algorithm group; and the false detection rate was 36·6% (95% CI 22·7–52·1) in the algorithm group and 22·7% (11·6–35·9) in the non-algorithm group. We identified 659 h in which seizures occurred (seizure hours): 268 h in the algorithm versus 391 h in the non-algorithm group. The percentage of seizure hours correctly identified was higher in the algorithm group than in the non-algorithm group (177 [45·3%; 34·5–58·3] of 391 h; difference 20·8% [3·6–37·1]). No significant differences were seen in the percentage of neonates with seizures given at least one inappropriate antiseizure medication (37·5% [95% CI 25·0 to 56·3] vs 31·6% [21·1 to 47·4]; difference 5·9% [–14·0 to 26·3]).

Interpretation ANSeR, a machine-learning algorithm, is safe and able to accurately detect neonatal seizures. Although the algorithm did not enhance identification of individual neonates with seizures beyond conventional EEG, recognition of seizure hours was improved with use of ANSeR. The benefit might be greater in less experienced centres, but further study is required.

Funding Wellcome Trust, Science Foundation Ireland, and Nihon Kohden.

Copyright © 2020 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

Introduction

Newborn infants can exhibit a range of unusual repetitive movements, not all of which are seizures.1 Recognition of seizures is vital because they are often a sign of an underlying neurological condition such as hypoxic ischaemic encephalopathy, stroke, or meningitis,2 and because treatment for non-seizure events exposes infants to unnecessary harmful drugs.3 The diagnosis of neonatal seizures is challenging for clinicians because most neonatal seizures are electrographic only, clinical signs can become uncoupled after medication, and, even when present, clinical signs can be subtle and hard to distinguish from the normal repertoire of neonatal movements.4–8 Amplitude-integrated electroencephalography (aEEG) is often used by neonatologists for seizure detection, but limitations have been reported.9 Continuous conventional EEG (cEEG) monitoring is the gold standard for the diagnosis of all seizures.10 Evidence suggests that regardless of the underlying cause, seizures themselves have a negative effect on neurodevelopment, adding to the importance of early recognition and treatment.11-16 Despite the availability of cEEG in some neonatal intensive care
Evidence before this study
During development of our machine-learning algorithm (Algorithm for Neonatal Seizure Recognition [ANSeR]) for neonatal seizure detection, we did a systematic review of the scientific literature. We searched PubMed for research articles published in English from inception to Oct 24, 2013, using the following search terms: “automated seizure detection algorithms (SDA) vs gold standard” (221 articles found, of which 27 articles were included), “automated SDA safety” (221 articles found, of which one article was included), and “amplitude-integrated EEG vs continuous EEG – interobserver agreement” (87 articles found, of which nine articles were included). We identified six different research groups that assessed the performance of different seizure detection algorithms (SDAs) using at least 100 h of electroencephalography (EEG) monitoring from at least ten neonates. All groups reported performance results on post-acquisition EEG recordings and not in real time in a clinical setting, except for one. This study compared the clinical care of two cohorts of neonates: one cohort using continuous amplitude EEG with SDA output and one cohort of neonates who were clinically monitored with 1 h of conventional EEG monitoring. The SDA detected 55% of seizures but, most importantly, they found that neonates in the SDA cohort did not receive additional doses of antiseizure medication, suggesting that it is safe to use in a real-time clinical setting. The same search was done again on April 25, 2020, to include the 2013–20 period, and we found 297 articles. During this period, only one recent study used a SDA in a neonatal clinical setting. The aim of this study was to assess the feasibility of a monitoring infrastructure proposed for an antiseizure drug trial. The health-care professionals interviewed found the SDA to be useful for seizure detection but with a high rate of false detections. We did not find any clinical trials assessing the impact of a SDA on neonatal seizure recognition by health-care professionals in real time in the neonatal unit.

Added value of this study
To the best of our knowledge, the current study is the first randomised, multicentre clinical investigation to assess the clinical impact of a machine-learning algorithm in real time on neonatal seizure recognition in a clinical setting. Although it did not reach our predefined target, using the ANSeR algorithm as a support tool increased the percentage of seizures correctly detected. However, in a post-hoc analysis of the difference between weekdays and weekends, the predefined target was exceeded. This finding might be explained by increased neuropsychology expertise available during the week in the participating neonatal centres. We also demonstrated that the use of ANSeR for seizure recognition was safe and did not result in an increased use of antiseizure medication.

Implications of all the available evidence
Our results support the potential benefits and safety for the use of a real-time neonatal seizure detection algorithm. We demonstrated an increase in the percentage of seizures recognised using our algorithm in centres that already had good neuropsychology support or had neonatologists who were comfortable with EEG interpretation. We suggest that the impact might be greater in less experienced neonatal units, although further studies are needed.

Methods
Study design and participants
This multicentre, randomised, two-arm, parallel, controlled study was done in eight NICUs across Ireland, the Netherlands, Sweden, and the UK. All neonates between the corrected gestational age of 36 and 44 weeks who were admitted to the NICUs of recruiting hospitals and required EEG monitoring because they had clinically suspected seizures or who were at high risk of seizures were screened for eligibility, and parents or guardians were approached for consent. If written informed consent was not obtained from at least one parent or guardian, the neonate was not included in the investigation.

The trial was a regulated clinical investigation of a medical device. The clinical investigation plan was...
approved by national competent authorities and local ethics committees of participating centres and adhered to all applicable local and national regulations.

Randomisation and masking
Eligible neonates were randomly assigned (1:1) to receive cEEG monitoring with the aid of ANSeR (algorithm group) or routine cEEG monitoring alone (non-algorithm group), which is considered to be standard of care in the participating NICUs. Block randomisation (with block sizes of two or four), stratified by recruiting hospital, was used to allocate neonates to each group. The randomisation list was generated by a biostatistician using the ralloc procedure in Stata and incorporated into the central web-based electronic system used for allocation. Because this study was an investigation of a medical device, the research personnel, clinical team, and neonates’ families were all aware of group allocation. The neurophysiologists who reviewed the EEGs for post-acquisition seizure annotation and the biostatistician who did the statistical analysis were masked to the group allocation.

Procedures
For the intervention group, the ANSeR software system ran on a bedside laptop linked to the Nihon Kohden Neurofax monitor (EEG-1200, Tokyo, Japan) and displayed the seizure probability trend in real time. An audible alarm sounded when a predefined probability threshold was breached (0.5), and a red marker was visible on the aEEG display when a possible seizure was detected by the algorithm. Training for the operation of the ANSeR software system was provided to all personnel involved. Clinical management, including interventions and treatments, of all included neonates was otherwise provided as per standard clinical practice of the recruiting hospitals as the study protocol did not include any instructions regarding seizure treatment.

Neonates in the non-algorithm group were monitored with cEEG using Nihon Kohden Neurofax (EEG-1200), Neurofax monitor (EEG-1200, Tokyo, Japan) and respiration monitoring and synchronised with the EEG recording. EEG was recorded for a minimum of 2 h and up to 100 h (to include the rewarming period for newborn infants receiving therapeutic hypothermia), or longer if clinically indicated, but the use of the algorithm was only evaluated for the first 100 h of monitoring.

cEEG recordings were reviewed in their entirety, and all EEGs were annotated for seizures twice by independent expert neurophysiologists. One expert (SRM) annotated all EEGs and the second annotation was provided by one of the other experts from the group (GBB, RMP, ACvH, and EP). All expert reviewers adhered to a review protocol specifying reviewing parameters (including montage, sensitivity, and time base) and limiting review periods to prevent fatigue. A neonatal electrographic seizure has been defined as at least 10 s of evolving, sudden, and repetitive stereotyped waveforms on at least one EEG channel. However, this definition is arbitrary, and we have previously shown that there is poor agreement between experts in seizures with a duration of less than 30 s. Therefore, to strengthen the experts’ annotation as a gold standard, in this investigation an electrographic seizure was confirmed if there was an overlap in annotation of 30 s between two expert reviewers.

Periods during which the two expert annotations overlapped were used to produce a final annotation (ie, the gold standard for seizure detection). For each neonate with seizures, summary measures of seizure burden—the total seizure burden in minutes (total accumulated seizure duration in the entire recording), maximum hourly seizure burden (the total seizure burden in the hour with the maximum seizure activity expressed in min/h), and median seizure duration in seconds—were calculated using the final annotations.

For each neonate, the gold standard was the final annotation based on two experts. A neonate was confirmed as having electrographic seizures (neonate with seizure) if there was at least one seizure with an overlap of 30 s between the two expert annotations. A seizure hour was confirmed if there was at least one confirmed electrographic seizure within that hour.

Figure 1: Trial profile
EEG=electroencephalography.
the form was annotated accordingly. If no seizures were noted in that hour, the form was marked as “no seizures recorded”. A neonate was considered identified by the clinical team as having seizures (neonate with seizure) if at least 1 h was marked on the seizure record form or at least one therapeutic dose of an antiseizure medication was given during the investigation. A seizure hour was considered to be identified by the clinical team if the seizure record form was marked or if at least 1 h was marked on the seizure record form or if one expert did not annotate any seizures in that hour or if at least one therapeutic dose of an antiseizure medication was given during that hour or the hour immediately after an electrographic seizure and the annotations on the seizure record form, which accounted for any missing documentation on the form.

### Outcomes

For the detection of neonates with seizures, the primary outcomes were sensitivity (percentage of seizure neonates correctly identified by the clinical team), specificity (percentage of non-seizure neonates correctly identified by the clinical team), and false detection rate (percentage of neonates classified as seizure neonates by the clinical team who did not have seizures).
For the detection of seizure hours, the primary outcomes were sensitivity (percentage of seizure hours correctly identified by the clinical team) and false detection rate (percentage of hours classified as seizure hours by the clinical team that were not seizure hours). The secondary outcomes were summary measures of seizure burden (total seizure burden, maximum hourly seizure burden, and median seizure duration) and number of inappropriate antiseizure medications given. Administration of an antiseizure medication was considered to be inappropriate if an antiseizure medication was given with no confirmed electrographic seizure in that hour or the hour before administration (which allows for time to prepare and administer the antiseizure medication). The post-hoc outcomes were sensitivity (percentage of seizure hours correctly identified by the clinical team) based on day of the week (weekdays, from Monday to Friday, and weekends, Saturday and Sunday) and time of the day (day shift, from 0800 h to 2000 h, and night shift, from 2000 h to 0800 h).

Statistical analysis
To demonstrate superiority of the intervention group in terms of true detections, a sample size of 33 neonates with seizures per group was necessary to detect an absolute difference of 25% in mean sensitivity between groups, assuming an SD of 35%, a power of 80%, a level of significance of 5%, and a two-tailed test. Assuming that 40% of the neonates who were monitored would have confirmed electrographic seizures, we estimated that 83 neonates per group (total n=166) would be required. The seizure status of each neonate was only confirmed after the evaluation of their EEG by a neurophysiologist. After 50% of the planned sample size was recruited, their EEGs were annotated and 25% had confirmed electrographic seizures. On the basis of this finding, the sample size requirement was increased to 132 neonates per group (total n=264) to account for the difference in estimated and actual proportion of neonates with seizures.

We describe continuous variables using median (IQR) and categorical variables using frequency (%). For each group (algorithm and non-algorithm), we calculated estimates of the primary and secondary outcomes and their corresponding 95% CIs. We also calculated differences between the groups (95% CIs) for each outcome. To account for stratified randomisation by hospital and the within-infant clustering of infant hours (for the detection of seizure hours), we calculated bias-corrected bootstrap 95% CIs (based on 100 000 iterations). For bootstrapping, neonates were divided into 16 clusters on the basis of their group allocation and the hospital to which they were admitted. For each iteration, we generated a bootstrap sample of neonates from each cluster (using simple random sampling with replacement), combined the bootstrap samples, and calculated the outcome for each group (algorithm and non-algorithm) and the

| Overall | Algorithm group | Non-algorithm group | Difference in
sensitivities (95% CIs*) |
|---------|-----------------|---------------------|---------------------------|
|         | Number of neonates with seizures | Number of seizure hours identified | Sensitivity† (95% CIs) | Number of neonates with seizures | Number of seizure hours identified | Sensitivity† (95% CIs) |         |
| Overall | 32 | 268 | 177 | 66.0% (53.8–77.3) | 38 | 391 | 177 | 45.3% (34.5–58.3) | 20.8% (3.6–37.1) |
| Day of the week | 27 | 197 | 125 | 63.5% (57.4–74.1) | 35 | 331 | 155 | 46.8% (35.9–59.6) | 16.6% (0.1–32.3) |
| Weekend (Saturday-Sunday) | 13 | 71 | 52 | 73.2% (44.7–87.3) | 10 | 60 | 22 | 36.7% (20.7–54.2) | 36.6% (4.4–64.3) |
| Time of the day | 25 | 128 | 91 | 71.1% (53.8–85.5) | 31 | 191 | 92 | 48.2% (38.6–59.6) | 22.9% (3.2–41.2) |
| Day shift (0800 h-2000 h) | 27 | 140 | 86 | 61.4% (49.6–72.7) | 31 | 200 | 85 | 42.5% (30.3–56.3) | 18.9% (1.1–35.9) |

*Bias-corrected. †Percentage of seizure hours correctly classified as seizure hours by the clinical team.

Table 3: Comparison of detection of seizure hours between algorithm and non-algorithm groups

![Figure 2: Percentage of seizure hours detected, by total seizure burden within the hour](image-url)

(day shift, from 0800 h to 2000 h, and night shift, from 2000 h to 0800 h).
difference in outcomes between the two groups. We calculated point estimates from the original data. We did post-hoc comparisons of detection of seizure hours between the two groups based on day of the week (weekdays, from Monday to Friday, and weekend, Saturday and Sunday) and time of the day (day shift, from 0800 h to 2000 h, and night shift, from 2000 h to 0800 h) using logistic regression models with an interaction term. For each outcome, neonates were analysed according to their randomisation group and neonates were excluded if they had missing data on that outcome. We did statistical analyses using Stata (version 15.0).

This study is registered with ClinicalTrials.gov, NCT02431780.

Role of the funding source
The funders of the study had no role in study design, data collection, data analysis, data interpretation, or the writing of the report. The corresponding author had full access to all the data and had final responsibility for the decision to submit for publication.

Results
Between Feb 13, 2015, and Feb 7, 2017, 132 neonates were randomly assigned to the algorithm group and 132 to the non-algorithm group. Six neonates (four from the algorithm group and two from the non-algorithm group) were excluded from the analysis (figure 1). Hence, 258 neonates (128 in the algorithm group and 130 in the non-algorithm group) were included in the study analysis. Neonates in both groups were similar in terms of clinical characteristics and EEG monitoring (table 1).

The percentage of neonates with electrographic seizures was similar in both groups (32 [25·0%] of 128 in the algorithm group and 38 [29·2%] of 130 in the non-algorithm group; table 2). The primary outcome of measures of diagnostic accuracy (sensitivity, specificity, and false detection rate) for recognition of a neonate with seizures were not significantly different between the two groups (table 2). Sensitivity was 81·3% (95% CI 66·7–93·3) in the algorithm group and 89·5% (78·4–97·5) in the non-algorithm group; specificity was 84·4% (95% CI 76·9–91·0) in the algorithm group and 89·1% (82·5–94·7) in the non-algorithm group; and the false detection rate was 36·6% (95% CI 22·7–52·1) in the algorithm group and 22·7% (11·6–35·9) in the non-algorithm group. In the algorithm group, all six neonates with seizures who were not identified by the clinical team had a total seizure burden of 40 min or less. In the non-algorithm group, three of the four neonates with seizures not identified had a total seizure burden of 40 min or less.

Overall, there were 659 h in which confirmed EEG seizures occurred (268 h in the algorithm group and 391 h in the non-algorithm group). The percentage of seizure hours identified was significantly higher in the algorithm group (177 [66·0%]; 95% CI 53·8–77·3) of 268 h in the algorithm group vs 177 [45·3%]; 34·5–58·3] of 391 h in the non-algorithm group; difference 20·8% [3·6–37·1]; table 3). For both groups, identification of seizure hours increased with increasing total seizure burden within the hour (figure 2).

Seven neonates from the algorithm group and six neonates from the non-algorithm group had no seizure hour correctly identified (figure 3). 177 (38·4%) of 461 of the seizure hours detected on the seizure record form were not marked as seizure hours by the EEG experts, indicating that, although seizures were
suspected by the clinical team, no electrographic evidence of seizures was observed (false detection). The false detection rate on the seizure record form did not differ between the groups (97 [39·3%] of 247 h in the algorithm group vs 80 [37·4%] of 214 h in the non-algorithm group; difference 1·9% [95% CI –14·0 to 18·6]).

No significant differences were found between the groups regarding the secondary outcomes of seizure characteristics (total seizure burden, maximum hourly seizure burden, and median seizure duration) and percentage of neonates with seizures given at least one inappropriate antiseizure medication (10·4% [95% CI 6·3 to 17·7] vs 4·3% [2·2 to 9·8]; difference 5·9% [–14·0 to 26·3]; table 4).

In the post-hoc comparison of detection of seizure hours between the two groups based on day of the week (weekday vs weekend), the interaction term was significant (p=0·038). Differences in the recognition of seizure hours between the algorithm group and non-algorithm group was greater at weekends than weekdays (table 3). For the time of the day comparison, the interaction term was not significant (p=0·535), indicating that differences in the recognition of seizure hours between the two groups did not depend on the time of the day.

Discussion

To the best of our knowledge, this multicentre, randomised investigation is the first to clinically evaluate a neonatal seizure detection algorithm in real time at the cot side in the NICU. We used a large dataset (13 827 h of cEEG) in a term neonatal population, including a proportion of neonates with seizures of various causes. Although the algorithm did not enhance the identification of individual neonates with seizures, the recognition of seizure hours increased significantly when the ANSeR algorithm was used as a support for seizure identification (66·0% in the algorithm group vs 45·3% in the non-algorithm group), and this increase was greater at weekends than during weekdays.

Over the past 20 years, several research teams have developed and validated seizure detection algorithms for the neonatal population,21,22,23,24,33,34 and these studies have been discussed in a review.35 ANSeR has a variable sensitivity threshold that can be set before use, giving the trade-off between seizure detection and false alarms. In a previous study,36 we determined that a threshold range between 0·5 and 0·3 was suitable for clinical use, giving a seizure detection range of 52·6–75·0% and false alarm rate range of 0·04–0·36 false alarms per h. In this study, we chose a fixed threshold of 0·5 to prioritise a very low false alarm rate on the basis that more frequent false detections might degrade confidence in the algorithm and increase the likelihood that the alarm was silenced for further detection, negating its primary function to alert staff at the time of a potential seizure.

Lawrence and colleagues37 used the Recognise algorithm in a prospective, randomised pilot study to assess the feasibility and clinical impact of continuous aEEG in a NICU environment. Neonates from a single centre were randomly assigned to a blinded group in which aEEG and the algorithm were not visible to the clinical team (20 neonates) or a visible group in which clinicians could interpret the aEEG recording supported by the algorithm (20 neonates). Of the 25 neonates with seizures, 12 were recognised by the clinical team (seven in the visible group and five in the blinded group). The real-time seizure detection rate of the algorithm was 55% (615 of 1116 seizures), with an increase to 73% for seizures longer than 30 s and a false alarm rate of 0·09 false alarms per h. 34 neonates had conventional EEG together with aEEG detecting 426 seizures (in ten neonates), of which 323 (76%) seizures were detected by the aEEG and 103 (24%) seizures were missed due to the limited EEG montage. Although the study37 used a seizure detection algorithm in a live setting, it did not constitute a clinical trial of the algorithm (and this was not the intention), as output of both the algorithm and EEG were unavailable to the blinded group. The other commercially available seizure detection algorithm, the Gotman algorithm, was tested on post-acquisition EEG recordings and not in real time in a clinical setting.38

The current multicentre, randomised investigation was powered to detect 25% superiority of the ANSeR algorithm in terms of true seizure detection, including neonates at risk of seizures from all causes. All of our recruiting hospitals routinely use aEEG, and some hospitals were also familiar with conventional EEG.

| Table 4: Secondary outcomes |
|---|
| **Algorithm group** | **Non-algorithm group** | **Difference (95% CI)** |
| **Neonates with seizures** | | |
| Number of neonates with seizures | 32 | 38 | |
| Seizure characteristics | | |
| Total seizure burden, min | 32·3 (17·0 to 45·1) | 46·8 (20·6 to 63·5) | –14·5 (–37·8 to 15·9) |
| Maximum seizure burden, min/h | 15·6 (11·0 to 23·5) | 12·9 (9·1 to 20·6) | 0·7 (–7·4 to 9·8) |
| Median seizure duration, s | 102 (91 to 182) | 108 (77 to 184) | –6 (–84 to 54) |
| At least one inappropriate antiseizure medication | 12 | 12 | |
| Percentage (95% CI) | 37·5% (25·0 to 56·3) | 31·6% (21·1 to 47·4) | 5·9% (–14·0 to 26·3) |
| **Neonates without seizures** | | |
| Number of neonates without seizures | 96 | 92 | 12 |
| At least one inappropriate antiseizure medication | 10 | 4 | 6 |
| Percentage (95% CI) | 10·4% (6·3 to 17·7) | 4·3% (2·2 to 9·8) | 6·1% (0·3 to 13·5) |

Data are median (bias-corrected 95% CI) unless otherwise stated. *Bias-corrected. †In the algorithm group, nine infants were given one inappropriate antiseizure medication; in the non-algorithm group, four infants were given one inappropriate antiseizure medication. ‡In the algorithm group, nine infants were given one inappropriate antiseizure medication and one was given two inappropriate antiseizure medications, one was given three inappropriate antiseizure medications. In the algorithm group, eight infants were given one inappropriate antiseizure medication, three were given two inappropriate antiseizure medications, one was given three inappropriate antiseizure medications. In the algorithm group, two infants were given one inappropriate antiseizure medication and one was given two inappropriate antiseizure medications; in the non-algorithm group, four infants were given one inappropriate antiseizure medication.
Regardless of previous experience, all clinical teams received training for the interpretation of the aEEG and cEEG, as well as the ANSeR algorithm. By using multichannel EEG for our algorithm, we increased seizure detection compared with previous reports of limited two-channel EEG monitoring.22 The current study showed no difference between the groups in diagnosis of a neonate with seizures, but this finding might be explained by the fact that all recruiting hospitals were experienced in EEG monitoring and interpretation.

Ten neonates with seizures were not identified by the clinical team: six in the algorithm group and four in the non-algorithm group. Of the six neonates in the algorithm group, the algorithm did not alarm for two. Both neonates were diagnosed with moderate hypoxic ischaemic encephalopathy, with a suppressed EEG background and low amplitude, localised central seizures, and a low total seizure burden (<40 min). The other four neonates were probably missed as a result of short seizure durations.

The percentage of seizure hours recognised by the clinical team was higher in the algorithm group than in the non-algorithm group. Although the absolute difference in sensitivities between groups (20·8% [95% CI 3·6–37·1]) was below the set threshold of 25%, this finding is important considering the association between seizure burden and adverse long-term outcomes.12–16 In addition to the ten neonates with seizures not identified, three others had no seizure hours correctly identified by the clinical team: one in the algorithm group (with seizures detected by the algorithm) and two in the non-algorithm group.

Consistent with the report previously discussed by Lawrence and colleagues,22 the seizure hour detection rate by the clinical staff increased with an increase in the seizure burden for both groups but remained superior in the algorithm group. Differences in hospital staffing between day and night shifts and between weekdays and weekends are documented in the literature.33–35 Therefore, we wished to examine the effect of time of day and day of the week on the performance of the algorithm in a post-hoc analysis. Although no difference was noted between day and night shifts, a significant difference was observed between weekdays and weekends, with seizures being less likely to be recognised during weekends without the support of the algorithm. This finding might be more reflective of the situation in NICUs with less EEG expertise than in NICUs with more experience in interpreting EEG and in which no EEG expertise exists readily. We found no significant difference between groups in terms of seizure burden and inappropriate use of antiseizure medication, indicating that the algorithm did not result in infants receiving unnecessary antiseizure medication, supporting our conservative sensitivity cutoff to limit false alarms.

Our investigation has some limitations. We analysed all seizures that had an overlap of 30 s between the two experts, which were used as the gold standard for seizure diagnosis. In doing so, we excluded seizures with a duration of less than 30 s from both groups, for which agreement between experts is poor.15 The analysis was done using seizure hour instead of looking at each individual seizure. This decision was pragmatic, as it would not have been feasible to ask hospital staff to record every single suspected seizure lasting 10 s or more. To ensure that the clinical investigation plan was acceptable to all NICU personnel, the clinical team was asked to mark the hour, rather than mark specific times for onset and duration of seizures. Because NICU environments can be very busy, we are aware that missing documentation does not necessarily mean that seizures were not recognised. To account for this, we also considered that seizures had been recognised if an antiseizure medication was given during that hour or the hour immediately after an electrographic seizure.

Attempts have been made in a growing number of clinical conditions to use large datasets to aid categorisation of patient phenotypes and to assign outcome risk or diagnosis.33–35 This current investigation is one of the first to move beyond a proof-of-concept into a real-time clinical investigation. Many years of signal processing analysis, using large datasets of neonatal EEG, allowed us to develop a reliable neonatal seizure algorithm.20–21,40 Newly developed machine-learning techniques have allowed us to make rapid developments in the reliability of the algorithm and to bring this technology to the cot side.26–40 We have shown that machine-learning techniques can be successfully and safely implemented into the clinical care of vulnerable neonates. We hope that this progress will encourage researchers in other fields of neonatal care to consider these techniques to solve real clinical problems.

In conclusion, this clinical investigation was the first to assess the performance of a machine-learning algorithm for neonatal seizure detection in real time and in the real-world setting of busy NICUs throughout Europe. Although all participating hospitals were experienced in neonatal EEG and the clinical teams were generally comfortable in interpreting the aEEG or cEEG, the support provided by the ANSeR algorithm still had a considerable effect on the seizure recognition rate. Our experience suggests that the benefit provided by the ANSeR algorithm might be greater if it was made available to centres with less experience of interpreting neonatal EEG at the cot side, but further research is required. The use of cEEG monitoring has increased considerably as therapeutic hypothermia has become standard practice, which has driven the need for accurate and timely interpretation. Many guidelines that seek to identify babies who might be suitable for therapeutic hypothermia recommend using EEG criteria including seizures, making accurate interpretation imperative. Future work on the probability setting of the algorithm and personalising it for each baby will probably improve performance.
We especially thank the families of all neonates included in this trial. University College Cork for sponsoring this clinical investigation. University College Cork, Cork, Ireland), and also the clinical teams at all the engineers involved in the development of the algorithm, particularly Karolinska University Hospital, Stockholm, Sweden). We thank all of Foundation Trust, London, UK), Jessica Colby-Milley (Rotunda Hospital, Nicola Openshaw-Lawrence (Homerton University Hospital NHS Trust), London, UK), Jean Conway, Taragh Kiely, and Denis Dwyer (INFANT Research Foundation Trust), London, UK), and have a patent A Method of Analysing an long-term neonatal electroencephalography monitoring: a multicentre experience. Arch Dis Child Fetal Neonatal Ed 2019; 104: F491–501. 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: F219–24. Glass HC, Glidden D, Jeremy R, Barkovich AJ, Ferriero DM, Miller SP. Clinical neonatal seizures are independently associated with outcome in infants at risk for hypoxic-ischemic brain injury. J Pediatr 2009; 155: 118–23. Boylan G, Burgeyone L, Moore C, O’Flaherty B, Rennie J. An international survey of EEG use in the neonatal intensive care unit. Acta Paediatr 2010; 99: 1350–55. Wurrell EC, Armstrong EA, Osman LD, Yager YJ. Prolonged seizures exacerbate perinatal hypoxic-ischemic brain damage. Pediatr Res 2001; 50: 445–54. Mulkey SB, Sweantgen CJ. Advancing neurologic care in the neonatal intensive care unit with a neonatal neurologist. J Child Neurology 2014; 29: 31–15. Clancy RR. The contribution of EEG to the understanding of neonatal seizures. Epilepsia 1996; 37 (suppl 1): S52–59. Gotman J, Flanagan D, Rosenblatt B, Bye A, Mizrahi EM. Evaluation of an automatic seizure detection method for the newborn EEG. Electroencephalogr Clin Neurophysiol 1997; 103: 363–69. Lawrence R, Mathur A, Nguyen The Tich S, Zempel J, Inder T. A pilot study of continuous limited-channel aEEG in term infants with encephalopathy. J Pediatr 2009; 154: 813–41.e1. Liu A, Hahn JS, Heldt GP, Coen RW. Detection of neonatal seizures through computerized EEG analysis. Electroencephalogr Clin Neurophysiol 2002; 12: 30–37. Smit LS, Vermeulen RJ, Fetter WP, Strijers RL, Stam CJ. Neonatal seizure monitoring using non-linear EEG analysis. Neuropediatrics 2004; 35: 329–35. Navakatiyakan MA, Colditz PB, Burke CJ, Inder TE, Richardson J, Williams CE. Seizure detection algorithm for neonates based on wave-sequence analysis. Clin Neurophysiol 2006; 117: 1190–203. Deburchgrave W, Chierian PJ, De Vos M, et al. Automated neonatal seizure detection mimicking a human observer reading EEG. Clin Neurophysiol 2008; 119: 2647–54. Mitta J, Glover J, Kutas P, et al. A multistage system for the automated detection of epileptic seizures in neonatal electroencephalography. J Clin Neurophysiol 2009; 26: 218–26. Sharpe C, Davis SL, Reiner GE, et al. Assessing the feasibility of providing a real-time response to seizures detected with continuous long-term neonatal electroencephalography monitoring.
J Clin Neurophysiol 2019; 36: 9–13.

29 Temko A, Thomas E, Marnane W, Lightbody G, Boylan GB. Performance assessment for EEG-based neonatal seizure detectors. Clin Neurophysiol 2011; 122: 474–82.

30 Temko A, Boylan G, Marnane W, Lightbody G. Robust neonatal EEG seizure detection through adaptive background modeling. Int J Neonatal Syst 2013; 23: 1350018.

31 Mathieson SR, Stevenson NJ, Low E, et al. Validation of an automated seizure detection algorithm for term neonates. Clin Neurophysiol 2016; 127: 156–68.

32 Clancy RR, Legido A. The exact ictal and interictal duration of electroencephalographic neonatal seizures. Epilepsia 1987; 28: 537–41.

33 Stevenson NJ, Clancy RR, Vanhatalo S, Rosén I, Rennie JM, Boylan GB. Interobserver agreement for neonatal seizure detection using multichannel EEG. Ann Clin Transl Neurol 2015; 2: 1062–11.

34 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: 1490–99.

35 Temko A, Lightbody G. Detecting neonatal seizures with computer algorithms. J Clin Neurophysiol 2016; 33: 394–402.

36 de Cordova PB, Phibbs CS, Schmitt SK, Stone PW. Night and day in the VA: associations between night shift staffing, nurse workforce characteristics, and length of stay. Res Nurs Health 2014; 37: 90–97.

37 Pauls LA, Johnson-Palens R, McGready J, Murphy JD, Pronovost PJ, Wu CL. The weekend effect in hospitalized patients: a meta-analysis. J Hosp Med 2017; 12: 760–66.

38 Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018; 15: 20170387.

39 Belle A, Kon MA, Najarian K. Biomedical informatics for computer-aided decision support systems: a survey. ScientificWorldJournal 2013; 2013: 769619.

40 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.