THE LANCET
Digital Health

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: McKee JL, Kaufman MC, Gonzalez AK, et al. Leveraging electronic medical record-embedded standardised electroencephalogram reporting to develop neonatal seizure prediction models: a retrospective cohort study. *Lancet Digit Health* 2023; **5:** e217–26.

McKee 2022 Supplemental Information

Table of Contents

Validation of the EMR algorithm

Our EMR-based algorithm identified 82/86 (95%) of the individuals with HIE from an independent clinical list and identified 68 additional individuals who met criteria for inclusion. The individuals not captured by this algorithm but identified on the clinical list were patients transferred from outside hospitals (n=2) or patients in whom the diagnosis of either HIE or hypothermia was incorrectly omitted from the diagnosis codes (n=2). Individuals incorrectly selected by our algorithm, including those with initial EEGs performed at outside hospitals or who did not actually undergo therapeutic hypothermia, were manually removed (n=70).

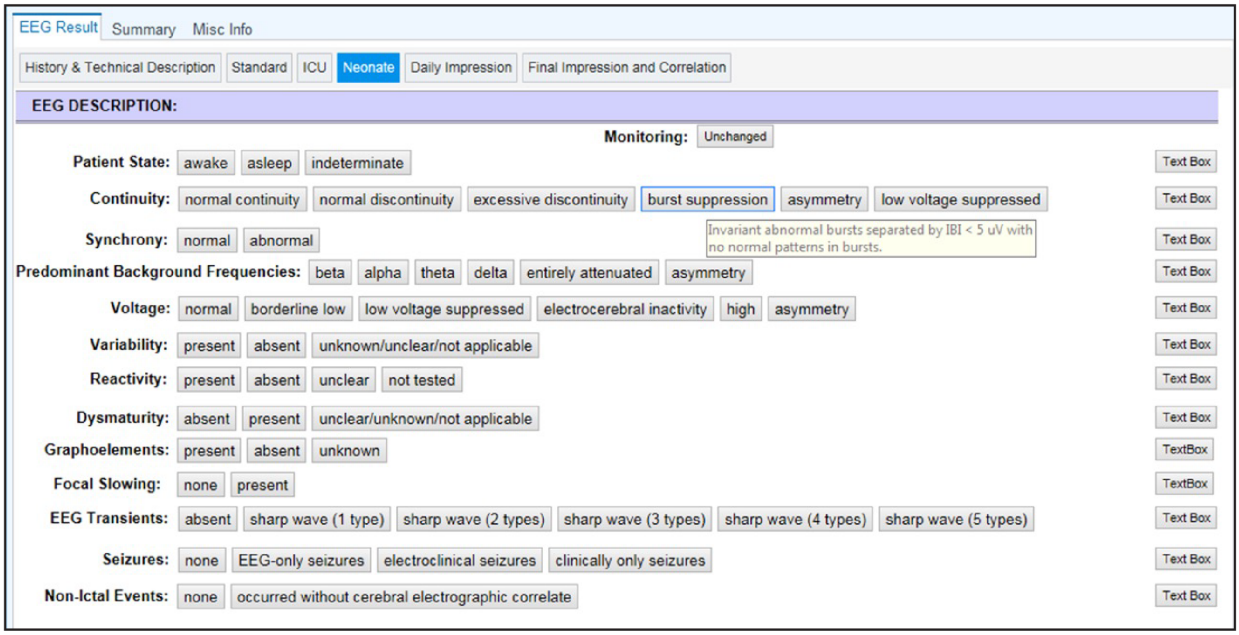

Figure S1: Neonatal EEG Template. A screenshot of the neonatal EEG reporting template is shown. EEG variables are derived from American Clinical Neurophysiology Society standardized EEG terminology, and hover boxes define terms to users.

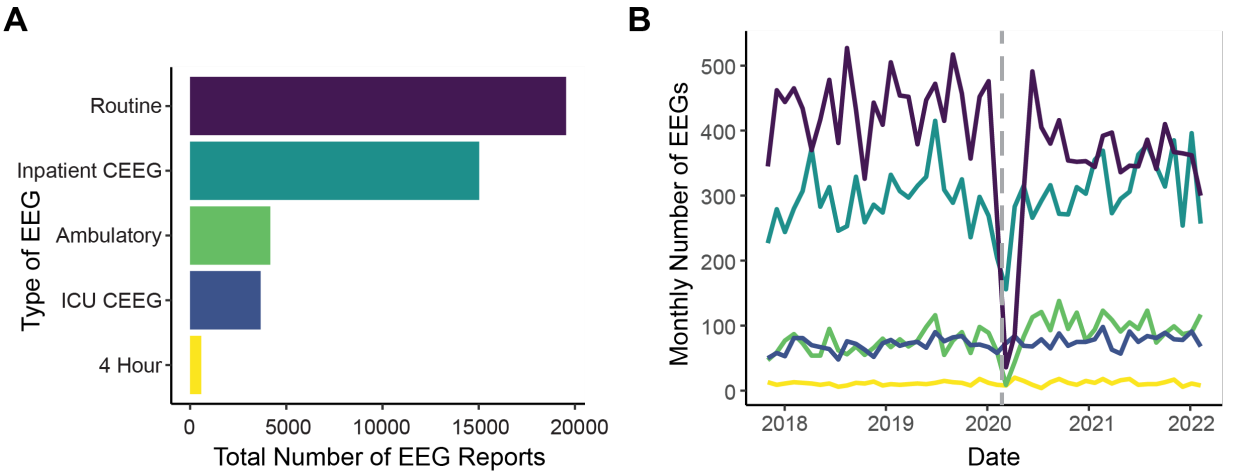

Figure S2: Breakdown of EEG types and accrual overtime. (**A**) >42,000 EEGs have been reported using the templated system, most of these being routine (<1 hour) EEGs or hospital-based long-term monitoring (CEEG). (**B**) Monthly numbers of EEGs reported using the novel system have remained roughly stable, except during the COVID-19 pandemic (dashed vertical line).

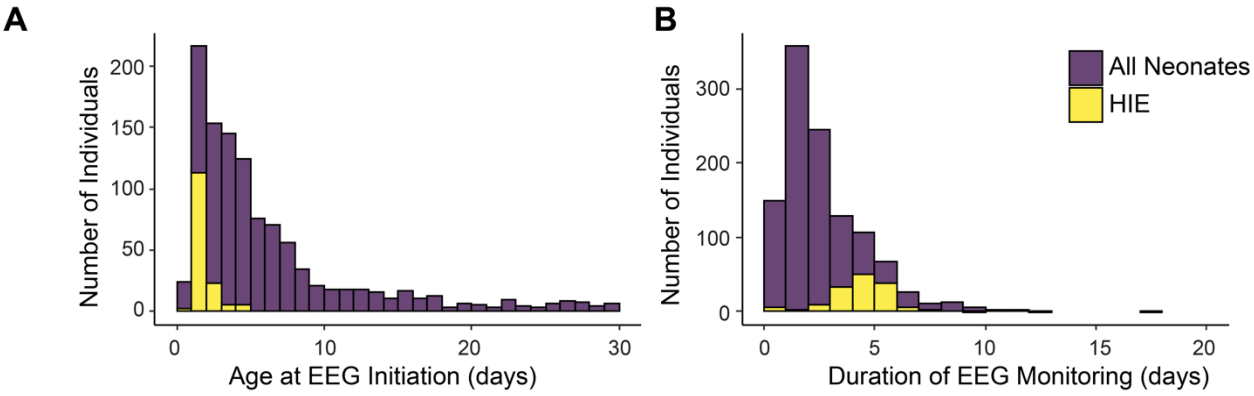

Figure S3: Population characteristics. Distributions of age at EEG initiation (**A**) and duration of continuous EEG monitoring (CEEG) (**B**) for the overall neonatal population (purple) and the subgroup of neonates with hypoxic-ischemic encephalopathy (HIE) (yellow). Neonates with HIE had CEEG initiated at younger ages and underwent longer duration CEEG than the overall cohort (see main text for details).

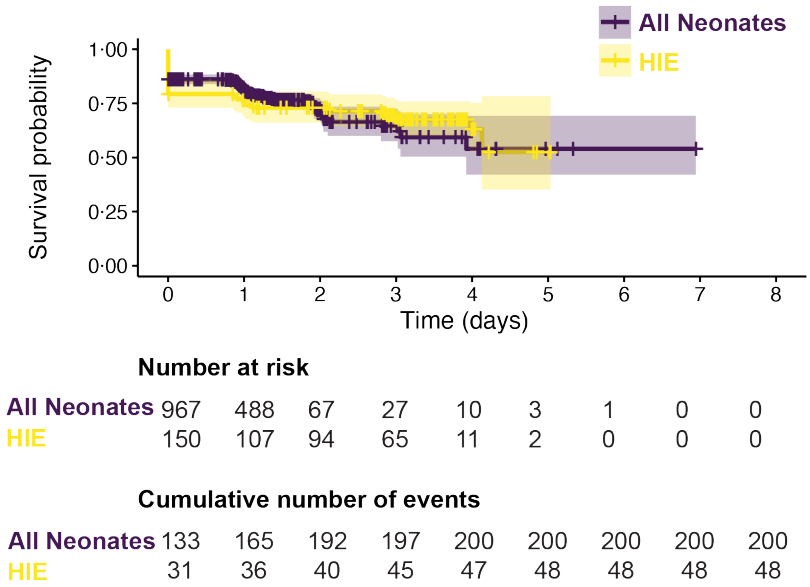

Figure S4: Kaplan-Myer Survival Analysis. Proportion of individuals with seizure-free survival is displayed for both the entire cohort (yellow) and those with HIE (blue). Individuals are censored when monitoring is discontinued (vertical marks).

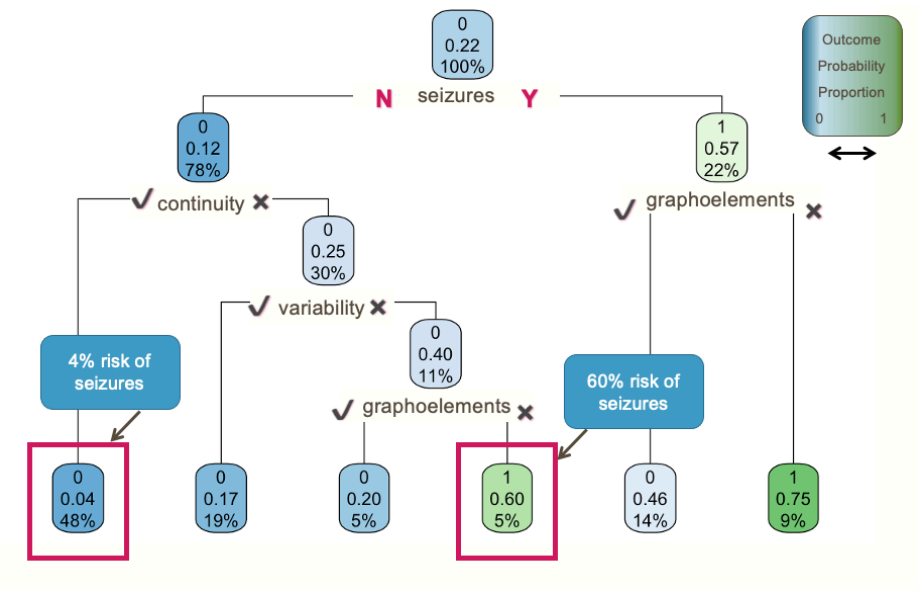

Figure S5: Example decision tree for HIE patients. The model is initiated with all patients and 22% risk of seizures (top). This tree then divides patients based on the presence or absence of seizures on day 1. For example, among patients who did not have seizures on day 1 (on the left), only had a 12% risk of seizures on subsequent days. Next, the model recursively splits the population based on the other features to create the branches, until it reaches a terminal leaf that is either homogenous in outcome or too small to split further. The highlighted group on the left accounts for 48% of the population, but only has a 4% risk of future seizures. However, the highlighted group on the right, while only representing 5% of the population has a 60% chance of future seizures.

Table S1: Model Methods

Key	Model Name	Model Type	Description
LR1	log_regress_caret	Logistic Regression	Default logistic regression model with k-fold cross-validation (k=10) using the caret package in R.
DT	regresstree_caret	Decision Tree	Default decision tree model with cross-validation (k=10) using the caret package in R.
RF1	random_model1	Random Forest	Default random forest model using the randomForest package in R.
RF2	random_model_mtry	Random Forest	Random Forest model, optimized for minimal OOB (out of bag) error using stepwise tuning of mtry (number of variables sampled at each split). Optimal mtry = 7.
RF3	random_model_opt	Random Forest	Random Forest model, optimized for minimal error rate. Parameters tested included mtry (range 1-10, by increments of 1), minimal node size (3-9, by increments of 2) and number of trees (250-500, by increments of 50). Optimal mtry = 4, optimal node size = 9, optimal number of trees = 500.
RF4	range_1	Random Forest	Random forest model using the ranger package in R optimized for OOB error rate. Parameters tested included mtry (range 1-10, by increments of 1), minimal node size (3-9, by increments of 2), sample size (0·55, 0·632, 0·7, 0·8), and number of trees (250-500, by increments of 50). Optimal mtry = 8, optimal node size = 3, optimal sample size = 0·632, optimal number of trees = 250.
RF5	h2o_1	Random Forest	Distributed Random Forest model using the H2O package in R. Parameters tested included mtry (range 1-10, by increments of 1),sample size (0·55, 0·632, 0·70, 0·80), and number of trees (200-500, by increments of 100). The model was optimized towards maximum AUCPR. Optimal mtry = 2, optimal sample size = 0·55, optimal number of trees = 400.
RF6	h2o_balanced	Random Forest	Distributed Random Forest model using the H2O package in R. Parameters tested included mtry (range 1-10, by increments of 1), sample size (0·55, 0·632, 0·70, 0·80), and number of trees (200-500, by increments of 100). In order to create balance, the model was stratified and the class balance default parameter was activated. The model was optimized towards maximum AUCPR. Optimal mtry = 2, optimal sample size = 0.80, optimal number of trees = 200. Cross-validation (k=10) was also implemented within the model.
RF7	h2o_custom_bal	Random Forest	Distributed Random Forest model using the H2O package in R. Parameters tested included mtry (range 1-10, by increments of 1), sample size (0·55, 0·632, 0·70, 0·80), and number of trees (200-500, by increments of 100). In order to create balance, the model was stratified and the class balance parameter was activated with "no subsequent seizures" undersampled at a rate of 0·5 and "subsequent seizures" sampled at a rate of 0·9. The model was optimized towards maximum AUCPR. Optimal mtry = 1, optimal sample size = 0·70, optimal number of trees = 400. Cross-validation (k=10) was also implemented within the model.
RF8	h2o_weighted_0.6068152_2.840491	Random Forest	Distributed Random Forest model using the H2O package in R. Parameters tested included mtry (range 1-10, by increments of 1), sample size (0·55, 0·632, 0·70, 0·80), and number of trees (200-500, by increments of 100). In order to create balance, the model was stratified and weighted in order to proportionally distribute points to "non-subsequent seizure" (0·61) and "subsequent seizure" (2·84) instances. The model was optimized towards maximum AUCPR. Optimal mtry = 1, optimal sample size = 0·55, optimal number of trees = 200. Cross-validation (k=10) was also implemented within the model.
RF9	h2o_weighted_0.5_1.5	Random Forest	Same as above model, aside from weighted metrics for "non-subsequent seizure" (0·5) and "subsequent seizure" (1·5). Optimal mtry = 1, optimal sample size = 0·55, optimal number of trees = 200.
RF10	h2o_weighted_0.5_2	Random Forest	Same as above model, except weights for "non-subsequent seizure" (0·5) and "subsequent seizure" (2·0). Optimal mtry = 1, optimal sample size = 0·55, optimal number of trees = 200.
RF11	h2o_weighted_0.5_3	Random Forest	Same as above model, except weights for "non-subsequent seizure" (0·5) and "subsequent seizure" (3·0). Optimal mtry = 1, optimal sample size = 0·55, optimal number of trees = 200.
RF12	h2o_weighted_0.5_4	Random Forest	Same as above model, except weights for "non-subsequent seizure" (0·5) and "subsequent seizure" (4·0). Optimal mtry = 1, optimal sample size = 0·55, optimal number of trees = 200.
RF13	h2o_weighted_0.5_5	Random Forest	Same as above model, except weights for "non-subsequent seizure" (0·5) and "subsequent seizure" (5·0). Optimal mtry = 1, optimal sample size = 0·55, optimal number of trees = 200.

RF14	h2o_weighted_0.5_10	Random Forest	Same as above model, except weights for "non-subsequent seizure" (0·5) and "subsequent seizure" (10). Optimal mtry = 1, optimal sample size = 0·55, optimal number of trees = 300.
RF15	h2o_weighted_0.5_15	Random Forest	Same as above model, except weights for "non-subsequent seizure" (0·5) and "subsequent seizure" (15). Optimal mtry = 1, optimal sample size = NA, optimal number of trees = 200.
LR2	log_regress_caret_wb	Logisitc Regression	Logistic regression model with cross-validation (k=10) using the caret package in R. Weights were added to proportionally distribute points to "non-subsequent seizure" (0·61) and "subsequent seizure" (2·84).
LR3	log_regress_caret_w3	Logisitc Regression	Same as above model, aside from weighted metrics for "non-subsequent seizure" (0·5) and "subsequent seizure" (3·0)
LR4	log_regress_caret_w5	Logistic Regression	Same as above model, aside from weighted metrics for "non-subsequent seizure" (0·5) and "subsequent seizure" (5·0)

The above table provides a list (not in rank order) of all models tested on our data. Models were chosen from common R machine learning packages, with the goal of moving from more classical statical methods such as logistic regression to more sophisticated machine learning algorithms such as decision trees and random forests. Stronger performing models were expanded and tested with more depth than weaker models, including tests of different weights. We did not intend for this to be an exhaustive search for the "optimal model" but rather aimed to explore how different types of models, and different weights, would affect performance.

Table S2: Day 1 EEG features

	All Neonates		Neonates with HIE			
EEG feature	**Number reported**	**Frequency of abnormalities**	**Number reported**	**Frequency of abnormalities**	**95% CI for OR***	**P-value***
Continuity	1283/1313 (97·7%)	847/1283 (66%)	146/150 (97·3%)	93/146 (63·7%)	0.62, 1.32	0·58
Voltage	1278/1313 (97·3%)	180/1278 (14·1%)	147/150 (98·0%)	42/147 (28·6%)	**1.61, 3.65**	**<0·001**
Variability	1289/1313 (98·2%)	193/1289 (15·0%)	145/150 (96·7%)	40/145 (27·6%)	**1.42, 3.25**	**<0·001**
Reactivity	1247/1313 (95·0%)	231/1247 (18·5%)	141/150 (94·0%)	37/141 (26·2%)	**1.02, 2.37**	**0·032**
Graphoelements	1293/1313 (98·5%)	159/1293 (12·3%)	146/150 (97·3%)	36/146 (24·7%)	**1.50, 3.57**	**<0·001**
Transients	1256/1313 (95·7%)	256/1256 (20·4%)	139/150 (92·7%)	22/139 (15·8%)	0.43, 1.19	0·22
EEG Seizures	1307/1313 (99·5%)	203/1307 (15·5%)	147/150 (98·0%)	31/147 (21·1%)	0.92, 2.25	0·097
EEG Impression	1292/1313 (98·4%)	989/1292 (76·5%)	148/150 (98·7%)	110/148 (74·3%)	0.59, 1.35	0·54

Number reported represents the number of first day EEG reports commenting on the feature of interest. Frequency of abnormalities is the number of first day reports specifying that the feature of interest was abnormal, out of all reports describing the feature. *Fisher's exact test comparing frequency of abnormalities in all neonates to those with hypoxic-ischemic encephalopathy (HIE). 95% confidence intervals for the odds ratio and p-values are provided.