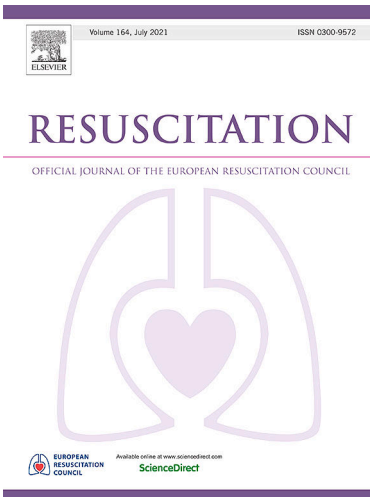
Journal Pre-proofs



Clinical paper

Predicting Neurological Outcome in Comatose Patients after Cardiac Arrest with Multiscale Deep Neural Networks

Wei-Long Zheng, Edilberto Amorim, Jin Jing, Wendong Ge, Shenda Hong, Ona Wu, Mohammad Ghassemi, Jong Woo Lee, Adithya Sivaraju, Trudy Pang, Susan T. Herman, Nicolas Gaspard, Barry J. Ruijter, Jimeng Sun, Marleen C. Tjepkema-Cloostermans, Jeannette Hofmeijer, Michel J. A. M. van Putten, M. Brandon Westover

PII: S0300-9572(21)00441-X
DOI: https://doi.org/10.1016/j.resuscitation.2021.10.034
Reference: RESUS 9260

To appear in: *Resuscitation*

Received Date: 31 August 2021
Revised Date: 19 October 2021
Accepted Date: 20 October 2021

Please cite this article as: W-L. Zheng, E. Amorim, J. Jing, W. Ge, S. Hong, O. Wu, M. Ghassemi, J. Woo Lee, A. Sivaraju, T. Pang, S.T. Herman, N. Gaspard, B.J. Ruijter, J. Sun, M.C. Tjepkema-Cloostermans, J. Hofmeijer, M. J. A. M. van Putten, M. Brandon Westover, Predicting Neurological Outcome in Comatose Patients after Cardiac Arrest with Multiscale Deep Neural Networks, *Resuscitation* (2021), doi: https://doi.org/10.1016/j.resuscitation.2021.10.034

Predicting Neurological Outcome in Comatose Patients after Cardiac Arrest with Multiscale Deep Neural Networks

Wei-Long Zheng, PhD[1]*, Edilberto Amorim, MD[2]*, Jin Jing, PhD[1], Wendong Ge, PhD[1], Shenda Hong, PhD[3], Ona Wu, PhD[4], Mohammad Ghassemi, PhD[5,6], Jong Woo Lee, MD, PhD[7], Adithya Sivaraju, MD[8], Trudy Pang, MD[9], Susan T. Herman, MD[10], Nicolas Gaspard, MD, Ph.D[11], Barry J. Ruijter, MD, PhD[12], Jimeng Sun, PhD[3], Marleen C. Tjepkema-Cloostermans, PhD[13,14], Jeannette Hofmeijer, MD, PhD[12,14], Michel J. A. M. van Putten, MD, PhD[12,13], M. Brandon Westover, MD, PhD[1]

[1] Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[2] Department of Neurology, University of California, San Francisco, San Francisco, CA, USA

[3] Department of Computer Science, University of Illinois at Urbana Champaign, Champaign, IL, USA

[4] Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[5] Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA

[6] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA

[7] Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA

[8] Department of Neurology, Yale School of Medicine, New Haven, CT, USA

[9] Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA

[10] Barrow Neurological Institute, Phoenix, AZ, USA.

[11] Department of Neurology, Université Libre de Bruxelles, Brussels, Belgium

[12] Department of Clinical Neurophysiology, University of Twente, Enschede, the Netherlands

[13] Departments of Neurology and Clinical Neurophysiology, Medisch Spectrum Twente, Enschede, the Netherlands

[14] Department of Neurology, Rijnstate Hospital, Arnhem, the Netherlands

* Contributed equally as co-first authors.

Keywords: Cardiac Arrest; EEG; Neurological Outcome; Deep Learning; Machine Learning

Corresponding Authors:

M. Brandon Westover, M.D. Ph.D.

Department of Neurology

Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114

mwestover@mgh.harvard.edu

Wei-Long Zheng, Ph.D.

Department of Neurology

Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114

weilonglive@gmail.com

Number of words in the Abstract:	247
Number of words in the manuscript:	2983
Number of figures/tables:	4/2

Abstract

Objective:

Electroencephalography (EEG) is an important tool for neurological outcome prediction after cardiac arrest. However, the complexity of continuous EEG data limits timely and accurate interpretation by clinicians. We develop a deep neural network (DNN) model to leverage complex EEG trends for early and accurate assessment of cardiac arrest coma recovery likelihood.

Methods:

We developed a multiscale DNN combining convolutional neural networks (CNN) and recurrent neural networks (long short-term memory [LSTM]) using EEG and demographic information (age, gender, shockable rhythm) from a multicenter cohort of 1,038 cardiac arrest patients. The CNN learns EEG feature representations while the multiscale LSTM captures short-term and long-term EEG dynamics on multiple time scales. Poor outcome is defined as a Cerebral Performance Category (CPC) score of 3-5 and good outcome as CPC score 1-2 at 3-6 months after cardiac arrest. Performance is evaluated using area under the receiver operating characteristic curve (AUC) and calibration error.

Results:

Model performance increased with EEG duration, with AUC increasing from 0.83 (95% Confidence Interval [CI] 0.79-0.87 at 12h to 0.91 (95%CI 0.88-0.93) at 66h. Sensitivity of good and poor outcome prediction was 77% and 75% at a specificity of 90%, respectively. Sensitivity of poor outcome was 50% at a specificity of 99%. Predicted probability was well matched to the observation frequency of poor outcomes, with a calibration error of 0.11 [0.09-0.14].

Conclusions:

These results demonstrate that incorporating EEG evolution over time improves the accuracy of neurologic outcome prediction for patients with coma after cardiac arrest.

Introduction

Coma is a common clinical presentation after successful resuscitation from cardiac arrest (CA) due to transient cerebral hypoxia and ischemia.[1,2] The electroencephalogram (EEG) after CA is highly abnormal, with improvement or deterioration of EEG patterns being predictive of which patients have the potential to eventually recover. Current guidelines recommend starting EEG monitoring early to rule out seizures and to identify signatures predictive of severe brain injury leading to poor outcomes.[3,4] Persistent EEG background suppression, burst suppression with identical bursts, and seizure-like, i.e. ictal interictal activity on a suppressed background predict poor outcomes with low false-positive rates. However, the large volume of continuous EEG data and limited availability of clinicians with expertise in EEG interpretation limit real-time applications of EEG for prognostication and management.

Machine learning (ML) approaches to EEG interpretation have the potential to address this problem. In addition, ML may increase EEG's prognostic value by not only accurately estimating the prognostic significance of individual patterns at a given point in time but also by leveraging the dynamics of continuous EEG data over the course of neurological recovery from coma.[5,6] Most ML studies in CA prognostication to date make predictions based on EEG patterns after a certain time window and do not leverage the evolution of the EEG data over time.[6–10] However, a recent study showed that a time-sensitive model outperformed baseline methods that were time-insensitive[6]. This approach used standard quantitative EEG metrics such as burst-suppression ratio, band power, and spike frequency. However, other potentially relevant characteristics of the EEG signal may be lost when only predefined features are used. We

hypothesized that a data-driven representation of EEG data might allow superior prognostication performance compared to approaches using hand-crafted features.

In this study, we developed a multiscale deep neural network that combines data driven features learned by a CNN with a long-short time memory recurrent neural networks (LSTM) to predict neurological outcome after CA. In this framework, EEG features are automatically learned from raw waveforms by the CNN, and temporal dependencies are modeled using LSTM. We propose a framework of multiscale CNN-LSTMs to capture both fine-grained information (EEG patterns within current time windows) and coarse-grained information (EEG evolution from the beginning to current time) in the EEG time series. The framework is designed to model short-term and long-term EEG dynamics during the time course of EEG monitoring as a real-time prediction model.

Materials and methods

Dataset. Through the International Cardiac Arrest EEG Consortium (ICARE), we assembled EEG dataset including 1,038 CA subjects from seven hospitals in Europe and the U.S. The seven hospitals were Medisch Spectrum Twente (Enschede, Netherlands), Rijnstate Hospital (Arnhem, Netherlands), Erasmus Hospital (Brussels, Belgium), Yale New Haven Hospital (New Haven, CT, USA), Brigham and Women's Hospital (Boston MA, USA), Beth Israel Deaconess Medical Center (Boston, MA, USA), and Massachusetts General Hospital (Boston MA, USA). EEG data were prospectively collected in all hospitals. The 19-channel EEG data were standardized by matching channel names, applying digital bandpass filters (0.5-30 Hz), and resampling to 100 Hz. Clinical data and outcome measures were prospectively collected in the two Dutch centers and retrieved from patient medical records in the remaining centers. The research protocol was

approved by the Institutional Review Boards of participating hospitals (Partners Healthcare IRB #2013P001024).

The ICARE dataset contains approximately 58,000 hours of clinical EEG data, demographic information (age, sex, and presence of a shockable rhythm), and functional neurological outcome from 3 to 6 months after CA. Good neurological outcome was defined as a CPC score of 1 or 2 (minimal to moderate neurologic disability), and poor outcome was defined as a CPC score of 3-5 (severe neurologic disability, persistent coma or vegetative state, or death) at 3 to 6-months[11,12]. Four institutions (MGH, BWH, YNH, and BIDMC) assessed best CPC scores retrospectively through chart review at 6 months and one (ULB) at 3 months. In these institutions, CPC scores were not reviewed for patients who achieved a good outcome (CPC 1-2) or died by hospital discharge. Two institutions recorded CPC scores prospectively through phone or in-person interview for surviving patients (Medisch Spectrum Twente and Rijnstate Hospital).

The inclusion criteria included non-traumatic cardiac arrest, age≥18 years, return of spontaneous circulation (ROSC), Glasgow Coma Scale score ≤8 on admission, and management with targeted temperature management (TTM). Exclusion criteria were acute cerebral hemorrhage or acute cerebral infarction. The TTM protocol in each institution starts as soon as possible after admission to the emergency room or intensive care unit with external cooling pads. Goal temperature (32-34 °C or 36 °C) is maintained for 24 hours, and there is gradual rewarming at 0.25-0.5 °C to 37 °C. Systematic neuromuscular blockade during the hypothermia and rewarming phase of TTM was only pursued in one institution (MGH), and the remaining hospitals used neuromuscular blockade as needed for shivering or other clinical indications. Commonly used sedatives and standard dosing ranges are propofol (25-80 mcg/kg/h), midazolam

(0.1 mg/kg/h), or fentanyl (25-200 mcg/h). Only one institution (ULB) used midazolam for sedation preferentially, with the remaining institutions using propofol.

Model development. Fig. 1 shows the framework of our proposed model, which was composed of three parts: A CNN, a multiscale LSTM, and a baseline demographics model. A CNN was built to automatically extract features from consecutive 10 seconds of preprocessed EEG waveforms, which compressed the information and reduced feature dimensions. We used a recent model architecture[13], which employed a neural architecture space search strategy to find a family of best models. The output dimension of the 61-layer CNN was 1,024. CNN features generated within consecutive 5-minute time windows (i.e. 30 outputs) were averaged as the inputs of LSTMs. The CNN was optimized to classify common pathologic EEG patterns in critically ill patients, including seizures and seizure-like rhythmic and periodic patterns (lateralized and generalized periodic discharges or rhythmic delta activity), as described in prior work[14]. In the present work, we use the CNN not for classification; rather, we used the activation pattern in the last hidden layer as a representation (feature vector) of the information contained in the EEG on a local (10 second) time scale. In cases of intermittent missing data (i.e. periods when EEG monitoring was temporarily interrupted), missing epochs were interpolated to values in the nearest available epochs.

Our multiscale CNN-LSTM model contains a set of LSTMs that aim to capture the dynamics of longitudinal EEG data on multiple time scales. We used bidirectional LSTMs to learn time dependencies between time steps in time series in both forward and backward directions. The lower scale step made predictions based on the most recent 6h time blocks with fine-grained inputs (the EEG feature sequences with a 5-min time resolution). In contrast, the upper scale focused on temporal evolution over long time scales. In our framework, the lower LSTMs focus

on fine-grained information, and the upper LSTMs deliver coarse-grained information (EEG evolution from the beginning to the current time with a lower time resolution). To avoid the vanishing gradient problem in model training during back-propagation through time, the whole feature sequence until the current time was down-sampled to a feature sequence with a fixed length (equivalent to 48h). Feature sequences with a shorter length were passed without down-sampling. Batch normalization and dropout were used to avoid overfitting during network training. The lower scale and upper scale LSTMs had two hidden layers with 50 and 40 neurons, respectively.

To incorporate clinical admission variables into the model, we developed a demographics model using a random forest classifier incorporating three clinical features: age, sex, and presence of a shockable rhythm. The outputs of the upper LSTMs, lower LSTM, and demographics model were averaged to obtain the final prediction probabilities of neurological outcome. We investigated model performance in the time range of 12-72h post-cardiac arrest with steps of 6h. For consecutive 6h time blocks, we built the models as described above. In order to leverage past information and obtain stable predictions, outputs of the combination models in prior time blocks were averaged as the final prediction probabilities at the current time.

Performance Evaluation. To quantify the stability of model performance, we used 5-fold cross validation and reported average performance and 95% confidence intervals (CI). The area under the receiver operating characteristic curve (AUC-ROC) and calibration error (absolute deviation from diagonal line) were used as evaluation metrics. We compared the performance of our proposed model with several baseline models on the same dataset utilizing nine hand-crafted quantitative EEG features (burst suppression ratio[15], Shannon entropy, δ (0.5-4 Hz), θ (4-7 Hz), α (8-15 Hz), β (16-31 Hz) band power, α/δ ratio, regularity[7], and spike frequency[16]). We utilized

several models incorporating these nine quantitative EEG features for comparison with our model: 1) Time-sensitive models: A) two types of benchmark deep neural networks Bi-LSTM and temporal convolutional network (TCN)[17]; B) Another time-dependent model called a sequence of generalized linear models (GLM) with Elastic Net regularization was proposed recently by our group, which allowed feature selection in the past and present feature sets[6]; C) Hidden Markov Models (HMM) are used in time series prediction by modeling states and their transition probabilities. 2) Time-insensitive model: A conventional time-independent benchmark classifier, a Random Forest, was also created to compare against prognostication performance of models that lack the ability to leverage temporal trends.

For reproducibility, computer codes to generate the figures and sample data are available at this link https://github.com/mghcdac/icare-dl.

Results

Prediction Performance. Poor outcome was observed in 665 (64%) subjects. Subject characteristics grouped by CPC scores are summarized in **Table 1**. **Fig. 2**a. shows that the proposed framework (combination of multiscale CNN-LSTM and demographics models) achieved the best performance among all models. The performance of the proposed model combining data-driven features, longitudinal EEG dynamics, and demographic information gradually improved over time with more EEG observations from an AUC of 0.83 [0.79-0.87] at 12h to an AUC of 0.90 [0.87-0.93] at 72 h post-cardiac arrest. The best performance was obtained at 66 h with an AUC of 0.91 [0.88-0.93]. The proposed model achieved longitudinal improvement in prediction performance by leveraging prognostic information of both EEG and clinical variables, which supports the value of multimodal prognostication approaches for developing robust models for outcome prediction after cardiac arrest.

The performance of time dynamic (Bi-LSTM, TCN, GLM with Elastic Net and HMM) and time-insensitive prediction models using EEG exclusively was similar early on, but the performance of time dynamic models increased over time. The multiscale Bi-LSTM had superior performance in comparison to other models, specifically after 36h post-cardiac arrest. The best performance using only EEG was obtained by multiscale Bi-LSTMs with a mean AUC of 0.89 [0.85-0.91] at 72h post-cardiac arrest. Despite the limited number of available clinical variables across centers, the demographics model achieved relatively good performance, with a mean AUC of 0.73 [0.69-0.76]. Moreover, combining demographics and EEG features improved prediction performance, which indicates that demographics and EEG variables are complementary for neurologic outcome prediction post-cardiac arrest.

Because this analysis was retrospective and EEG monitoring duration was determined based on clinician's judgment and patient clinical course, the duration of EEG recordings varied across individuals (**Fig. 2**b). **Fig. 2**c shows the ROC curves across time intervals. At 66 h after CA, for predicting good outcome, the model's sensitivity was 39%, 60%, and 77% at specificity thresholds of 99%, 95%, and 90%, respectively; whereas specificity was 23%, 49%, and 68% at sensitivity thresholds of 99%, 95%, and 90%. For predicting poor outcome, sensitivity was 50%, 66%, and 75% at specificity thresholds of 99%, 95%, and 90%; whereas specificity was 22%, 48%, and 67% at sensitivity thresholds of 99%, 95%, and 90%.

We evaluated model generalizability by training the model on data from all hospitals except one holdout hospital, then testing the model on the held out institutions' data as validation. Results are shown in **Table 2**. We see that the model performance varies between hospitals. The model consistently performed very well over time for some hospitals (e.g., Hospital # 3 and # 5) with AUC values >90%. Model performance was worst for Hospital # 4, with a best AUC of 76%.

Model performance for Hospitals # 1 and # 6 increased over time from AUCs 83% early after CA to AUCs >90% later 54 hours after CA.

Calibration Risk. We evaluated model calibration by comparing prediction probabilities of poor outcomes with the proportion of patients who had poor outcomes. The calibration error was calculated as the absolute deviation from the diagonal line (perfect calibration). Calibration curves at different time intervals and corresponding calibration errors are shown in **Fig. 2**d. Although model calibration deteriorated somewhat over time, models remained well calibrated.

Individual-level Analysis. We evaluated model performance for individual patients and CPC scores. **Fig. 3**a shows that the predicted poor outcome probabilities over time from our models for individual patients were stable and higher for patients with poor outcomes. Moreover, prediction performance improved progressively over time as more EEG data became available.

We compared prediction probabilities of poor outcome over time for individual CPC scores in **Fig. 3**b. The mean prediction probabilities for individual CPC scores followed the ordering of CPC scores, with the CPC 5 group having the highest mean prediction probabilities and the CPC 1 group the lowest. The mean prediction probabilities across CPC 1-5 scores were 0.43, 0.43, 0.65, 0.74, and 0.76, respectively.

Visualization. Fig. 4 illustrates some prediction examples for subjects with different CPC scores, showing multi-taper EEG spectrograms and samples of the corresponding raw EEG signals. The first patient shown had a good outcome with a CPC score of 1 and had continuous EEG with normal amplitudes early during recovery. After 36h, there was an increase in power across low frequencies. Model probability outputs were consistently low for poor outcome prediction. The CPC 2 patient had a continuous EEG early on as well, however, EEG amplitudes

were lower and epileptiform discharges emerged later. The patient with CPC 3 had isoelectric EEG at 12-24h post-cardiac arrest, followed by epileptiform discharges after 36h. Poor outcome probabilities increased from 64% at 12h to 85% at 72h. The subject with CPC 4 had burst-suppression initially followed by an isoelectric EEG after 24h. Prediction probabilities for poor outcome, in this case, remained >80% across all time blocks. The last patient shown with a poor outcome (CPC 5) had increasing epileptiform discharges superimposed on a suppressed background with corresponding poor outcome probabilities >90% throughout.

Discussion

Various approaches have been proposed to predict neurologic outcomes in post-cardiac arrest coma in the literature. The Cerebral Recovery Index (CRI) method and its variants extracted several quantitative EEG features in the first 24 h after CA and trained random forest classifiers for prediction[7,8,18]. Maximal AUC was achieved in the original CRI paper at 18 h (0.94) using a hand-crafted parametric model with 5 QEEG features[7]; at 12 h (0.92) in the second CRI using a random forest model employing 9 QEEG features[8]; and at 12 h (0.94) in the third CRI employing 44 features in a random forest model[18]. These methods focused on favorable and unfavorable EEG patterns early after CA without considering the temporal evolution of the EEG. Recently, the same group trained a convolutional neural network to predict outcomes at 12 and 24 h after CA[9]. The best performance with an AUC of about 0.90 was obtained at 12 h. Although the performance of these studies were relatively high (~0.90 AUC), the data used for evaluation were more homogenous and only 5-min artifact-free epochs at specific time points were used and analyzed. A recent time-sensitive model was proposed to progressively select critical features in prior time intervals using a GLM with Elastic Net[6] model. Comparisons with the CRI model and several baseline models showed superior performance with the best AUC of 0.83 at 72 h.

Important differences between studies include: 1) The current data set is larger and more heterogeneous, coming from seven hospitals and three different countries. Our results show that model performance varies across institutes (**Table 2**). Possible reasons include differences in patient characteristics, care practices, and decision-making regarding withdrawal of care. 2) Model training and validation strategies differ across studies. 3) Model evaluation practices differ across studies. In our study, we used model calibration in addition to AUC for evaluation. Calibration provides a measure of a model's ability to provide clinically relevant probabilistic estimates of risk, which can be done at the individual patient-level and across all predicted probabilities. Overall, the present work's contributions thus include technical innovations with more complete use of the data, utlilization of temporal information, and increased automation; increased rigor of the investigation of model generalizability; and increased rigor and scale of model validation.

Our approach has several limitations. First, clinical information incorporated in our model was limited. We showed that including age, sex, and initial cardiac rhythm improved outcome prediction performance; however it is likely that additional clinical and ancillary information could improve performance even further.[4,19–21] Second, although the prediction performance of our multiscale model was superior to other models, the interpretability of deep neural networks is limited when compared to other machine learning approaches using interpretable clinical features. Third, comparisons of the proposed prognostic method with other interpretable standard prognostication approaches needs further investigation. Fourth, physicians caring for these patients were not blinded to demographics or EEG data used to build these models, therefore our results might be influenced by self-fulfilling prophecies due early withdrawal of life-sustaining therapies. Future efforts in machine learning for neurological prognostication in cardiac arrest

must focus on methods to support debiasing algorithm predictions as models are developed using patient outcomes that were censored by withdrawal of life-sustaining therapies. These biases are lessened by guidance from the European Resuscitation Council and American Heart Association about neurological prognostication practices in the participating hospitals, however variability in neurological prognostication among medical centers is a well-known barrier for building prognostic models that are robust and generalizable.[22]

Conclusion

Multiscale CNN-LSTM models predict potential for neurological recovery from coma following cardiac arrest by capturing short- and long-term EEG dynamics and demographic information.

Acknowledgements

This work was supported in part by grants from NIH-NINDS (1K23NS090900, 1K23NS119794, 1R01NS102190, 1R01NS102574, 1R01NS107291, T32HL007901, T90DA22759, T32EB001680, 1K23NS119794), American Heart Association (postdoctoral fellowship and 20CDA35310297), Neurocritical Care Society (NCS research training fellowship), Society of Critical Care Medicine-Weil research grant, CURE Epilepsy Foundation (Taking Flight Award), Massachusetts Institute of Technology-Philips Clinician Award, and the Dutch Epilepsy Fund (Epilepsiefonds; NEF 14-18).

Author Contributions

Drs. Zheng, Amorim, and Westover contributed to conception and design of the study. Drs. Zheng, Amorim, Hong, and Westover contributed to analysis of data. Drs. Zheng, Amorim, Jing and

Westover contributed to preparing the figures. Drs. Amorim, Ghassemi, Lee, Pang, Herman, Sivaraju, Gaspard, Hofmeijer, van Putten, Ruijter, Tjepkema-Clostermans, and Westover contributed to data acquisition. Drs. Zheng, Amorim, Jing, Ge, Hong, Wu, Ghassemi, Lee, Pang, Herman, Sivaraju, Gaspard, Ruijter, Sun, Tjepkema-Cloostermans, Hofmeijer, van Putten, and Westover contributed to data collection, drafting and revising the text.

Competing interests

Dr. Westover is co-founder of Beacon Biosignals. Dr. van Putten is founder of Clinical Science Systems. Dr. Lee is co-founder of Soterya, Inc. Neither company contributed funding nor played any role in the study. All other authors declare no potential competing interests.

Figures and Tables

Table 1 Subjects characteristics, grouped by CPC scores.

Table 2 Model performance for individual institutions (AUC, 95% confidence intervals)

Fig. 1 The proposed model architecture. The framework contains three components: CNNs, multiscale LSTMs and the demographics model. CNNs were used to extract features for every 10-s EEG segments. The outputs of CNN features were averaged as the inputs of Bi-LSTMs. The fine-grained Bi-LSTMs made predictions based on the most recent 6-h EEG features while the coarse-grained Bi-LSTMs aimed at the snapshots of EEG evolutions from the beginning to the current time. The modeling framework for EEG has two important novelty: EEG feature learning in a data-driven way and short-term and long-term time dynamic modeling. The

demographics model made predictions based on the clinical variables age, sex and ventricular fibrillation. The outputs of all models were fused by averaging to obtain the final prediction probabilities of neurological outcomes.

Fig. 2 Model performance of different models in outcome prediction. a, Mean AUC values of different models within each 6-hour time interval. The proposed method (red line) performed best, exhibiting consistent improvement in performance with more observations (from mean AUC of 0.83 at 12 h to mean AUC of 0.90 at 72 h). **b**, Numbers of patients with EEG available with respect to time after cardiac arrest. **c**, Mean ROC curves at different time intervals. Shaded areas indicate the standard errors in 5-fold cross validation. **d**, Calibration curves at different time intervals. The numbers are calibration errors (deviations from the diagonals).

Fig. 3 Model performance of individual patients and CPC scores. a, Individual prediction probabilities. Each row shows the output probabilities for one patient over consecutive 6h blocks, the darker the color, the higher the predicted probability of poor outcome. **b**, Predicted probabilities over time, grouped by final CPC scores. The overall mean predicted probabilities were consistent with the expected order of CPC scores.

Fig. 4 Illustration of prediction results on sample patients. Each row illustrates the mean multi-taper spectrogram and EEG waveforms in multiple time blocks. At the bottom of each spectrogram, prediction probabilities of the model for the corresponding EEG segments are shown. The time length of EEG snapshots was 10 s while the spectrograms spanning a 5-min time window are shown. The proposed model made good predictions based on the EEG patterns. Continuous EEG contributes to lower prediction probabilities of poor outcomes while

epileptiform discharges with flat background contribute to high prediction probabilities of poor outcomes.

Reference

1. Callaway CW, Donnino MW, Fink EL, et al. Part 8: Post-cardiac arrest care: 2015 American Heart Association guidelines update for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation*. 2015. doi:10.1161/CIR.0000000000000262
2. Sandroni C, Cariou A, Cavallaro F, et al. Prognostication in comatose survivors of cardiac arrest: An advisory statement from the European Resuscitation Council and the European Society of Intensive Care Medicine. *Intensive Care Med*. 2014;40(12):1816-1831. doi:10.1007/s00134-014-3470-x
3. Ruijter BJ, Tjepkema-Cloostermans MC, Tromp SC, et al. Early electroencephalography for outcome prediction of postanoxic coma: A prospective cohort study. *Ann Neurol*. 2019;86(2):203-214. doi:10.1002/ana.25518
4. Oddo M, Rossetti AO. Early multimodal outcome prediction after cardiac arrest in patients treated with hypothermia. *Crit Care Med*. 2014;42(6):1340-1347. doi:10.1097/CCM.0000000000000211
5. Lee S, Zhao X, Davis KA, Topjian AA, Litt B, Abend NS. Quantitative EEG predicts outcomes in children after cardiac arrest. *Neurology*. 2019;92(20):E2329-E2338. doi:10.1212/WNL.0000000000007504
6. Ghassemi MM, Amorim E, Alhanai T, et al. Quantitative Electroencephalogram Trends Predict Recovery in Hypoxic-Ischemic Encephalopathy. *Crit Care Med*. 2019;47(10):1416-1423. doi:10.1097/CCM.0000000000003840
7. Tjepkema-Cloostermans MC, van Meulen FB, Meinsma G, van Putten MJAM. A Cerebral Recovery Index (CRI) for early prognosis in patients after cardiac arrest. *Crit Care*. 2013;17(5). doi:10.1186/cc13078
8. Tjepkema-Cloostermans MC, Hofmeijer J, Beishuizen A, et al. Cerebral recovery index: Reliable help for prediction of neurologic outcome after cardiac arrest. *Crit Care Med*. 2017;45(8):e789-e797. doi:10.1097/CCM.0000000000002412
9. Tjepkema-Cloostermans MC, da Silva Lourenço C, Ruijter BJ, et al. Outcome Prediction in Postanoxic Coma With Deep Learning. *Crit Care Med*. 2019;47(10):1424-1432. doi:10.1097/CCM.0000000000003854
10. Amorim E, van der Stoel M, Nagaraj SB, et al. Quantitative EEG reactivity and machine learning for prognostication in hypoxic-ischemic brain injury. *Clin Neurophysiol*. 2019;130(10):1908-1916. doi:10.1016/j.clinph.2019.07.014
11. Wijdicks EFM, Hijdra A, Young GB, Bassetti CL, Wiebe S. Practice parameter: Prediction of outcome in comatose survivors after cardiopulmonary resuscitation (an

evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*. 2006;67(2):203-210. doi:10.1212/01.wnl.0000227183.21314.cd

12. Booth CM, Boone RH, Tomlinson G, Detsky AS. Is This Patient Dead, Vegetative, or Severely Neurologically Impaired? Assessing Outcome for Comatose Survivors of Cardiac Arrest. *J Am Med Assoc*. 2004;291(7):870-879. doi:10.1001/jama.291.7.870

13. Radosavovic I, Kosaraju RP, Girshick R, He K, Dollar P. Designing Network Design Spaces. In: ; 2020. doi:10.1109/cvpr42600.2020.01044

14. Jing J, d'Angremont E, Ebrahim S, et al. Rapid annotation of seizures and interictal-ictal-injury continuum EEG patterns. *J Neurosci Methods*. 2020. doi:10.1016/j.jneumeth.2020.108966

15. Brandon Westover M, Shafi MM, Ching SN, et al. Real-time segmentation of burst suppression patterns in critical care EEG monitoring. *J Neurosci Methods*. 2013;219(1):131-141. doi:10.1016/j.jneumeth.2013.07.003

16. Jing J, Sun H, Kim JA, et al. Development of Expert-Level Automated Detection of Epileptiform Discharges during Electroencephalogram Interpretation. *JAMA Neurol*. 2020;77(1):103-108. doi:10.1001/jamaneurol.2019.3485

17. Bai S, Kolter JZ, Koltun V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. March 2018. http://arxiv.org/abs/1803.01271. Accessed January 14, 2020.

18. Nagaraj SB, Tjepkema-Cloostermans MC, Ruijter BJ, Hofmeijer J, van Putten MJAM. The revised Cerebral Recovery Index improves predictions of neurological outcome after cardiac arrest. *Clin Neurophysiol*. 2018;129(12):2557-2566. doi:10.1016/j.clinph.2018.10.004

19. Youn CS, Callaway CW, Rittenberger JC. Combination of initial neurologic examination, quantitative brain imaging and electroencephalography to predict outcome after cardiac arrest. *Resuscitation*. 2017;110:120-125. doi:10.1016/j.resuscitation.2016.10.024

20. Bevers MB, Scirica BM, Avery KR, Henderson G V., Lin AP, Lee JW. Combination of Clinical Exam, MRI and EEG to Predict Outcome Following Cardiac Arrest and Targeted Temperature Management. *Neurocrit Care*. 2018;29(3):396-403. doi:10.1007/s12028-018-0559-z

21. Kim JH, Kim MJ, You JS, et al. Multimodal approach for neurologic prognostication of out-of-hospital cardiac arrest patients undergoing targeted temperature management. *Resuscitation*. 2019;134:33-40. doi:10.1016/j.resuscitation.2018.11.007

22. May TL, Lary CW, Riker RR, et al. Variability in functional outcome and treatment practices by treatment center after out-of-hospital cardiac arrest: analysis of International Cardiac Arrest Registry. *Intensive Care Med*. 2019. doi:10.1007/s00134-019-05580-7

Competing interests

Dr. Westover is co-founder of Beacon Biosignals. Dr. van Putten is founder of Clinical Science Systems. Dr. Lee is co-founder of Soterya, Inc. Neither company contributed funding nor played any role in the study. All other authors declare no potential competing interests.

Author Contributions

Drs. Zheng, Amorim, and Westover contributed to conception and design of the study. Drs. Zheng, Amorim, Hong, and Westover contributed to analysis of data. Drs. Zheng, Amorim, Jing and Westover contributed to preparing the figures. Drs. Amorim, Ghassemi, Lee, Pang, Herman, Sivaraju, Gaspard, Hofmeijer, van Putten, Ruijter, Tjepkema-Clostermans, and Westover contributed to data acquisition. Drs. Zheng, Amorim, Jing, Ge, Hong, Wu, Ghassemi, Lee, Pang, Herman, Sivaraju, Gaspard, Ruijter, Sun, Tjepkema-Cloostermans, Hofmeijer, van Putten, and Westover contributed to data collection, drafting, and revising the text.

Table 3. Subjects characteristics, grouped by CPC scores.

CPC group	CPC 1	CPC 2	CPC 3	CPC 4	CPC 5
Number of patients	303	70	31	17	617
Age (years)	57 (15)	56 (15)	66 (11)	54 (21)	62 (16)
Female sex (%)	29	24	35	47	32
Shockable Rhythm (VFib/VT, %)	71	67	42	41	31
EEG start time (h)	17 (14)	16 (16)	16 (13)	20 (6)	20 (17)
EEG duration (h)	52 (33)	63 (44)	69 (51)	99 (60)	53 (40)
Out-of-hospital CA (N/A) *	232 (21)	50 (6)	17 (4)	14 (0)	439 (43)
TTM (N/A) *	261 (34)	61 (7)	26 (5)	11 (2)	514 (64)

VFib: ventricular fibrillation; VT: ventricular tachycardia; TTM: targeted temperature management; EEG start time (h) is relative to time of cardiac arrest. All numbers related to age and EEG expressed as mean (standard deviation). *For the number of out-of-hospital CA patients and TTM, we didn't have all information available from different hospitals.

Table 2. Model performance for individual institutions (AUC, 95% confidence intervals)

Time Interval	18 h	24 h	30 h	36 h	42 h
# 1	0.83 [0.72,0.94]	0.86 [0.78,0.95]	0.84 [0.75,0.93]	0.86 [0.78,0.94]	0.87 [0.78,0.95]
# 2	0.83 [0.74,0.93]	0.78 [0.69,0.87]	0.84 [0.77,0.91]	0.82 [0.75,0.90]	0.87 [0.81,0.93]
# 3	0.93 [0.87,0.99]	0.95 [0.91,0.99]	0.95 [0.90,0.99]	0.95 [0.91,0.99]	0.93 [0.88,0.97]
# 4	0.67 [0.56,0.78]	0.68 [0.58,0.78]	0.72 [0.63,0.81]	0.76 [0.66,0.85]	0.71 [0.60,0.83]
# 5	0.88 [0.83,0.92]	0.89 [0.85,0.93]	0.91 [0.87,0.94]	0.91 [0.88,0.95]	0.93 [0.89,0.96]
# 6	0.83 [0.73,0.93]	0.87 [0.78,0.95]	0.89 [0.81,0.97]	0.89 [0.82,0.96]	0.90 [0.83,0.96]
Time Interval	48 h	54 h	60 h	66 h	72 h
# 1	0.87 [0.79,0.95]	0.90 [0.83,0.97]	0.91 [0.84,0.98]	0.91 [0.83,0.99]	0.91 [0.82,0.99]
# 2	0.87 [0.80,0.93]	0.87 [0.81,0.94]	0.86 [0.80,0.93]	0.88 [0.81,0.95]	0.87 [0.79,0.95]
# 3	0.91 [0.86,0.96]	0.92 [0.87,0.97]	0.94 [0.89,0.98]	0.96 [0.93,1.00]	0.97 [0.94,1.00]
# 4	0.61 [0.47,0.75]	0.55 [0.37,0.72]	0.54 [0.33,0.74]	0.53 [0.32,0.73]	0.60 [0.39,0.81]
# 5	0.92 [0.89,0.96]	0.91 [0.87,0.95]	0.91 [0.87,0.96]	0.91 [0.87,0.96]	0.90 [0.84,0.95]
# 6	0.90 [0.83,0.96]	0.90 [0.83,0.97]	0.93 [0.88,0.99]	0.94 [0.88,0.99]	0.98 [0.94,1.00]