Predicting Neurological Recovery from Coma After Cardiac Arrest: The George B. Moody PhysioNet Challenge 2023

Matthew A Reyna[1,*], Edilberto Amorim[2,3,*], Reza Sameni[1], James Weigle[1], Andoni Elola[4], Ali Bahrami Rad[1], Salman Seyedi[1], Hyeokhyen Kwon[1], Wei-Long Zheng[5], Mohammad M Ghassemi[6], Michel JAM van Putten[7,8], Jeannette Hofmeijer[7,9], Nicolas Gaspard[10,11], Adithya Sivaraju[10], Susan T Herman[12], Jong Woo Lee[13], M Brandon Westover[14,15,16**], Gari D Clifford[1,17,**]

[1]Department of Biomedical Informatics, Emory University, USA
[2]Department of Neurology, Weill Institute for Neurosciences, University of California, San Francisco, USA
[3]Department of Neurology, Massachusetts General Hospital, USA
[4]Department of Electronic Technology, University of the Basque Country UPV/EHU, Spain
[5]Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[6]Department of Computer Science and Engineering, Michigan State University, USA
[7]Clinical Neurophysiology Group, University of Twente, The Netherlands
[8]Department of Neurology and Clinical Neurophysiology, Medisch Spectrum Twente, The Netherlands
[9]Department of Neurology, Rijnstate Hospital, The Netherlands
[10]Department of Neurology, Yale School of Medicine, USA
[11]Service de Neurologie, Hôpital Universitaire de Bruxelles and Université Libre de Bruxelles, Belgium
[12]Department of Neurology, Barrow Neurological Institute Comprehensive Epilepsy Center, USA
[13]Department of Neurology, Brigham and Women's Hospital, USA
[14]Department of Neurology, Beth Israel Deaconess Medical Center, USA
[15]McCance Center for Brain Health, Massachusetts General Hospital, USA
[16]Harvard Medical School, USA
[17]Department of Biomedical Engineering, Georgia Institute of Technology, USA
[*]Co-First Authors [**]Co-Senior Authors

Abstract

The George B. Moody PhysioNet Challenge 2023 invited teams to develop algorithmic approaches for predicting the recovery of comatose patients after cardiac arrest.

A patient's prognosis after the return of spontaneous circulation informs treatment, including the continuation or withdrawal of life support. Brain monitoring with an electroencephalogram (EEG) can improve the objectivity of a prognosis, but EEG interpretation requires clinical expertise. The algorithmic analysis of EEGs can potentially improve the accuracy and accessibility of prognoses, but existing work is limited by small and homogeneous datasets.

The PhysioNet Challenge 2023 contributed to addressing these problems. It introduced the International Cardiac Arrest REsearch consortium (I-CARE) dataset, which is a large, multi-center collection of EEGs, other physiological data, and clinical outcomes, with over 57,000 hours of data from 1,020 patients from seven hospitals. It required teams to submit their complete training and inference code to improve the reproducibility and generalizability of their research. A total of 111 teams participated in the Challenge, contributing diverse approaches from academic, clinical, and industry participants worldwide.

1. Introduction

Survival rates for cardiac arrest are generally low. Brain ischemia is common in individuals who survive initial resuscitation, and most survivors who are admitted to an in-

tensive care unit (ICU) are comatose [1]. During the first few days following cardiac arrest, physicians are typically asked for a patient's prognosis. This prognosis influences the patient's subsequent care, with a good prognosis frequently resulting in continued treatment and a poor prognosis frequently leading to the withdrawal of treatment and death. However, some patients also recover after poor prognoses, leading to concerns that poor prognoses may, in some cases, be self-fulfilling prophesies.

Electroencephalography can improve the objectivity of prognoses after cardiac arrest. A number of brain activity patterns in an electroencephalogram (EEG), including reduced voltage, burst suppression, seizures, and seizure-like patterns, are associated with patient outcomes [2]. Moreover, the evolution of these patterns over time may provide additional prognostic information [3–6]. However, the interpretation of a continuous EEG is a laborious task that requires neurological expertise, limiting the accessibility of EEG-informed prognoses.

The automated analysis of EEG data has the potential to improve the accuracy and accessibility of such prognoses, especially in environments with limited access to expert neurologists. However, the small and homogeneous datasets in most studies of algorithmic EEG interpretation are unsuitable for the development of generalizable machine learning algorithms.

The George B. Moody PhysioNet Challenge 2023 (formerly the PhysioNet/Computing in Cardiology Challenge) sought to address these issues by inviting teams to develop automated approaches for coma prognostication after cardiac arrest on a large international database with over 57,000 hours of EEG recording data from 1,020 patients from seven hospitals.

2. Methods

2.1. Challenge Data

The 2023 Challenge used the International Cardiac Arrest REsearch consortium (I-CARE) dataset [7]. I-CARE compiled a large international dataset from comatose patients after cardiac arrest. This dataset includes 1,020 patients from 7 hospitals:

1. Rijnstate Hospital, Arnhem, The Netherlands
2. Medisch Spectrum Twente, Enschede, The Netherlands
3. Erasme Hospital, Brussels, Belgium
4. Massachusetts General Hospital, Boston, USA
5. Brigham and Women's Hospital, Boston, USA
6. Beth Israel Deaconess Medical Center, Boston, USA
7. Yale New Haven Hospital, New Haven, USA

The data collection was approved by the institutional review boards of the respective hospitals.

The I-CARE dataset includes EEG, electrocardiogram (ECG), electromyogram (EMG), and electrooculogram (EOG) recordings; basic demographic (age, sex, hospital) and clinical information (time to return of spontaneous circulation (ROSC), in-hospital or out-of-hospital cardiac arrest, presence of a shockable rhythm, targeted temperature management (TTM)); and patient outcomes (Cerebral Performance Category (CPC) scores).

The CPC scores follow a five-point scale: (1) good recovery, (2) moderate disability, (3) severe disability, (4) unresponsive wakefulness syndrome (previously known as a persistent vegetative state), and (5) death [8]. CPC scores of 1 and 2 are generally considered to be good or favorable outcomes, and CPC scores of 3, 4, and 5 are poor or unfavorable outcomes.

The data collection practices varied between hospitals and between different patients from the same hospital. However, all patients had recordings with the following 19 EEG channels: Fp1, Fp2, F7, F8, F3, F4, T3, T4, C3, C4, T5, T6, P3, P4, O1, O2, Fz, Cz, and Pz. Nearly all patients had basic demographic and clinical information.

The data provided to the Challenge participants were unchanged from the data provided by the hospitals except to encode the recording data as 16-bit signed integers for Waveform Database (WFDB) format, to consistently name equivalent channels from different hospitals, and to remove protected health information (PHI) by grouping ages above 89 as a single age of "90".

We included data for 60% of the patients in a public training set and sequestered data for 10% of the patients in a hidden validation set and data for the remaining 30% of the patients in a hidden test set. The training set was released at the beginning of the Challenge, and the validation and test sets were used to evaluate the Challenge entries and were not released during the Challenge. The split of the dataset into training, validation, and test sets approximately preserved the univariate distributions of the variables and labels. To better assess the generalizability of the algorithms, we excluded data from one hospital from the training and validation sets and only included data for these patients in the test set.

2.2. Challenge Objective

The goal of the 2023 Challenge was to use longitudinal EEG recordings and other collected data to predict good and poor patient outcomes for comatose patients after cardiac arrest. We asked the Challenge participants to develop open-source algorithms that use these data to provide the probability of a poor outcome for these patients.

2.2.1. Challenge Timeline

This year's Challenge was the 24th George B. Moody PhysioNet Challenge [9]. As in previous years, the Challenge had an unofficial phase and an official phase. The

unofficial phase (February 10, 2023 to April 24, 2023) introduced the teams to the Challenge. We publicly shared the Challenge objective, training data, example algorithms, and evaluation metric and invited the teams to submit their code for evaluation, scoring at most five entries from each team on the hidden validation set. Between the unofficial and official phases, we took a hiatus (April 25, 2023 to June 8, 2023) to improve the Challenge. The official phase (June 9, 2023 to August 31, 2023) continued the Challenge. We updated the Challenge data, example algorithms, and evaluation metric and again invited teams to submit their code for evaluation, scoring at most ten entries from each team on the hidden validation set. Notably, we released a smaller version of the data (40.3 GB) with fewer channels and truncated recording for the unofficial phase, and we released a larger version of the data (2.63 TB) with more channels and full recordings for the official phase.

After the end of the official phase, each team chose a single entry from their team for us to evaluate on the test set. The winners of the Challenge were the teams with the best scores on the test set. We announced the results at the end of the Computing in Cardiology (CinC) 2023 conference, where the teams presented, defended, and published their work. Only teams that presented and published their work at the conference were eligible for rankings and prizes. We will publicly release the algorithms after the end of the Challenge and the publication of these papers.

2.2.2. Challenge Evaluation

The evaluation metric for the 2023 Challenge measured the rate at which teams made poor prognoses for patients with poor outcomes at a low rate of incorrectly making poor prognoses for patients with good outcomes.

For each patient, we asked teams to provide a probability of a poor outcome. We defined a positive case as a poor outcome, i.e., a CPC score of 3, 4, or 5, and a negative case as a good outcome, i.e., a CPC score of 1 or 2. The choice of a decision threshold δ determines the numbers of true positive, false positive, false negative, and true negative cases, i.e., TP_θ, FP_θ, FN_θ, TN_θ, respectively. Let

$$\mathrm{FPR}_\theta = \frac{\mathrm{FP}_\theta}{\mathrm{FP}_\theta + \mathrm{TN}_\theta} \tag{1}$$

be the false positive rate (FPR) for an algorithm at a decision threshold of θ, i.e., the fraction of patients with good outcomes but poor algorithmic prognoses at a decision threshold of θ. For hospital h, let $\theta_{\alpha,h}$ be the largest value of θ such that $\mathrm{FPR}_\theta \leq \alpha = 0.05$. We define the total numbers of true positive, false positive, false negative, and true negative cases, respectively, across all hospitals as $\mathrm{TP}_\alpha = \sum_h \mathrm{TP}_{\theta_{\alpha,h}}$, $\mathrm{FP}_\alpha = \sum_h \mathrm{FP}_{\theta_{\alpha,h}}$, $\mathrm{FN}_\alpha = \sum_h \mathrm{FN}_{\theta_{\alpha,h}}$, and $\mathrm{TN}_\alpha = \sum_h \mathrm{TN}_{\theta_{\alpha,h}}$, respectively, at a FPR of $\alpha \leq 0.05$ for each hospital. The Challenge score is the true positive rate (TPR)

$$\mathrm{TPR}_\alpha = \frac{\mathrm{TP}_\alpha}{\mathrm{FP}_\alpha + \mathrm{FN}_\alpha}, \tag{2}$$

across all hospitals at FPR of $\alpha \leq 0.05$ at each hospital. The team with the highest value of the Challenge score at 72 hours after ROSC won the Challenge.

We strictly limited the FPRs based on clinical needs. While both false positive and false negative predictions are problematic, the withdrawal of life support from patients who could recover with continued treatment is much more serious than prolonging care for patients who ultimately do not recover. Therefore, professional societies generally recommend that prognostic tests operate with FPRs of less than or equal to 5% [10, 11].

We focused on the predictions at 72 hours after ROSC to allow teams to observe trends in the recordings over time but to require them to offer prognoses within clinically relevant time frames. The algorithms to make predictions at 12, 24, 48, and 72 hours after ROSC, but we only used the scores at 72 hours after ROSC to determine the winners.

3. Challenge Results

A total of 111 teams submitted 982 algorithms during the Challenge, including 107 teams with 179 successful entries and 194 unsuccessful entries during the unofficial phase and 76 teams with 269 successful entries and 313 unsuccessful entries during the official phase. After the end of the official phase, we attempted to score one entry from each team on the hidden test set. A total of 58 teams had a successful entry on the test set, and 36 teams had a successful entry and met the other Challenge criteria. Table 1 summarizes the highest-ranked teams. Team summaries, additional scores, and the full Challenge criteria for rankings are available on [12].

Rank	Team name	Training set score	Validation set score	Test set score
1	AIRhythm	0.992	0.657	0.792
2	ComaToss	1.000	0.612	0.787
3	TUD_EEG	0.958	0.687	0.718

Table 1: The three teams with the highest Challenge score (2) on the test set; higher scores are better, and only ranked teams are shown.

4. Discussion

The large size of the Challenge data limited the number of successful participants and the types of successful approaches. Clinically-informed feature engineering and other data reduction and summarization techniques were

important for completing training and inference on these data within the computational resource constraints.

The algorithms had lower performance on the hidden validation and test sets than on the publicly available training set. These performance changes reflect the difficulty of generalizing to unseen data. The similar performance on the validation and test sets suggests that the algorithms generalized well to data from a hospital that was only represented in the test set and may generalize well to new data.

5. Conclusions

This year's Challenge explored the potential for algorithmic prognostication of neurological recovery for comatose patients following cardiac arrest. We asked the Challenge participants to design working, open-source algorithms for the predicting the risk of poor outcomes from EEG and other data. The development of such prognostic algorithms has the potential to improve clinical decision support in the critical hours after cardiac arrest.

Acknowledgements

This research is supported by the National Institute of General Medical Sciences (NIGMS: 2R01GM104987-09); the National Institute of Biomedical Imaging and Bioengineering (NIBIB: R01EB030362); National Institute of Neurological Disorders and Stroke (NINDS: R01NS102190, R01NS102574, R01NS107291, RF1NS120947, R01NS126282); the National Heart, Lung, and Blood Institute (NHLBI: R01HL161253); the National Institute on Aging (NIA: RF1AG064312, R01AG073410, R01AG073598); the National Science Foundation (NSF: grant 2014431); the National Center for Advancing Translational Sciences of the National Institutes of Health (NCATS: UL1TR002378); MCIN/ AEI/10.13039/501100011033/; FEDER through grant PID2021-122727OB-I00; the Basque Government through grant IT1717-22 and the University of the Basque Country (UPV/EHU: COLAB20/01); as well as AliveCor, Amazon Web Services, the Gordon and Betty Moore Foundation, and MathWorks under unrestricted gifts. MvP is a co-founder and medical advisor of Clinical Science Systems. MBW is a co-founder and holds equity in Beacon Biosignals and receives royalties for authoring Pocket Neurology from Wolters Kluwer and Atlas of Intensive Care Quantitative EEG by Demos Medical. GDC has financial interests in AliveCor and Mindchild Medical and holds a board position in Mindchild Medical. None of these entities influenced the design of or provided data for this year's Challenge. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the above entities.

References

[1] Rundgren M, Westhall E, Cronberg T, Rosen I, Friberg H. Continuous amplitude-integrated electroencephalogram predicts outcome in hypothermia-treated cardiac arrest patients. Critical Care Medicine 2010;38(9):1838–1844.

[2] Hirsch LJ, Fong MW, Leitinger M, LaRoche SM, Beniczky S, Abend NS, et al. American Clinical Neurophysiology Society's standardized critical care EEG terminology: 2021 version. Journal of Clinical Neurophysiology Official Publication of the American Electroencephalographic Society 2021;38.

[3] Amorim E, Rittenberger JC, Zheng JJ, Westover MB, Baldwin ME, Callaway CW, et al. Continuous EEG monitoring enhances multimodal outcome prediction in hypoxic-ischemic brain injury. Resuscitation 2016;109:121–126.

[4] Hofmeijer J, Beernink TM, Bosch FH, Beishuizen A, Tjepkema-Cloostermans MC, van Putten MJ. Early EEG contributes to multimodal outcome prediction of postanoxic coma. Neurology 2015;85(2):137–143.

[5] Zheng WL, Amorim E, Jing J, Ge W, Hong S, Wu O, et al. Predicting neurological outcome in comatose patients after cardiac arrest with multiscale deep neural networks. Resuscitation 2021;169:86–94.

[6] Ruijter BJ, Tjepkema-Cloostermans MC, Tromp SC, van den Bergh WM, Foudraine NA, Kornips FH, et al. Early electroencephalography for outcome prediction of postanoxic coma: a prospective cohort study. Annals of Neurology 2019;86(2):203–214.

[7] Amorim E, Zheng WL, Ghassemi MM, Aghaeeaval M, Kandhare P, Karukonda V, et al. The international cardiac arrest research consortium electroencephalography database. Critical Care Medicine 10 2023;.

[8] Jennett B, Bond M. Assessment of outcome after severe brain damage: a practical scale. The Lancet 1975;305(7905):480–484.

[9] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[10] Forstmeier W, Wagenmakers EJ, Parker TH. Detecting and avoiding likely false-positive findings – a practical guide. Biological Reviews 2017;92(4):1941–1968.

[11] Healy B, Khan A, Metezai H, Blyth I, Asad H. The impact of false positive COVID-19 results in an area of low prevalence. Clinical Medicine 2021;21(1):e54.

[12] George B. Moody PhysioNet Challenge 2023. https://physionetchallenges.org/2023/. Accessed: 2023-10-20.

Address for correspondence:

Matthew A Reyna

DBMI, 101 Woodruff Circle, 4th Floor East, Atlanta, GA 30322

matthew.a.reyna@emory.edu