Supplemental Material

Auditory stimulation and deep learning predict awakening from coma after cardiac arrest

Florence M. Aellen[1,2], Sigurd L. Alnes[1,2], Fabian Loosli[1], Andrea O. Rossetti[3], Frédéric Zubler[4], Marzia De Lucia[5], Athina Tzovara[1,2,6,7]

Author affiliations:

[1] Institute of Computer Science, University of Bern, Bern, Switzerland

[2] Zentrum für Experimentelle Neurologie, Department of Neurology, Inselspital University Hospital Bern, Bern, Switzerland

[3] Neurology Service, Department of Clinical Neurosciences, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

[4] Sleep-Wake-Epilepsy Center Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Switzerland

[5] Laboratory for Research in Neuroimaging (LREN), Department of Clinical Neurosciences, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland

[6] Sleep Wake Epilepsy Center - NeuroTec, Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

[7] Helen Wills Neuroscience Institute, University of California Berkeley, CA, USA

Evaluation of outcome prediction results for patients in a 'gray zone'

In the results presented in the main manuscript we imitate as closely as possible a 'real life' situation where a CNN would be used in a clinical practice. One could envision training the network with all available data, and then using it to predict outcome in new patients, as they arrive in the intensive care unit, without setting explicit ratios of determinate vs. 'gray zone' patients in the train or test sets. The main manuscript reports aggregate results across

train/validation/test sets (Figure 3) for ‘gray zone’ patients, as the distribution of the confidence scores of the network can still be informative about whether these patients are perceived as particular cases or outliers by the network.

Here, we perform two control analyses to evaluate systematically the generalizability of the trained networks on ‘gray zone’ patients:

1. First, instead of randomly splitting patients to train/test/validation, we now curate this split, so that some of the ‘gray zone’ patients will be part of the 10-fold train/validation datasets, and a fixed (but high) number of ‘gray zone’ patients will be part of the test set.
2. Second, we trained one network using exclusively patients with determinate outcomes for train/validation, and kept all ‘gray zone’ patients as a test set, to evaluate generalization of results.

1. Training neural networks with a curated patient split resulting in a test set containing only ‘gray zone’ patients

We trained a neural network where the test set only contained patients from the ‘gray zone’. 27 ‘gray zone’ patients were randomly selected for the test set, and the remaining 21 ‘gray zone’ patients were randomly split between the train and validation sets. This resulted in a train set of 80 patients, a validation set of 27 patients and a test set of 27 patients, exclusively part of the ‘gray zone’. This split was repeated in a cross validation, and contained the same patient numbers per set as the network that is reported in the main manuscript.

With this approach, on the test set of ‘gray zone’ patients, we obtained an AUC of 0.64 ± 0.01, a PPV of 0.86 ± 0.02 and NPV of 0.43 ± 0.01 (Supplemental Table 1 and Supplemental Figure 1). These results are very close to the values obtained with the networks presented in the main manuscript, without a curated patient split (Table 1, results on the test

set: AUC: 0.70 ± 0.04, PPV: 0.83 ± 0.03 and NPV: 0.57 ± 0.04), and suggest an unbiased and robust outcome prediction on ‘gray zone’ patients, which were previously unseen by the network.

2. Training of a neural network with all ‘gray zone’ patients in the test set

We additionally trained one single network, where the test set contained all ‘gray zone’ patients (N=48) and the train and validation sets contained the rest of the patients (N=86 patients in total, or 64 for train and 22 patients for validation).

In this case, we obtained an AUC of 0.67, PPV of 0.85 and NPV of 0.46 on the test set of all ‘gray zone’ patients (Supplemental Table 2 and Supplemental Figure 2). The PPV and AUC are slightly lower than the ones obtained with curating the splits of train/validation/test sets (Supplemental Table 1), but are well in line with the main results of our manuscript, that our approach is primarily sensitive to predicting survival.

One caveat of this control analysis is that we had a smaller train set of only 64 patients, compared to the original train set of 80 patients, which can result in less accurate training of the network and therefore less strong outcome prediction. Importantly, this new network was trained without any of the ‘gray zone’ patients and therefore performs, as expected, worse than a network which was trained with at least some ‘gray zone’ patients, and has been exposed to some of their characteristics.

Network performance for different types of auditory stimulation

For the analysis reported in the main manuscript, we only considered EEG responses to standard or duration deviant sounds, following previous work [1]. This was based on the implicit assumption that the network's prediction mostly captures broad patterns of EEG responses to auditory stimulation, rather than their identity. However, the experimental paradigm also included EEG responses to location and pitch deviant sounds. We thus additionally evaluated the performance of the trained neural network on all four types of auditory stimulation (standard, duration, location and pitch deviants, Supplemental Table 3 for the mean number of extracted trials per sound type).

We then evaluated the performance of the trained network on trials where patients were exposed to standard and duration deviant sounds, separately, and we then extended the analysis to the EEG responses to location and pitch deviant sounds. The AUC score for patients of the test set was at very similar levels for EEG responses to any of the four auditory stimuli (Supplemental Table 4), around 0.70, in accordance to what reported in the main manuscript, combining EEG responses to standard and duration deviant sounds (Table 1, Mean AUC over the test set). The distribution of single-trial predictions for EEG responses to the four different sounds was also very consistent (Supplemental Figure 3 for the distribution of single-trial predictions for one exemplar survivor).

These results suggest that the outcome prediction of the network mainly relies on characteristics of the EEG response to sounds in coma, and not specifically their identity.

Supplemental References

1. Alnes SL, Lucia MD, Rossetti AO, Tzovara A. Complementary roles of neural synchrony and complexity for indexing consciousness and chances of surviving in acute coma. NeuroImage 2021;245:118638.

Supplemental Table 1. Prediction of outcome for networks trained on determinate and 'gray zone' patients and tested on 'gray zone' patients only.

		All	Hypothermia	Normothermia
Mean over 10 folds	**AUC Train**	0.81 ± 0.01	0.83 ± 0.01	0.77 ± 0.02
	AUC Validation	0.77 ± 0.02	0.78 ± 0.03	0.74 ± 0.04
	AUC Test	0.64 ± 0.01	0.60 ± 0.03	0.72 ± 0.02
	PPV Train	0.90 ± 0.01	0.93 ± 0.01	0.82 ± 0.03
	PPV Validation	0.90 ± 0.03	0.93 ± 0.03	0.71 ± 0.11
	PPV Test	0.86 ± 0.02	0.79 ± 0.04	0.96 ± 0.03
	NPV Train	0.71 ± 0.01	0.67 ± 0.02	0.71 ± 0.02
	NPV Validation	0.67 ± 0.03	0.63 ± 0.03	0.67 ± 0.04
	NPV Test	0.43 ± 0.01	0.41 ± 0.01	0.43 ± 0.04
Single fold	**AUC Train**	0.82	0.83	0.79
	AUC Validation	0.77	0.75	0.80
	AUC Test	0.61	0.51	0.79
	PPV Train	0.94	0.96	0.91
	PPV Validation	1.00	1.00	1.00
	PPV Test	0.86	0.67	1.00
	NPV Train	0.70	0.62	0.76
	NPV Validation	0.63	0.62	0.67
	NPV Test	0.40	0.36	0.50

Prediction of outcome on networks trained with a curated cross-validation patient split, where the test set contained only 'gray zone' patients. First, we report the mean ± standard error over all ten trained folds, as well as the performance of an exemplar fold. We also report the AUC, PPV and NPV, with respect to survival, of the train, validation and test sets, for all patients and separately for the sub-cohorts of patients treated with hypothermia and normothermia.

Supplemental Table 2. Prediction of outcome for a network trained and validated with determinate outcome patients, and tested on all 'gray zone' patients.

	AUC	PPV	NPV
Train	0.85	0.93	0.77
Validation	0.82	0.83	0.80
Test	0.67	0.85	0.46

Prediction results refer to one single value, as only one fold was trained. We report the AUC, PPV and NPV scores for the train, validation (patients with determinate outcome) and test ('gray zone' patients) sets.

Supplemental Table 3. Mean number of EEG responses across patients per type of auditory stimulation.

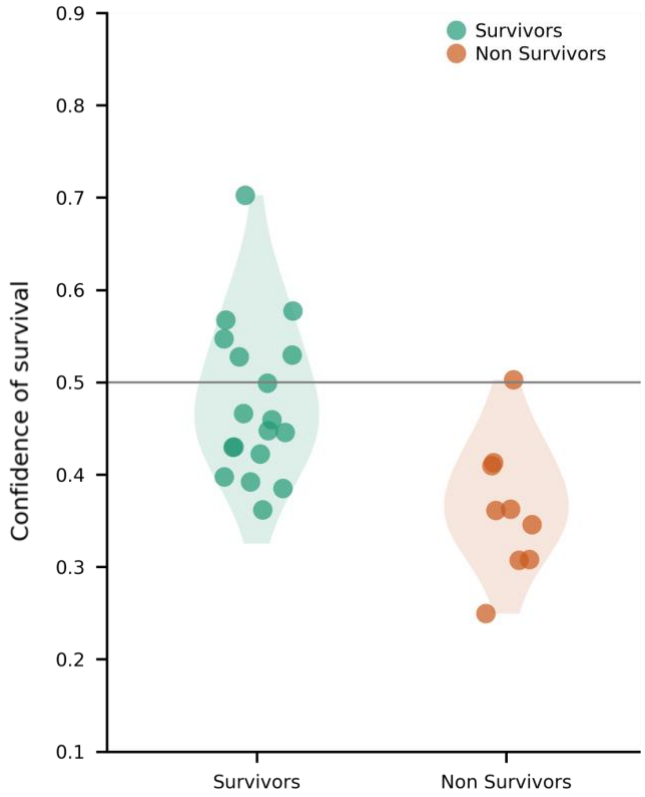
Standard	Duration	Location	Pitch
204.76 ± 8.66	142.46 ± 1.42	141.84 ± 1.48	141.40 ± 1.48

The table reports the mean ± standard error number of trials across patients in our dataset.

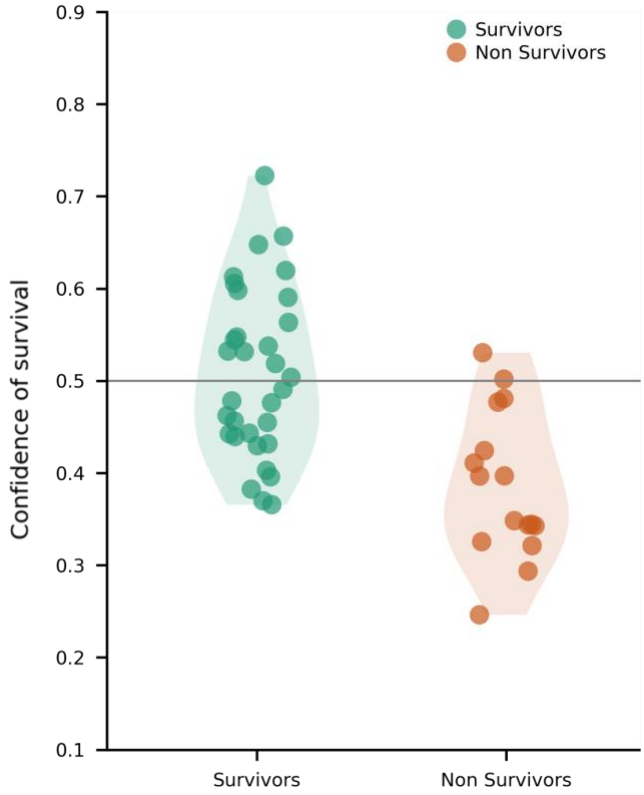
Supplemental Table 4. Mean AUC per stimulation type for patients in the test set.

Standard	Duration	Location	Pitch
0.698 ± 0.035	0.701 ± 0.035	0.702 ± 0.037	0.709 ± 0.038

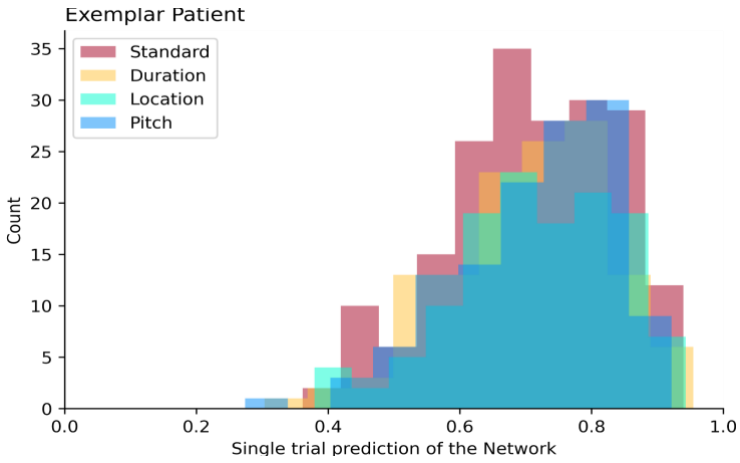
The table reports the mean ± standard error of the AUC of predicting outcome, per type of auditory stimulation for patients in the test set.



Supplemental Figure 1. Confidence of survival assigned by the network for patients in a 'gray zone' that were part of the test set. Confidence scores of survival for the test set, for a network where the test set only contained 'gray zone' patients. The network was trained with determinate outcome and gray zone patients, via a curated cross validation.



Supplemental Figure 2. Confidence of survival assigned by the network for patients in the test set including all 'gray zone' patients. This network was trained with determinate outcome patients only. This figure depicts patients of a 'gray zone' that were only used for testing the network.



Supplemental Figure 3. Distribution of the network's predictions across trials for one exemplar patient. The network's predictions were largely overlapping for the four sound types included in our experimental protocol. The x axis depicts the network's output per single trial, ranging between 0 (non survivor) and 1 (survivor). The data correspond to an exemplar survivor that was correctly classified by the network.