PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Development of a Point of Care System for Automated Coma Prognosis – A Prospective Cohort Study Protocol
AUTHORS	Connolly, John; Reilly, James; Fox-Robichaud, Alison; Britz, Patrick; Blain-Moraes, Stefanie; Sonnadara, Ranil; Hamielec, Cindy; Herrera-Díaz, Adianes; Boshra, Rober

VERSION 1 - REVIEW

REVIEWER	Francesco Lolli University of Florence, Italy
REVIEW RETURNED	19-Feb-2019

GENERAL COMMENTS	I think the the study as has limitation as prognostic study, to be further considered.
	The coma "can be caused by a variety of clinical conditions" as stated, but I could not find any indication on how is reflected in the study method and account for. Inclusion criteria just define the coma state. Exclusion criteria exclude open-head injury, intracranial pathology requiring neurosurgical interventions and other common conditions or treatments such as sedation (medically-induced coma), limiting the general interpretation of the results. Another limitation is the missed patients for withdrawing life-sustenance, not more specified as % occurring in the area. A high percentage would certainly bias the results.
	A coma will last months, and a 30-day window is very narrow. Patients should be followed clinically, and possibly instrumentally, for at least 6 months to ensure the prognostic value of the results.
	The study is exploratory and includes many electrophysiological tests. Can the author select a gold standard test? In my opinion, somatosensory evoked potentials have shown a useful prognostic value.
	Sedation level is not a binomial variable. Many drugs are used in a different dosage and different concentrations or time as a routine. This affects neurophysiological tests. Where is this heterogeneity considered?
	Sixty-four EEG electrodes recording are useful, but not easily applied in neurosurgical patients. How to test and consider these patients?

REVIEWER	Athena Demertzi
	University of Liege, Belgium
REVIEW RETURNED	05-Mar-2019

GENERAL COMMENTS	Review Comments
	Summary: This is a clinical research protocol which aims at predicting consciousness recovery after coma in an automated way using electrophysiology markers. The authors will include 50 patients in acute coma state and will perform all-day recordings at 5 points across a month. They will use different types of evoked and event-related potentials as well as biomarkers from resting state. By means of machine learning they aspire to introduce the pipeline in the clinical setting.
	Merits: The aims of the study are highly important to address medical and ethical issues in comatose conditions. Please, see my comments below on which dimensions I feel that the protocol is not ready enough to be directly applied, mainly due to feasibility reasons, especially of the analysis and implementation part.
	Abstract: Add aims; define GCS acronym; what does it mean that the testing session is shifted? Please, rephrase automatic (i.e. reflex) to automated analysis. Mention what kind of ERPs you will be using
	Strengths: what is a complete as opposed to an incomplete hierarchical investigation? Before performing automated predictions, the authors need to add the step of feature extraction.
	Introduction:
	Please, explain the GCS in details since it's an important measure. What is the time cascade of spontaneous recoveries after coma? How long it lasts in average? Does it depend on etiology?
	Before mentioning the value of the cognitive ERPs, mind to mention the somesthetic EPs leading to a N20 and how this contributes to the emergence of coma. Boveroux et al (Reanimation 2008) provide an insightful decision tree on the good/bad outcome prediction after anoxic coma.
	For the machine learning paradigms, the authors do not mention the recent work from King JR; Engemann & Raimondo; Sitt et al who have provided automated feature extraction and class prediction. These studies are performed in patients with disorders of consciousness, yet methodologically can provide unique insights. As it seems, most of the methods are overlapping, in that case please mention those dimensions and how you will use them in your study.
	Methods:
	What is the recruitment rate of comatose conditions in the Hospital? How did the authors choose the time intervals for the assessments? How did the authors estimate the n=50?
	What is the estimated time depth of the study considering the big number of patients (e.g. for how many years do they expect the study to take place?)
	Could the authors team up with another center to aim at a multicentric study, which can be more efficient in raising the desired number, limit the evaluation time, and reduce costs?

	Is there an open-source dataset available of this kind that the authors might use as a contingency plan?
	Have the authors performed acquisitions yet? In how many patients do they have preliminary results? What is their opinion about 24h recording and patient comfort with the active electrodes?
	Define the acronym "MLAEPs"
	The control group is not optimal, the authors might want to collect data from deeply anesthetized healthy controls which appear closer to a comatose state at least phenomenologically
	Will the order of the tests within each block be counterbalanced? Please describe the GOS.
	Please provide more details on the ML part (feature extraction will follow a feature engineering process, and will be dependent on the user? If yes, mind a blinded process for feature extraction, which classifier and why (please define LFS acronym, it seems it's part of the feature selection and not classification)? With which parameters? How will validation be performed).
	Which biomarkers will the authors consider from resting acquisitions?
	Please, define what you mean by "automated" given the involved of users in the whole process including feature extraction.
	Recommendation: Given my concerns about the feasibility of the study I recommend the authors to revise the protocol to meet the above-mention points.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Francesco Lolli

The coma "can be caused by a variety of clinical conditions" as stated, but I could not find any indication on how is reflected in the study method and account for.

We are attempting to capture as much of the population as possible. We will be recording the different etiologies and accounting for their differences and influence on results. We are not able to draw an inference on the actual numbers expected, but we intend to investigate to what extent our proposed method for predicting coma emergence is independent of etiology.

Inclusion criteria just define the coma state. Exclusion criteria exclude open-head injury, intracranial pathology requiring neurosurgical interventions and other common conditions or treatments such as sedation (medically-induced coma), limiting the general interpretation of the results.

☐ Open-head injuries are in the exclusion criteria to avoid a heterogeneous distribution of electrode densities and other difficulties with EEG setup. We will test these patients but due to montage differences, these data will be stored separately from those patients for whom full data can be obtained.

☐	The reviewer raises a good point regarding intracranial pathologies. However, we have specified that only "those requiring neurosurgical interventions in the past 72 hours" will be excluded. Many of these pathologies, such as severe traumatic brain injury, subarachnoid hemorrhage, intracranial tumors and cerebral edema require high sedation levels and some restrictions on moving patients. After 3 days we should be able to test them if there are no further complications.

☐	Our study is not interested in medically-induced comas. The sedatives involved in those directly affect EEG/ERP results and is a significant confound (assured negative). We argue that it will not help our data collection. However, the numbers of these types of patients will be recorded to indicate how much this category of patient contributes to the overall population.

Another limitation is the missed patients for withdrawing life-sustenance, not more specified as % occurring in the area. A high percentage would certainly bias the results.

The reviewer makes a good point. In a multicentre cohort study in six Canadian level-one trauma centres (in which the Hamilton General Hospital was included), most deaths of patients with traumatic brain injury (70.2%) were associated with withdrawal of life-sustaining therapy, ranging from 45 % to 86.8% across centres.[1] Based on this high percentage, we will be testing patients who are clinically judged to have a poor outcome. We will be testing them at and after the point of life-support withdrawal when possible. That will provide us with improved heterogeneity in our data, as it is unlikely that patients show MMN after support withdrawal (cases of no emergence).

A coma will last months, and a 30-day window is very narrow. Patients should be followed clinically, and possibly instrumentally, for at least 6 months to ensure the prognostic value of the results.

According to Giacino and colleagues (2014), coma typically resolves within 2 weeks and is followed by emergence either into Unresponsive wakefulness syndrome (UWS, also known as Vegetative State), Minimally conscious state (MCS), or better. In severe cases, a coma may last as long as months, but this is not very common. We are focused strictly on coma (not UWS or MCS) with the primary goal of the present study to predict emergence vs. expiry. However, we are also capturing the clinical state upon emergence (e.g., UWS, MCS etc), which would enable further studies into prediction of specific emergence states utilizing the same set of collected data.

A sentence was incorporated in the Introduction section (Background and Rationale) to define properly what coma is and its typical duration.

The study is exploratory and includes many electrophysiological tests. Can the author select a gold standard test? In my opinion, somatosensory evoked potentials have shown a useful prognostic value.

We agree with the reviewer that somatosensory evoked potentials are a useful prognostic tool. However, in this study we will be focusing on the auditory system to make inferences about

prognosis. We would like to note that there is no current gold standard test. One of the reasons for the study is to advance the field to develop a gold standard measure.

Sedation level is not a binomial variable. Many drugs are used in a different dosage and different concentrations or time as a routine. This affects neurophysiological tests. Where is this heterogeneity considered?

That is correct, and that is why we are altering it to include minimal levels of sedation and consult with experts to ensure they have minimal effect on the EEG signal. For instance, high Propofol levels are fully excluded due to their known effects on cortical network activation and EEG connectivity.[2–4] Therefore, we have considered that Propofol treatments should be stopped at least 24 hours prior to the recording session. Essentially, we ensure that sedation is not the only reason they are in coma (not medically-induced).

Sixty-four EEG electrodes recording are useful, but not easily applied in neurosurgical patients. How to test and consider these patients?

They are indeed difficult to apply; however, our lab is highly trained in the application of high-density EEG on patients in a hospital setting (operating rooms and ICUs). For a consistent data collection procedure and access to enough data to draw strong conclusions on captured ERPs (topographical, connectivity-based, and others), we choose the sixty-four EEG. Once a formal model is created and feature selection is applied, a lower-density EEG specialized for this application may be considered.

Reviewer: 2

Reviewer Name: Athena Demertzi

Abstract: Add aims; define GCS acronym; what does it mean that the testing session is shifted? Please, rephrase automatic (i.e. reflex) to automated analysis.Mention what kind of ERPs you will be using

All the points have been addressed in the revised manuscript.

Strengths: what is a complete as opposed to an incomplete hierarchical investigation? Before performing automated predictions, the authors need to add the step of feature extraction.

We assume that a complete hierarchical investigation includes all levels of information processing, taking into account short-latency evoked potentials (BAEPs, MLAEPs) that helps to estimate the integrity of ascending auditory pathways, but also cognitive and late-latency ERPs components such as the MMN, P3, and N400. An incomplete investigation, in our opinion, would be targeting only on either BAEPS, late ERPs components, or resting state.

With respect to feature extraction, space was limited in the strengths and limitations section for the inclusion of further detail. Feature selection, extraction, along with other procedures, are an integral part of the machine learning process and are explained in detail later in the paper.

Introduction:

Please, explain the GCS in detail since it's an important measure.

Incorporated into the new manuscript.

What is the time cascade of spontaneous recoveries after coma? How long it lasts in average? Does it depend on etiology?

Coma state usually lasts a few weeks and transitions to either Unresponsive wakefulness syndrome (UWS, also known as Vegetative State) or Minimally conscious state (MCS).[5] In severe cases a coma may last as long as months, but this is not very common. The outcome of coma is also related to the etiology and it seems to be independent of the physical signs, depth of coma or length of coma.[6] In a meta-analysis estimated with the data of 548 comatose and low responsive patients, the prognosis was worst for patients with anoxia or metabolic encephalopathy and best for trauma or brain surgery.[7] We will be accounting the etiology of the patients to determine whether our proposed method for predicting coma emergence is independent of etiology.

 We incorporated some of these sentences into the revised manuscript.

Before mentioning the value of the cognitive ERPs, mind to mention the somesthetic EPs leading to a N20 and how this contributes to the emergence of coma. Boveroux et al (Reanimation 2008) provide an insightful decision tree on the good/bad outcome prediction after anoxic coma.

The N20 does have some value in predicting poor outcome in anoxic coma. Although there are other views of its overall utility. The larger issue is that our paradigms are concentrating on the auditory responses and in particular the MMN which has proven to be the most useful tool in predicting positive outcomes. The MMN's greatest problem is low sensitivity and our early papers indicate our approach improves that short-coming. Our expansion of the cognitive elements seems to us to be the more effective approach. The reviewer is correct that the N20 has a place in coma research. However, introducing somatosensory methodology to record bilateral N20 (that is the usual "marker" for using the response to identify a likely negative outcome) is simply not feasible in our current environment.

For the machine learning paradigms, the authors do not mention the recent work from King JR; Engemann & Raimondo; Sitt et al who have provided automated feature extraction and class prediction. These studies are performed in patients with disorders of consciousness, yet methodologically can provide unique insights. As it seems, most of the methods are overlapping, in that case please mention those dimensions and how you will use them in your study.

Thank you to the reviewer for some excellent references. They will be very helpful during next steps once the prototype is complete and next steps are in order to achieve proper scalability and online execution. We have included some of these references into the new manuscript.

Methods:

What is the recruitment rate of comatose conditions in the Hospital?

Approximately 5% of the patients present to the emergency department with an altered mental state and 1% of the admissions is due to coma (Kanich et al. 2002).[8] We estimate similar proportion of comatose patients in the emergency department at Hamilton General Hospital.

How did the authors choose the time intervals for the assessments?

The time intervals were chosen to maximize the number of data points being captured while extending the time-window to 1 month. Patients in coma tend to recover less than a month after injury (Bates, 2001). Moreover, the initial date of testing is typically several days after admission, this provides us with a time restriction before a patient is likely to emerge. Thus, we designed the time-windows to have a highest density initially before tapering off in frequency to cover the 5th test at the end of 30-days.

How did the authors estimate the n=50?

This was an estimate made by the clinical experts on the team to account for 2 full years of data collection. Provided the listed inclusion/exclusion criteria and other clinical circumstances preventing data collection, an estimate of 2 patient recruitments per month was made. That was rounded to 50 total.

What is the estimated time depth of the study considering the big number of patients (e.g. for how many years do they expect the study to take place?)

The study is expected to be in the data-collection stage for two years total. Data analysis and reporting is expected to commence 6-months into data collection, extending 1 year after termination of data collection. We incorporated the duration of data collection into the revised manuscript

Could the authors team up with another center to aim at a multicentric study, which can be more efficient in raising the desired number, limit the evaluation time, and reduce costs?

The research team has strong connections in the Hamilton General Hospital that facilitates timely and responsive consenting and EEG testing. Due to logistical reasons, the study is limited to this site. Based on the projected estimates, the authors believe one site to be adequate for the purposes of the present study. However, follow-up studies for validation of the electrophysiological paradigms and the generated prognosis device will require expansion to additional sites.

Is there an open-source dataset available of this kind that the authors might use as a contingency plan?

To the best of the authors' knowledge, no open-source datasets currently present provide the breadth of ERPs/protocols, the EEG electrode density, or the number of patients required for a sound implementation of a machine learning prognostic tool. Thus, data collection is critical for the success of the presented protocol.

Have the authors performed acquisitions yet? In how many patients do they have preliminary results? What is their opinion about 24h recording and patient comfort with the active electrodes?

Several pilot studies were conducted in the same setting resulting in some publications.[4,9] Since the date of submission, an additional subject was also recruited. We have no concerns regarding patient comfort. Primarily, the EEG cap does not introduce any layer of discomfort that is additive to a person in coma. Based on our experience, gel must be applied again at least every 6-8 hours to ensure a good signal during 24h recording. After emergence, testing is only conducted for a couple of hours and is respectful of the patient's condition at the time of testing.

Define the acronym "MLAEPs"

Done

The control group is not optimal, the authors might want to collect data from deeply anesthetized healthy controls which appear closer to a comatose state at least phenomenologically.

The reviewer raises a great point. Our group is actively involved in a study that relates coma and anesthesia.[4] However, we argue that recruitment is to be made significantly harder and might not be possible provided restrictions by the local ethics board. Moreover, while deep anesthesia is more akin to coma, it introduces a confounding variable. Several anesthetics are known to impact both ERPs and resting-state EEG (e.g., Propofol as utilized in the article mentioned above). This would impose severe limitations on our ability to capture an elicited effect of cognitive function that we attempt to train a model to detect. We agree that the control group is not an optimal one, but it permits the validation of our electrophysiological paradigms and facilitates a preliminary step towards the realization of the study's main goal.

Will the order of the tests within each block be counterbalanced?

No. Provided the continuous loop of paradigms, we argue that counterbalancing is not pertinent to the results of this study. Prior pilot work has showed no evidence of carry-over effects, likely due to the 10 minute-resting state between each two paradigms. Moreover, we argue that if there is an observable carry-over effect, it would be beneficial in detecting cognitive function in comatose patients. Ultimately, the decision to utilize static ordering was to simplify the application of a tool that will be executing for 24-hours with multiple interruptions that require resuming from a well-known stopping point.

Please describe the GOS.

This was added to the revised manuscript

Please provide more details on the ML part (feature extraction will follow a feature engineering process, and will be dependent on the user? If yes, mind a blinded process for feature extraction, which classifier and why (please define LFS acronym, it seems it's part of the feature selection and not classification)? With which parameters? How will validation be performed).

Which biomarkers will the authors consider from resting acquisitions?

The reviewer's comments have all been addressed in the revised version. We have explained both feature extraction and feature selection processes more clearly. Features will be selected to be independent of the subject. Local Feature Selection (LFS) is both a feature selection and classification algorithm. The parameters associated with the LFS algorithm are determined automatically .[10] We made some changes into the new manuscript to address the validation methods that could be performed. With respect to biomarkers (features) from resting EEG, we will use a variety of spectral features and connectivity metrics as described in the revised manuscript.

Please, define what you mean by "automated" given the involved of users in the whole process including feature extraction.

"Automated" here entails the finalized tool that enables coma-prognosis, not the process by which to build one. The automation doesn't take place until full realization of the prognostic tool. Automation indicates a continuous monitoring of a comatose patients' EEG that signals whether an ERP (or other resting-state marker) is present or not without requiring an EEG/ERP expert on site.

VERSION 2 – REVIEW

REVIEWER	Francesco Lolli
	University of Florence, Italy
REVIEW RETURNED	10-May-2019

| GENERAL COMMENTS | I find that the points previously addressed were considered. The corresponding sections were amended and clarified and the manuscript can be considered for publication. |