Reviews

# Reviewer #1

**Methods**

This is a retrospective cross-sectional study of large number of patients. I find it excellent that the physicians' evaluations are taken into account in validation of the results.

1. EA burden; if I understood correctly, the presence of EA was analyzed from 2 s EEG periods in a 6 hours time windows, correct? Both the mean and maximum EA fraction was calculated. EA burden was divided into four levels - mild, moderate, severe and very severe. While I understand that this classification is based on the amount of observed EA, I would think that the overall condition of the patient, age, underlying illness etc. affect the "tolerance" of the brain to EA - for example, "moderate" EA of 25-50% could be mild for one patient but severe to another. Comments?

We appreciate the reviewer’s concern about the heterogeneity of the effect of EA burden on the outcome. We account for this apriori by matching on patient’s covariates such as age, disease history, diagnosis, etc. A similar stratification of EA burden was also studied by the following papers in the literature: [citations]

Secondly, the possible spread of EA and type of EA is likely to be associated with the outcome. Have you considered to take this into account in your analysis?

We agree with the reviewer that we summarized different EA categories into a single category. There is often inter-rater disagreement in categorizing Seizure/GPD/LPD/GRDA/LRDA where the agreement rate ranges from 40% to 60% [Figure 8 in Ge et al., 2021]. EA into different classes which complicates the analysis further. Analysis based on different EA classes is a future research direction.

Reference:

Ge, W., Jing, J., An, S., Herlopian, A., Ng, M., Struck, A.F., Appavu, B., Johnson, E.L., Osman, G., Haider, H.A. and Karakis, I., 2021. Deep active learning for interictal ictal injury continuum EEG patterns. *Journal of neuroscience methods*, 351, p.108966.

2. Covariates: Drug responsiveness

As the authors point out, patients respond differently to ASM. For the PK model, the half-lives of ASM were obtained from drug databases and fixed. How long has the critical care lasted in these patients, and do you think those pharmacokinetics of the critically ill patient does not significantly change during the intensive treatment period?

[Brandon/Sahar] need help

1. We agree that there are some gradual changes in both PK and PD parameters over the course of treatment. This is primarily due to organ changes, for e.g., changes in liver or kidney function, or changes in patient’s weight.
2. However, we believe that these variations within PK are small and slowly varying.
3. In a future study, drug dose exploration to estimate PK and PD parameters simultaneously For example, in propofol one can do a ramp [cite].
4. There is a statistical challenge to adaptive changes in PK and PD.

For the PD model, the Hill coefficient and the ED50 represented the patient's drug responsiveness. Usually, the ASM medication of a critically ill patient with ongoing EA is changed/added quite frequently, and the combination/interactions of different antiepileptics (as well as varying kidney and liver function of critically ill patients) affect profoundly their elimination/efficacy. How altering combinations of medication affect the results of PK/PD model used for the patient?

[Brandon/Sahar] need help

Statistical response: large number of parameters - careful controlled trials - response surfacer modeling [cite]

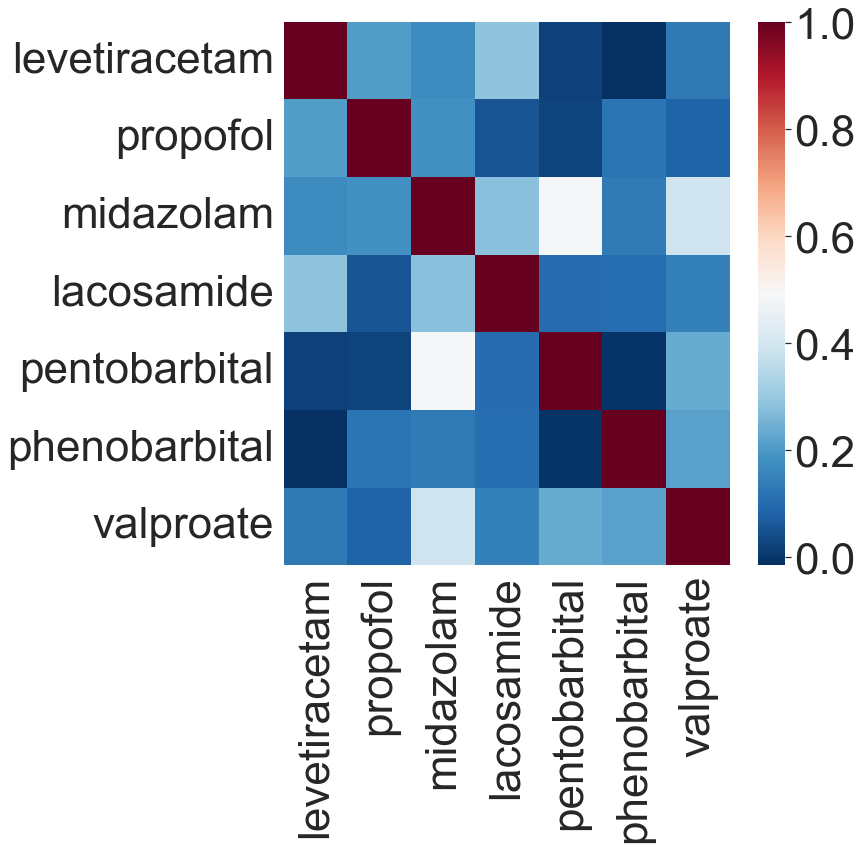
Medical justification:

1. Summarize the change in ASM and the number of simultaneous ASM and the frequency. Histogram:
   1. per patient distribution of ASM (how many patients get n ASM?),

|  |  |
| --- | --- |
|  | **Num. of Patients** |
| **levetiracetam** | 539 |
| **propofol** | 437 |
| **midazolam** | 99 |
| **lacosamide** | 76 |
| **pentobarbital** | 5 |
| **phenobarbital** | 24 |
| **valproate** | 4 |

|  | **levetiracetam** | **propofol** | **midazolam** | **lacosamide** | **pentobarbital** | **phenobarbital** | **valproate** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **levetiracetam** | 539 | 285 | 73 | 64 | 3 | 13 | 28 |
| **propofol** | 285 | 437 | 80 | 36 | 4 | 16 | 19 |
| **midazolam** | 73 | 80 | 99 | 20 | 4 | 6 | 13 |
| **lacosamide** | 64 | 36 | 20 | 76 | 2 | 4 | 10 |
| **pentobarbital** | 3 | 4 | 4 | 2 | 5 | 0 | 1 |
| **phenobarbital** | 13 | 16 | 6 | 4 | 0 | 24 | 1 |
| **valproate** | 28 | 19 | 13 | 10 | 1 | 1 | 41 |

* 1. within the patient? (how many 6hr epochs have a combination of ASMs)
  2. In any given hour and given ASM = 1, then what is the probability that next window has ASM=1, ASM > 1 and ASM < 1

1. Most of the ASMs are Keppra, lacosomide, and propofol for which there is no known interactions in lit [cite].
   1. heatmap of correlation of drugs are used together
   2. The ones that do interact are not common in our study cohort

**Results**

3. The message is clear for the main results -the monotonically increasing risk of poor outcome with increasing EAmax and increased of risk of poor outcome with Emean above the moderate level. In Fig 3 the effect is shown to be pronounced if the patient has CNS infection or toxic metabolic encephalopathy. Were these the only CNS pathologies that showed such pronounced effect? One would suspect that any simultaneous or prior major CNS pathology, such as degenerative brain disease or any serious acute brain lesion would behave similarly?

CNS Infection

[Brandon/Sahar: What do you think about the common about other CNS pathologies? Say it’s interesting and worthy of further discussion?]

4. Interpretable matched group analysis: The validity of the effect estimation is dependent on the validity of the matching groups. This is the major point, and I struggled to understand how to evaluate pros and cons of this method Could you elaborate this more in the MS (and not only in Suppl. material)?

Thank you so much for the comment. We have added more descriptions of the interpretability of the matched groups. To delineate further, the medical doctors evaluated a random subset of matched groups using the “doctors’ notes”. Note that, doctors’ notes or the information in them were not explicitly used by our method for matching. This allowed doctors to reason about the existence of any unobserved confounder that could corrupt the analysis and following inference.

5. Nevertheless, although it was quite difficult to evaluate all possible caveats of MALTS, the possibility to evaluate the results by looking back at the raw data unit by unit appears very rational. However, I found some of the doctors' estimates puzzling (Table 2, matched group 1, last patient); patient was admitted due to generalized convulsive seizures multiple times per day, but between seizures was cognitively normal. All three evaluated poor prognosis (mRS 4 or worse) in 80-100%? Does that mean that they were expecting a bad outcome with 80-100% chance? Cannot be or they know something I don't. On the other hand, the lower panel states that "patients with refractory epilepsy have a very good chance of recovering" - which is true- please clarify.

[Brandon/Sahar need your help]

**Discussion**

6. Clinical Implications. The authors present two primary implications for the treatment of EA:

- intense bursts of EA burden (captured by EAmax), even if relatively brief (6 hours) lead to worse outcomes

- sustained periods of EA < 50% has minimal effect, but EA ≥ 50% causes a worse outcome.

From clinical point of view, the first one probably does not change much the treatment practice of critically ill patient; intense EA, even for a quite short time, usually elicits vigorous attempts to control it, whereas the need to react to longer lasting but less intense EA is more difficult decision. However, patients' underlying disease, age, prognosis etc. has a major role in these decisions.

Thank you for the comment. We agree. We have added your discussion into our discussion section.

7. Results in context. Although several prior studies have shown that EA is associated with worse outcomes, the criticism has been directed at the lack of adjustment of the effect of aggressive ASM. While I understand that anesthetics such as propofol and tiopenthal maybe harmful, I do not understand why antiepileptic treatment -particularly the newer ones such as levetiracetam-should lead to poorer outcome. But as long as such discussion continues, this kind of analysis is necessary and clearly shows that intense/prolonged EA results in poor outcome if not treated adequately.

Thank you for your comment!

# Reviewer #2

I find this study well conducted from a methodological point of view. Analyses are very detailed and important sensitivity analyses have been carried out. There are some aspects that need to be addressed.

**Major comments**

- "Confidence intervals are derived via bootstrapping". Please elaborate on how matching was dealt with when bootstrapping.

Thank you for your question. We added a description of this in the supplementary material of the paper. We used 50 bootstrapped samples to estimate the variance. For each bootstrapped sample, we learned a new distance metric and created new match groups using the learned metric. The final matched group that was analyzed by the doctors was the union of the match groups across all bootstrap runs.

- The authors use two definitions of exposure, Emean and Emax. It would be informative to see the correlation between the two (including a bivariate plot) to understand how much the mean is influenced by extreme values. On the other hand, would not be reasonable in this case to use the median instead of the mean as a measure of central tendency as it's not influenced by extremes?

Thank you for your comment. We added a plot showing the correlation between Emean and Emax across patients.

- "Therefore, the missingness pattern can be considered as not influencing our results" and "Selection bias was checked by whether data was missing at random". I would be more cautious with such statements. Missing at random (MAR) and missing not at random (MNAR) cannot be tested from the data. One can test missing completely at random (MCAR) vs MAR/MNAR ([doi.org/10.2307/2290157](http://doi.org/10.2307/2290157)) but not the other two individually. There are some solutions (see for example [doi.org/10.1191/0962280206sm44](http://doi.org/10.1191/0962280206sm44) and [arxiv.org/abs/2105.12921](http://arxiv.org/abs/2105.12921)) but what the authors did is not one of them. I think the authors should address the missing data issue more thoughtfully.

Thank you for your comment. As you state, we are not doing a test of MCAR vs MAR/MNAR but rather a confirmation that the excursion criteria is not unduly creating a biassed patient cohort. We have included some more detail on this in section XX of the appendix.

Other comments

- "For statistical efficiency and interpretability, we bin the EA burden into 4 levels". It seems such a broad statement. In what way discretizing a continuous variable is beneficial to "efficiency" or "interpretability"? I would argue that there's a loss of efficiency as you have 4 - 1 = 3 coefficients instead of 1. As for interpretability, that depends on the functional form of the relationship.

From a matching standpoint, one can think of a continuous treatment as having “uncountably infinite counterfactuals” and finding an appropriate match becomes more difficult as the number of categories increases while simultaneously increasing the chances of a violation in the positivity assumption that is required for causal inference. Thus, we determined that our 4 levels provided a good balance between statistical estimability and resolution in EA levels.

- Can the authors please replace Table S3 with a 4 by 4 table that shows the frequencies of combinations of Emax and Emean, plus marginals?

Thank you, we added a 4x4 table showing the frequency of Emax and Emean.

- EEG pre-processing and artefact detection in supplementary. Please provide reference(s) for criteria.

[Haoqi]

- In supplementary, please include paragraph or section to define \*all\* the symbols for variables used in the document. Currently some of those symbols appears without definition at first occurrence.

Thank you, we added a paragraph defining all the symbols.

- Figure S11. Please use same range (e.g., from 0.525 to 0.705) for scale of both plots.

Thank you, we updated the figures to match the scale.

# Reviewer #3

I have several methodological comments that need to be addressed.

**Automated EEG analysis and annotation**

This is a clear strength of the study, which allows to avoid the pitfalls of human EEG interpretation (inter-rater and intra-rater inconsistency).

Why exclude all EEG recordings <2h? Recent studies have shown that using for instance the 2HELPS2 score (Struck et al., JAMA Neurol 2017) or TERSE algorithm (Struck et al., Ann Neurol 2018) allows to determine the minimal duration of EEG recording to exclude further EAs with a reasonable false negative rate.

Thus, in selected patients, a EAs-free 1 or 2h recording could be acceptable and included, while in others a recording of intermediate duration (>2h but shorter than 2HELPSB or TERSE would have required) and a low burden EAs might actually not reflect the real (and future) exposure

I would suggest the authors to select all EEG that fulfil 2HELPS2B/TERSE-predicted minimal duration. Data to use these tools (GCS, seizure at presentation/epilepsy, early EAs) are already available.

Thank you for your comment. However, the 2HELPS2 paper only included patients that had >6 hours of EEG recordings, and the TERSE paper included those with >24 hours of recording and no interruptions greater than 2 hours. Based on this, we believe that it is not unreasonable to exclude those with less than 2 hours of EEG. In addition, a large focus of this paper is on the Anti-seizure medications that were given. Based on our discussions with the clinicians, it seemed that anti-seziure medications at MGH were monitored and or adjusted around every 3-6 hours. Including those with <2 hours would be too short for this.

From the Methods, it is unclear if BIPDs were included. Were they?

Thank you, due to the briefness of BIPD patterns, our automatic EEG classifier was not designed to identify BIPDs. It is possible that some patients exhibiting BIPDs are included in our cohort, and would thus be a good direction for future study.

Current definitions of electrographic SE include duration >10 min or hourly burden >20% (the latter based on 1 study). Could the authors provide a more granular analysis of the relation between EAs burden and probability of good outcome (Figure 2), especially for lower burden (0-25%). It would very interesting to see where the inflexion point is (20% or below)?

This will be an ideal extension of our work. However, due to sample size limitations, our study cannot answer this question.

**Confounders**

The authors already included many confounders but a few important are still missing.

For instance, the presence of secondary neurological injury on top of the primary admission diagnosis (ICP crisis in TBI, hematoma expansion in [B, rebleeding or DCI in SAH, brain edema from large ischemic stroke, ICH in patients with liver failure, PRES in patients with renal failure, stroke in septic shock, etc.) needs to be accounted for as it might be a cause of both EAs and worse outcome.

The other very important variable to account for is any decision to withhold or withdraw treatment, which could occur in response to an EEG with EAs (anoxic cases, for instance).

More generally, it would be interesting to see if the EAs burden-outcome relationship is similar across broad etiological categories (acute brain injury, acute systemic illness, seizure/epilepsy-related disorder, anoxic brain injury).

[Brandon help]

**Treatment intervention**

How many patients with EAs and acute systemic illness received non-sedating AEDs? I find hard to believe that many of them received such medications. How is it possible then to model and infer the effects of AEDs in this subgroup. Similarly, it is likely that sedating AEDs (propofol, midazolam) would have been administered for sedation purposes (and at sedating, not anti-seizure, doses) in most of these patients, again limiting the ability to model their anti-seizure effect. For instance, it would seem impossible to me to infer lack of response to a drug that was not administered to achieve that response.

Thank you for your comment. We have added a table including counts of the number of patients with EA (n=??) and acute system illness (n=??) who are receiving non-sedating AEDs.

We have added it as a limitation in the Discussion section that the sedating AEDs (propofol, midazolam) were indeed administered for either anti-seizure or sedation purpose. However, we do not have that detailed information of the intention of the treatment, therefore the inferred effect may not be specific to anti-seizure effects.

**Pharmacokinetic/pharmacodynamic modeling**

This is the part of the work that raises the most skepticism.

How were co-medications and organ failure accounted for? This is currently not clear to me. For instance, the half-lives that were used correspond to an individual with normal drug metabolism. This was certainly not the case for many patients in the cohort (e.g., those with renal or hepatic failure, or shock).

We did not account for interaction during co-medications.

The pharmacodynamic model was estimated for each patient individually. The estimated parameters are conditioned on patient specific characteristics as well as patient’s medical history. This allowed us to account for organ failure such as kidney or liver failure.

For pharmacokinetics, while our approach do not explicitly account organ failure, we do account for organ failure in the matching procedure where we explicitly look for matches between similar patients that have organ failure.

For some drugs, especially sedating ones (propofol, midazolam) their half-live increases with prolonged exposure.

We acknowledge that the sedating drugs can be better modelled using three-compartment model, rather than one-compartment model, to enable such adaptive half-life [Cite the propofol PK adaptive modeling paper]. We have put this important point into the limitation part of the Discussion section.

For others, such as levetiracetam or phenobarbital, the serum/plasma concentration live does not reflect the concentration of the drug at the level of its end target, the brain.

We agree the serum/plasma concentration is not equal to the brain concentration. However, the combination of PK (only models serum/plasma concentration) + PD (correct the imprecision in PK by fitting to EA burden, a reflection of the brain) can alleviate this problem.

In a nutshell, I found the modelling oversimplistic and unrealistic. I understand it is a model but it can hardly reproduce the complexity of PK/PD in critically ill patients. Since it is an important part of the work (not accounting for this substantially affects the EAs burden-outcome relationship; Supp Fig 8b), it should be more carefully performed.

[Brandon help]

# Reviewer #4

Some specific minor concerns are listed below:

**Introduction**

Epileptiform activity, as well as ictal-interictal continuum, are relatively vague terms that are not well known outside the ICU EEG literature. It would be useful to state specifically which EEG patterns are the subject of this analysis (LRDA, GPD, LPD) in the introduction and distinguish them from other patterns such as sporadic interictal discharges.

We added the clarification and definition of to the introduction of the paper

**Methods**

It is curious why phenytoin/fosphenytoin not included in the PK/PD models as it commonly used IV ASM. Is it simply due to local practice patterns?

Thank you for your comment. In the patient cohort we study, the frequencies of drug administered are as follows:

|  |  |
| --- | --- |
|  | **Num. of Patients** |
| **levetiracetam** | 539 |
| **propofol** | 437 |
| **midazolam** | 99 |
| **lacosamide** | 76 |
| **pentobarbital** | 5 |
| **phenobarbital** | 24 |
| **valproate** | 4 |

As these counts for these drugs were much lower than other ASMs, we excluded them from the study.

We have also confirmed with our local neurologists that it is a local pattern.

[Brandon / Alice Lam: is this a local practice pattern?]

It is not clear ED50s were estimated in the cases when multiple medications are employed simultaneously or in short succession

The ED50s for multiple drugs were estimated simultaneously however our model did not account for drug interactions. This modelling choice was made for two reasons. First, looking at the top 5 most common drug combinations given in our cohort, [see table below], to the best of our knowledge, we are unaware of substantial interactions between these drugs that would lead to a large change in ED50s [This claim has to be checked]. Second, concerns the feasibility of estimating interaction parameters. Fitting a model with an interaction effect for ED50 would grow the number of parameters at least quadratically, which makes consistent estimation of the interaction effects challenging.

# Reviewer #5

**Main criticism and questions:** The idea and study set up seems intuitive at first, but the presentation is confusing and complicated for a non-mathematical clinician. There are omissions in background, methods and results that include important points: eg it is not specified that these are all critically ill patients; mechanical ventilation rate with use of propofol and midazolam for sedation rather than for EA (and its potential implications) is not addressed. The discussion is similarly difficult to follow, and with the limitations, the clinical implication seems an overstep.

We have clarified that all patients were patients admitted to ICU (defined as critically ill) in Line ??. We have modified the flowchart figure to indicate they are solely ICU patients.

We have added it as a limitation in the Discussion section that the sedating AEDs (propofol, midazolam) were indeed administered for either anti-seizure or sedation purpose. However, we do not have that detailed information of the intention of the treatment, therefore the inferred effect may not be specific to anti-seizure effects.

The discussion, limitation, and clinical implications have been polished by another professor in clinical epilepsy research, [Dr. Lawrence Hirsch or Dr. Jennifer Kim], to make it more readable for the clinical audience.

Criticism in detail:

Introduction:

- A hypothetical RCT could treat or not treat EA - please correct the suggestion to "induce EA"

Corrected.

- The last two sentences of the introduction seem to be methods

[Defer to Cynthia]

- Please explain better the importance of drug responsiveness, especially in the setting of purely observational data where we cannot know whether time might have played a role in resolution of EA

Thank you for your comment. We agree that there can be heterogeneity in drug responsiveness across patients. We account for this by estimating fitting a patient-specific pharmacodynamic (PD) model where parameters such as ED50 or Hill coefficients capture a patient’s drug responsiveness. We adjust for this by matching patients with similar PD parameters along with medical history and diagnosis. Further, the fitted PD model also can be used to estimate a counterfactual EA burden when no drugs are administered (this is analogous to letting EA resolve itself without ASMs)

Methods:

- You included EEG by reports only (without verification/standardization of reads) since 2011 - with ongoing standardization for ICU-EEG nomenclature (eg Hirsch 2013: 2012 version) - how did you mitigate reporting changes?

[Brandon]

- While apparent that not all ASM could be captured, EA may often be treated with an enteral ASM - with significant effect (eg pregabalin eg PMID: 29719278, or gabapentin, fycompa)

As these counts for these drugs were much lower than other ASMs, we excluded them from the study (Please see Table). We have also confirmed with our local neurologists that the common use of leveciracetem and propofol and the uncommon use of other ASMs including gabapentin and fycompa is a local pattern in this cohort.

- How did you ensure that propofol, midazolam were not for sedation but for EA?

We have added it as a limitation in the Discussion section that the sedating AEDs (propofol, midazolam) were indeed administered for either anti-seizure or sedation purpose. However, we do not have that detailed information of the intention of the treatment, therefore the inferred effect may not be specific to anti-seizure effects.

- Overall, the methods section is difficult to follow, a graphic might be helpful

Moved a graphic illustrating the pipeline from the appendix to the paper.

- Are these solely ICU patients?

Yes. These were solely ICU patients admitted between 20##-20##. We have modified the flowchart figure to indicate they are solely ICU patients. We have added a line in the paper [where?] to emphasise this.

Results

- No comparison of missing data was provided for the group with missing covariates

A comparison of missing data for the group with missing covariates was performed in the appendix in Figure S?. We have now moved this from the appendix to the body.

Discussion

- the discussion starts with a clinical implication that is deducted - how about summarizing the hard data of what you found, first?

We have included several lines at the beginning starting with a summary of the numerical results before going into clinical implications.

- You point out that treatment of EA varies widely - you present a single center study and deduct a clinical implication about treatment. This oversteps - first, would have to verify that findings are similar in other institutions where treatment patterns may be different

We have toned down the language and added limitations concerning the single-center observational nature of our analysis.

- The summary sentence is the most clear in the discussion

Minor: use of present tense throughout is confusing

Changed to past tense.

# Reviewer #6

**MAJOR:**

Uncertainty estimates are not emphasized sufficiently in the reporting of the results. For example, confidence intervals are not provided for the 16.7% shift in Emax and the 11.2% shift in Emean reported as the main results (the closest thing is the error bars for individual group means in Figures 2A and 2B). In addition, no uncertainty estimates are given for the purported treatment effect variation reported in Figure 3; in the absence of this information, and given the size of the error bars in Figures 2A and 2B, I'm not convinced that any of the reported variation (e.g. the claim that "sex does modify the risk" p. 6 third full paragraph) is anything besides sampling noise.

To account for this, we have included a description of the bootstrapping procedure that was employed to get the estimates.

I am not seeing how PK and PD parameters were obtained for patients without ASM. Based on the Pharmacological Modeling section in the supplement, the model appear to require ASM dose as a predictor but all these subjects will have a dose of zero. Were PK and PD parameters indeed obtained for non-ASM users? If so how? If not, how did the matching metric deal with PK/PD parameters only being available for some of the groups to be matched?

Thank you for the question. Indeed, it is impossible to estimate PKPD parameters for patients without ASM. However, we do not have any patients who were not administered with any ASM (and this is expected because these are patients in ICU where ASM treatments are very common). Similar to EA, we binned each administered ASMs into two groups: (a) minimal or low and (b) significant. The bins were chosen using the estimated median ED50 - where “minimal” is defined as any drug dose that is less than half of ED50, else it is considered “significant”. If all the ASMs administered for each patient were in “minimal” category then we characterize that patient’s overall ASM regime as “untreated”, else the patient is “treated”.

Table 2 suggests that not all of the 4x2 conditions of interest are contained in each matched set. I am not understanding how counterfactual estimation works for matched sets that are missing the desired condition; e.g. if a matched set is missing a non-ASM patient with low (< 25%) EA burden, does it simply make no contribution to the average for the low EA burden groups in Figures 2A and 2B? More broadly, it would be useful to understand the general size and composition of the matched sets created. For example, I do not see the number of matched sets formed reported anywhere.

Thank you for this question. The size of the matched sets is indeed an important aspect to consider. To this end, we have added in Appendix # a table illustrating the sizes of the matched sets that were made.

While I really like the idea of having neurologists validate matched sets using chart data, I don't find the interpretation of Table 2 on p. 6 very focused or complete. In particular:

- In order for "validation" to be meaningful, it is important for the reader to have a sense for what both success and failure would look like. It's not clear from this discussion what the authors would have viewed as a failure of validation, or what the neurologists might have viewed as a "problematic source of confounding."

We have clarified that:

Successful validation is: (1) clinician has a qualitative and subjective judgement of being similar; and (2) the outcome prognosis and EA propensity are similar, unless the reason can be explained in the patient’s summary.

Failed validation is: the reviewing clinician finds an obvious difference in either outcome prognosis and EA propensity that cannot be explained by the covariates used for matching and patient’s summary.

The "problematic source of confounding" viewed by neurologists include disease conditions or genetic background that are not included in the dataset.

- The authors emphasize that some matched sets are more closely matched than others. While this may be useful to know, it doesn't seem particularly related to the stated goal of validation, and I'm not sure how it is supposed to shape our understanding of the study results as a whole.

Thank you for your question. The tightness of the matched groups can be understood as an indication of how distant one would have to look to find a similar patient which received a different treatment. A matched group that is abnormally large can be an indication that the patients within the matched group may not be comparable to each other. Thus, we explain in further detail in section # of the appendix how we dealt with this issue via pruning of the matched groups.

- The discussion focuses on similarities/differences within each matched set on APACHE scores, age, and presence of acute neurological injuries, all things that could have been examined without reference to charts and that the matching algorithm seems to have been able to work with directly. In contrast, there is no discussion of the rough estimates of EA propensity and outcome risk generated by the clinicians, which would seem to be the natural quantity to use for validation (since it was not available to the matching algorithm).

Thank you for your astute comments. Due to this and other questions about the goals of the chart review, we have rewritten parts of the chart review section to elucidate this.

**MINOR:**

Typo p. 5, 3rd full paragraph: "their estimated chance of experiencing a high EA burden and the chance of a poor outcome and chance"

Fixed.

The legend to Table 2 claims that the tightness ordering produced by the chart review "matches an automatic measure of tightness produced by MALTS." This measure should be explained and reported for the matched sets in question, in the Supplement at least.

Along with the matched groups, MALTS also returns the diameter of the matched groups. We have added this measure of tightness to each matched group in Table 2.

I assume that the "potential outcome" label on the y-axis of Figures 2A-B refers to counterfactual probability of poor outcome in the absence of ASM, but a more informative label would help a lot. I also assume that the dotted line in each plot is the study-wide average but this should be clarified as well.

Thanks for the comment. The dotted line is *not* the study-wide observed average, rather it is the baseline potential outcome if the patients experienced EA burden was “mild”. We specifically used potential outcomes instead of counterfactuals because some of the potential outcomes are actually observed and are not counterfactuals.

Were the subgroups in Figure 3 created in advance based solely on scientific judgment or chosen in some way from the data?

The features to search for interesting subgroups were chosen apriori based on patients’ hisotry and diagnosis. However, the exact subgroups were identified via post-hoc analysis of conditional average potential outcomes after the estimation.

In the "Generalizability" section of the Discussion, I would appreciate a comment on population to which it is hoped that results will generalize. Is there any evidence to suggest the selected patients at Mass Gen are representative of the broader population of interest? For examples did any of the multiple studies referenced have notably distinct clinical populations but find similar patterns?

Added a comment on the population we hope to generalize to.

Possible typo: in the last cell of Table 2, the neurologists' note states, "all of the patients have at least a moderately high chance (>=60%) of a good neurological outcome." However, the first column of the table appears to show that all patients have a poor outcome probability of at least 40% under all the neurologists' judgments.

Thank you for this catch, this was indeed a typo. It is now fixed.

Supplement page 5 second full paragraph: Several notations are used without introduction, notably Wbar and delta, and the process by which quantities are averaged within/across matched groups to estimate outcome probabilities is not clear from the verbal description. Please provide full definitions and consider giving a formula for the conditional probability estimates.

We have provided the full definition of every notation, including Wbar and delta. We have clarified that ?, ?, and ? terms are obtained from averaging within/across matched group. We have provided the formula for conditional probability estimates.

# Reviewer #7

**MAJOR**

1. While the MALTS algorithm develops and perhaps improves the ideas behind the k-nearest neighbors and perhaps provides a more understandable "black box" approach to analyzing this data, it does not, as far as my understanding goes, solve the issue of time-varying confounding. Especially when patients are undergoing continuous EEG monitoring and the treatment is tailored based on this monitoring, confounders have values that change over time. Therefore, I am not entirely convinced that the causal effects estimated are accurate. I of course understand that the data is cross-sectional and not longitudinal, but this limitation has more far-reaching consequences and should be discussed, especially in terms of the certainty that can be attributed to these results.

Thank you for your astute point. Indeed the question of time-varying confounding is a long running and difficult problem. In order to study this problem in a real life, finite data scenario, we thus make the assumption that the relationship between ASM and EA activity is directly governed by the mechanistic model and other external factors such as age and sex affect outcome of discharge probability in ways that can be controlled for via matching. Thus, we conceive of both the EA and ASMs not as time-varying confounders but rather as exposures and are thus not interested in the difficult problem of integrating out time-varying confounders.

2. Delirium is also a cause of decreased GCS and a very difficult issue in the ICU, were there any steps undertaken to delineate patients with delirium from patients with decreased consciousness of organic causes? I could not find this in the analysis.

We agree delirium is an important factor in ICU. In the updated patient characteristics table, we have added the number of patients who had experienced delirium at baseline before adimtting to ICU [haoqi]. The information about the cause of delirium is not documented. We also acknowledged the limitation of not considering delirium in the Discussion section.

3. While the MALTS has been the subject of some research up to this point, it is an approach that one can hardly call robustly validated. Given that it has not been used in this way on large datasets, would the authors still argue that these results are sufficiently robust or is this report another step in improving the MALTS? In fact, it would be highly relevant if the authors could discuss the dificulties in implementing the MALTS algorithm in this data, in order to understand the caveats of the analysis better.

Thank you so much for the question. MALTS was particularly chosen for this analysis because: (1) it is a flexible non-parametric matching method that can estimate conditional average effects, and (2) unlike other state-of-the-art methods, MALTS is also auditable and interpretable where doctors can directly look at the raw units matched group to validate the analysis. The existing MALTS package available on GitHub is developed for scenarios with two treatment arms: 0 and 1. In our case, we have 8 treatment arms. Extending the framework for n-ary treatment was fairly straightforward where in we add additional constraints on matched groups having at least one unit from each treatment arm of our interest i.e. units with minimal ASM administration and different degrees of EA burden.

4. In the Limitations section, the detection of EA through the authors' deep learning algorithm is not discussed as a potential limitation. What is the chance of false positives and how could these have influenced results?

We have added the discussion of the deep learning algorithm as a potential limitation. The accuracy (according to the paper, although not measured as false positive) for each EA type is: seizure=30% (human inter-rater agreement (IRA)=42%), LPD=40% (IRA=58%), GPD=61% (IRA=61%), LRDA=30% (IRA=40%), GRDA=40% (IRA=40%), and Other=75% (IRA=75%) [Ref: [doi.org/10.1016/j.jneumeth.2020.108966](https://doi.org/10.1016/j.jneumeth.2020.108966)].

Importantly, we have merged into two classes: EA = seizure+LPD+GPD+LRDA and non-EA = GRDA+Other, based on the clinician’s opinion. Therefore the binary false positive, as indicated in the confusion matrix in Figure 9 in the Supplementary Material, is 19%.

To study the influence of the EA deep learning algorithm, we have already performed sensitivity analysis where we assume there is a systematic shift the EA burden prediction. In Figure 10 in the Supplemenatry Material we show that the result is robust to the shift.

5. While Table 2 is very relevant, the way it is presented at this point, it is quite difficult to understand at a first glance. It took several read-throughs and back-and-forth rounds between the text and the Table for me to understand what is going on there. Could the legend be written using more detail, so the table in itself is easy to understand?

Thank you for the comment, we have added a legend to table 2.

**Minor**

6. A STROBE checklist should be included as a supplement, for ease of use

Thanks. We added it to the supplementary.

7. It would be useful if the authors mentioned in the end of the Discussion what the ideal data set in order to investigate this clinical issue would look like, to stimulate international cooperation and data-sharing on this topic in the future.

[Defer to Brandon / Cynthia / Alex]