Supplementary Appendix for *Why Interpretable Causal Inference is Important for High Stakes Medical Decision Making in Neurology and How to Do It*

-

March 15, 2022

Appendix A Data Summary

Table 2: Covariates (C) being matched

Variable	Value
Age, year, median (IQR)	61 (48 – 73)
Male gender, n (%)	475 (47.7%)
Race	
Asian, n (%)	33 (3.3%)
Black / African American, n (%)	72 (7.2%)
White / Caucasian, n (%)	751 (75.5%)
Other, n (%)	50 (5.0%)
Unavailable / Declined, n (%)	84 (8.4%)
Married, n (%)	500 (50.3%)
Premorbid mRS before admission, median (IQR)	0 (0 – 3)
APACHE II in first 24h, median (IQR)	19 (11 – 25)
Initial GCS, median (IQR)	11 (6 – 15)
Initial GCS is with intubation, n (%)	415 (41.7%)
Worst GCS in first 24h, median (IQR)	8 (3 – 14)
Worst GCS in first 24h is with intubation, n (%)	511 (51.4%)
Admitted due to surgery, n (%)	168 (16.9%)

Cardiac arrest at admission, n (%)	79 (7.9%)
Seizure at presentation, n (%)	228 (22.9%)
Acute SDH at admission, n (%)	146 (14.7%)
Take anti-epileptic drugs outside hospital, n (%)	123 (12.4%)
Highest heart rate in first 24h, /min, median (IQR)	92 (80 – 107)
Lowest heart rate in first 24h, /min, median (IQR)	71 (60 – 84)
Highest systolic BP in first 24h, mmHg, median (IQR)	153 (136 – 176)
Lowest systolic BP in first 24h, mmHg, median (IQR)	116 (100 – 134)
Highest diastolic BP in first 24h, mmHg, median (IQR)	84 (72 – 95)
Lowest diastolic BP in first 24h, mmHg, median (IQR)	61 (54 – 72)
Mechanical ventilation on the first day of EEG, n (%)	572 (57.5%)
Systolic BP on the first day of EEG, mmHg, median (IQR)	148 (130 – 170)
GCS on the first day of EEG, median (IQR)	8 (5 – 13)
History	
Stroke, n (%)	192 (19.3%)
Hypertension, n (%)	525 (52.8%)
Seizure or epilepsy, n (%)	182 (18.3%)
Brain surgery, n (%)	109 (11.0%)
Chronic kidney disorder, n (%)	112 (11.3%)
Coronary artery disease and myocardial infarction, n (%)	160 (16.1%)
Congestive heart failure, n (%)	90 (9.0%)
Diabetes mellitus, n (%)	201 (20.2%)
Hypersensitivity lung disease, n (%)	296 (29.7%)
Peptic ulcer disease, n (%)	50 (5.0%)
Liver failure, n (%)	46 (4.6%)
Smoking, n (%)	461 (46.3%)
Alcohol abuse, n (%)	231 (23.2%)
Substance abuse, n (%)	119 (12.0%)
Cancer (except central nervous system), n (%)	180 (18.1%)
Central nervous system cancer, n (%)	85 (8.5%)
Peripheral vascular disease, n (%)	41 (4.1%)
Dementia, n (%)	45 (4.5%)
Chronic obstructive pulmonary disease or asthma, n (%)	139 (14.0%)

Leukemia or lymphoma, n (%)	22 (2.2%)
AIDS, n (%)	12 (1.2%)
Connective tissue disease, n (%)	47 (4.7%)
Primary diagnosis	
Septic shock, n (%)	131 (13.2%)
Ischemic stroke, n (%)	85 (8.5%)
Hemorrhagic stroke, n (%)	163 (16.4%)
Subarachnoid hemorrhage (SAH), n (%)	188 (18.9%)
Subdural hematoma (SDH), n (%)	94 (9.4%)
SDH or other traumatic brain injury including SAH, n (%)	52 (5.2%)
Traumatic brain injury including SAH, n (%)	21 (2.1%)
Seizure/status epilepticus, n (%)	258 (25.9%)
Brain tumor, n (%)	113 (11.4%)
CNS infection, n (%)	64 (6.4%)
Ischemic encephalopathy or Anoxic brain injury, n (%)	72 (7.2%)
Toxic metabolic encephalopathy, n (%)	104 (10.5%)
Primary psychiatric disorder, n (%)	35 (3.5%)
Structural-degenerative diseases, n (%)	35 (3.5%)
Spell, n (%)	5 (0.5%)
Respiratory disorders, n (%)	304 (30.6%)
Cardiovascular disorders, n (%)	153 (15.4%)
Kidney failure, n (%)	65 (6.5%)
Liver disorder, n (%)	30 (3.0%)
Gastrointestinal disorder, n (%)	18 (1.8%)
Genitourinary disorder, n (%)	34 (3.4%)
Endocrine emergency, n (%)	28 (2.8%)
Non-head trauma, n (%)	13 (1.3%)
Malignancy, n (%)	65 (6.5%)
Primary hematological disorder, n (%)	24 (2.4%)

A.1 Anti-Seizure Medications

Six drugs were studied: propofol, midazolam, levetiracetam, lacosamide, phenobarbital, and valproate. Propofol and midazolam are sedative antiepileptic drugs (SAEDs) which are given as continuous infusion, while the others are non-sedative antiepileptic drugs (NSAEDs) which are given as bolus. Only the period when there is EEG recording is used. The dose is normalized by body weight (kg). We use the half-lives from the literature (see Table 3) for calculating the drug concentrations $D_{i,t,j}$ in the blood using the PK model.

Table 3: Half life for the anti-seizure medications used in the PD modeling.

Drug	Half Life
Propofol	20 minutes
Midazolam	2.5 hours
Levetiracetam	8 hours
Lacosamide	11 hours
Phenobarbital	79 hours
Valproate	16 hours

A.2 Binning of EA Burden

For statistical efficiency and interpretability, we bin the EA burden (e) into 4 levels – mild, moderate, severe, very severe – see Table 4.

Table 4: Binning of EA burden into 4 levels

EA Burden	**Mild**	**Moderate**	**Severe**	**Very Severe**
$E_{\max}$ or E_{mean}	0 to 0.25	0.25 to 0.5	0.5 to 0.75	0.75 to 1
Number of patients with $E_{\max}$	272	130	107	451
Number of patients with E_{mean}	661	134	88	77

A.3 Summary of Notation

Table 5: Primary table of notations.

Symbol	Description
C_i	Vector pre-admission covariates such as age, vital signs, and medical history
$W_{i,t}$	Sequence of ASMs administered during their stay in the hospital
$D_{i,j,t}$	Blood concentration of ASM j at time t
$E_{i,\max}$	Worst 6 hour epoch of EA burden within a 24 hour period
$E_{i,\text{mean}}$	Average amount of time a patient experiences EA in a 24 hour period
Y_i	Binarized post-discharge outcome (0 if mRS ≤ 3 and 1 if mRS > 3)
$Y_i(e, w)$	Potential outcome if EA burden is e and total ASMs administered is w

Appendix B Methodology

Let us describe how we applied our framework to analyze the EA data to obtain the results. We divide the estimation pipeline into three stages (Figure 6):

1. In the *first stage* (Section C), we calculate $E_{\max}$ and E_{mean}. To do this, we need to first identify segments of the EEG signal containing seizure-like EA behavior. Doing this using human annotators would be extremely time consuming, so we use a convolutional neural network (CNN) trained on human annotators' classifications of 10 second windows into non-EA and EA in a semi-supervised fashion [1–3]. We use the predictions to compute EA time series (Z_t^ω). $E_{\max}$ and E_{mean} are computed directly from Z_t^ω. Details are in the appendix in Section C.

2. In the *second stage* (Section B.1), we fit a personalized pharmacokinetic/pharmacodynamic (PK/PD) model to each patient's response to ASM [4].

3. In the *third stage* (Section B.2) we combine the pre-admission covariates, such as baseline demographic datas and data related to the nature and severity of the present illness, and the PK/PD parameters estimated in the second stage, to adjust for potential confounding and to estimate the potential outcomes of interest. We learn a distance metric to create high-quality matched groups using an *interpretable and accurate* matching method, Matching After Learning to Stretch for EA effect estimation MALTS, [5].

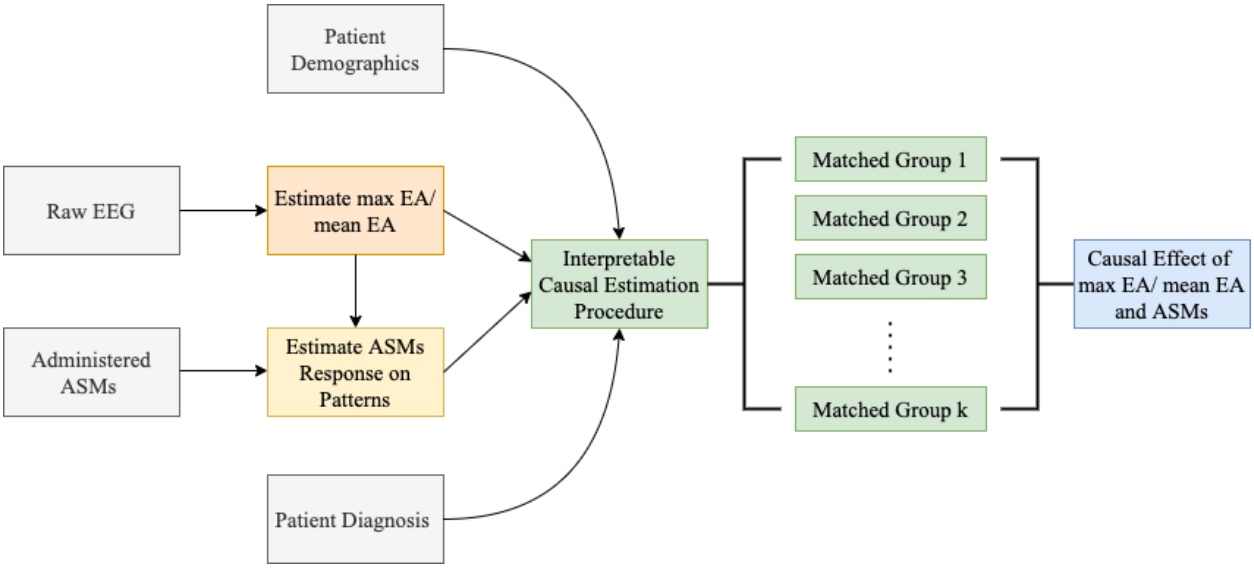

Figure 6: The overall analysis framework, consisting of three parts (indicated by different colors): EA burden computation, individual PK/PD modeling, and MALTS matching and effect estimation.

B.1 Mechanistic Pharmacological Model

Doctors dynamically modify the type and dosage of ASM using the current EA observation, previous treatment, and patient's responsiveness to these treatments. This cyclical relationship potentially confounds the relationship between EA and a patient's final outcome. The heterogeneity in a patient's responsiveness to ASMs can be due to a variety of factors such as past medical history, current medical conditions, age, etc. However, the infrequency of some rare medical conditions makes it difficult to learn a nonparametric model of drug response that incorporates all relevant medical factors. To account for this, we leveraged the domain knowledge from pharmacology and use a one-compartment Pharmacokinetic/Pharmacodynamic (PK/PD) mechanistic model to estimate drug response as a function of ASM dose. The parameters of the PK/PD model can be interpreted as high-dimensional propensity scores that summarize a patient's responsiveness to a drug regime, such that any two patients with similar PK/PD parameters will exhibit similar responses under identical drug regimes.. To account for the effect of past medical history and current medical conditions on drug responsiveness, these factors and the parameters from the PK/PD model are controlled for via a matching procedure as described in Section B.2.

We use a single-compartment PK model to estimate the bloodstream concentration $D_{i,t,j}$ of ASM j in

patient i at time t (drug PK), and Hill's PD model[4] to estimate a short-term response to drugs:

$$\frac{dD_{i,t,j}}{dt} = -\frac{1}{\kappa_j} D_{i,t,j} + W_{i,t,j} \ , \tag{1}$$

$$Z_{i,t} = 1 - \sum_j \frac{D_{i,t,j}^{N_{i,j}}}{D_{i,t,j}^{N_{i,j}} + ED_{50,i,j}^{N_{i,j}}} \ . \tag{2}$$

Here κ_j is the average half-life of the drug (see Appendix 3 for half-lives), $W_{i,j,t}$ is the body weight-normalized drug administration rate in units of mg/kg/h, $N_{i,j}$ represents how responsive the patient is to drug j, and $ED_{50,i,j}$ is the dosage required to reduce the patient's EA burden by 50%. Since $N_{i,j}$ (the Hill coefficient) is constrained to be non-negative, a positive correlation between drug concentration and EA burden results in an $N_{i,j}$ value of 0. The PD parameters were fit using *scipy*'s nonlinear least squares function. The estimated PD parameters reflect wide heterogeneity across patients as well, and indicate clearly which patients responded well to ASMs (shown in Figure 7).

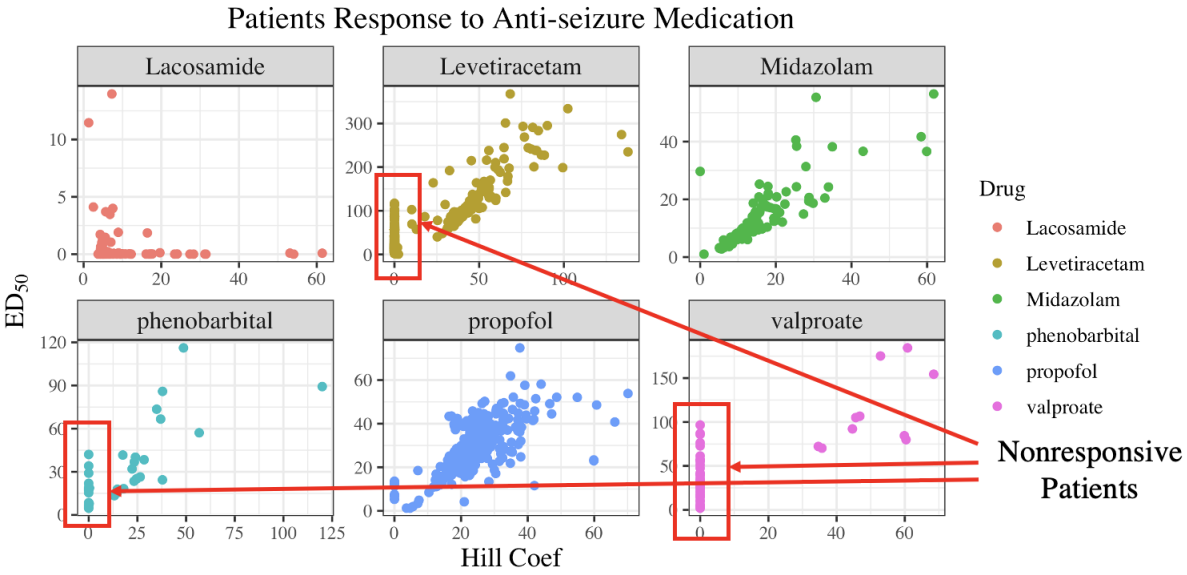

Figure 7: Hill coefficient vs. ED_{50} for the six drugs. Each point is a patient. The non-responsive patients with Hill coefficient of zero are highlighted.

B.2 Interpretable-and-Accurate Causal Inference

In this section, we discuss the causal inference method used to estimate the potential outcomes. Given the stakes involved and the high level of noise in the data, we chose an interpretable-and-accurate causal inference method, MALTS, to estimate cause-effect relationships. MALTS is an *honest* matching method that learns a distance metric using a subset of data as training set. Further, the learned metric is used to

produce high-quality matched groups on the rest of the units (also called as estimation set). These matched groups are used to estimate heterogeneous causal effects with high accuracy. Previous work on MALTS shows that it performs on-par with contemporary black-box causal machine learning methods while also ensuring interpretability[5,6].

The conventional objective function of MALTS, described in[5], was designed to estimate the contrast of potential outcomes under binary "treatment." In this paper, we adapt it to estimate conditional average potential outcomes for n-ary "treatment." For our problem there are 4×2 "treatment" arms – four levels of EA burden crossed by whether or not drugs were administered. We construct the matched group G_i for each patient i by matching on $X_i = [\{C_{i,j}\}_j, \{N_{i,j}\}_j, \{ED_{50,i,j}\}_j]$ - the vector of pre-admission covariates and PD parameters. We estimate $Pr[Y(e,\delta) = 1 | X = X_i]$ by averaging the observed outcomes for units in the matched group G_i with $E_{\max}$ equals e and $\overline{W}$ equals δ. We use an analogous estimator for E_{mean}.

MALTS' estimates of the conditional average potential outcome are *interpretable* because it is computed with the units in the matched groups. These matched groups can be investigated by looking at the raw data to examine their cohesiveness. One might immediately see anything that may need troubleshooting, and easily determine how to troubleshoot it. For instance, if the matched group does not look cohesive, the learned distance metric might need troubleshooting. Or, processing of the EEG signal might need troubleshooting if the max EA burden values do not appear to be correct. Or, the PK/PD parameters might need troubleshooting if patients who appear to be reacting to drugs quickly are matched with others whose drug absorption rates appear to be slower, when at the same time, the PK/PD parameters appear similar. We will demonstrate this with a matched group analysis in the next section.

Appendix C Extracting EA Patterns from EEG

Expert Labeling of EEG Signals. The EEG signals of 1309 patients at Massachusetts General Hospital who met the inclusion criteria were recorded from September 2011 to February 2017. Of these, 82 randomly selected patients had their EEG signals re-referenced into 18 channels via a standard double banana bipolar montage[7] to create a time-frequency representation of a patient's neurological state. These time-frequency representations were then segmented by domain experts using the labeling assistance tool *NeuroBrowser*[3] to identify occurrences of EA patterns. These 82 patients served as the training set for a semi-supervised procedure to create an neural network to automatically identify EA patterns.

Neural Network Based Labeling of EEG Signals. For the cEEG signal labeling procedure, the time-frequency representation was split into 10-second sliding windows with an 8-second overlap. These windows were then converted into an 8-bit color image and used as inputs to the recursive convolutional neural network DenseNet[8]; a Hidden Markov model was added to smooth the outputs[9]. By treating this as

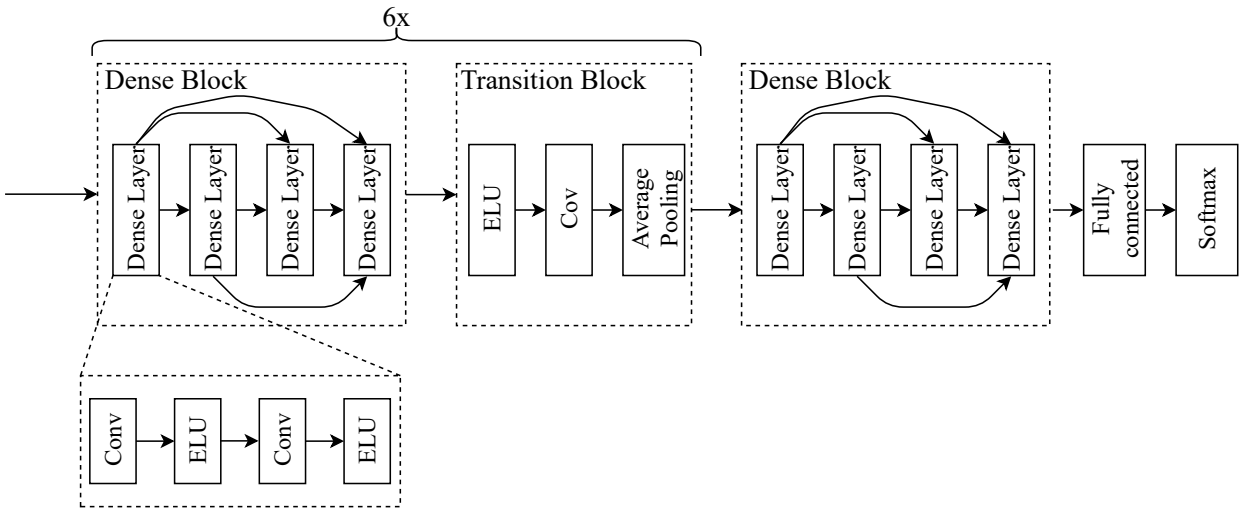

Figure 8: Structure of the DenseNet for automatic EA labeling.

an image classification problem, this closely mimics the procedure performed by the domain experts using *NeuroBrowser*. DenseNet classified each 10-second window as either normal brain activity or one of 4 types of common EA patterns: (1) generalized periodic discharges (GPD), (2) lateralized periodic discharges (LPD), (3) lateralized rhythmic delta activity (LRDA) and (4) Seizure (Sz), as defined by the American Clinical Neurophysiology Society[10]. The trained automatic EA annotator demonstrated accuracy for Seizure at 39% (human inter-rater agreement 42%), GPD at 62% (62%), LPD at 53% (58%), LRDA at 38% (38%), GRDA at 61% (40%), and normal brain-activity/artifact at 69% (75%), therefore, closely matching human performance up to the level of uncertainty one would get from interrater reliability studies.

Operationalizing DenseNet. We used DenseNet with 7 blocks (Figure 8). Each block included 4 dense layers. Each dense layer is comprised of 2 convolutional layers and 2 exponential linear unit (ELU) activations. In between each dense block was a transition block consisting of an ELU activation, a convolutional layer, and an average pooling layer. There were 6 transition blocks in total. The last two layers of DenseNet were a fully connected layer followed by a softmax layer. The loss function includes Kullback-Leibler divergence inversely weighted by the class ratio to account for imbalance among the EA classes. After fitting, it was observed that DenseNet's classifications were much more volatile than the original data, with predictions abruptly changing from normal brain activity to EA patterns. This highlighted a limitation of traditional EEG classification from images, as the images were fed independently with no context about neighboring images beyond the 10-second window given. To correct for this volatility, the results of DenseNet were smoothed using a Hidden Markov Model. To smooth to a similar degree as the human labeled data, the probabilities of the transition matrix were fit on the 82 human-labeled patients. These probabilities were then used as the hidden state to smooth the output from DenseNet. We made the HMM first order due to

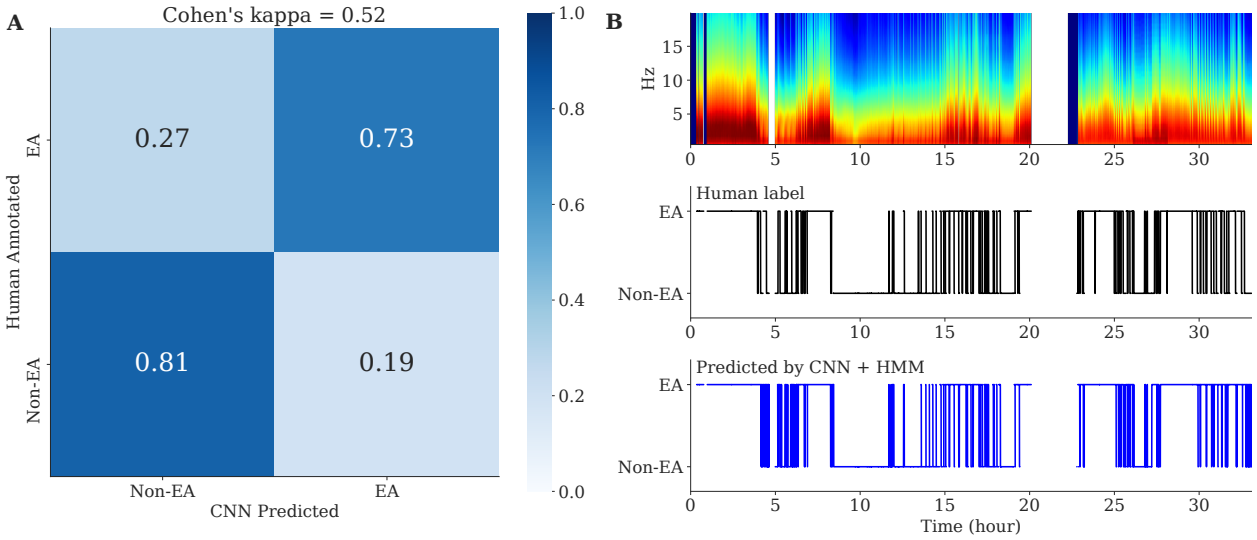

Figure 9: (A) Confusion matrix for the CNN prediction vs. human annotation, where each row represents the fraction of 2-second segments classified into EA (seizure/GPD/LPD/LRDA) or Non-EA (GRDA/other/artifact). The overall Cohen's kappa is 0.52. (B) The top panel shows the spectrogram of the EEG signal of one example subject; the middle panel shows EA patterns annotated by a human expert for every 2 second interval. The bottom panel shows the EA pattern annotated by the CNN followed by HMM smoothing.

precedent of first order HMMs providing good smoothing for other EEG problems [11].

The results of the automatic EA annotator resulted in accuracy for Seizure at 39% (human inter-rater agreement 42%), GPD at 62% (62%), LPD at 53% (58%), LRDA at 38% (38%), GRDA at 61% (40%), and others/artifact at 69% (75%). Therefore matching human performance. We further combined the classification into binary classes, EA (seizure/GPD/LPD/LRDA) vs. non-EA (GRDA/other/artifact) (Figure 9) to reduce the chance of error since these patterns are intrinsically on a continuous spectrum.

Appendix D Sensitivity to the definition of EA burden

Throughout the analysis, the summaries of EA burden, $E_{\max}$ and E_{mean} are quantized into four equally sized groups . This is done in accordance with clinician recommendations. In this section we evaluate the sensitivity of our analysis to these decisions. Specifically, we consider $E_{\max} \in \{[0, \rho_1), [\rho_1, 0.5), [0.5, \rho_2), [\rho_2, 1.0]\}$ where the analysis in the paper specifies $\rho_1 = 0.25$ and $\rho_2 = 0.75$. The interpretation of these parameters is as follows: the *mild* EA burden category allows for no more than $100 \times \rho_1$ percent of a six hour window to be spent with EA and the *very severe* EA burden category allows for no less than $100 \times \rho_2$ percent of a six hour window to be spent with EA. By varying these parameters we redefine which individuals are considered mild versus very severe EA during the analysis.

From sensitivity analysis to definition of EA burden, we observe following (see Figure 10):

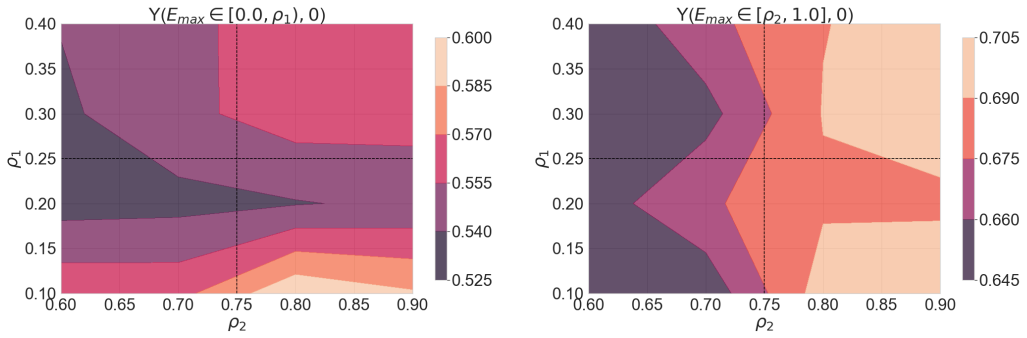

Figure 10: Sensitivity to quantization of EA burden into four levels. ρ_1 is the boundary between mild and moderate EA burden and ρ_2 is the boundary between severe and very severe EA burden. The contour plot shows estimated average potential outcomes – $Y([0, \rho_1), 0)$ and $Y([\rho_2, 1], 0)$ – for a range of ρ_1 and ρ_2. We find that the gradient of contours is more or less flat and the estimates do not change by a large amount as the sensitivity parameters change.

- The potential outcome under mild EA burden ($\mathbb{E}[Y([0.0, \rho_1), 0)]$) is mildly sensitive to changes in ρ_2 which is expected. Further, we observe that the gradient of the same with respect to ρ_1 is relatively flat, and $\mathbb{E}[Y([0.0, \rho_1), 0)]$ is bounded between 0.525 and 0.6 for $\rho_1 \in [0.1, 0.4]$.
- Analogously the potential outcome under mild EA burden ($\mathbb{E}[Y([\rho_2, 1.0], 0)]$) is mildly sensitive to changes in ρ_1 and its gradient with respect to ρ_2 is relatively flat, and $\mathbb{E}[Y([0.0, \rho_1), 0)]$ is bounded between 0.645 and 0.705 for $\rho_1 \in [0.6, 0.9]$.
- The point estimates of $\mathbb{E}[Y([0.0, \rho_1), 0)]$ are always strictly less than the point estimates of $\mathbb{E}[Y([\rho_2), 1.0]]$

Appendix E Missingness Pattern

To check for possible selection bias, we compared the discharge mRS in patients with different missing conditions in Figure 11 where some of them were excluded in this cohort. We used the Mann-Whitney U test (nonparametric t-test) to compare the medians, since mRS does not follow a normal distribution.

The results show that the medians of discharge mRS in patients with EEG, versus that in patients without EEG, are not significantly different; similarly, the medians in patients with both EEG and drug data, versus that in patients without EEG or drug data, are not significantly different neither. Therefore the missingness pattern can be considered as not influencing our results, hence the selection bias is negligible.

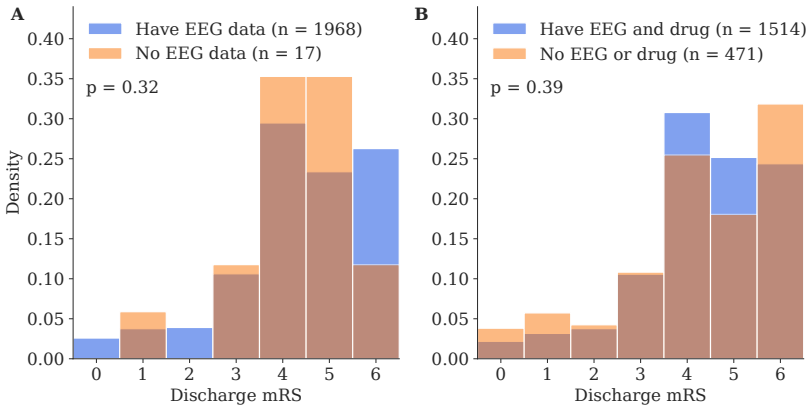

Figure 11: (A) The histogram of patients' discharge mRS (possible values are 0,1,2,3,4,5,6). The two subsets that are compared are patients who have EEG data (n = 1968) vs. patients who do not have EEG data (n = 17). To make the subsets comparison, the y-axis shows the density instead of the count. The p-value is from the Mann-Whitney U test of the two subsets. (B) Similar to A, but for patients who have EEG and drug data (n = 1514) vs. patients who do not have EEG or drug data (n = 471).

Appendix F Robustness to causal assumptions

In providing our estimate of average potential outcome, our causal approach makes several important assumptions including: 1) pre-admission covariates and PD parameters are both potential sources of confounding and thus need to be controlled for 2) the post-discharge outcome, Y, is directly affected by **both** the level of EA burden, E_{max}/E_{mean} and the presence of Anti-seizure medications $\overline{W}$. In this section, we demonstrate how the estimation of potential outcome can vary with these assumptions.

F.1 Assumption 1): The need to control for pre-admission covariates and PD parameters

Previously, it was posited that pre-admission covariates such as age and diagnosis and PD parameters could be large sources of confounding in the estimation of average potential outcomes. In this section, we investigate this assumption by having MALTS create matched groups based on fewer and fewer factors and comparing the resulting average potential outcomes.

The left side of Figure 12 shows the estimated average potential outcome when MALTS controls for only one, albeit important, variable, age. The results do not to show a monotonic relationship between EA burden and average potential outcome. When matching on all pre-admission covariates but no PD parameters using MALTS, while the monotonic relationship between EA burden is average potential outcome is now clear, the uncertainty in the estimates and the shape of the trend differs. In particular, without adding in the information from the ASM's PK/PD models, one tends to underestimate the probability that a patient would leave the hospital impaired or dead.

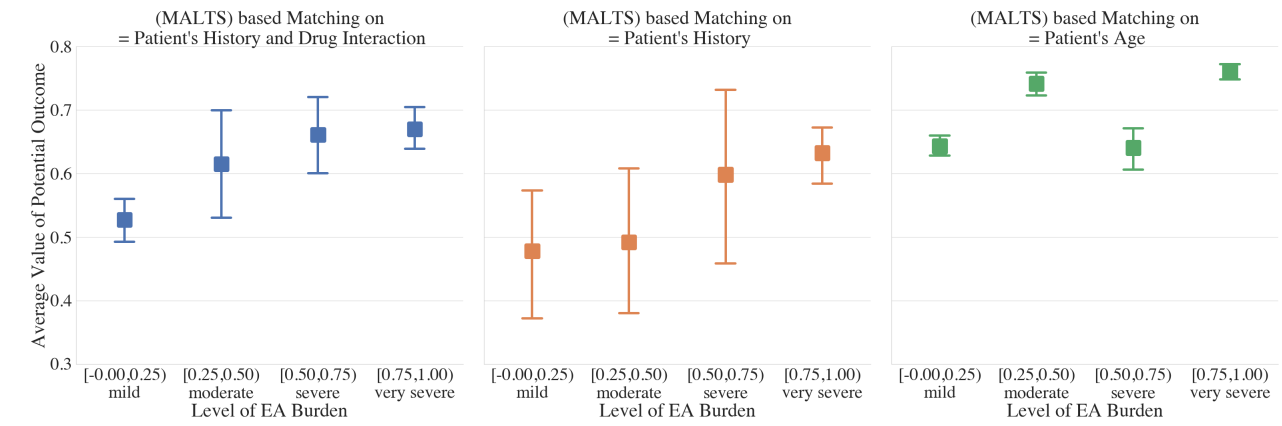

Figure 12: Estimated average potential outcome for different levels of $E_{\max}$ by matching on (left) all pre-admission covariates and PD parameters, (middle) all pre-admission covariates, and (right) only age of the patients.

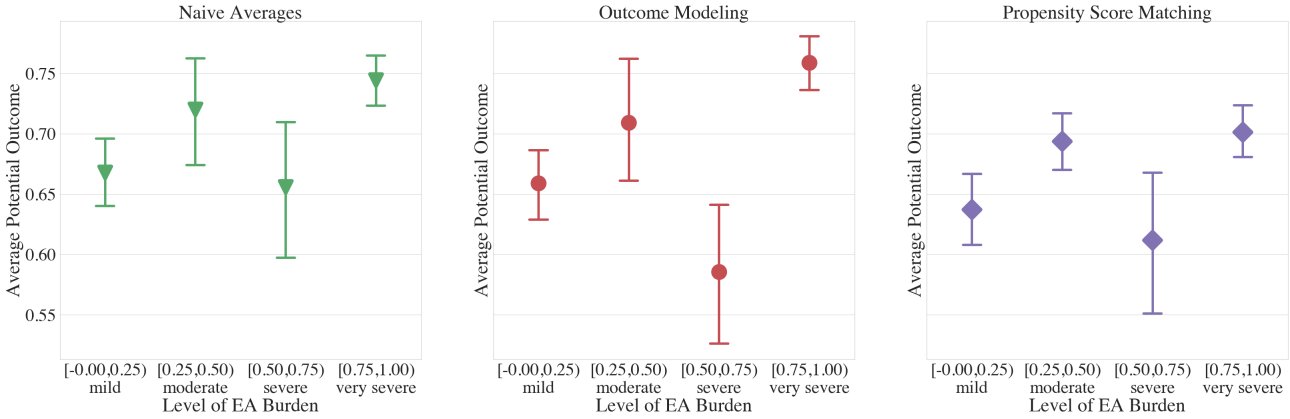

Figure 13: Estimated average potential outcomes computed using (left) Naive Average approach, (middle) Outcome modeling approach, and (right) Propensity Score matching.

F.2 Assumption 2): Post-discharge outcome are a function of the level of EA burden and the presence of Anti-seizure medications

In this section, we compare our method, which posits that both the level of EA burden and the presence of Anti-seizure medications are the only two causal factors with a "Naive Average" approach which posits that EA burden is the **only** causal factor, an "Outcome Modeling" approach that treats all of the factors in our study as having a direct causal effect on the outcome, and a Propensity score approach, which performs a causal estimation, albeit under differing assumptions.

In the Naive Average approach, at each EA burden, $\frac{1}{3}$ of the data is left out and the probability of leaving the hospital impaired is computed on the remaining $\frac{2}{3}$ of the data. This procedure is repeated 15 times and the mean and standard deviation of the replicates are report as the left-most figure in Figure 13. The choice

of 15 and $\frac{2}{3}$ was done to match as closely as possible the 15 replicates and 2:1 training to testing ratio that was used by MALTS.

In the outcome modeling approach, which takes up the middle of Figure 13, we perform a logistic regression where we regress the post-discharge outcome against EA burden, the presence of anti-seizure medications, and all of the factors that MALTS matched such as pre-admission covariates and PD parameters. Note that this approach makes the assumption that there are no interactions between the regressors, which goes contrary to our understanding of the treatment procedure, as factors such as age and diagnosis have a known interaction with a patient's response to anti-seizure medications. Like the naive averages approach, we perform 15 replicates of a logistic regression with the same 2:1 train/test used in the Naive Averages approach and MALTS approach.

On the right of Figure 13, we have the average potential outcome computed with a common approach to causal estimation, propensity score matching. Unlike MALTS which matches together patients directly on their covariates, propensity score matching is based on matching together patients based on a their probability of being within the treatment or control arm. This makes the stronger assumption that the probability of being within the treatment or control arm can be modeled parametrically, in this case as using a logistic regression.

The results of these three approaches all yield similar results, showing an approximately sinusoidal relationship between EA burden and average potential outcome. This differs from the original MALTS result in the top left of Figure 12 which shows a clear monotonic relationship between EA burden and average potential outcome. As MALTS is the only method that takes a causal approach without making the strong parametric assumptions in propensity score matching, this seems to hint that perhaps the lack of control for confounding variables has been throwing off the regression based approaches to analyzing the damage caused by EA burdens.

Appendix G Sensitivity Analysis for Unobserved Confounding

In this section, we study how sensitive our inferences are to unobserved confounding. In particular, we study the sensitivity to an unobserved confounder that correlates patients' post-discharge outcome with E_{max}. We would like to see if the presence of an unobserved confounder we failed to control for could have biased our inferences. We can encode the effect of an unobserved confounder using a selection bias function $q(e)$ with sensitivity parameter ψ. This approach is similar to the one proposed in[12]. We parameterize $q(\cdot)$ as a logarithmic function of e.

$$\begin{aligned} q(e) &= \mathbf{E}[Y_i((e,0))|E_{max,i} = e, \overline{W}_i = 0] - \mathbf{E}[Y_i((e,0))|E_{max,i} \neq e, \overline{W}_i = 0] \\ &= \psi ln(e+1) \end{aligned}$$

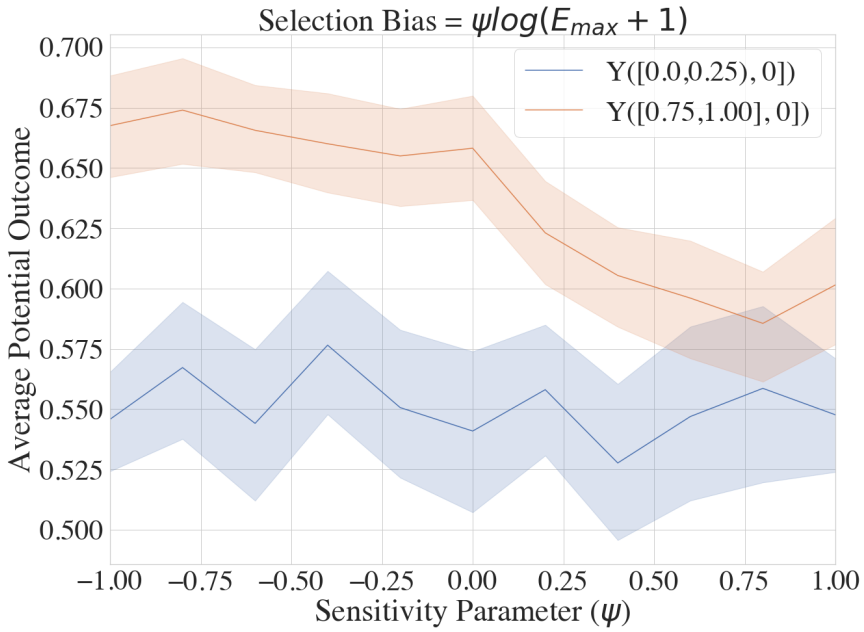

Figure 14: Sensitivity to unobserved confounding The results show that even at very high levels of selection bias, the effect of EA burden is not lost, indicating a degree of robustness in our results.

When ψ is positive (negative), this indicates that patients with observed *bad* (*good*) outcomes also have high observed EA burden. This parametric form also assumes that a patient with low $E_{\max}$ is affected less by an unobserved confounder U compared to a unit with higher $E_{\max}$ with the marginal increase tapering off as the $E_{\max}$ increases. This is congruent with the neurologist's intuition that a perfectly healthy individuals with normal brain activity will be affected less by an unobserved confounder U.

To perform the sensitivity analysis, we apply the following debiasing to the observed outcome and re-estimate the average potential outcomes:

$$Y_i^{debiased} \quad = \quad Y_i - q(E_{\max,i})(1 - P(E_{\max,i}|X = X_i)).$$

If the unobserved confounding does not large impact on the estimation of average potential outcome, then the estimated potential outcome under very severe EA burden ([0.75,1.0]) will be more than average potential outcome under mild EA burden ([0.0,0.25)).

Our sensitivity analysis found that point estimate of potential outcome under very severe EA burden is always worse than the potential outcomes under mild EA burden for a range of sensitivity parameter ψ between [-1,1]. We further find that our inference is statistically significant for a wide range of ψ: $-1.0 \leq \psi \leq 0.50$. The sensitivity highlights that the conclusions from our study and analysis are insensitivity to

high levels of unobserved confounding.

References

[1] Wendong Ge, Jin Jing, Sungtae An, Aline Herlopian, Marcus Ng, Aaron F. Struck, Brian Appavu, Emily L. Johnson, Gamaleldin Osman, Hiba A. Haider, Ioannis Karakis, Jennifer A. Kim, Jonathan J. Halford, Monica B. Dhakar, Rani A. Sarkis, Christa B. Swisher, Sarah Schmitt, Jong Woo Lee, Mohammad Tabaeizadeh, Andres Rodriguez, Nicolas Gaspard, Emily Gilmore, Susan T. Herman, Peter W. Kaplan, Jay Pathmanathan, Shenda Hong, Eric S. Rosenthal, Sahar Zafar, Jimeng Sun, and M. Brandon Westover. Deep active learning for interictal ictal injury continuum EEG patterns. Journal of Neuroscience Methods, 351:108966, 2021. ISSN 0165-0270.

[2] Sahar F Zafar, Eric S Rosenthal, Jin Jing, Wendong Ge, Mohammad Tabaeizadeh, Hassan Aboul Nour, Maryum Shoukat, Haoqi Sun, Farrukh Javed, Solomon Kassa, et al. Automated annotation of epileptiform burden and its association with outcomes. Annals of neurology, 90(2):300–311, 2021.

[3] J Jing, J Dauwels, T Rakthanmanon, E Keogh, SS Cash, and MB Westover. Rapid annotation of interictal epileptiform discharges via template matching under dynamic time warping. Journal of Neuroscience Methods, 274:179–190, 2016.

[4] Archibald Vivian Hill. The mode of action of nicotine and curari, determined by the form of the contraction curve and the method of temperature coefficients. The Journal of Physiology, 39(5):361–373, 1909.

[5] Harsh Parikh, Cynthia Rudin, and Alexander Volfovsky. MALTS: Matching after learning to stretch. arXiv 1811.07415, 2020.

[6] Harsh Parikh, Cynthia Rudin, and Alexander Volfovsky. An application of matching after learning to stretch (MALTS) to the ACIC 2018 causal inference challenge data. Observational Studies, 5:118–130, 01 2019. doi: 10.1353/obs.2019.0006.

[7] Selim R Benbadis. Introduction to sleep electroencephalography. Sleep: A Comprehensive Handbook, pages 989–1024, 2006.

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

[9] Wendong Ge, Jin Jing, Sungtae An, Aline Herlopian, Marcus Ng, Aaron F Struck, Brian Appavu, Emily L Johnson, Gamaleldin Osman, Hiba A Haider, et al. Deep active learning for interictal ictal injury continuum EEG patterns. Journal of Neuroscience Methods, 351:108966, 2021.

[10] Lawrence J Hirsch, Michael WK Fong, Markus Leitinger, Suzette M LaRoche, Sandor Beniczky, Nicholas S Abend, Jong Woo Lee, Courtney J Wusthoff, Cecil D Hahn, M Brandon Westover, et al. American clinical neurophysiology society's standardized critical care eeg terminology: 2021 version. Journal of Clinical Neurophysiology, 38(1):1–29, 2021.

[11] Haoqi Sun, Jian Jia, Balaji Goparaju, Guang-Bin Huang, Olga Sourina, Matt Travis Bianchi, and M Brandon Westover. Large-scale automated sleep staging. Sleep, 40(10), 2017.

[12] Matthew Blackwell. A selection bias approach to sensitivity analysis for causal effects. Political Analysis, 22(2):169–182, 2014.