BORROWING FROM YOUR NEIGHBORS: THREE STATISTICAL TECHNIQUES FROM NONTRADITIONAL SOURCES

Kentaro J. Hoffman

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics and Operations Research in the Department of Statistics and Operations Research.

Chapel Hill
2022

Approved by:

Kai Zhang

Cynthia Rudin

Richard Smith

Jan Hannig

Chuanshu Ji

Yufeng Liu

ABSTRACT

KENTARO J. HOFFMAN: Borrowing from your Neighbors: Three new statistical techniques from non-traditional sources.
(Under the direction of Cynthia Rudin and Kai Zhang)

From Generalised Fiducial Inference to Causal Inference, the past few years have seen a rising tide of new statistical paradigms calling into question our previous approaches of learning from data. This thesis will follow in this movement and demonstrate how these newer paradigms allow us to perform analyses that would be difficult to perform using conventional approaches. In the first chapter, we show how Dempster-Shafer and Fiducial Inference can be used as an alternative approach to the conventional Neyman-Pearson hypothesis testing paradigm through the inclusion of an "unknown" class into the testing procedure. This not only allows for tests with in-built robustness estimates, but allows for a natural analysis of the effects of adversarial attacks on hypothesis tests. In the second chapter, we demonstrate how interpretable causal inference combine with differential equation modeling gives users a powerful new approach to answering causal questions about patients exhibiting epileptiform activity. Finally, we combine the Empirical Mode Decomposition, which pioneered a signal decomposition that makes far fewer assumptions than traditional Fourier or Wavelet decompositions, with statistical techniques to allow for more accurate signal identification and cleaning.

ACKNOWLEDGEMENTS

I am extremely grateful to my advisors Dr. Kai Zhang and Dr. Cynthia Rudin for serving as a source of inspiration and mentorship throughout the long road of the PhD. My work with both of you has turned me into a better person and a better researcher. In addition, I would like to give a thanks to my committee. Dr. Richard Smith, your careful attention to my writing markedly improved the quality of this thesis. Dr. Chuanshu Ji and Yufeng Liu, thank you for your valuable time and insightful comments. And finally, a special thanks to Dr. Jan Hannig, who helped me turn my thesis around and made research enjoyable and something I want to pursue further.

Next, in rough chronological order, I would like to thank all who left an indelible mark on me for the better. First, I am indebted to the hardworking care of my mother, and the love of science installed in by my father. Without both, I would not have been able to finish my degree. Aunt Miki, for your care and love. Cole Land, for showing me wisdom far beyond what should be possible. Isabel Boraels, for showering me in endless treats and template to stay forever young. My high school history teachers, for instilling a love of our past which will never fade. Sahithi Cherukuri, for being my constant companion and favorite attack dog/defender. Keiko Kaplan, for all your fatherly advice, help navigating adulthood, and bad puns. Jesse and Michael Peirce, for the consistent world to crash in on Friday night and screaming about our favorite nerdy-ass voice actors. The rest of 4th South for being my forever friends and helping me leave my shell. Matt Adrianowycz, for our probabilistic, liquid adventures. Rudy Guerra, for patiently listening to my early statistical journey even though I was bad at taking your advice. Nicolas Hengartner, for showing me that statistics can be fun and exciting. Sabrina Civale, for your patient, loving care, resulting in the most important present I have ever received. Hongshen Liu, for showing me the ropes. Benjy Leinwand, for being my first friend in grad school and helping carry each other through the struggles of first year. Eyes Robson, for widening my narrow perspective. Kevin O'Connor, for helping me get my steps in while developing the future of statistics. Sāmapriya Basu, for showing me the limits of bad faith. Peter/Prabhanka, for lunch and nothing else. Pavlos

Zouboulouglou, for showing me how to be fashionable. Nate Pham, for you and your wife's endlessly welcome company. Dhruv Patel, for relational advice. Xinyuan Niu, for teaching me something new every time we talk. Miheer Dewaskar, for showing me how to be an open mided scientist. Minji Kim, for being a great office mate and future scientist. The rest of the cast of the Basement who gave me community and listening to my constant attempts to give advice. Harsh Parikh, for your ingenious work on MALTS and being a scientific role model. Mikhal Ben-Joseph, for being my best student who will easily overtake me someday, if not already. Roubin Gong, for helping me take my first steps into the wider world as an independent researcher. And finally, Bob & Others, for giving me the tools to get through the toughest times. I hope to see all of you again someday and show you how i've changed.

-Kentaro Hoffman, April 10, 2022

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In this thesis, we demonstrate how theoretical developments in three non-statistical field, Reliability Engineering, Differential Equations, and Signal Processes can be adapted to create new statistical methodology. Such cross disciplinary work has a long history. One of the most famous statisticians of the 20th century, John Tukey, was known for saying, "The best thing about being a statistician is that you get to play in everyone's backyard" (Lin et al., 2014). While pithy, this quote illustrates that statistics, at its core, is an interdisciplinary subject. An interdisciplinary approach begets new problems and new solutions. For example, the 2010's saw the rise of single-cell RNA sequencing which pushed the boundaries of higher dimensional statistics through a new demand for sparser methods (Efron and Hastie, 2016). Other times, new theoretical developments in other subjects have important implications for existing statistical problems. Take, for example, the development of Dempster-Shafer Inference. While it rose to prominence in the Reliability Engineering and Artificial Intelligence communities (Dempster, 1967) for its novel approach to decision making under uncertainty, it has had an important influence on the development of robust inferential methods (Gelman, 2006) and Fiducial Inference (Hannig, 2009). Another important example is the development of causal inference. Statistics and Computer Science have each proposed a powerful paradigm for causal inference, the Rubin Causal Model (Splawa-Neyman et al., 1990) and Do-Calculus (Pearl, 1995) respectively, each of which have influenced and challenged the other.

In this thesis, we build on these cross-paradigm conversations of the past and demonstrate how different statistical and non-statistical paradigms can be applied to problems, new and old, to create powerful new approaches which address weaknesses in existing approaches. In the first project, we study one of the oldest problems in statistical inference: multinomial hypothesis testing. Being one of the oldest distributions in statistics (Bernoulli, 1713), the multinomial plays

an important role in application areas which rely on categorical data. Two such areas are the fields of Differential Privacy and Adversarial Attacks. In both domains, it is important not only to be able to perform statistical inference, but to include uncertainty estimates of the inferences which incorporate the unique noise structure of these problems. A primary tool for such analyses is Dempster-Shafer Inference, but it was only in the past decade that reliable methods for inferring multinomial parameters under the Demspter-Shafer paradigm have become feasible (Zhang and Liu, 2014; Jacob et al., 2021).

In our investigation, we develop a testing procedure outside of the Neyman-Pearson paradigm to perform a Dempster-Shafer hypothesis test for multinomial data. Our method closely matches the results for a corresponding frequentist Chi-squared test and provides additional information about the epistemic uncertainty in the test. Additionally, it provides several surprising connections between Dempster-Shafer, Frequentist, and Fiducial inferential procedures which hint at new way of relating the three approaches.

In our second project, we investigate the causal effects of epileptiform activity on critically ill patients. Epileptiform activity is an irregular proto-seizure-like brain activity that can be identified from a patient's EEG (Hirsch et al., 2021). Despite epileptiform activity being common in critically ill patients suffering from brain injuries (Lucke-Wold et al., 2015), there is yet to be a causal analysis of epileptiform activity which controls for the various anti-seizure medications a patient received. Ignoring anti-seizure medications is especially problematic as anti-seizure medications are administered by a physician to directly lower the level of epileptiform activity, creating a powerful confounding factor. Furthermore, due to the temporal aspect of these drug treatments and low signal to noise ratio, analyzing the damage that epileptiform activity causes requires the creation of new statistical methodology.

This is accomplished through a new framework which merges interpretable causal inference and differential equation-based Pharmacokinetic/Pharmacodynamic modeling. This method retains the strengths of both approaches, allowing us to make well founded causal claims, while leveraging known drug mechanisms to make accurate drug response estimates even in the presence of high background noise and frequently missing EEG segments. Using this framework, we have identified a causal effect between epileptiform activity and reduced adverse patient outcomes, as well as identified the patients that are at increased risk for epileptiform activity.

Finally, in our third project, we bring to attention how the mathematical properties of a new signal decomposition technique, the Ensemble Empirical Mode Decomposition can be utilized to design a new change point and signal cleaning technique. Specifically, while it lacks the large-sample properties of Fourier and Wavelet decompositions, the Ensemble Empirical Mode Decomposition is adept at modeling non-linear signals with relatively few basis functions. This low dimensional representation presents a new opportunity as some techniques, such as change point detection, perform much better at lower dimensions than higher ones. By combining the low-dimensionality of the representation with appropriate low-dimensional statistical techniques, we demonstrate a signal cleaning technique, `LCDSC` is able to match the performance of competing signal cleaning techniques and serves as a powerful tool for cleaning infrasound acoustic waves.

Chapter 2

Dempster-Shafer Hypothesis Testing for Multinomial Data

2.1 Introduction

Dempster-Shafer (DS) Inference was developed in the 1960-70s by Arthur Dempster (Dempster, 1968) and Glen Shafer (Dempster and Shafer, 1976) as a prior-free approach to statistical inference. Historically, DS Inference has played an important role in the development of artificial intelligence (Yager and Liu, 2008) and reliability engineering (Sentz and Ferson, 2002), leading to the development of the field of "imprecise probability." Despite its popularity in other fields, DS Inference has had a smaller impact on the statistical world in which it originated. It has been hypothesized that it failed to take off due to technical difficulties in interpreting beliefs and plausibilities (Pearl, 1988), computational burden, and lack of conventional long-run frequentest properties under repeated sampling (Martin et al., 2010).

However, in the last several years, there has been a renewed interest from the statistical community in other forms of prior-free forms of posterior inference which incorporate epistemic uncertainty. This has come from closely related topics such as Jan Hannig's Generalized Fiducial Inference (Hannig, 2009), Confidence Distributions (Xie and Singh, 2013), and faster computational techniques using Gibbs Sampling on polytopes (Jacob et al., 2021). We build on this rising wave and demonstrate that one can perform a wide variety of Dempster-Shafer Hypothesis Tests which are computationally fast and model epistemic uncertainty: in particular, we focus on the uncertainty that can result from adversarial attacks. Moreover, we will show through classical large sample bounds and simulations that there exists a close connection between our form of DS hypothesis testing and the classical frequentist testing paradigm. The connection is close enough that we can approximate frequentist results using our DS approach. Finally, we will demonstrate how our approach gives unique insights into the dimensionality of a hypothesis test as well as models the effects of adversarial attacks on multinomial data.

2.2 Multinomial Data Generation via *Dirichlet-DSM*

In a Multinomial Data generation scheme, we observe n, k-dimensional multinomial observations $(n_1, ...n_k)$ with unknown but fixed probabilities $\mathcal{P} = (p_1, ...p_k)$ such that $\sum_{i=1}^{k} p_i = 1$

$$n_1, ..., n_k \sim \text{ Multinomial } (n, p_1, ..., p_k).$$

To perform DS inference, it is first necessary to define the data generation process that yielded $(n_1, ..., n_k)$. This can be done through prior insight into the data generation scheme, or it can be chosen to facilitate the inference of the parameters. In line with the second reason, we will employ an insightful data generation process described in Lawrence et al. (2009) as the *Dirichlet-DSM* process. In this process, we first define an unknown permutation π of $\{1, ...k\}$. We then define I_a, $a \in \{1, ..., k\}$ as an interval of length p_a whose location is determined by the process:

$$I_a = \begin{cases} [0, p_a) & \text{if } \pi(1) = a \\ [\sum_{i=1}^{\tilde{a}-1} p_{\pi(i)}, \sum_{i=1}^{\tilde{a}-1} p_{\pi(i)} + p_a) & \text{if } \pi(1) \neq a \end{cases}$$

where $\tilde{a} = \pi^{-1}(a)$. Regardless of each permutation π, each segment I_a is of length p_a. Using these latent intervals, it is posited that each n_i is generated according to the equation:

$$n_i = a \textit{ iff } W_i \in I_a (i = 1, ...n)$$

where $(W_1, ...W_n)$ are n independent, uniform random variables on [0,1].

While this data generation scheme is fairly complex compared to others such as *simplex-DSM* seen in Dempster (1966), *interval-DSM* in Hannig (2009), and *Simplex-DS* in Jacob et al. (2021), it holds two distinct advantages. First, unlike *interval-DSM*, this inference procedure for *Dirichlet-DSM* is not sensitive to the order of the categories. Further, unlike the *simplex-DSM* and *Simplex-DS*, *Dirichlet-DSM* has a relatively simple expression for its posterior estimation which requires no expensive acceptance-rejection or Gibbs sampling. For this reason we will be demonstrating our hypothesis testing procedure on data generated using *Dirichlet-DSM*, although much of the intuition could also be applicable to these other approaches.

2.2.1 A Recipe for Multinomial DS Inference using *Dirichlet-DSM*

To infer $\mathcal{P}$ from data generated by the *Dirichlet-DSM* scheme, we employ a latent parameter approach to estimate the posterior random set. This is equivalent to a reformulation of the method described in Lawrence et al. (2009), except we emphasize the Fiducial perspectives of this approach to connect this into related developments from Generalized Fiducial Inference. To start, we first generate a $k+1$ dimensional vector of latent parameters $\mathcal{Z} = (Z_0, \mathcal{Z}_{-0})$

$$(Z_0, \mathcal{Z}_{-0}) = (Z_0, Z_1, ...Z_k) \sim \text{ Dirichlet } (1, n_1, ...n_k) \tag{2.1}$$

Given fixed instances of $(Z_0, \mathcal{Z}_{-0})$ the latent parameters, $\mathbf{z} = (z_0, z_{-0})$ from our Dirichlet distribution, we will define the posterior random set, $\hat{\mathcal{P}}(z)$, as a polytope with edges:

$$\boldsymbol{v}_j = (z_1, ...z_{j-1}, z_j + z_0, z_{j+1}, ...z_k), \forall j \in \{1, ...k\}.$$

It is because these polytope edges come from a Dirichlet distribution that this may be referred to as a *Dirichlet* DSM. Any element of $\hat{\mathcal{P}}(z)$ can equivalently be written as the set:

$$\hat{\mathcal{P}}(z) = \{z_{-0} + \boldsymbol{\theta}^T z_0 | \boldsymbol{\theta} = (\theta_1, ..., \theta_k), \sum_{j=1}^{k} \theta_j = 1,\ 0 \leq \theta_j \leq 1, \text{ for } j = 1, ..., k\}$$

or in vector form:

$$\hat{\mathcal{P}}(z) = \begin{pmatrix} Z_1 + \theta_1 Z_0 \\ \vdots \\ Z_k + \theta_k Z_0 \end{pmatrix}.$$

When referring to the *distribution* of posterior random sets, we will denote this as:

$$\hat{\mathcal{P}}(\mathcal{Z}) = \{\mathcal{Z}_{-0} + \boldsymbol{\theta}^T Z_0 | \boldsymbol{\theta} = (\theta_1, ..., \theta_k), \sum_{j=1}^{k} \theta_j = 1, \theta_j \geq 0 \text{ for } j = 1, ..., k\}. \tag{2.2}$$

To demonstrate that $\hat{\mathcal{P}}(\mathcal{Z})$ provides a reasonable estimate of $\mathcal{P}$, we can show that these random sets asymptotically converge to $\mathcal{P}$.

Theorem 1. *$\hat{\mathcal{P}}(\mathcal{Z})$ converges almost surely to $\mathcal{P}$.*

Proof. First, recall that we can write the distribution of our posterior random set as

$$\hat{\mathcal{P}}(\boldsymbol{\mathcal{Z}}) = \begin{pmatrix} Z_1 + \theta_1 Z_0 \\ \vdots \\ Z_k + \theta_k Z_0 \end{pmatrix}$$

for any fixed $\theta_1, ...\theta_k$ on a $k+1$ dimensional simplex. This random vector lies on the probability space $(\Delta^k, \mathcal{B}(\Delta^k), P)$ where Δ^k is the $k+1$ dimensional simplex and $\mathcal{B}(\Delta^k)$ is the borel sigma algebra. To make the relationship with this space clear, for this proof, we will use the notation:

$$\hat{\mathcal{P}}(\mathcal{Z})(\mathbf{w}) = \begin{pmatrix} Z_1(w_1) + \theta_1 Z_0(w_0) \\ \vdots \\ Z_k(w_k) + \theta_k Z_0(w_0) \end{pmatrix}$$

where $\mathbf{w} = (w_1, ..., w_k) \in \Delta^k$. To show almost sure convergence, it suffices to show that for all $\epsilon > 0$:

$$P(\limsup_{n\to\infty}\{\mathbf{w} \in \Delta^k : ||\hat{\mathcal{P}}(\boldsymbol{\mathcal{Z}})(\mathbf{w}) - \mathcal{P}(\mathbf{w})||_1 \geq \epsilon) = 0$$

To demonstrate this, we employ the triangle inequality, the union bound, and the fact that $|\theta_i| \leq 1$ to get the bound:

$$P(\limsup_{n\to\infty}\{\mathbf{w} \in \Delta^k : ||\hat{\mathcal{P}}(\mathcal{Z})(w) - \mathcal{P}||_1 \geq \epsilon\}) \tag{2.3}$$

$$= P(\limsup_{n\to\infty}\{\mathbf{w} \in \Delta^k : \left\| \begin{pmatrix} Z_1(w_1) + \theta_1 Z_0(w_0) \\ \vdots \\ Z_k(w_k) + \theta_k Z_0(w_0) \end{pmatrix} - \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix} \right\|_1 \geq \epsilon\}) \tag{2.4}$$

$$= P(\limsup_{n\to\infty}\{\mathbf{w} \in \Delta^k : \sum_{i=1}^{k} |Z_i(w_i) - p_i + \theta_i Z_0(w_0)| \geq \epsilon\}) \tag{2.5}$$

$$\leq P(\limsup_{n\to\infty}\{\mathbf{w} \in \Delta^k : \sum_{i=1}^{k} |Z_i(w_i) - p_i| \geq \frac{\epsilon}{2}\}) + P(\limsup_{n\to\infty}\{\mathbf{w} \in \Delta^k : \sum_{i=1}^{k} |\theta_i Z_0(w_0)| \geq \frac{\epsilon}{2}\}) \tag{2.6}$$

$$\leq \sum_{i=1}^{k} P(\limsup_{n\to\infty}\{\mathbf{w} \in \Delta^k : |Z_i(w_i) - p_i| \geq \frac{\epsilon}{2k}\}) + P(\limsup_{n\to\infty}\{\mathbf{w} \in \Delta^k : |Z_0(w_0)| \geq \frac{\epsilon}{2k}\}) \tag{2.7}$$

Thus, it suffices to show that $Z_i \overset{a.s}{\to} p_i$ for $i \in \{1, ...k\}$ and $Z_0 \overset{a.s}{\to} 0$ with respect to n.

To see this, due to the aggregation property of the Dirichlet (Ng et al., 2011):

$$Z_0 \sim Beta(1, n) \tag{2.8}$$

$$Z_{-0} \sim Dirichlet(n_1, ..., n_k). \tag{2.9}$$

For the first summand, any Beta distribution can be written as the ratio of independent Gammas which in turn can be written as a sum of Exponentials (Casella and Berger, 2001):

$$Z_0 = \frac{U}{U+V} = \frac{U}{U + \sum_{i=1}^{n} V_i} = \frac{U/n}{(U + \sum_{i=1}^{n} V_i)/n}.$$

Here $U \sim Gamma(1,1) = Exp(1)$, and $V \sim Gamma(n,1)$, $V_1, ...V_n \sim Exp(1)$. Then, by the strong law of large numbers $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} V_i}{n} = 1$ and $\lim_{n\to\infty} \frac{U}{n} = 0$. Furthermore, by the continuous mapping theorem, the numerator goes to 0 and the denominator goes to 1. Therefore, by the continuous mapping theorem:

$$Z_0 \overset{a.s}{\to} 0.$$

As for the second summand, like the Beta, we can write any Dirichlet as a ratio of independent Gammas using $U_i \sim Gamma(n_i, 1)$ for $i \in \{1, ...k\}$, $U = \sum_{i=1}^{k} U_i \sim Gamma(n, 1)$, and $V_i \sim Exp(1)$. (Ng et al., 2011):

$$Dirichlet(n_1, ..., n_k) = \left(\frac{U_1}{U}, ..., \frac{U_k}{U}\right).$$

Each $\frac{U_i}{U}$ for $i \in \{1, ...k\}$ can be written as a ratio of gamma distributions:

$$\begin{aligned}
\frac{U_i}{U} &= \frac{U_i}{U_i + \sum_{j\neq i} U_j'} \\
&= \frac{\sum_{j=1}^{n_i} V_i}{\sum_{j=1}^{n_i} V_i + \sum_{j\neq i} \sum_{k=1}^{n_j} V_k'} \\
&= \frac{\sum_{j=1}^{n_i} (V_i)/n}{\sum_{j=1}^{n_i} (V_i)/n + \sum_{j\neq i} \sum_{k=1}^{n_j} (V_k')/n}.
\end{aligned}$$

Here, $V_i^{'}$ is an independent copy of V_i. Focusing just on the denominator, their sum is equivalent to:

$$\sum_{j=1}^{n_i}(V_i)/n + \sum_{j\neq i}\sum_{k=1}^{n_j}(V_k^{'})/n = \sum_{j=1}^{n} V_i/n.$$

Which by the strong law of large numbers, converges almost surely to 1.

As for the numerator, $(n_1, ..., n_k)$ comes from a multinomial so each count n_i $\forall i \in \{1, ...k\}$ is marginally a Binomial(n, p_i) distribution. Combine this with Wald's lemma and we get:

$$\lim_{n\to\infty} E(\sum_{j=1}^{n_i} V_i/n) = \lim_{n\to\infty} E(n_i)E(V_i)/n = \lim_{n\to\infty} \frac{np_i}{n} = p_i.$$

So the whole fraction goes to p_i almost surely. Thus, by the continuous mapping theorem:

$$Z_{-0} \xrightarrow{a.s} \mathcal{P}$$

which completes the proof. □

2.3 Point Estimation with DS Inference

Before we describe hypothesis testing, we must make a few comments about how we will be performing point estimation of our posterior random sets, $\hat{\mathcal{P}}(z)$, and how this connects to point estimation in other forms of inference. Recall that $\hat{\mathcal{P}}(z)$ is a polytope that describes the region of feasible parameter estimates. To emphasize this fact and differentiate it from a point estimate $\hat{\mathcal{P}}$, we will alternatively denote $\hat{\mathcal{P}}(z)$ using $\Delta(z)$ and $\hat{\mathcal{P}}(Z)$ using $\Delta(Z)$. Using a test statistic $T(\hat{\mathcal{P}}, \mathcal{P}_0)$ and fixed $(n_1, ..., n_k)$, one can compute an **upper**, **mean**, and **lower test statistic** for each $\Delta(z)$ as follows:

$$T_{upper}(z) = \sup_{\hat{\mathcal{P}}\in\Delta(z)} T(\hat{\mathcal{P}}, \mathcal{P}_0) \tag{2.10}$$

$$T_{mean}(z) = T(E_{\Delta(z)}\hat{\mathcal{P}}, \mathcal{P}_0) \tag{2.11}$$

$$T_{lower}(z) = \inf_{\hat{\mathcal{P}}\in\Delta(z)} T(\hat{\mathcal{P}}, \mathcal{P}_0). \tag{2.12}$$

$E_{\Delta(z)}$ in this case represents taking the expectation with respect to the uniform measure upon $\Delta(z)$.

By marginalizing over $\mathcal{Z}$, this turns the upper, mean and lower statistics into **upper**, **mean**, and **lower distributions**:

$$T_{upper} = \sup_{\hat{\mathcal{P}} \in \Delta(\mathcal{Z})} T(\hat{\mathcal{P}}, \mathcal{P}_0) \tag{2.13}$$

$$T_{mean} = T(E_{\Delta(\mathcal{Z})}\hat{\mathcal{P}}, \mathcal{P}_0) \tag{2.14}$$

$$T_{lower} = \inf_{\hat{\mathcal{P}} \in \Delta(\mathcal{Z})} T(\hat{\mathcal{P}}, \mathcal{P}_0). \tag{2.15}$$

Assuming that T defines a valid distance metric, $T_{upper}(z)$ represents the largest possible test statistic one could create on $\Delta(z)$, $T_{lower}(z)$ represents the smallest, and T_{mean} represents the average case. Note that since by definition for any fixed z:

$$T_{upper}(z) \geq T_{mean}(z) \geq T_{lower}(z).$$

This implies that:

$$T_{upper} \succcurlyeq T_{mean} \succcurlyeq T_{lower}$$

where $\succcurlyeq$ defines a stochastic ordering.

2.3.1 Comparison to other common point estimators

The creation of the upper, mean, and lower test statistics contrasts with the frequentist hypothesis testing paradigm. Under a frequentist paradigm, one often derives a point estimate, $\hat{\mathcal{P}}_{Freq}$, usually using the maximum likelihood equation, and evaluates the asymptotic distribution of the point estimate plugged into the test statistic:

$$T_{Freq} = T(\hat{\mathcal{P}}_{Freq}, \mathcal{P}_0)$$

where $\mathcal{P}_0$ is the parameter value for the null hypothesis. However, in spite of their seeming differences, one can connect the frequentist test statistic with that of the DS through the choice of summary statistic used on the posterior random set.

Theorem 2. *For fixed observations $(n_1, ..., n_k)$ from a multinomial distribution and test statistic T, T_{mean} is equal to $T_{Freq} = T(\hat{\mathcal{P}}_{Freq}, \mathcal{P}_0)$ where $\hat{\mathcal{P}}_{Freq} = (\frac{n_1+1/k}{n+k}, ..., \frac{n_k+1/k}{n+k})$*

Proof. Since:

$$T_{mean} = T(E_{\Delta(Z)}\hat{\mathcal{P}}, \mathcal{P}_0)$$
$$T_{Freq} = T(\hat{\mathcal{P}}_{Freq}, \mathcal{P}_0)$$

and $\mathcal{P}_0$ is fixed, it suffices to show that:

$$E_{\Delta(Z)}\hat{\mathcal{P}} = \hat{P}_{Freq}.$$

To see this, note that we can write:

$$\begin{aligned}
E_{\Delta(Z)}(\hat{P}) &= E_Z E_{\Delta(z)}(\hat{P}) \\
&= E_Z E_{\Delta(z)} z_{-0} + \theta_i z_0 \\
&= E_Z z_{-0} + \frac{1}{k} z_0 \\
&= E_Z(z_{-0}) + \frac{1}{k} E_Z(z_0) \\
&= \frac{n_i}{n+1} + \frac{1}{k}\frac{1}{n+1} \\
&= \frac{n_i + 1/k}{n+1} \\
&= \hat{\mathcal{P}}_{Freq}.
\end{aligned}$$

□

Note that at this point, T_{mean} is the bayes estimator under MSE of when the $(p_1, ..., p_k)$ has a prior distribution of Dirichlet$(1/k, ..., 1/k)$ (Lehmann and Casella, 1998). This illustrates that the frequentist approach of choosing an arbitrary point estimate is equivalent to choosing an arbitrary way to summarize the posterior random sets $\hat{\mathcal{P}}(Z)$. To the best of our knowledge, we are unaware of other results which connect various forms of inference via the summarization of DS random sets.

2.4 Hypothesis Testing with DS Inference

With our upper, lower, and mean point estimates and their distributions defined, we are ready to define how we perform a level α-hypothesis test with our DS approach.

Definition 2.4.1. Level α one-sided DS Hypothesis Test

Let us have two competing hypotheses H_0 and H_1 about k-dimensional parameter $\mathcal{P}_0 \in \Delta^{k-1}$ such that:

$$H_0 : \mathcal{P} = \mathcal{P}_0, H_1 : \mathcal{P} \neq \mathcal{P}_0.$$

Using $\mathcal{P}_0$ and point estimate, $\hat{\mathcal{P}}_{observed}$ based on fixed counts $(n_1, ..., n_k)$, we compute tail probabilities for the upper and lower distributions:

$$\pi_{upper} = P(T_{lower} \leq T(\hat{P}_{observed}, \mathcal{P}_0)) \tag{2.16}$$

$$\pi_{lower} = P(T_{upper} \leq T(\hat{P}_{observed}, \mathcal{P}_0)). \tag{2.17}$$

We then conclude based on the tail probabilities:

$$\begin{cases} \text{Reject } H_0 & \text{If } \pi_{upper} \leq \alpha \\ \text{Fail to Reject } H_0 & \text{If } \pi_{lower} > \alpha \\ \text{Unknown} & \text{If } \pi_{lower} \leq \alpha \text{ but } \pi_{upper} > \alpha \end{cases}$$

As an example, let us perform a DS goodness of fit test using the chi-squared test statistic. The data comes from a Multinomial(n, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and the null hypothesis is that all three classes are equally likely. The number of elements observed in each class are $(n_1, n_2, n_3) = (3, 2, 5)$. In such a case, our upper and lower test statistics are based on the chi-squared statistic:

$$T_{upper}(\mathcal{Z}) = \sup_{(\hat{p}_1, \hat{p}_2, \hat{p}_3) \in \Delta(\mathcal{Z})} \sum_{i=1}^{3} \frac{(10\hat{p}_i - 10/3)^2}{10/3}$$

$$T_{lower}(\mathcal{Z}) = \inf_{(\hat{p}_1, \hat{p}_2, \hat{p}_3) \in \Delta(\mathcal{Z})} \sum_{i=1}^{3} \frac{(10\hat{p}_i - 10/3)^2}{10/3}$$

where elements of $\Delta(\mathcal{Z})$ are:

$$\hat{\mathcal{P}}(\mathcal{Z}) = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \hat{p}_3 \end{pmatrix} = \begin{pmatrix} Z_1 + \theta_1 Z_0 \\ Z_2 + \theta_2 Z_0 \\ Z_3 + \theta_3 Z_0 \end{pmatrix}$$

for fixed $\theta_1 + \theta_2 + \theta_3 = 1$ and $0 \leq \theta_i \leq 1$ for $i \in \{1, 2, 3\}$ and:

$$(Z_0, Z_1, Z_2, Z_3) \sim Dirichlet(1, 3, 2, 5).$$

In Figure 2.1, we plot out 100 realizations of posterior random sets as well as the the null point, $\mathcal{P}_0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and point estimate $(\frac{n_1+1}{n+3}, \frac{n_2+1}{n+3}, \frac{n_3+1}{n+3}) = (\frac{4}{13}, \frac{3}{13}, \frac{6}{13})$. Note how most of the polytopes are clustered around the point estimate which is some distance from the null point.

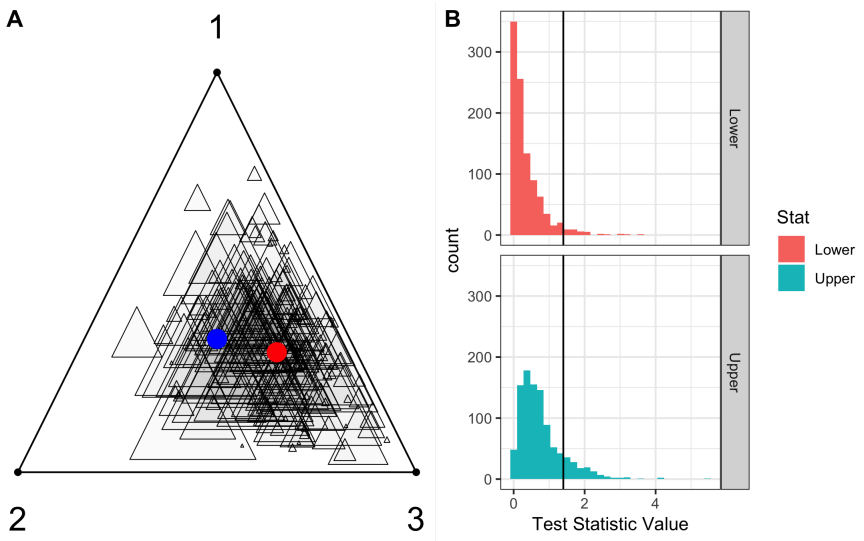


Figure 2.1: **A)** 100 randomly chosen posterior random sets from a 3 dimensional test of uniformity. Note how the polytopes are centered around the point estimate (red). 100 was chosen for visibility reasons. **B)** 1000 simulations of the upper and lower test statistic and a vertical line representing $\mathcal{P}_0$. For this data, π_{lower} is 0.12 and π_{upper} is 0.034. Notice how since the lower test statistic is stochastically smaller than the upper, π_{lower} is larger than π_{upper}.

In the figure 2.2, we have summarized 1000 posterior random sets into their upper and lower chi-squared test distributions. As is, the lower tail probability is 0.034 while the upper tail probability is 0.12, so at the $\alpha = 0.05$ level, we conclude that we lack the power to make a conclusion either way. More precisely, if we performed a hypothesis test using our most optimistic upper test statistic, we would reject while our conservative lower test statistic would fail to reject.

Having reasonable posterior estimates leading to opposite conclusions is not a sign of reliability so we conclude that we cannot make a conclusion either way. Now, if we increase our sample size from n=10 to n=100 and observe $(n_1, n_2, n_3) = (30, 20, 50)$, we now have much more evidence that the null hypothesis of each class being equally likely is incorrect. Note that our posterior random sets are now defined by:

$$(Z_0, Z_1, Z_2, Z_3) \sim Dirichlet(1, 30, 20, 50).$$

The width of each random set is $Z_0 \sim Beta(1, 100)$, rather than $Z_0 \sim Beta(1, 10)$ as in the previous example. Correspondingly, in Figure 2.2, we can see in part A) that the polytopes are smaller and more clustered around our point estimate. This corresponds with our intuition that a larger sample size should result in more concentrated posterior estimates. Now none of the 1000 posterior random sets are close to the null point. Thus, π_{lower} and π_{upper} are both < 0.001 meaning that we reject our null hypothesis with confidence.

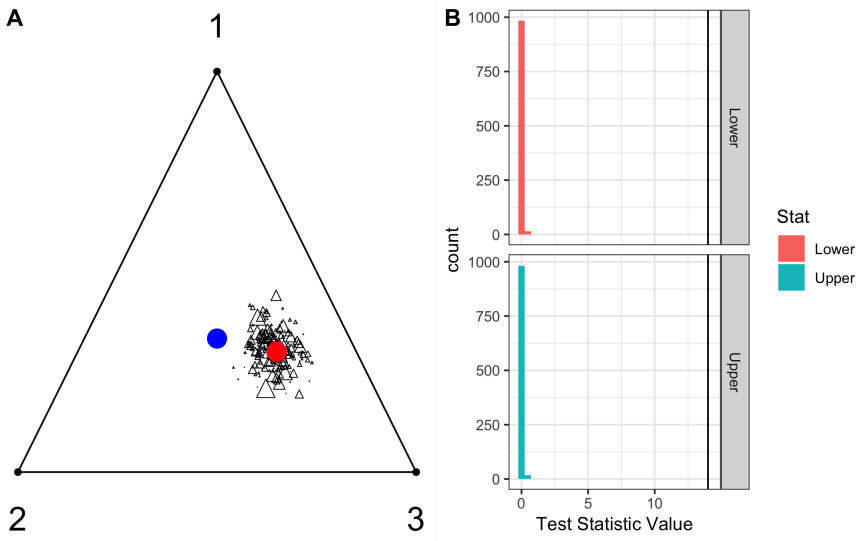


Figure 2.2: **A)** 100 randomly chosen posterior random sets. Note how the polytopes are more tightly centered around the point estimate than in figure 2.1. **B)** 1000 simulations of the upper and lower test statistic and a vertical line representing $\mathcal{P}_0$. With the larger sample size, both π_{lower} and π_{upper} are < 0.0001.

As we saw in the previous example, increasing our sample size made us go from an uncertain conclusion to a certain one. This turns out to be not only a feature of the previous example, but holds for any continuous, bounded test statistic.

Theorem 3. *Consider a k-dimensional DS hypothesis as described above based on a continuous and bounded test statistics, $T(\hat{\mathcal{P}}, \mathcal{P}_0)$, where there exists an M such that $\sup_{p_1,p_2\in\Delta^k} |T(p_1, p_2)| \leq M$. As the sample size increases, the probability of a DS hypothesis test returning an unknown result goes to 0.*

Proof. To make the relationship between the posterior random sets, $\hat{\mathcal{P}}$, sample size, n, and an element in the complement of the null set $w \in N^c$ clear, we will denote the posterior random set based on a sample of size n as $\hat{\mathcal{P}}_n(w)$.

As we saw in Theorem 1, for any fixed n, $\hat{\mathcal{P}}_n(w) \xrightarrow{a.s} \mathcal{P}$, by the continuous mapping theorem, we may claim that:

$$T(\hat{\mathcal{P}}(Z)_n(w), \mathcal{P}_0) \xrightarrow{a.s} T(\mathcal{P}, \mathcal{P}_0).$$

Moreover, by the boundedness of our measure on our compact set, Egorov's theorem states that a.s convergence implies almost uniform convergence (Beals, 2004). And since almost uniform convergence of a bounded function on a compact set allows for the interchange of supremum and limits, (Rudin, 1953) we may claim:

$$\lim_{n\to\infty} \sup_{w\in N^c} T(\hat{\mathcal{P}}(Z)_n(w), \mathcal{P}_0) = \sup_{w\in N^c} \lim_{n\to\infty} T(\hat{\mathcal{P}}(Z)_n(w), \mathcal{P}_0) = \sup_{w\in N^c} T(\mathcal{P}, \mathcal{P}_0) = T(\mathcal{P}, \mathcal{P}_0)$$

$$\lim_{n\to\infty} \inf_{w\in N^c} T(\hat{\mathcal{P}}(Z)_n(w), \mathcal{P}_0) = \inf_{w\in N^c} \lim_{n\to\infty} T(\hat{\mathcal{P}}(Z)_n(w), \mathcal{P}_0) = \inf_{w\in N^c} T(\mathcal{P}, \mathcal{P}_0) = T(\mathcal{P}, \mathcal{P}_0)$$

for any $w \in N^c$.

To finish the proof, recall that we return an "unknown" conclusion when $\pi_{upper} \leq \alpha$ but $\pi_{lower} > \alpha$. However, as T_{upper} and T_{lower} converge almost surely to a constant, $T(\mathcal{P}, \mathcal{P}_0)$, they must convergence in distribution as well. Since convergence in distribution is, by definition convergence in CDFs, by direct consequence, the CDFs of T_{upper} and T_{lower} converge to a Heaviside function centered at $T(\mathcal{P}, \mathcal{P}_0)$. Now, if we recall that we have defined π_{upper} and π_{lower} as quantiles of T_{upper} and T_{lower} :

$$\pi_{upper} = P(T_{lower} \leq T(\hat{P}_{observed}, \mathcal{P}_0)) \tag{2.18}$$

$$\pi_{lower} = P(T_{upper} \leq T(\hat{P}_{observed}, \mathcal{P}_0)), \tag{2.19}$$

and the CDFs of T_{upper} and T_{lower} converge to a Heaviside function centered at $T(\mathcal{P}, \mathcal{P}_0)$, thus π_{upper} must be equal to π_{lower}. Since π_{upper} and π_{lower} are equal, it is not possible for one to be larger than the other, thus completing the proof. □

2.5 DS and Fiducial Interpretations

Now that we have defined what we claim to be a DS hypothesis test, we show how this test fits into the DS paradigm and the Fiducial paradigm. In addition, demonstrate additional connections between our method and the frequentist paradigm.

2.5.1 DS Interpretation

To summarize our current procedure, the *Dirichlet-DSM* data generation scheme, we are given observations $(n_1, ..., n_k)$ which arise from some unknown parameter $\mathcal{P}$, unknown random variables $\mathcal{U} = (U_1, ...U_n)$, and unknown permutation π. Moreover, through the inference procedure, we can generate Dirichlet random variables $(Z_0, ..., Z_k)$ which define a posterior random set to estimate $\mathcal{P}$. Finally, these random sets are then summarized using various summaries and test statistics, tail probabilities are computed, and hypotheses are either rejected, not rejected, or unknown. To see how this procedure fits into the DS framework, we will first abstract this procedure in the spirit of Fraser's structural inference (Fraser, 1968) and illustrate how our tail probabilities define proper DS upper and lower probabilities.

First, let us write our data generation scheme in terms of a Fraser inspired *randomized structural equation*:

$$\mathcal{X} = G(\mathcal{P}, \mathcal{U}|\pi).$$

here G is our deterministic data generating equation, $\mathcal{P}$ are our parameters of interest, $\mathcal{U}$ is the random component and π is an unknown permutation. Unlike Fraser's structural equation framework, we will not be assuming a group structure and will allow for aspects of the data generating equation to contain randomized components via π. We will further expand upon the implications of this randomized data generating equation in the Fiducial Inference interpretation section.

Using this, we can define the subset of dependent uniform random variables $\mathcal{U} = (U_1, ...U_n)$ that could have generated $(n_1, ...n_n)$ for some $\mathcal{P} = (p_1, ...p_k)$ as:

$$\mathcal{R}_x = \{\mathcal{U} \in [0,1]^n : \exists \mathcal{P} \in \Delta^k \text{ and } \exists \pi \in \mathcal{S}(\{1, ..., k\}) \text{ st } (n_1, ...n_n) = G(\mathcal{P}, \mathbf{u}|\pi)\}.$$

Here, $\mathcal{S}(\{1, ..., k\})$ is the symmetry group on the set $\{1, ..., k\}$ which is the set of all permutations, π. Given $\mathbf{u} \in \mathcal{R}_x$, there exists a non-empty "feasible" set on the parameter space $\mathcal{F}(\mathcal{U}) \subset \Delta^k$ that could have been used to generate the data:

$$\mathcal{F}(\mathcal{U}) = \{\mathcal{P} \in \Delta^k : (n_1, ...n_k) = G(\mathcal{P}, \mathbf{u}|\pi), \exists \pi \in \mathcal{S}_{\{1,...,k\}}\}.$$

While this object may seem daunting, in the case of *Dirichlet-DSM*, $\mathcal{F}(\mathcal{U})$ has a very simple expression. $\mathcal{F}(\mathcal{U})$ is directly equal to our previously described polytopes with Dirichlet vertices, what we have been calling $\Delta(Z)$. Moreover, as we can see in our definition of $\Delta(Z)$, there is no dependence on which permutation π was used to generate the data. The ability to decouple the class randomization from the inference is one of the method's greatest strengths.

We now introduce how these random sets yield valid DS upper and lower probabilities. Consider a continuous test statistic $T(\hat{\mathcal{P}}, \mathcal{P}_0)$ which defines a measurable map $T(\cdot, \mathcal{P}_0) : (\Delta^{k-1}, \mathcal{M}_{\Delta^{k-1}}) \to (\Omega, \mathcal{M}_\Omega)$, where $\mathcal{M}_\Delta^{k-1}$ is the usual Borel algebra on the $k-1$ dimensional simplex and Ω is a one-dimensional Polish space, most commonly $\mathbb{R}$ or $\mathbb{N}$. Given a measurable posterior random set, $\Delta(Z)$, via continuity of T, $T(\Delta(Z))$ as well as $\sup T(\Delta(Z))$ and $\inf T(\Delta(Z))$ are all continuous measurable sets in $\mathcal{M}_\Omega$. To perform a hypothesis test, the user now provides a measurable set $\Sigma \subset \mathcal{M}_\Omega$ that corresponds to the hypothesis of interest. For the previous example in section 2.4, we tested if the parameters were uniform, $(p_1 = p_2 = p_3 = 1/3)$ with chi-squared test statistic $T(\mathcal{P}, \mathcal{P}_0) = \sqrt{\frac{\sum_{i=1}^3 (np_i - n/3)^2}{n/3}}$. With the test of uniformity, our corresponding set of interest is: $\Sigma = \{T(\mathcal{P}, \mathcal{P}_0) \leq T(\hat{\mathcal{P}}, \mathcal{P}_0)_{\alpha\%}\}$ where $T(\hat{\mathcal{P}}, \mathcal{P}_0)_{\alpha\%}$ is the α quantile of $T(\hat{\mathcal{P}}, \mathcal{P}_0)$.

DS theory presumes the existence of a mass function m such that:

$$m : 2^\Omega \to [0,1]$$

such that:

$$m(\emptyset) = 0$$

and

$$\sum_{S \in 2^{\Omega}} m(S) = 1.$$

As we can see, any probability measure can be written as a mass function by letting every measurable set be equal to its probability and setting unmeasurable sets to 0:

$$m(A) = \begin{cases} P(A) & \text{if } A \in \mathcal{M}_{\Omega} \\ 0 & \text{if } A \notin \mathcal{M}_{\Omega} \end{cases}.$$

Moreover, for any set $S \subset 2^{\Omega}$ (as opposed to any measurable set), the belief of a set S is defined as the mass of sets which are contained within S:

$$bel(S) = \sum_{A|A \subset S} m(A).$$

The plausibility of S is defined as the mass of sets that intersect S:

$$pl(S) = \sum_{A|A \cap S \neq \emptyset} m(A).$$

Due to this construction:

$$bel(A) \leq P(A) \leq pl(A).$$

With the above defined, we now have the tools to show that π_{upper} and π_{lower} are valid belief and plausibility functions for the one-sided hypothesis test previously described:

Theorem 4. *When T defines a distance metric, π_{upper} and π_{lower} define valid belief and plausibility functions for S where $S = [T(\mathcal{P}, \hat{\mathcal{P}})_{\alpha\%}, \infty)$.*

Proof. First, since T defines a distance metric, if for a given Z:

$$T_{upper}(Z) \leq T(\mathcal{P}, \hat{\mathcal{P}})_{\alpha,}$$

this implies that:

$$S = [T_{upper}(Z)_\alpha, \infty) \subset [T_{upper}(Z), \infty).$$

Now, if we let m be the counting measure:

$$\begin{aligned} bel(S) &= \sum_{[T_{upper}(Z),\infty)|[T_{upper}(Z),\infty)\subset S} m([T_{upper}(Z), \infty)) \\ &= \sum_{T_{upper}(Z)\leq T(\mathcal{P},\hat{\mathcal{P}}_{\alpha\%})} m([T_{upper}(Z), \infty)) \\ &= P(T_{upper} \leq T(\mathcal{P}, \hat{\mathcal{P}})_{\alpha\%,}) \\ &= \pi_{upper} \end{aligned}$$

Likewise using the same argument, we can show that $pl(S) = \pi_{lower}$. □

Based on this theorem, we can view the belief function as the probability we observe a posterior set where *every* element have a T value less than the $\alpha\%$ cutoff. This belief function is what we have been referring to as our lower p-value. On the other hand, the plausibility function is the probability that we observe a posterior set where there exists *at least* one element who's T value is less than α.

2.5.2 Fiducial Interpretation

In terms of Fiducial Inference, we will be showing that the previously defined "feasible" set $\Delta(Z)$ defines what is known as a Generalized Fiducial Quantity. Recall that given fixed latent parameters z from a Dirichlet Distribution, our inference procedure results in a polytope $\Delta(z)$ of parameters values for which there exists uniform random variables $(U_1, ...U_k)$ that could have generated the observed data. As above, we write the set of parameters that could have feasibly generated $(n_1, ..., n_k)$ as:

$$\mathcal{F}(\mathcal{U}) = \{\mathcal{P} \in \Delta^k : (n_1, ...n_k) = G(\mathcal{P}, \mathbf{u}|\pi), \exists \pi \in \mathcal{S}_{\{1,....,k\}}\}.$$

However, unlike DS, which directly works with the random posterior sets, $\mathcal{F}(\mathcal{U})$, in the Generalized Fiducial Scheme, when $\mathcal{F}(\mathcal{U})$ contains multiple elements one selects an element

according to some possible random rule V. Mathematically, we say that for any measurable set $S \in \Delta^k$ of feasible parameters, one selects a possibly random element $V(S)$ with support $\overline{S}$ where $\overline{S}$ in the Closure of S. In our case, our random set is $\mathcal{F}(\mathcal{U})$ from which we choose either the supremum, infimum, or the average value of the polytope. As supremum, infimum, and average value are all random rules that lie within the closure, our upper, lower, and mean test distributions define valid Fiducial Generalized Quantities.

2.6 Connection to Frequentist Hypothesis Testing

Earlier, we defined the tail probabilities for T_{upper} and T_{lower} as π_{upper} and π_{lower}. Some readers may have noticed that this resembles the notation commonly used for p-values. This similarity is no coincidence. For common test statistics such as the chi-squared statistic, it is possible to show that: 1) at finite samples, our p-values are **relatively close** to the frequentist p-values based with the same test statistic, 2) that our p-value have the **same** asymptotic coverage as a frequentist test, and 3) π_{min} and π_{max} not only bound the DS π_{mean}, but the frequentist π_{freq}. We will start with the first two claims:

Theorem 5. *For a k-dimensional test of uniformity:*

$$H_0 : (p_1, ..., p_k) = \left(\frac{1}{k}, ..., \frac{1}{k}\right) H_1 : (p_1, ..., p_k) \neq \left(\frac{1}{k}, ..., \frac{1}{k}\right)$$

based on fixed observations $(n_1, ...n_k)$*, using the usual chi-squared test statistic and the point estimate* $(\frac{n_1+1/k}{n+1}, ..., \frac{n_k+1/k}{n+1})$*:*

$$P(|\pi_{freq} - \pi_{mean}| \geq \alpha) \leq \frac{n(k-1)}{(a/2)k(n+1)^2} + \frac{kn^2 + (k-1)n - k\sum_{i=1}^{k} n_i^2}{(a/2)k(n+1)^2(n+2)}$$

where n is the sample size, $a > 0$*, and* π_{freq} *is the p-value from a chi-squared test and* π_{mean} *is the p-value from our DS test.*

Proof. Like previously, our *Dirichlet-DSM* assumes that:

$$(n_1, ..., n_k) \sim Multinomial(n, p_1, ..., p_k)$$
$$(Z_0, ...Z_k) \sim Dirichlet(1, n_1, ..., n_k).$$

Using the point estimate, $\hat{p}_{Freq} = (\frac{n_1+1/k}{n+1}, ..., \frac{n_k+1/k}{n+1})$, π_{freq} and π_{mean} become:

$$\pi_{freq} = P_*(||\hat{p}^*_{Freq} - p_0||_2 \geq ||\hat{p}_{Freq} - p_0||_2) \tag{2.20}$$
$$\pi_{mean} = P_{\Delta(Z)}(E_\theta||\Delta(Z) - \hat{p}_{Freq}||_2 \geq ||\hat{p}_{Freq} - p_0||_2). \tag{2.21}$$

Here $\hat{p}^*_{Freq}$ refers to a frequentist point computed from an independent realization of a Multinomial(n,$(\frac{1}{k}), ..., \frac{1}{k})$) distribution. By rearranging equation 2.20, one can see that π_{freq} is indeed equal to the p-value one generates using the chi-squared statistic.

$$P_*(||\hat{p}^*_{Freq} - p_0||_2 \geq ||\hat{p}_{Freq} - p_0||_2) = P_*(||n\hat{p}^*_{Freq} - np_0||_2 \geq ||n\hat{p}_{Freq} - np_0||_2) \tag{2.22}$$
$$= P_*\left(\sum_{i=1}^{k} \frac{(n\hat{p}^*_{Freq,i} - np_{0,i})^2}{np_{0,i}} \geq \sum_{i=1}^{k} \frac{(n\hat{p}_{Freq,i} - np_{0,i})^2}{np_{0,i}}\right) \tag{2.23}$$

As π_{freq} and π_{mean} are tail probabilities for the same quantity, $||\hat{p}_{Freq} - p_0||_2$, we will demonstrate that $||\hat{p}^*_{Freq} - p_0||_2$ and $||\Delta(Z) - \hat{p}_{Freq}||_2$ have similar levels of concentration in their tails.

We will first start with $||\hat{p}^*_{Freq} - p_0||_2$. In scalar form:

$$||\hat{p}^*_{Freq} - p_0||_2^2 = \sum_{i=1}^{k} (\hat{p}^*_{Freq,i} - p_{0,i})^2 = \sum_{i=1}^{k} \left(\frac{n^*_i + 1/k}{n+1} - 1/k\right)^2.$$

As $\hat{p}^*_{Freq}$ is an unbiased estimator of p_0:

$$\begin{aligned}
E(\hat{p}^*_{Freq}) &= E\left(\frac{n^*_1 + 1/k}{n+1}, ..., \frac{n^*_k + 1/k}{n+1}\right) \\
&= \left(\frac{E(n^*_1) + 1/k}{n+1}, ..., \frac{E(n^*_k) + 1/k}{n+1}\right) \\
&= \left(\frac{n/k + 1/k}{n+1}, ..., \frac{n/k + 1/k}{n+1}\right) \\
&= \left(\frac{1}{k}, ..., \frac{1}{k}\right).
\end{aligned}$$

This has variance:

$$\begin{aligned}
Var(\hat{p}^*_{Freq,i}) &= Var\left(\frac{n^*_i + 1/k}{n+1}\right) \\
&= \frac{1}{(n+1)^2} Var(n^*_i) \\
&= \frac{np_i(1-p_i)}{(n+1)^2} \\
&= \frac{n\frac{1}{k}\frac{k-1}{k}}{(n+1)^2} \\
&= \frac{n(k-1)}{k^2(n+1)^2}.
\end{aligned}$$

From this, we can get a Chebyshev bound of the form:

$$\begin{aligned}
P(||\hat{p}_{Freq} - p_0||_2^2 > a^2) &\leq \frac{E(||\hat{p}_{Freq} - p_0||_2^2)}{a^2} \\
&\leq \frac{E(||\hat{p}_{Freq} - E(\hat{p}_{Freq})||_2^2)}{a^2} \\
&\leq \frac{E(\sum_{i=1}^{k}(\hat{p}_{Freq,i} - E(\hat{p}_{Freq,i}))^2)}{a^2} \\
&\leq \frac{\sum_{i=1}^{k} Var(\hat{p}_{Freq,i})}{a^2} \\
&\leq \frac{\sum_{i=1}^{k} \frac{n(k-1)}{k^2(n+1)^2}}{a^2} \\
&\leq \frac{\frac{n(k-1)}{k(n+1)^2}}{a^2} \\
&\leq \frac{n(k-1)}{a^2 k(n+1)^2}.
\end{aligned}$$

We now switch our focus to:

$$E_\theta||\Delta(Z) - \hat{p}_{Freq}||_2^2 = E_\theta \sum_{i=1}^{k} \left(Z_i + \theta_i Z_0 - \frac{n_i + 1/k}{n+1} \right)^2$$

For this section, $Z_i + \theta_i Z_0$ is a function of two independent random variables Z and θ, we must delineate which variable we are taking expectation with respect to. We will let E_Z and E_θ refer to expectations over Z and θ respectively. Like previously, $E_Z E_\theta(Z_i + \theta Z_0) = \frac{n_i + 1/k}{n+1}$ so:

$$\begin{aligned}
E_Z E_\theta(Z_i + \theta_i Z_0) &= E_Z \left(Z_i + \frac{1}{k} Z_0 \right) \\
&= E_Z(Z_i) + \frac{1}{k} E(Z_0) \\
&= \frac{n_i}{n+1} + \frac{1}{k} \frac{1}{n+1} \\
&= \frac{n_i + 1/k}{n+1}.
\end{aligned}$$

Which has component-wise variances:

$$\begin{aligned}
Var_Z E_\theta(Z_i + \theta Z_0) &= Var_Z(Z_i + \frac{1}{k} Z_0) \\
&= Var_Z(Z_i) + \frac{1}{k^2} Var(Z_0) + \frac{2}{k} Cov(Z_i, Z_0) \\
&= \frac{\frac{n_i}{n+1}(1 - \frac{n_i}{n+1})}{n+2} + \frac{1}{k^2}(\frac{\frac{1}{n+1}(\frac{n}{n+1})}{n+2}) + \frac{2}{k}\frac{-\frac{n_i}{n+1}\frac{1}{n+1}}{n+2} \\
&= \frac{n_i(n+1-n_i)}{(n+1)^2(n+2)} + \frac{n}{k^2}\frac{1}{(n+1)^2(n+2)} + \frac{2}{k}\frac{-n_i}{(n+1)^2(n+2)}.
\end{aligned}$$

Thus by the same Chebychev inequality:

$$\begin{aligned}
P_Z(E_\theta||\Delta(Z) - \hat{p}_{Freq}||_2^2 > a^2) &\leq \frac{Var_Z(E_\theta \sum_{i=1}^k (Z_i + \theta Z_0))}{a^2} \\
&\leq \frac{Var_Z(E_\theta \Delta(Z)}{a^2}) \\
&\leq \frac{Var_Z(E_\theta \sum_{i=1}^k Z_i + \theta_i Z_0)}{a^2}) \\
&= \frac{Var_Z \sum_{i=1}^k (Z_i + \frac{1}{k} Z_0)}{a^2} \\
&= \frac{\sum_{i=1}^k Var(Z_i + \frac{1}{k} Z_0)}{a^2} \\
&= \sum_{i=1}^k \frac{n_i(n+1-n_i)}{a^2(n+1)^2(n+2)} + \frac{n}{k^2}\frac{1}{a^2(n+1)^2(n+2)} + \frac{2}{k}\frac{-a^2 n_i}{(n+1)^2(n+2)} \\
&= \frac{n^2 + n - \sum_{i=1}^k n_i^2}{a^2(n+1)^2(n+2)} + \frac{n}{k}\frac{1}{a^2(n+1)^2(n+2)} + \frac{2}{k}\frac{-n}{a^2(n+1)^2(n+2)} \\
&= \frac{k(n^2 + n - \sum_{i=1}^k n_i^2) - n}{a^2 k(n+1)^2(n+2)} \\
&= \frac{kn^2 + (k-1)n - k\sum_{i=1}^k n_i^2}{a^2 k(n+1)^2(n+2)}.
\end{aligned}$$

Then using the union bound, we can create the bound:

$$\begin{aligned}
P(|\pi_{freq} - \pi_{mean}| \geq \alpha) &\leq P(|\pi_{freq} + \pi_{mean}| \geq \alpha) \\
&\leq P(\pi_{freq} \geq \alpha/2) + P(\pi_{mean} \geq \alpha/2) \\
&\leq \frac{n(k-1)}{(a/2)k(n+1)^2} + \frac{kn^2 + (k-1)n - k\sum_{i=1}^k n_i^2}{(a/2)k(n+1)^2(n+2)}
\end{aligned}$$

□

A direct implication of this finite sample bound is that asymptotically, π_{freq} becomes equal to π_{mean}.

Using theorems 5, we have demonstrated that:

$$\pi_{min} \leq \pi_{mean} \leq \pi_{max} \tag{2.24}$$

$$\pi_{min} \approx \pi_{freq} \tag{2.25}$$

From this, one may suspect that:

$$\pi_{min} \leq \pi_{freq} \leq \pi_{max}$$

Indeed, we observe through simulation that this holds, regardless of whether the null or alternate hypothesis is true. In Figure 2.3, we show the results for a test-of-uniformity for a 4-dimensional multinomial distribution:

$$Multniomial\left(q, \frac{1-q}{3}, \frac{1-q}{3}, \frac{1-q}{3}\right)$$

Where $q \in [0.05, 0.45]$, n is 100, and the test statistic is the chi-squared test statistic. When q is 0.25, we have the null hypothesis being true. From the results in Figure 2.3 we can observe that at all values of q, the average upper and lower power is well bounded the average power of a frequentist test while the mean power very closely matches the Frequentist p-value. From this, we now have our two main results, 1) *upper and lower DS p-values can be used as a reasonable bound for frequentist p-values* and 2) *mean p-values allow for the creation of the oft-difficult DS testing procedure that controls the asymptotic Type I error level.* This result is particularly exciting as one of the largest disadvantages of DS is the inability for its testing procedures to guarantee long-running type I error rates, which is something our two main results directly address. In our case, we can say that the correct Type I error rate can be controlled through π_{mean}, while using the gap between π_{upper} and π_{lower} to do the uncertainty analysis that DS is adept at.

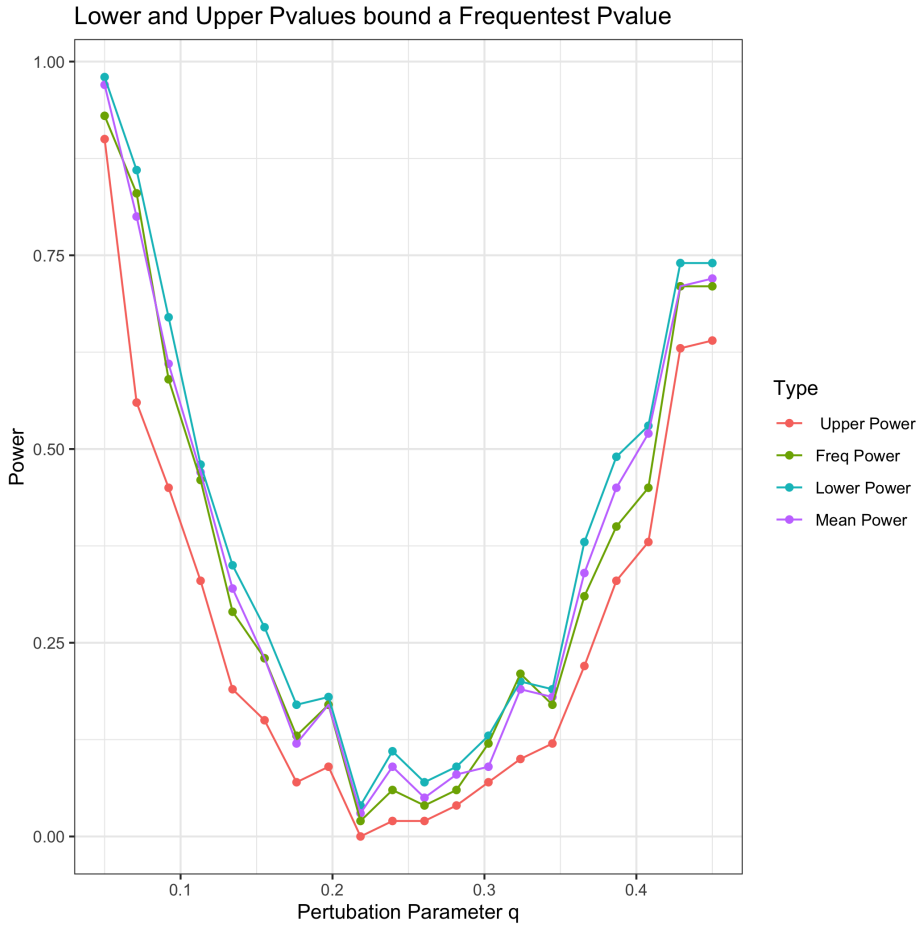


Figure 2.3: Comparison of power to reject the null hypothesis for a variety of perturbation parameters (q). Regardless of choice of q, the Upper p-value and Lower p-value serve as a good bound on the Frequentist p-value and the Mean p-value.

2.7 The Advantages of DS Multinomial Hypothesis Testing

Up to now, we have focused our attention on defining DS Multinomial Hypothesis testing as well as noting its connections to other forms of inference. While the surprising connection between frequentist and DS hypothesis testing is insightful, we would now like to pivot to demonstrating features of our DS hypothesis testing procedure and the powerful properties though an investigation of DS's robust insight into testing and DS's useful high-dimensional properties.

2.7.1 The Unknown Class gives insight into the difficulty of a test

DS has has a long history in the engineering community providing a way to comprehend the level of epistemic uncertainty in a hypothesis test through the inclusion of an "unknown" option. In figure 2.4, we demonstrate the utility of this unknown class through a simulated 4-class Multinomial goodness-of-fit test. The alternate hypothesis had the probabilities $(\frac{2}{6}, \frac{1}{6}, \frac{2}{6}, \frac{1}{6})$. The test statistic used was the chi-squared statistic and the sample size, n, ranged from 4 to 256. At each sample size, 1000 tests were performed and the upper and lower p-values were generated using 100 polytopes. The null distribution of the Frequentist test was simulated from 1000 draws from a multinomial distribution with equal probabilities. No multiple testing correction was performed.

In the first plot in figure 2.4, we can see that while both tests reach the correct 0.05 level as the sample size increases, they have very different behavior at low samples sizes. In the frequentist test, when the dimension to sample size ratio is one, the frequentist test rejects too few tests. This inability to reject is a well-documented high-dimensional phenomenon (Balakrishnan and Wasserman, 2018) not just for multionmials, but for a large class of hypothesis tests (Wainwright, 2019). However, in the DS test, when the sample size is too small, there is a substantial possibility that the hypothesis test will return with an "Uncertain" conclusion. Unlike the frequentist test which fails to reject nearly all, this unknown class is giving additional information to the user on the difficulty of making a conclusion at the current sample size. We see this additional information at work in Figure 2.5. Once again, at low sample sizes, the frequentist test does not give indication if we are failing to reject the null because the null is true or because

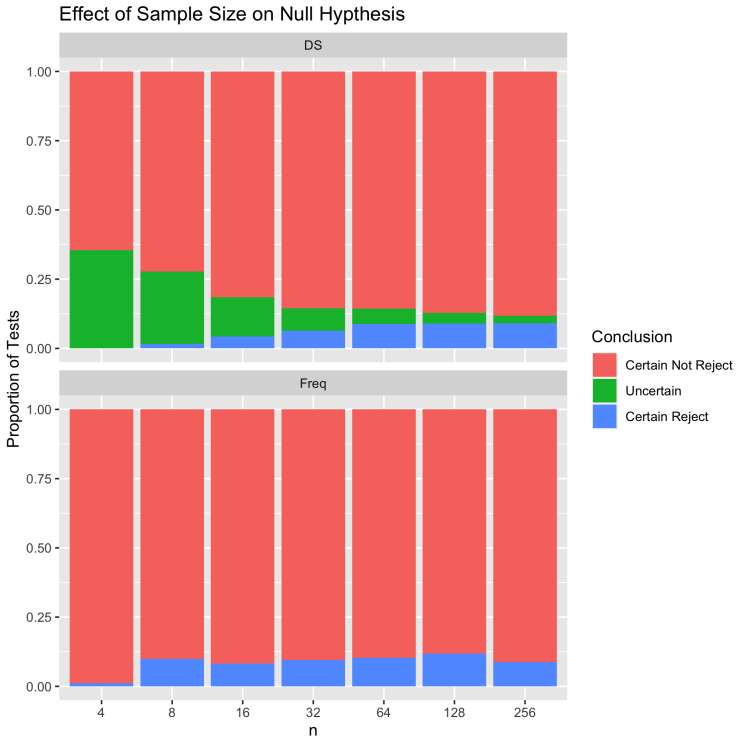


Figure 2.4: Results of Frequentist and DS hypothesis test of uniformity when the null is true. At low sample sizes, the Frequentist test gives a deceptively low rejection rate while the DS properly indicates the difficulty of this problem by having a large probability for an "Uncertain" result.

we lack the sample size to make any conclusions. However, if we look at the DS test, the large probability of a test returning an "Uncertain" result indicates that our issue is in the sample size.

2.7.2 DS can model Adversarial Attacks on a Test

In addition to providing a novel characterization of power, our approach can also employ a technique called "weakening" to evaluate the the effect an adversarial attack would have on a hypothesis. Weakening is a common practice in reliability engineering when one gives additional weight for the unknown class either to satisfy long-running frequentist properties (Martin et al., 2010) or to account for potential uncertainties in the observed data (de Campos and Benavoli, 2011). For our DS model, we will perform weakening by artificially increasing the width of our polytopes. Formally:

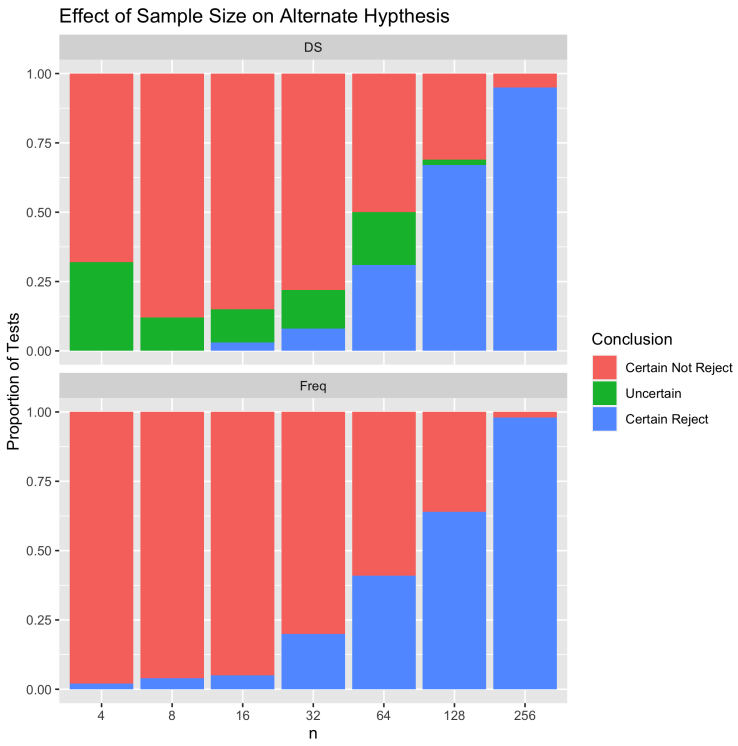


Figure 2.5: Results of Frequentist and DS hypothesis test of uniformity when the null is false. Once again, with the frequentist test, at low sample sizes, it is unclear if the alternate is correct or if the sample size is lacking, while the DS makes this clear through the unknown class that the sample size is lacking.

Definition 2.7.1 (α-Weakened Hypothesis Test)**.** An α-Weakened Hypothesis test is the same as the original DS hypothesis testing procedure except the Dirichlet parameters are drawn from:

$$(Z_0, ..., Z_k) \sim Dirichlet(1 + \alpha, x_1, ..., x_k)$$

where $\alpha > 0$.

Recalling that Z_0 directly controls the width of the random polytopes that serves as our posterior estimates, one can see how this directly increases the gap between the upper and lower test statistics. Since our DS testing procedure returns an "unknown" when the upper and lower test statistics disagree, this weakening directly leads to increasing the possibility of a test returning an 'unknown" result.

As a demonstration of this, in Figure 2.6, we added weakening to a 4-class multinomial hypothesis test with the same alternate hypothesis as the previous simulation. The sample size is 128 and once again, no multiple testing correction was performed. From the results in the figure, we can see that increasing the amount of weakening leads to a dramatic increase in the probability of an uncertain result with a weakening of around 10-20 leading to almost all tests resulting in an unknown result.

In addition to allowing for the addition of user-specified uncertainty, this particular form of weakening comes with an easy-to-understand interpretation. For demonstration purposes, let us consider a 3-dimensional hypothesis test where the observed counts were (n_1, n_2, n_3) with total count $n^* = n_1 + n_2 + n_3$. Since:

$$(Z_0, Z_1, Z_2, Z_3) \sim Dirichlet(1, n_1, n_2, n_3)$$

component wise, the Z_i's have the distributions:

- $Z_0 \sim Beta(1, n + 1)$
- $Z_1 \sim Beta(n_1, n + 1 - n_1)$
- $Z_2 \sim Beta(n_2, n + 1 - n_2)$
- $Z_3 \sim Beta(n_3, n + 1 - n_3)$

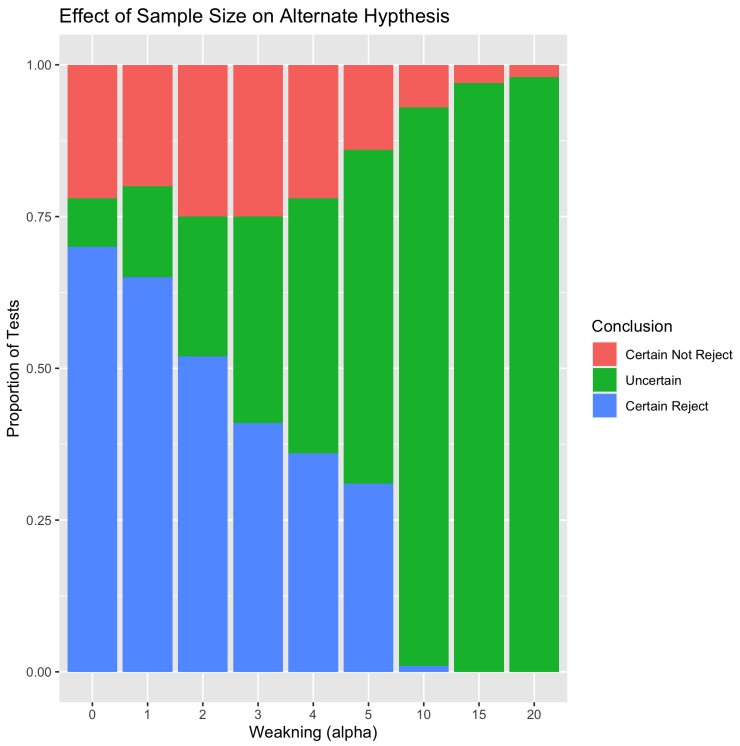


Figure 2.6: Effect of weakening on a DS test of uniformity when the null is false. As the number of adversarial samples (represented by alpha) increases, the conclusions become increasingly more muddled.

As described previously, Z_0 is then added to each of the other Zs to define the edges of the polytope:

- $(Z_1 + Z_0, Z_2, Z_3)$
- $(Z_1, Z_2 + Z_0, Z_3)$
- $(Z_1, Z_2, Z_3 + Z_0)$

So Z_0 controls the width of our polytope while Z_1, Z_2, Z_3 controls the location of the polytope.

Now, let us say that an additional α data points has been added to our dataset. We are unclear which of the 3 classes it has been added to, so we act as though it was as likely to have been added to any of the 3 class. This would result in observed counts of $(n_1 + \frac{\alpha}{3}, n_2 + \frac{\alpha}{3}, n_3 + \frac{\alpha}{3})$ and $n = n_1 + n_2 + n_3 + \alpha$. Component wise, our Z_i's will become:

- $Z_0 \sim Beta(1 + \alpha, n + 1 + \alpha)$
- $Z_1 \sim Beta(n_1, n + 1 + \alpha - n_1)$
- $Z_2 \sim Beta(n_2, n + 1 + \alpha - n_2)$
- $Z_3 \sim Beta(n_3, n + 1 + \alpha - n_3)$

Thus, our width, Z_0, has been increased proportional to the addition of α datapoints. Therefore, α- *weakening is analogous to the effect of adding* α *datapoints to any of the classes.* If we previous confidently rejected or failed to reject before weakening, but then became unknown after α-weakening, this indicates that α datapoints is sufficient to contradict our previous results. Combine this with the previous assertions that π_{lower}, π_{upper} serve as a reasonable bound on the behavior of π_{freq} and we can also make reasonable claims about the behavior of π_{freq} under α-weakening.

Returning to Figure 2.6, this simulation indicates that if 20 datapoints were added to the original 128, it is possible to have the hypothesis reject (hence why $\pi_{lower} \leq 0.05$) and not reject (hence by $\pi_{upper} \geq 0.05$). Stated another way, with 20 additional datapoints, you can make a hypothesis test return whichever result you would like.

2.8 Our DS approach compares favorably to *Simplex-DS*

Finally, we will demonstrate that our method compares well not only to frequentest methods, but also to other DS approaches for multinomials, in particular, one known as *Simplex-DS* (Jacob et al., 2021). *Simplex-DS* determines the edges of the polytope through an resourceful combination of a Gibbs sampler and shortest-path algorithm on a graph. While this approach offers a theoretically interesting representation using shortest path algorithm on graphs, this algorithm has some drawbacks in terms of scalability. For example, generating 1000 MCMC samples a 7-Dimensional multinomial takes approximately 1 minute. However, converting the Gibbs sampler's results into a convex polytope takes nearly 20 times as long as running the Gibbs sampler. This is due to the large number of vertexes in the polytopes that result from the Simplex method. The simulation resulted in polygons with an average of 62 and a max over 250 edges placing a large computational burden on the simplex method.

Time Running the Gibbs Sampler	Time converting MCMC to convex polytope
2.6 seconds	53.6 seconds

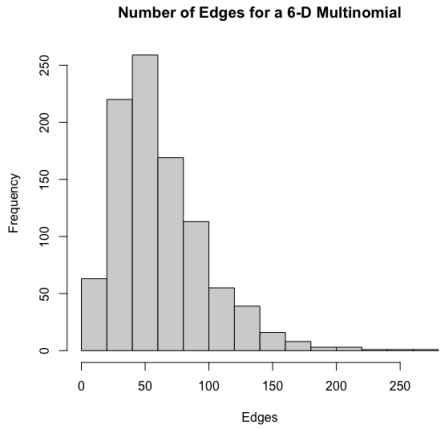


Figure 2.7: (Top): Runtime comparison of Gibbs sampler vs convex polytope computation. (Bottom): The number of edges in 1000 Convex polytopes from the Simplex method.

On the other hand, our Dirichlet DS has a number of clear computational advantages. First, independent sampling from Dirichlet$(1, z_1, \ldots, z_k)$ can be done directly using methods built into most software packages. Additionally the Dirichlet DS polytope is always a simplex, so for $k = 7$ it yields a convex polytope with 8 vertexes, instead of a maximum of 250. This makes the Dirichlet

DS to scale to much larger problems, e.g. multinomial distributions with thousands of categories, which is important for testing independence at high dimensions and/or high resolutions.

Next, many estimators of the multinomial proportions have the following invariance property: If we merge two categories, the estimate of the merged proportion is the sum of the proportions that are being merged. Investigation of Dempster (1966) reveals that Dirichlet DS has this invariance property. Consequently, inference on proportions for categories in which we have observations is not influenced by addition or deletion of empty categories. As seen in Section 4.1 of Jacob et al. (2021), simplex DS does not have this invariance property.

Finally, we have demonstrated that the effect of adversarial attacks can be simply included into the Dirichlet model through the weakening parameter. It is not clear if simplex DS could be weakened to accommodate missing observations.

In terms of performance, we consider a series of tests of independence. First, 100 data sets of sample size $n = 30$ are generated under either the following null or alternate hypotheses:

$$H_0 \sim \text{Beta}(1,1)^2, \quad H_1 \sim \text{Beta}(1,2)^2. \tag{2.26}$$

Each of these data sets is then discretized into 2×2, 3×3, and 6×6 contingency tables and each table is tested for independence using the simplex DS, the Dirichlet DS, and the classical χ^2 tests. To generate the p-values for the test of independence, both the simplex and Dirichlet DS generate 200 polytopes with a burn-in of 300 for the former. The purpose of the low sample size ($n = 30$) in this simulation is to demonstrate that as the resolution k and the number of multinomial categories k^2 increases with sample size held constant, the uncertainty indicated between the gap between the upper and lower p-values increases.

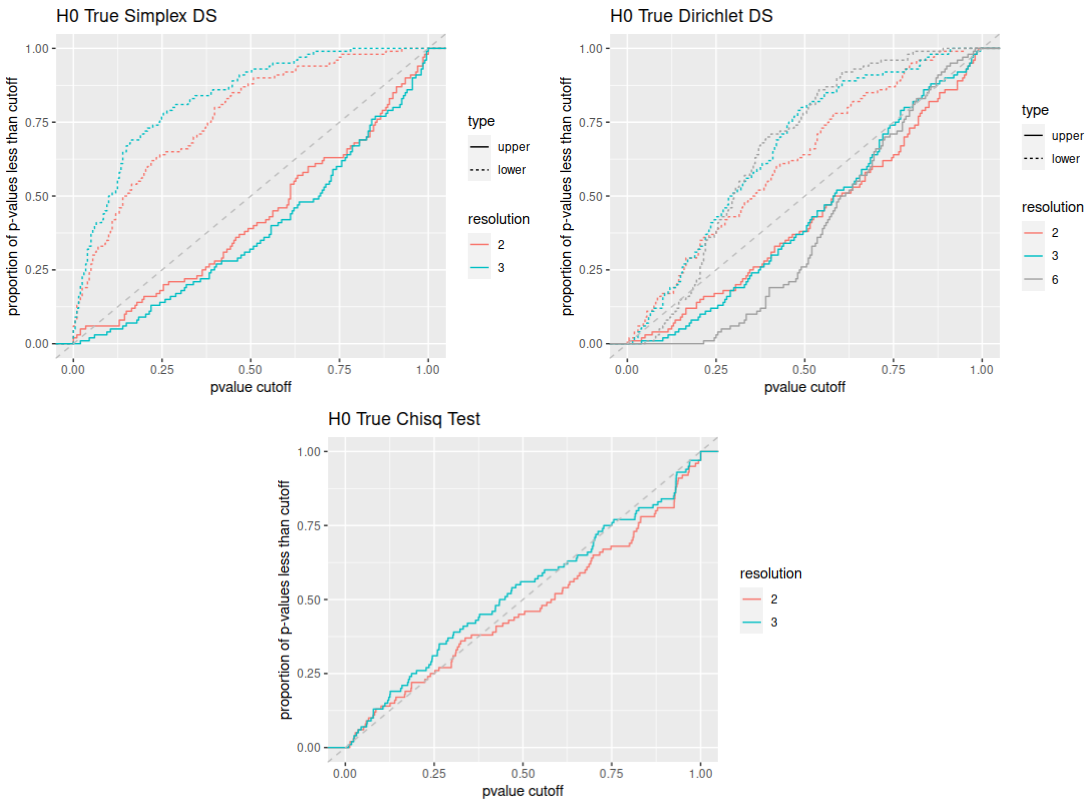


Figure 2.8: Empirical CDFs of the upper and lower p-values for H_0 analyzed using three tests: (Top Left): Simplex DS, (Top Right) Dirichlet DS, and (Bottom): χ^2. The x-axis is the nominal p-value, the y-axis is proportion of p-values below p-value cutoff.

In Figure 2.8 we present plots of the Empirical CDFs of the upper and lower p-values under the assumption that H_0 is true. Well calibrated p-values follow a uniform distribution which CDF is represented by the 45° line. As expected, we see that p-values empirical CDFs from the χ^2 test closely follow this dotted line. The upper p-values for both DS tests are below the dotted line, showing that these p-values are conservative, i.e., sub-uniform. Next, we see that while the upper p-values for the Dirichlet and simplex method behave similarly, the lower p-values of the simplex method are more skewed towards rejecting. Consequently, Dirichlet DS has a much smaller gap between the upper and lower p-values than the simplex DS. Finally, we remark that there are no p-values for the 6x6 simplex method as the computation timed out after 2hr without producing a simplex.

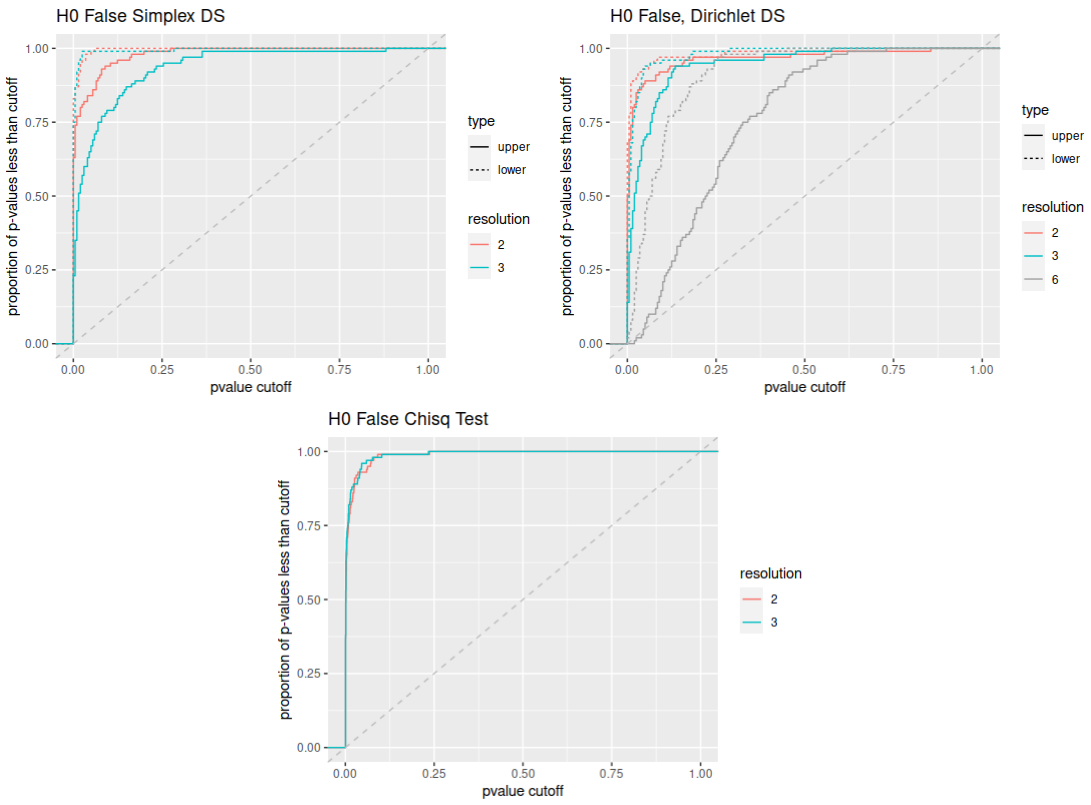


Figure 2.9: Empirical CDFs of the upper and lower p-values for H_0 analyzed using three tests: (Top Left): Simplex DS, (Top Right) Dirichlet DS, and (Bottom): χ^2. The x-axis is the nominal p-value, the y-axis is proportion of p-values below the cutoff.

In Figure 2.9 we show empirical CDFs based on data generated under the alternate hypothesis in (2.26). All three tests correctly lean towards rejecting the null hypothesis. In terms of the power of the lower p-value, the simplex method performs similarly to the Dirichlet method. However, the empirical CDFs of upper p-values for the simplex method is lower than their corresponding empirical CDFs for the Dirichlet DS empirical CDFs. This indicates that for the Dirichlet method has more power to reject H_0. In addition, we can see the gaps between lower and upper p-value plots increase as the resolution increases in both the simplex and Dirichlet DS.

As for runtime comparisons, the difference is substantial. Generating one polytope under Dirichlet DS at the 3×3 level takes approximately 2 seconds while a similar polytope takes nearly 30 seconds under Simplex DS. The difference in runtime comparison gets larger with the 6×6 level, where the Dirichlet DS is still under 5 seconds while the Simplex DS timed out after at least 2 hours.

Method	Runtime to generate one polytope
Our Method for 2×2	1.76 seconds
Our Method for 3×3	2.15 seconds
Our Method for 6×6	4.15 seconds
Simplex-DS 2×2	5.31 seconds
Simplex-DS 3×3	29.60 seconds
Simplex-DS 6×6	Timed out at $>$ 2hr

Table 2.1: Runtime for Our Method vs DS Simplex

2.9 Conclusion

Although Dempster-Shafer inference is not as common of an approach as Bayesian or Frequentist inference, DS Inference has played an important role in statistics, making us reconsider how inference should be performed and how to draw conclusions from data. This problem has been no different. Our DS inference procedure demonstrated that by including a third "unknown" class into hypothesis tests, this can give users new insights into the diminishing effect that dimension has on power in multinomial tests. On the theoretical side, our DS infernece has shown that there exist a strong connection between the pvalues we get from our DS approach and traditional frequentest pvalues, addressing a long-standing weakness in other forms of DS inference. And in application, our approach hints at a direct connection between the theoretical technique of weakening and that of adversarial attacks, giving us a new tool by which we can address this increasingly important statistical problem. From this, we hypothesize that DS still has more to contribute to the statistical world and its insights will be useful for years to come.

Chapter 3

Interpretable Causal Inference for Critically Ill Seizure Patients

3.1 Introduction

Caring for critically ill patients is extremely challenging: the decisions are high stakes, there are difficult causal questions (will this patient respond to available drugs?), and decisions about drug dosage are entangled with observations that physicians are making about the patient over time.

Experiments (clinical trials) on such patients are difficult, observational datasets are noisy and small, and there may be potential important variables, such as drug absorption rates, and the severity of the patient's condition, which typically are not recorded in a database. Ignoring these variables can lead to biased estimates of treatment effects, a naïve statistical analysis is doomed to fail, and the use of black box models in either analysis or decision making could easily lead to erroneous conclusions and cause harm. Ideally, we need an *interpretability-centered* framework for these types of high-stakes causal analyses: a physician should be able to verify the quality of every single step in the analysis, from how a current patient compares to past patients (case-based reasoning), how drug absorption and response is modeled, and an understanding of the relative importance of variables.

This paper introduces a general framework that can help estimate heterogeneous causal effects from high-dimensional patient data with complex time-series interactions, low signal-to-noise ratio and where treatments are not randomly assigned. Each step of the framework is designed to be interpretable. Importantly, we leverage established interpretable pharmacokinetic-pharmacodynamic (PK/PD) models to describe personalized clinical-decision-physiological-response interactions, allowing us to identify individuals who might react similarly to treatments. We learn a flexible distance metric on the space of covariates to perform matching for estimating the medium- and long-term causal effects of both the clinical

decisions and physiological responses; the matched group we construct for each patient can be validated, or possibly, criticized. In the context of medical data, this validation can be performed via a chart review that provides a qualitative assessment of the matches in terms of information that was not directly used in the matching procedure.

Using this framework, we perform the first causal analysis of a common form of potentially harmful electrical activity in the brain known as "epileptiform activity" (Hirsch et al., 2021). EA is common to critically ill patients suffering from brain injury (Lucke-Wold et al., 2015), cancer (Lee et al., 2013), organ-failure (Boggs, 2002), affecting more than half of patients who undergo electroencephalography (Gaspard et al., 2013). Prolonged EA is associated with increased in-hospital mortality, and survivors often suffer from a functional and cognitive disability (Ganesan and Hahn, 2019; Rossetti et al., 2019; Kim et al., 2018). While there is a growing body of literature indicating that EA is *associated* with poor outcomes (Oddo et al., 2009), there is still a debate as to whether (a) EA is part of a causal pathway that worsens a patient's outcomes and thus requires aggressive treatment, *or* (b) the worsened outcomes are due to mechanisms other than EA such as the side-effects of medications or the inciting medical illness, with EA occurring as an epiphenomenon. (Chong DJ, 2005; Rubinos et al., 2018; Osman et al., 2018; Johnson and Kaplan, 2017; Tao et al., 2020; Cormier et al., 2017).

However, the study of EA suffers from a variety of limitations. First, a hypothetical clinical trial studying EA would need to randomly induce EA in patients while limiting their drug treatments, which is neither plausible nor ethical. Second, as it requires a physician to order an EEG and trained technologist to monitor the device, sample sizes for EA datasets tend be to no larger than a few hundred to a thousand and of limited time windows. Worst of all, what complicates the study of EA is its complex interactions with anti-seizure medications (ASM). Medical caregivers administer ASMs based on patients' EA, and in turn, EA is affected by ASMs. Therefore, this creates an entanglement (see Figure 3.1) between the EA (treatment) and ASMs (confounder), potentially obscuring the true causal effect of EAs.

The study of EA has been a case where scientists have been using *predictive* models to answer a *causal* question despite strong confounding factors. Researchers have used regression models to adjust for the patient's medical history and demographic factors (Payne et al., 2014; De Marchis et al., 2016; Zafar et al., 2018; Muhlhofer et al., 2019), and have interpreted the resulting

regression coefficient for EA as the causal effect of EA on a patient's outcome. While this approach is appealing for its simplicity and is widely used, *it is not appropriate to interpret regression coefficients as causal in the presence of strong confounding interactions.* Using conventional prognostic modeling approaches can put one at the risk of misinterpreting the association between high levels of ASM, EA, and poor outcomes as causal even if no causal link exists.

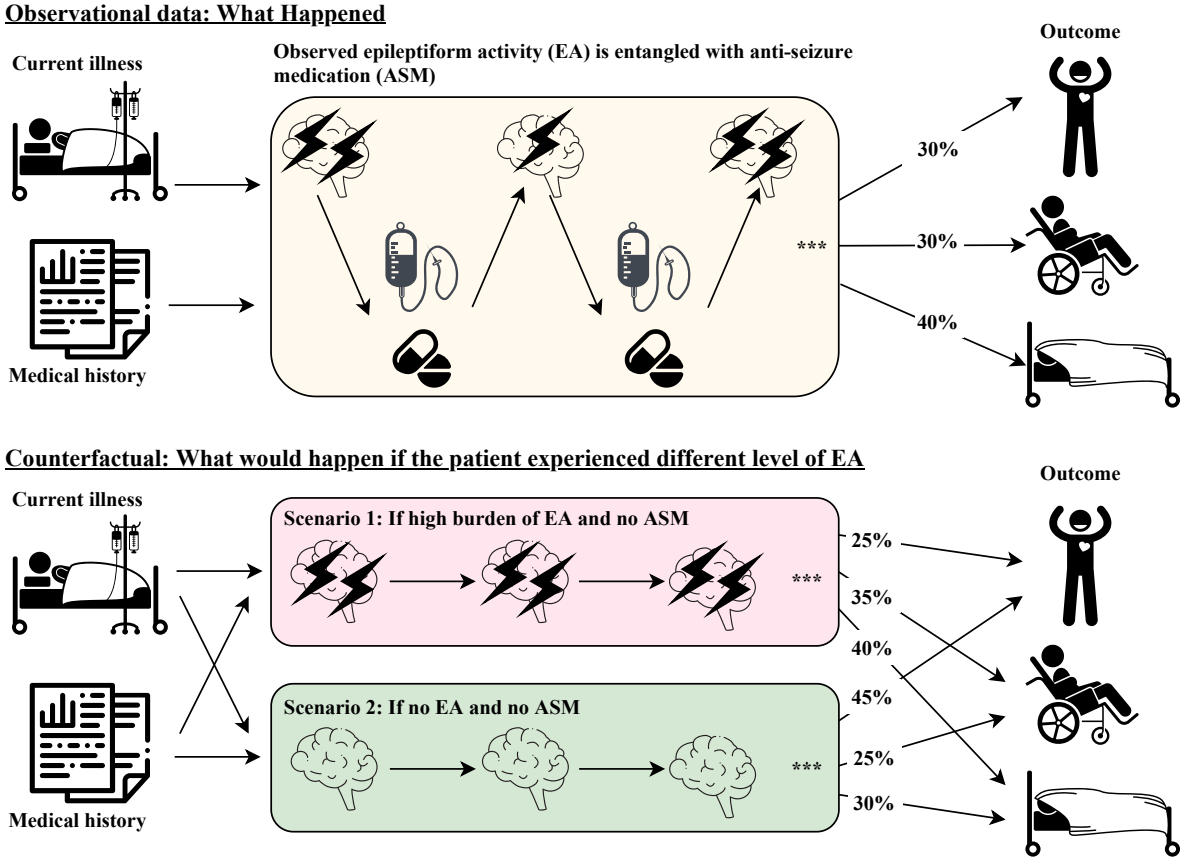


Figure 3.1: *Upper:* Illustration showing that observed epileptiform activity (EA) and treatment decision form a feedback loop, that is also influenced by current illness and medical history (left). The entire time-series of EA and ASM influence patient outcomes. Possible outcomes include return to normal health, disability, or death at the time of hospital discharge (right). *Lower:* Our goal is to estimate the effect of EA on patient outcomes. The effect is obtained by comparing the patient outcome across counterfactual scenarios. Scenario 1 is where every patient in this cohort had certain (high) level (or burden) of EA but no ASM is given; Scenario 2 is where every patient had no EA and also no ASM is could given. (Note that the probabilities given here are illustrative, and not taken from data.) [Credit to Dr. Brandon Westover for the figure]

Our framework is different than other approaches in that it aims to tightly match patients on *all known relevant confounding factors* such as medical history and diagnosis, pharmacological

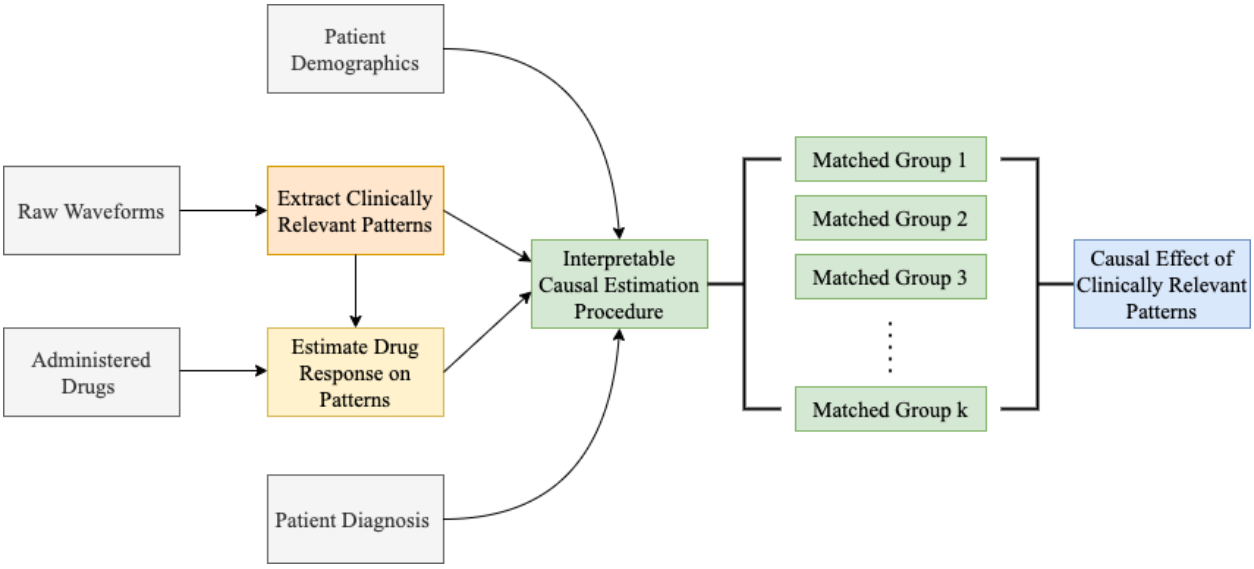


Figure 3.2: Flowchart demonstrating our framework for interpretable inference of causal effects

characteristics and demographics. We adjust for important pharmacokinetic/pharmacodynamic (PKPD) parameters to better characterize individualized responses to anti-seizure medications; this mechanistic information helps compensate for our not-large sample size and limited EEG observation time window. The interpretability of our framework also gives important medical insights into the EA process which are easier for practicing clinicians to incorporate.

We can thus finally provide the first high quality causal analysis of EA. We find that higher EA burden indeed leads to worse neurologic outcomes (Figure 3.4), in a way that depends on the intensity (max burden over 6 hours) and duration (average burden over 24 hours) of EA. Specifically, those with a max burden between 0.75 to 1 are on average 13.4% more likely to be discharged with a poor outcome (as defined by modified Rankin Scale of Burn, 1992) than those with max burden less than 0.25. Additionally, we find that patients with central nervous system infection or toxic metabolic encephalopathy are affected by EA more than the average level in this cohort. Importantly, the validity of the estimate is supported by a detailed clinical chart review of the matched groups, which could only be accomplished because of the interpretability of our framework.

3.2 Framework

The general framework is shown in Figure 3.2. The first step of our framework is the identification of physiological phenomena that might affect long-term health outcomes. Importantly, these phenomena are frequently not recorded directly, and instead relevant patterns must be extracted from raw waveforms. Examples include monitoring blood pressure and serial blood cultures in patients with sepsis; heart rhythms, blood pressure, oxygen levels, and serial blood electrolyte levels in patients with life threatening arrhythmias like atrial fibrillation or atrial flutter; urine output, body weight, and blood electrolytes in patients with acute kidney failure; intracranial pressure and brain tissue oxygen levels in patients with severe traumatic brain injury; or, as in the example that we analyze in this paper, detecting EA from EEG signals. *Our framework focuses on estimating the long-term effects of these patterns.* However, the raw waveform data rarely exists in settings without clinical interventions: we must control for the effects of interventions, for example, in the medical scenarios mentioned above: effects of medications to increase blood pressure and antibiotics given in sepsis; medications to abort arrhythmias and raise blood pressure in patients with atrial fibrillation/flutter; electrolyte infusions, diuretic drugs, and hemodialysis given to patients with acute kidney failure; high concentration saline or surgical treatments given to reduce intracranial pressure in patients with brain trauma; or the amount of antiseizure medication given how well it was absorbed in patients treated for EA.

As the goal is to identify the long-term effects of observed patterns, a patient who was never treated is not comparable to a patient who was. Thus, we combine the patient's demographic variables (e.g., age, weight) and patient characteristics within a pharmacodynamic/pharmacokinetic model to estimate drug response parameters for each patient. The patient data, including drug response parameters, are all used for high-quality matching; each patient is matched *almost exactly* to past patients with similar characteristics, medical history, and estimated drug response parameters. Almost-exact matching (Parikh et al., 2018) matches patients directly on potential confounders (not, for instance, on proxies such as the propensity score). The matched groups permit case-based reasoning, and allow us to estimate the effects of both seizure-like activity and drugs meant to reduce seizure-like activity on patient

outcomes. In addition to these matched groups being almost-exact, domain experts can perform chart review for each patient's matched groups to evaluate their quality. As these charts contain not only the quantitative factors used for matching but also qualitative information such as doctors' notes, they allows for a holistic assessment that might lead to unobserved confounding.

3.3 The Causal Study of EA

As discussed, EA affects more than half of critically ill patients on EEG (Gaspard et al., 2013), and understanding its effects can help prevent severe brain damage. In what follows, we outline our approach to EA analysis following the framework discussed above.

Patient Cohort Our study is a retrospective cross-sectional analysis of patients admitted to the Massachusetts General Hospital (MGH) between September 2011 and February 2017. Institutional review boards at MGH, Duke University, and University of North Carolina at Chapel Hill approved the retrospective analysis without requiring written informed consent. Inclusion criteria included (1) admission to the hospital, (2) monitoring with continuous electroencephalography (EEG) for more than 2 hours, and (3) availability of drug administration data from the hospital's electronic records. Patients who had poor quality of EEG signal for more than 30% of the total recording length or those missing discharge outcome were excluded from the study. For patients with multiple visits to the hospital, we only analyzed their first visit. A flowchart of the full patient selection procedure can be seen in Figure 3.3. The final cohort contained 995 critically ill patients.

For each patient, we collected a variety of variables about their medical history including demographics (gender, marital status, and age), clinical factors (substance abuse, history of seizures or epilepsy, chronic kidney disease, etc.), and what disease(s) they were diagnosed with (cancer, subarachnoid hemorrhage , or central nervous system infection). As this information concerns factors that are fixed before admission to the hospital for treatment, these are referred to as the *pre-admission variables*. A full list of these pre-admission variables can be found in Table 4.2.

Outcomes of Interest Once a patient has stabilized (or passed away), they are discharged from the hospital. The level of disability at discharge is quantified on a 0 to 6 ordinal scale called the

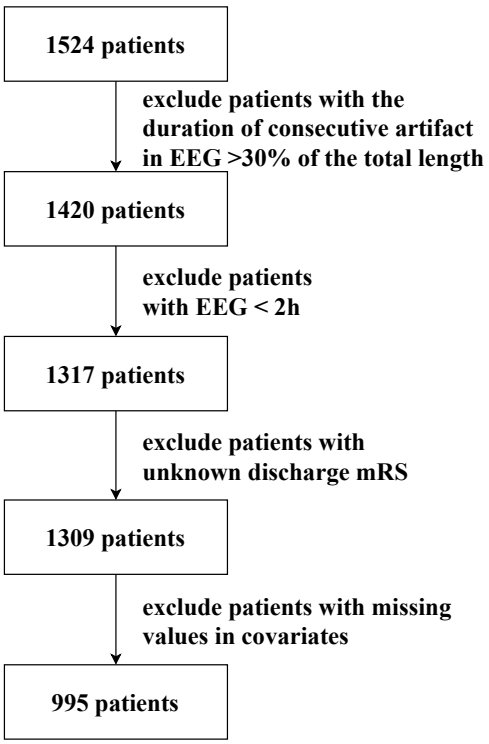


Figure 3.3: Data flowchart showing the preprocessing of patients

Modified Rankin Scale (mRS). In the literature the post-discharge outcome is frequently binarized into those with (mRS$\geq$4) and without (mRS$\leq$3) serious disabilities (Zafar et al., 2018). Our work also uses this binarized Modified Rankin Scale as the outcome of interest, with Y equal to 1 representing a patient discharged with serious disabilities or death and 0 representing a patient without serious disabilities.

Complex Time Series Interactions: Drug treatments and EA After treatment is started, patients are kept under close observation including frequent visits by physicians and nurses, and continuous brain monitoring using electroencephalography (EEG). Based on these observations, physicians update a patient's treatment by adjusting the types and doses of anti-seizure medications (ASMs). This observation-treatment cycle results in: (1) a univariate time series of the average proportion of time the i-th patient experienced EA in the past ω hours($\{Z_{i,t}^{\omega}\}_{t=1}^{T}$) based on an EEG sampling rate of 2 seconds and (2) a 6-dimensional vector time-series ($\{W_{i,t}\}_{t=1}^{T}$) representing the dose of 6 most commonly used ASMs (Lacosamide, Levetiracetam, Midazolam, Phenobarbital, Propofol, and Valproate) received by i-th patient at time-step $0 \leq t \leq T$. We use $\omega = 6$ hours as it is a reasonable amount of time to observe the effects of the

ASMs on EA and for physicians to adjust a patient's ASM regimen (Garoud et al., 2006). Details on how EA signals were identified in the EEG recordings can be found in Appendix **??**.

Clinically Relevant Summaries of EA Burden Over Time. We summarize the EA time series $\{Z^6_{i,t}\}_{t=1}^T$ in two clinically relevant ways, which we refer to as an *EA burden*:

1. *Mean EA burden* ($E_{i,\text{mean}}$) measures the average proportion of time a patient experiences EA in the first 24 hour recording period.

2. *Max EA burden* ($E_{i,\max}$) measures the 6 hour sliding window with the highest proportion of EA within the first 24 hour recording period.

The former measures the prevalence of EA while the second summary provides insights into the most intense periods of EA over a short period of time. By quantifying EA burden in these two different ways, we seek to separately understand the potential harm caused both by brief periods of intense EA and by prolonged periods of less intense EA burden.

Estimands of Interest. We would like to estimate the degree to which untreated epileptiform activity (of different intensities) can cause worse neurological outcomes. The potential outcomes of interest are a function of the full time series of EA burden and drug exposures $Y(\{E_{i,t}, W_{i,t} : t = 1, \ldots, T\})$ and we make the simplifying assumption that they vary only according to the clinically relevant summaries of EA burden and and whether drugs are present or absent. That is, we say that $Y(\{E_{1,t}, W_{1,t} : t = 1, \ldots, T\}) = Y(\{E_{2,t}, W_{2,t} : t = 1, \ldots, T\})$ if $E_{\max}(\{E_{1,t} : t = 1, \ldots, T\}) = E_{\max}(\{E_{2,t} : t = 1, \ldots, T\})$ and $\overline{W}_1 = \overline{W}_2$, where $\overline{W}_i = \mathbf{1}\left[(\sum_t \sum_j W_{i,t,j}) > 0\right]$. Thus, $\overline{W}_i = 1$ if any drugs are administered otherwise $\overline{W}_i = 0$. Our estimand of interest is the probability a patient is discharged with severe disability if the patient has EA burden (either $E_{\max}$ or E_{mean}) equal to e and was not treated with ASMs. This can be represented as:

$$Pr\left[Y_i(E_{i,\max} \in e, \overline{W}_i = 0) = 1\right] \text{ and } Pr\left[Y_i(E_{i,\text{mean}} \in e, \overline{W}_i = 0) = 1\right]. \tag{3.1}$$

Here, Y_i is the binarized post-discharge outcome, e is the binned EA burden with $e \in \{\text{mild}, \text{moderate}, \text{severe, very severe}\}$ and $W_{i,t} = 0\ \forall t$ indicates that no ASMs were ever administered. We are interested in estimating the potential outcome when there are no

administered ASMs because this allows us to disentangle the effects of EA on outcome from the effect of drugs. For interpretability, we bin EA burden (e) into 4 levels – mild (0% to 25%), moderate (25% to 50%), severe (50% to 75%), very severe (75% to 100%) – see Table 4.4 in the the Appendix for the number of patients in each category. The choice of cutoffs was influenced by animal models which showed that an EA burden of 50% serves as an important indicator of when EA begins to damage the brain (Trinka et al., 2015). A sensitivity analysis to these choices is provided in Appendix 4.5.

The variables we Control for: Pre-admission Covariates and Drug-response Covariates. In the ASM observation-treatment procedure, we observed two large sources of potential confounding. First, those with different diagnoses and patient characteristics may receive more or less ASM treatment from physicians, potentially confounding the estimated harm caused by EA with the harm due to diagnosis or patient characteristics. To address this, a collection of 70 pre-admission covariates that could potentially influence ASM treatment were selected by a group of practicing neurologists and were controlled for via the matching algorithm, Matching After Learning To Stretch (MALTS).

A second source of potential confounding comes from a patient's drug response. Due to differing past medical history, current medical conditions, age, and other factors, some patients respond well to some ASMs while other patients respond less. This in turn, can directly affect the amount and number of ASMs that a patient receives and their final outcome. To account for this, we modeled each patient's response to ASM drugs via a one-compartment Pharmacokinetic/Pharmacodynamic (PK/PD) model, and controlled for each patient's drug responsiveness parameters using MALTS.

3.4 Result: EA Burden have a Direct Causal Effect on Survival

With the EA data summarized as above, our framework can now provide the first causal analysis of the effect of seizure-like activity on the possibility of severe brain damage.

Average Effect of Max EA Burden on Patient Outcomes. Figure 3.4(a) illustrates our first main result: those with higher levels of E_{max} are at higher risk of poor neurologic outcomes. Moreover, the risk of a poor outcome increases monotonically as the EA burden increases,

(a) $Pr[Y(E_{\max} = e, \overline{W}_{i,t} = 0) = 1]$ (b) $Pr[Y(E_{\text{mean}} = e, \overline{W}_{i,t} = 0) = 1]$

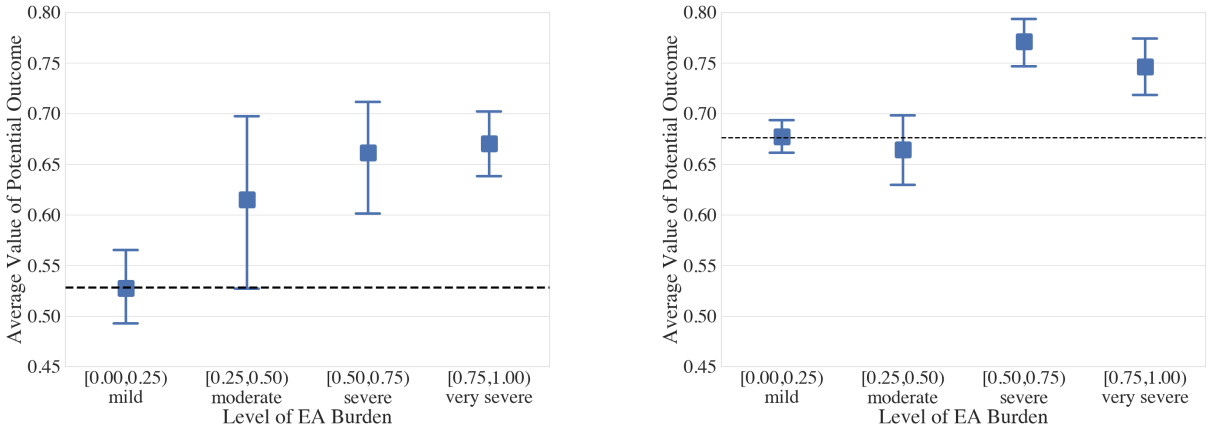


Figure 3.4: The probability of a poor outcome mRS for either Mild, Moderate, Severe, or Very Severe EA burden. EA Burden is quantified as (Left): Maximum EA in a 6-hour moving average window; (Right): Mean EA in a 6-hour moving average window. In both scenarios, an increase in EA burden leads to a worse outcome for the patient. Outcome worsens monotonically for $E_{\max}$, whereas for E_{mean}, there is a jump at approximately 0.5. In both plots the horizontal line represents the baseline median average potential outcome for mild case. Note that these baselines need not be equal due to the marginalization over $\overline{W}_{i,t}$.

culminating in *an average increase of* 16.7% *in probability of a poor outcome when a patient's untreated EA burden increases from mild (0 to 0.25) to very severe (0.75 to 1).*

Average Effect of Mean EA Burden on Patient Outcomes. Figure 3.4(b) shows our other main result: those with higher levels of E_{mean} are also at higher risk of being discharged with poor outcomes. However unlike E_{max}, the risk caused by increasing E_{mean} spikes up when a patient goes above even a moderate EA burden, $[0.25, 0.50)$. Our results indicate that *severe and very severe prolonged EA burden (over 24 hours) increase the risk of worse outcome by* 11.2% as compared to mild or moderate prolonged EA burden.

Heterogeneity in Effects for Max EA Burden. While increases in EA burden tend to lead to worse outcomes overall, we found also that there is significant heterogeneity in the size of the effect due to each patient's pre-admission covariates. We can quantifying the relative change in outcome from a very severe max EA as the ratio of expected outcomes for those with high EA

burden over the expected outcome of those with low EA burden minus one:

$$\text{Average Effect of EA} = \frac{Pr\left[Y(E_{\max} \geq 0.75, \overline{W} = 0) = 1\right] - Pr\left[Y(E_{\max} < 0.25, \overline{W} = 0) = 1\right]}{Pr\left[Y(E_{\max} < 0.25, \overline{W} = 0) = 1\right]}. \tag{3.2}$$

Thus if the average effect of EA is zero, a very severe max EA is no worse than a mild max EA while an average effect of EA of one would represent a 100% increase in the probability of a bad outcome. Based on this relative effect, we observe that *those with central nervous system infections or with toxic metabolic encephalopathy are at higher risk of a worse outcome if they had a large increase in E_{max} burden.* We conjecture that this may be the result of a central nervous system infection and EA leading to a higher inflammatory response in the patient, potentially leading to or exacerbating damage to the brain. Figure 3.5(a) uses a decision tree to break down the population into subpopulations with differing conditional average treatment effects.

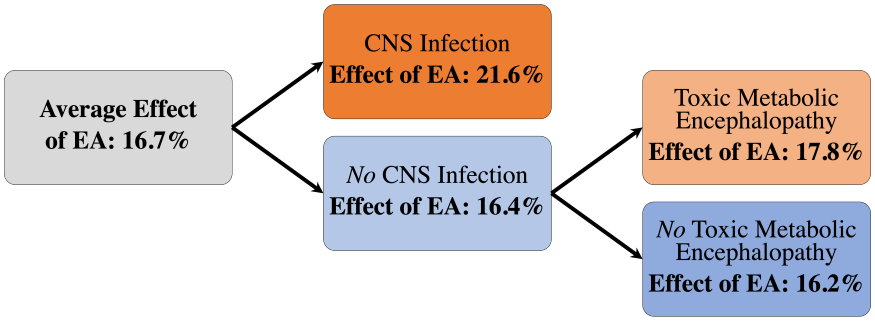
We further examined race and gender as possible effect modifiers of EA burden. Figure 3.5(b) shows that race does not seem modify the risk from increases in E_{max}. By contrast, sex does appear to modify the risk: male patients appear to be more susceptible to very severe $E_{\max}$ worsening the chances of recovery compared to female patients (see Figure 3.5(c)).

3.5 Interpretable Matched Group Analysis

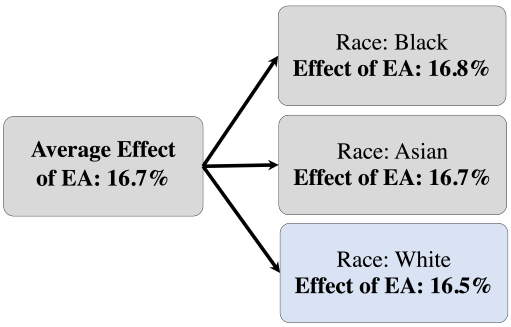
In this section, we provide an assessment of the quality of the matched groups. These types of analyses determine trust of the causal conclusions.

3.5.1 Stretch Coefficients Give Insight into the Matching Process

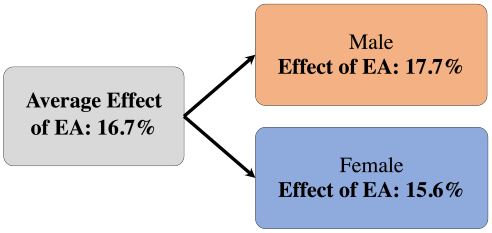
Through visualizing the stretch coefficients, one can gain insight into the relative importance of variables in the MALTS matching procedure. For max EA burden, one can see in Figure 3.6 that two medical scoring systems were both heavily weighted, with the iGCS Score being the most important variable and APACHE II score being the third most important variable. These two scoring systems capture a patient's level of consciousness and severity of illness. When considering that age and systolic blood pressure were the second and fourth most important variables, this shows that our matched groups essentially must consist of individuals that agree on these medical scores and biomarkers representing overall health and current level of neurologic



(a) Recursive partitioning of covariate space and respective relative effects of very severe EA



(b) Effect by Race



(c) Effect by Sex

Figure 3.5: Heterogeneity in the average effect of EA, stratified by: (a) Recursive partitioning on the entire covariate space using Gini splitting to find the most important splits; (b) Partitions the space according patients' race. The remaining race classes (other, undisclosed, and missing) are rare representing 0.5%, 5%, and 8.4% of the total population. (c) Partitions the space across to patients' gender. Orange coloring in the boxes implies that the subgroup experiences a larger average estimated causal effect of EA on neurologic outcomes than the cohort mean, and blue implies a smaller causal effect. Subgroups in orange fare worse as a result of a higher EA burden. [Credit to future Dr. Harsh Parikh for the figure]

impairment. In Figure 3.6, one can also see that the three least important variables to match on are Hill coefficients and ED_{50} parameters from one of the anti-seizure medications. This stands in contrast with the ED_{50} parameter for Propofol, which was one of the top five *most* important variables. This presents an interesting result: perhaps Propofol, a potent intravenous anesthetic drug used to treat seizures, including information about how it is prescribed, may be much more important in understanding the effect of seizure burden than its fellow anti-seizure medications, most of which are less potent.

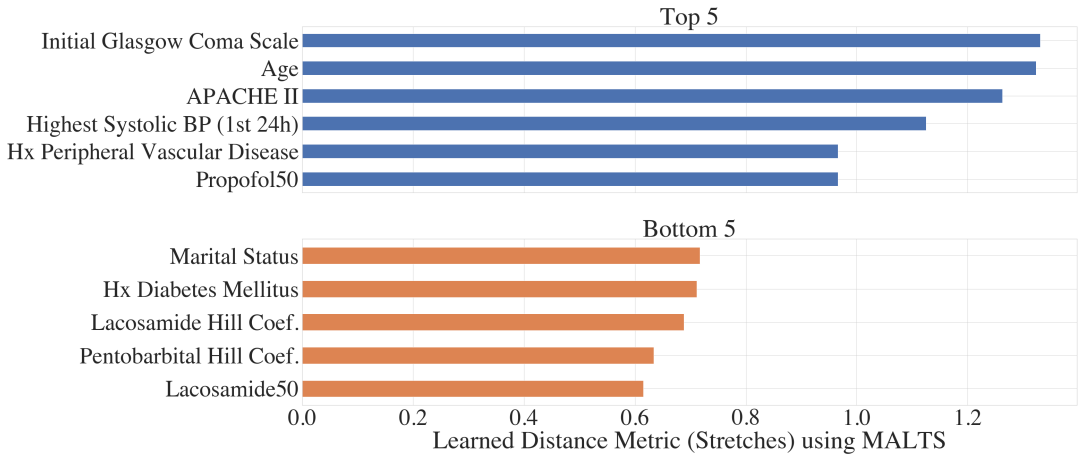


Figure 3.6: The top and bottom 5 variables, based on the average stretching weights in MALTS, when we are studying the effect of the maximum EA burden $E_{\max}$. BP = blood pressure; Coef = coefficient.; Lacosamide50 = concentration of Lacosamide that reduces EA burden by 50%; Propofol50 = concentration of propofol that reduces EA burden by 50%; Hx = History.

3.5.2 Matched Groups are Validated by Neurologist's Chart Review

To ensure the validity of our causal conclusions, it is crucial that the matching process does not overlook major unobserved confounding factors. Inspired by similar approaches in the social sciences (Hasegawa et al., 2019), one can check for unobserved confounders by having a domain expert perform a post-facto analysis of the matched groups. If a domain expert who has access to all of a patient's medical information finds the patients in each matched group to be qualitatively comparable, this gives us confidence that we are controlling for all relevant sources of confounding.

This approach to considering unobserved confounding is well suited for medical data. In addition to factors that are easy to quantify, such as APACHE II scores, it is common for a patient to have a large volume of qualitative information along with quantitative data in the form of doctor's notes and documentation. As doctors' chart reviews are not restricted to quantitative information, this ensures that we are checking for qualitative and quantitative sources of unobserved confounding.

For our matched groups analysis, three practicing neurologists, Chart Reviewers 1, 2, and 3 (CR 1-3), were sent 3 randomly chosen matched groups and asked to perform a manual chart review of the selected patients. Based on these charts, the neurologists were asked to independently make a qualitative analysis of the patients within the matched groups and to

report their outcome prognosis (chance of a poor neurologic outcome) and likelihood of experiencing a high EA burden.

From the results of the post-hoc analysis, as in Table 3.1, the three neurologists found no problematic sources of confounding, therefore validating our causal effect estimate. Moreover, from the reviewer's qualitative analyses, we can observe which factors each matched group was matched tightly on. For example, group three is tightly matched with all patients having similar APACHE II scores and all but one having relatively good prognoses. This contrasts with group one, where patients are tightly matched on acute neurological injuries at the cost of a tighter match on APACHE II score. Viewing what is tightly matched in each group provides a holistic evaluation of which factors have been properly controlled for, such as age, and which factors are either unimportant or lack the sample size to tightly match upon, such as many of the less common diseases.

3.6 Discussion

We presented four main points in this work: (1) First, we developed a novel framework that combines mechanistic modeling with a distance metric learning-based matching method to adjust for complex time-series confounders. (2) Second, we have provided, for the first time (to the best of our knowledge), an estimate of the causal effect of epileptiform activity (EA) on post-discharge outcomes in patients with critical illness. We find that higher EA burden indeed leads to worse neurologic outcomes (Figure 3.4), in a way that depends on the intensity (max burden over 6 hours) and duration (average burden over 24 hours) of EA. (3) Third, our results provide insights into individualized potential outcomes. For example, we show that patients with central nervous system infection or toxic metabolic encephalopathy are affected by EA to a higher extent compared to the average level in this cohort. (4) Finally, we leveraged the interpretability of our approach to validate our matched patients via chart review with the help of three neurologists. The general consensus in the chart review found that the matches were of high quality, matching together patients with similar prognoses.

Clinical Implications. Our findings have two primary implications for treatment of EA: (1) First, treatment should be based on EA duration as well as intensity. We find that intense

periods of EA burden (max EA), even if relatively brief (6 hours) lead to worse outcomes. By contrast, sustained periods of EA (mean EA burden) show a binary relationship with outcome: EA $< 50\%$ has minimal effect, but EA $\geq 50\%$ causes worse outcome. This suggests that interventions should put higher priority on patients with mean EA burden higher than 50%, while treatment intensity should be low and conservative when EA intensity is low. (2) Second, treatment policies should be based on admission profile, because the potential for EA to cause harm depends on age, past medical history, reason for admission, and other characteristics. For example, as our results suggest, patients with central nervous system infection or toxic metabolic encephalopathy should be monitored more closely with more robust treatment. By contrast, current treatment protocols used in hospitals tend to be generic, recommending that treatment be tailored based on the intensity or duration of EA (e.g., more aggressive treatment for status epilepticus), but providing little guidance on how to take other patient characteristics into account. As a result, treatment approaches vary widely between doctors. This suggests an opportunity to improve outcomes by personalizing treatment approaches.

Results in context. Our work builds on prior results demonstrating associations between EA, treatments, and neurologic outcomes. Oddo et al. (2009) studied a cohort of 201 ICU patients where 60% had sepsis as a primary admission diagnosis. They found that EA (seizures and periodic discharges) were associated with worse outcomes, after performing a regression adjustment for age, coma, circulatory shock, acute renal failure, and acute hepatic failure. However, these authors did not adjust for treatment with ASM, including phenytoin (given to 67% of patients), levetiracetam (62% of patients), lorazepam (57% of patients), and four other drugs. Tabaeizadeh et al. (2020) found that the maximum daily burden of EA/seizures, together with their discharge frequency, are associated with higher risk of poor outcome (mRS at hospital discharge 4–6) in 143 patients with acute ischemic stroke. However, they did not control for ASMs which were given to 83% of patients. Lack of adjusting for drug use is also found in the pediatric literature on EA (Ganesan and Hahn, 2019). Not adjusting for treatment is problematic because a growing number of studies suggest that aggressive treatment with ASM may be harmful. One example is the use of therapeutic coma for status epilepticus, where anesthetics such as pentobarbital or propofol are used to temporarily place the brain into a state of profoundly suppressed activity to stop EA while giving treatments of the underlying cause of EA

to take effect. Recent evidence shows that use of therapeutic coma is associated with worse outcomes, including a recent retrospective study of 467 patients with incident status epilepticus of Marchi et al. (2015) which found that therapeutic coma was associated with poorer outcome, higher prevalence of infection, and longer hospital stay (Lin et al., 2017; Rossetti et al., 2005). However, because more aggressive treatment is reserved for more severely ill patients, these studies have also come under criticism for failing to adequately adjust for the type and severity of medical illness, and for the burden of epileptiform activity. Adequately adjusting for these factors has been challenging before now because of the complex interactions and feedback loops involved. However, without adjusting for these factors, it remains unclear whether the association between EA and poor outcomes is due to over-treatment, the underlying illness, or the direct effects of EA. Without an answer to this question, it has remained unclear whether current treatment approaches are helping or hurting patients.

We addressed this gap by introducing an analytic approach that is able to simultaneously account for the entwined and time-varying effects drug and EA burden, and their interactions with patient characteristics. One key component of our approach is adjusting for patients' pharmacodynamic (PD) parameters to account for heterogeneity among patients. Critically ill patients can be different in many ways including measured and unmeasured variables. PK/PD parameters provide a way to quantify the dynamics of the propensity of experiencing EA. The PK/PD parameters are important to take into account since they create spurious correlations impacting both the propensity of having high EA burden and the clinical outcome. By accounting for PK/PD parameters, we were able to adjust for exposure to anti-seizure drugs, such as phenytoin and pentobarbital, where the medications themselves may worsen outcomes. Because prior studies did not disentangle the potential harmful effects of EA and seizures from anti-seizure drugs, the field remained worried but uncertain. Another key innovation is our application of an advanced methodology designed specifically for causal inference using observational data. In the prior studies cited above, multiple regression was used to adjust for potential confounding from medical history and patient diagnosis, but not drug response. The nature of observational data and multivariate regression (model misspecification) have made it impossible to establish a causal link between seizures and other EA vs. clinical outcome. The matching approach in MALTS, being a causal inference method, achieves both the flexibility of being free of model

misspecification (non-parametric) and the interpretability of the learned weights, therefore creating less biased estimates of the causal effects. With this new approach, we are able to provide, for the first time, credible estimates of how much harm EA causes and in which types of patients. Moreover, MALTS comes with the additional advantage that one can easily perform post-hoc analyses of the matched groups, ensuring that the causal claims are accounting for potential unobserved confounders that an expert may be able to identify.

Our approach has several limitations that could be improved in future work. When evaluating the EA burden, it would be worthwhile in future work to consider the subtype of EA (GPD/LPD/LRDA), discharge frequency for periodic discharge patterns, the morphological features (such as seizure with/without triphasic waves), and the spatial extent of EAs. We currently do not have high quality human labels at the necessary resolution to pursue these tasks. On the other hand, the automatic EA annotator, based on a trained convolutional neural network, although not perfect, achieves similar inter-rater reliability as that of experts for the six normal/EA/seizure patterns (Ge et al., 2021a). To reduce this uncertainty in this study, we grouped these EA patterns into binary EA (seizure/GPD/LPD/LRDA) vs. non-EA categories (GRDA/normal/artifact). The definition of EA burden is also relatively coarse compared to those defined by Ganesan and Hahn (2019). The PK/PD model can be further improved by including more mechanistic or physiological detail, such as a context sensitive half-life for propofol (Hughes et al., 1992).

In summary, our results present a data-driven statistical causal inference approach to quantify the harm of EA in ICU. We not only confirm that EA burden (adjusted for ASM) are indeed harmful and worsen patients' neurologic outcomes, but careful analysis illustrates that there exist important subgroups of patients that are more affected by EA. Based on this, a future direction is to learn an interpretable optimal treatment policy for EA burden to improve patient outcomes.

Table 3.1: Three randomly chosen matched groups. We include doctors' prognosis and their qualitative estimate of risk of high EA burden. The notes column presents neurologists' remarks during chart review, based on medical notes not used for matching. The last column contains neurologists' notes on quality of matched groups: the third group is the tightest, followed by the second and the first. Crucially, this ordering matches an automatic measure of tightness produced by MALTS.

Age	Gender	APACHE-II	Doctor's Poor Outcome Prognosis	Doctor's Estimate of High EA	Patient Summary	Matched Group Analysis
				(a) Matched Group 1		
57	Male	15	CR1, CR2: 40% - 60% CR3: 20% - 40%	CR1, CR2, CR3: **40% - 60%**	History of coronary artery disease and tongue cancer, Admitted due to cardiac arrest. On arrival to the hospital, he is comatose.	All three reviewers noted that the patients were similar in age with high risk of EA due to acute neurological injuries (ruptured brain aneurysm, brain tumor). However, as coma patients following a cardiac arrest tend to carry a worse prognosis, while patients with refractory epilepsy have a very good chance of recovering, the range of APACHE-II scores (6-15) is broad, this group is not tightly matched.
54	Male	7	CR1, CR3: 40% - 60% CR2: 60% - 80%	CR1: 40% - 60% CR2: 60% - 80% CR3: 20% - 40%	History of rectal cancer. Admitted due to a ruptured brain aneurysm. On arrival to the hospital, he is mildly confused.	
57	Female	4	CR1: 80% - 100% CR2: 60% - 80% CR3: 40% - 60%	CR1, CR2: 60% - 80% CR3: 40% - 60%	With a brain tumor, admitted due to a seizure. Brain tumor has grown larger and is causing swelling in the brain	
59	Male	6	CR1, CR2, CR3: **80% - 100%**	CR1: 60% - 80% CR2, CR3: 80% - 100%	With epilepsy, admitted due to generalized convulsive seizures multiple times per day. Between seizures, he is cognitively normal.	
				(b) Matched Group 2		
62	Female	3	CR1, CR2: 80% - 100% CR3: 60% - 80%	CR1, CR3: 60% - 80% CR2: 80% - 100%	History of a benign brain tumor (meningioma), complicated by epilepsy, treated with anti-seizure medication. Admitted due to recurrent generalized convulsive seizures.	Four of the patients are similar in age. The other patient is much younger, but her history of severe chronic illness makes her comparable to the other patients. All patients have relatively high risk for seizures / EA (risk factors: brain tumor, brain blood vessel malformation, brain hemorrhage, brain tumor, and epilepsy). Based on data available at hospital admission, the three patients with history of epilepsy or relatively static neurological injury (treated AVM with minor hemorrhage, remote meningioma, refractory epilepsy) have good short term prognosis compared to those with large intra-parenchymal tumor with cerebral edema and midline shift.
28	Female	3	CR1, CR3: 60% - 80% CR2: 80% - 100%	CR1: 0% - 20% CR2: 80% - 100% CR3: 60% - 80%	With multiple previous episodes of severe pneumonia and a large brain blood vessel malformation (arteriovenous malformation) that has caused focal seizures. Admitted due to pain and bleeding in the right eye.	
54	Male	7	CR1, CR3: 40% - 60% CR2: 60% - 80%	CR1: 40% - 60% CR2: 60% - 80% CR3: 20% - 40%	With rectal cancer in the past, admitted to the hospital because of a ruptured brain aneurysm. On arrival to the hospital he is mildly confused.	
64	Male	6	CR1: 60% - 80% CR2: 80% - 100% CR3: 40% - 60%	CR1, CR3: 60% - 80% CR2: 80% - 100%	History of seizures, admitted due to generalized convulsive seizure. At admission, there is a large frontal brain tumor.	
59	Male	6	CR1, CR2, CR3: **80% - 100%**	CR1: 60% - 80% CR2, CR3: 80% - 100%	With epilepsy, admitted due to generalized convulsive seizures multiple times per day. Between seizures, he is cognitively normal.	
				(c) Matched Group 3		
48	Male	4	CR1, CR3: 60% - 80% CR2: 80% - 100%	CR1, CR3: 40% - 60% CR2: 80% - 100%	Past headaches led to discovery of a brain tumor, admitted to have the brain tumor surgically removed.	The age range of these patients is relatively broad, however they have similar APACHE II scores, and similar levels of consciousness on arrival to the hospital. All of the patients in the group have a similar risk for seizures based on their baseline information: tumors or bleeding in the brain or (suspected) seizures. Based on data available at the time of admission, all of the patients have at least a moderately high chance ($\geq$60%) of a good neurological outcome.
61	Female	5	CR1, CR2: 80% - 100% CR3: 40% - 60%	CR1, CR2: 80% - 100% CR3: 60% - 80%	Epilepsy since childhood. Free of seizures for several years. Admitted since she began having generalized convulsive seizures again.	
49	Male	4	CR1, CR3: 60% - 80% CR2: 80% - 100%	CR1, CR2: 80% - 100% CR3: 40% - 60%	Unexplained adult-onset blindness, neuropathy, vertigo, problematic coordination, and migraines. Admitted due to suspected seizures, where he suddenly develops tunnel vision and becomes unresponsive for 5 minutes.	
55	Male	4	CR1: 60% - 80% CR2: 80% - 100% CR3: 40% - 60%	CR1, CR3: 60% - 80% CR2: 80% - 100%	Past hepatitis C and a benign neck tumor, admitted due to balance problems and a fall. Brain imaging shows bleeding around the brain.	
55	Male	4	CR1: 60% - 80% CR2, CR3: 80% - 100%	CR1, CR3: 40% - 60% CR2: 80% - 100%	Past hypertension, diabetes, recent back surgery, and a fever. Admitted due to unable to talk. No evidence of stroke, suspect he has had a seizure	
42	Male	3	CR1, CR2: 60% - 80% CR3: 40% - 60%	CR1, CR2, CR3: **60% - 80%**	History of lung cancer spread to brain, causing occasional seizures. Admitted due to difficulty to wake up after nap. Several brain tumors are bleeding.	
61	Male	3	CR1, CR2: 60% - 80% CR3: 40% - 60%	CR1, CR2, CR3: **60% - 80%**	History of lung cancer spread to liver. Admitted due to confusion. A new mass in brain. Suspect that cancer has spread to brain and may cause seizure.	

Chapter 4

Local Change Point Detection and Signal Cleaning

4.1 Introduction

The Ensemble Empirical Mode Decomposition (EEMD) and the preceding Empirical Mode Decomposition method have become important techniques for the decomposition of nonlinear and non-stationary signals in fields including medicine (Liu et al., 2012; Lozano et al., 2016), hydrology (Wang et al., 2015), seismology (Wang et al., 2012), and mechanical engineering (Chen and Cui, 2016; Zheng et al., 2017). A reason for their success has been the EEMD's ability to create data-adaptive, rather than predefined, basis functions called Intermediate Mode Functions (IMFs). These adaptive basis functions can be non-stationary and nonlinear, making them ideal for complex signals that are not as natural to express in Fourier or Wavelet bases.

However, this data-adaptive nature of the EEMD's basis functions can make it hard to know *a priori* in which basis function a signal may end up. For instance, consider a chirp signal linearly increasing in frequency perturbed with white noise. When decomposed by EEMD, we can see in Figure 4.1 that the signal glides between IMFs 8-6. Common EEMD signal cleaning techniques such as those used in (Wu et al., 2019; Hotradat et al., 2019; Chen et al., 2019; Lei and Zuo, 2009; Huimin et al., 2017; Li et al., 2016; Liu, 2015; Gaci, 2016) first decompose the signal into its base IMF functions, but then treat the entire length of an IMF as either signal or noise. However, in this example, due to the increasing frequency of the chirp signal, no basis function is consistently signal or noise. To properly clean this signal, a more nuanced technique that is able to identify subsections of IMF as signal or noise is necessary. In this paper, we provide a novel example of an EEMD signal cleaning technique, Local Change Detection and Signal Cleaning (`LCDSC`), that is able to identify and clean subsections of EEMD signals. Moreover, we show how this technique can improve the identification of acoustic shock waves.

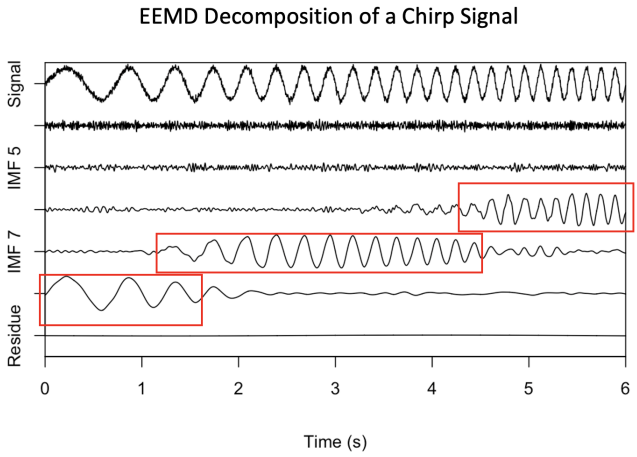


Figure 4.1: A chirp with white noise decomposed by EEMD. The boxed-in areas identify when each basis function is picking up the sinusoidal signal. Notice how the increasing frequency of the sinusoid makes it such that no basis function picks up the signal for the entire duration.

4.2 Local Change Point Detection and Signal Cleaning

4.2.1 EEMD

The Ensemble Empirical Mode Decomposition (EEMD) was invented by Wu and Huang (2009) as a novel technique for analyzing nonlinear and non-stationary time signals. The EEMD procedure uses iteratively computed, adaptive filters to decompose a signal $X(t)$ into basis functions:

$$X(t) = \sum_{j=1}^{n} IMF_j(t) + r(t).$$

Here, $IMF_j(t)$ is the j-th basis function, which is referred to as an Intermediate Mode Function (IMF), and $r(t)$ is the residual. As the EEMD is a numerical algorithm, it requires a stopping criteria to select the correct number of basis functions. While many stopping criteria exist for the EEMD family of numerical algorithms, one of the most common, called S-stoppage, results in the remainder term $r(t)$ becoming a monotonic or a constant function (Huang et al., 2003). In such cases, the resulting $r(t)$ can easily be subtracted from the original signal $X(t)$ to create a decomposition with no residual term. Thus, for the purposes of this paper, we will employ S-stoppage and assume that either $X(t)$ has an $r(t)$ of 0, or $X(t)$ has had its remainder

subtracted out resulting in:

$$X(t) = \sum_{j=1}^{n} IMF_j(t). \tag{4.1}$$

While this additive decomposition resembles the traditional Fourier decomposition, they differ in the properties of the basis functions that are generated. In a Fourier decomposition, the basis functions are orthogonal sinusoidals at fixed amplitudes and frequencies. In the EEMD, each IMF has a time-varying amplitude and time-varying frequency while still being orthogonal to one another (Huang and Shen, 2014). To evaluate how each IMF's amplitude and frequency are changing over time, it is common practice to compute each IMFs instantaneous amplitude and instantaneous frequency using a Hilbert Transform (Huang and Shen, 2014). Given the j-th IMF, IMF_j, the Hilbert Transform of IMF_j is:

$$H(IMF_j)(t) = \int_{-\infty}^{\infty} \frac{IMF_j(t)}{t-\tau} d\tau$$

From this, we can generate the analytic signal of IMF_j using:

$$\tilde{IMF}_j(t) = IMF_j(t) + jH(IMF_j)(t)$$

where j is the complex conjugate. Expressing this via polar coordinates using Euler's Identity, this results in:

$$\tilde{IMF}_j(t) = \tilde{a}_j(t) \exp(j\phi(t))$$

where $\phi(t)$ is the phase. Finally, from this analytic signal, we can extract the instantaneous amplitude for IMF_j, $a_j(t)$ and instantaneous $w_j(t)$ frequency using the identity (Huang and Shen, 2014):

$$a_j(t) = |\tilde{a}_j(t)|$$
$$w_j(t) = \frac{d\phi_j}{dt}(t)$$

Thus, given equation 4.1, we can represent each basis function of $X(t)$ as:

$$\begin{aligned} X(t) &= \sum_{j=1}^{n} IMF_j(t) \\ &= \sum_{j=1}^{n} a_j(t) e^{i\phi_j(t)}. \end{aligned}$$

IMFs also come with several crucial properties. By definition, an IMF is a nonlinear oscillatory function that satisfies the requirements (Huang et al., 1998):

1. For each IMF, the number of local extrema and zero crossings must differ by at most one.

2. Let $g_{j,max}(t)$ and $g_{j,min}(t)$ be smooth functions connecting the local maxima and minima of the j-th IMF. These functions are commonly referred to as the upper and lower envelope of $X(t)$. At time point t, the mean of the upper envelope of the j-th IMF, $g_{j,max}(t)$, and the lower envelope, $g_{j,min}(t)$, is zero:

$$g_{j,max}(t) + g_{j,min}(t) = 0.$$

These properties will serve an important role in designing our change point detection and signal cleaning algorithm.

4.2.2 Additive Local Noise Model

In performing local change point detection, we will assume that the observed signal $X(t)$ for $t \in [1, T]$ consists of the true underlying signal we would like to extract, $S(t)$, which only occurs during the set of times $A \subset [1, T]$ and independent, background white noise $R(t)$ which occurs throughout the entire duration of the signal. To ensure measurability and identifiablility, A is a finite union of finite intervals and $S(t)$ is a deterministic function. Moreover, as the EEMD is well posed for non-linear and non-stationary signals, $S(t)$ will be smoothly changing in amplitude and frequency. This will contrast with $R(t)$ which will be independent white noise. Assuming an additive relationship between signal and noise, this gives us the data generating process:

$$X(t) = S(t)\mathbb{1}_A(t) + R(t),$$

4.2.3 Change Point Detection of the IMFs

Under the additive local noise model, the goal of signal cleaning is to recover the true signal $S(t)$ by first estimating the interval A, or when the true signal is occurring, and then performing a signal cleaning on $X(t)$ for $t \in A$ to recover $S(t)$. To identify when changes are occurring in $X(t)$, we first decompose $X(t)$ into its constituent IMFs and then perform a change point detection procedure on each IMF. From a statistical perspective, identifying change points entails finding the set of time points $S_j = \{\tau_1^{(i)}, ..., \tau_{n_j}^{(i)}\}$ such that:

$$\begin{aligned} f(IMF_j(t_1)) \neq f(IMF_j(t_2)), \\ \forall t_1 \in [\tau_k^{(i)}, \tau_{k+1}^{(i)}], \\ \forall t_2 \in (\tau_{k+1}^{(i)}, \tau_{k+2}^{(i)}], \\ \forall k \in [1, ..., n_j - 2]. \end{aligned}$$

Here, $f(IMF_j(t))$ represents the distribution of IMF_j at time t. However, as the distribution of each IMF is generally unknowable *a priori* outside of well-known distributions such as white noise (Wang et al., 2013), this quickly becomes a very difficult problem. To make this more tractable, we utilize several of the properties of IMFs and the additive local noise model to construct a more feasible change point detection problem. According to our additive local noise model

$$X(t)^2 = \begin{cases} (S(t) + R(t))^2 & \text{If } t \in A \\ R(t)^2 & \text{If } t \notin A. \end{cases}$$

Combining this with the statistical independence between $R(t)$ and $S(t)$ assumed in the additive local noise model, this implies that

$$E[X(t)^2] = \begin{cases} E[S(t)^2] + E[R(t)^2] & \text{If } t \in A \\ E[R(t)^2] & \text{If } t \notin A. \end{cases} \tag{4.2}$$

Thus, when we are in interval A, there is an increase in expected power in $X(t)$ (power being $X(t)^2$). In addition, the orthogonality of each IMF directly implies that an increase in power in

$X(t)$ must lead to a corresponding increase in at least one of the constituent IMFs. Therefore, to identify when we are experiencing a signal, it suffices to search for IMFs that are showing increased power.

Additionally, as each IMF has a mean of zero with respect to its envelope, an increase in power in an IMF implies an increase in the variance in that IMF

$$E[IMF(t)^2] = E[(IMF(t) - E(IMF(t))^2] = Var(IMF(t)).$$

Therefore, we have reduced the complex problem of detecting an arbitrary change in a signal down to detecting changes in an IMF's variance.

To identify when an IMF is experiencing an increase in variance, we employ techniques from a well-developed branch of statistics, change point detection. A commonly used approach in the statistical estimation of change points is to minimize an objective function of the form

$$\min_{m} \min_{\tau_1,\ldots\tau_{m-2}} \sum_{i=1}^{m-1} L(X_{\tau_{i-1}:\tau_i-1}, X_{\tau_i:\tau_{i+1}-1}, X_{\tau_{i+1}-1} : \tau_{i+2}) + \beta D(m),$$

where τ_0 is 1 and τ_m is the length of the signal. In addition, m is the number of change points, τ_i is the location of the i-th change point, β is a constant, L is a function that decreases when τ is a true change point, and $D(m)$ is a penalization function that increases with the number of change points selected. By balancing L and $D(m)$, the objective seeks to select the correct number and locations of changes in variance.

As for choice of L, we can once again employ the properties of the EEMD. Wu and Huang (2014) identified that the power of $IMF_j(t)$ (power in this case referring to $a_j(t)^2$) is approximately normally distributed. As $a_j(t)^2$ controls the variance of the IMF, a natural choice for L is to base it on the likelihood ratio test for the change in variance in normal distributions (Inclan and Tiao, 1994)

$$L(X_{\tau_{i-1}:\tau_i-1}, X_{\tau_i:\tau_{i+1}-1}, X_{\tau_{i+1}-1} : \tau_{i+2}) = \frac{C_{\tau_i}}{C_{\tau_{i+1}-1}} - \frac{\tau_i - \tau_{i-1}}{\tau_{i+1} - 1 - \tau_{i-1}}$$

Here C_{τ_i} is the cumulative normalized second moment, $\sum_{k=\tau_{i-1}+1}^{\tau_i}(X(k) - \overline{X_{\tau_i}})^2$ and $\overline{X_{\tau_i}}$ is the cumulative mean, $1/(\tau_i - \tau_{i-1} + 1)\sum_{k=\tau_{i-1}+1}^{\tau_i} X(k)$. As for $\beta D(m)$, this is a penalization term

that combines some function of the number of change points, $D(m)$, with a constant, β. Some of the most popular penalty terms include Akaike's Information Criterion (βm) (Akaike, 1974) and Bayesian Information Criterion ($m\log(n)$) (Schwarz, 1978) (n is the total signal length). For our uses, the newer Modified Bayesian Information Criterion
$(-1/2(3m + \log(n) + \sum_{i=1}^{m+1} \log(\tau_i - \tau_{i-1}))$ (Zhang and Siegmund, 2007) is preferred for its combination of solid theoretical justification and real world performance.

4.2.4 Hypothesis Test and Sparse Basis Selection

Once the change points algorithm has identified the points where the each IMF has undergone a change, we must determine how we are to clean each signal segment. We propose a simple hypothesis-test-based algorithm that is able to automatically create sparsely cleaned IMFs based on changes in power. As equation 4.2 demonstrated that during the presence of the underlying signal, there should be an increase in variance relative to before and after, we are interested in the hypothesis test:

$$H_0 : \sigma^2_{during} \leq \gamma * \max(\sigma^2_{before}, \sigma^2_{after})$$
$$H_1 : \sigma^2_{during} > \gamma * \max(\sigma^2_{before}, \sigma^2_{after}),$$

where σ^2_{before} is the variance of the previous interval, σ^2_{during} is the variance of the current interval, σ^2_{after} is variance of following interval, and γ is assumed to be greater than or equal to 1. By rearranging the alternate hypothesis, $\gamma > \frac{\sigma_{during}}{\max(\sigma_{before}, \sigma_{after})}$, we can see that γ serves as a measure of how much the ratio of variances much increase to be considered significant. Setting $\gamma = 1$ tests if there has been any statistically significant increase in variance.

As for the test statistic, leaning on the fact that when there is Gaussian noise, the variance follows a normal distribution (Huang and Shen, 2014), we employ the F-statistic to detect changes in variance

$$F_{before/during} = \frac{\gamma \max(S^2_{before}, S^2_{after})}{S^2_{during}},$$

where S_{before} is the sample variance used to estimate σ_{before}. $F_{before/during}$ is compared against the F distribution with degrees of freedoms, $df_1 = n_{during}$, $df_2 = max(n_{before}, n_{after})$ (where n_{during} is the length of the during interval) to determine the p-value and thus significance.

As this process involves performing a hypothesis test at every potential change point, across every IMF, this can quickly lead to a large number of tests being performed for the same goal: identifying a significant segment. This large number of tests can lead to the multiplicity issue where one or more spurious false positives may occur. To perform these tests so they collectively have an α $(1 > \alpha > 0)$ probability of a false positive (which is known as the the Family-Wise Error Rate), we employ the multiple testing correction method, Holm-Bonferroni method (Holm, 1979). If the p-value is significant after the Holm-Bonferroni correction, then we can claim that the interval contains the desired signal. If not, the interval does not contain the true signal and is cleaned by setting it to 0. If an IMF only has one change point (and thus cannot have a before, during, and after interval), then then $\max(\sigma^2_{before}, \sigma^2_{after})$ is replaced with σ^2_{after}. If there are no change points, then the entire IMF is set to 0. As this performs a Local Change point Detection and Signal Cleaning, we refer to this as `LCDSC`.

4.3 Simulation

4.3.1 Simulation 1: Doppler Signal

To demonstrate this signal cleaning procedure, we take a synthetic example where a Doppler signal is hidden in the midst of Gaussian white noise. We will refer to this as the Local Doppler example. For Simulation 1, we will use a Local Doppler of length 2500 with the Doppler occurring during the middle of the signal:

$$X(t) = \begin{cases} R(t) & \text{if } t < 1000, t > 1500 \\ S(t) + R(t) & \text{if } 1000 \leq t \leq 1500. \end{cases}$$

$S(t)$ is the Doppler signal from Donoho and Johnstone (1994).

$$S(t) = 7(t(1-t))^{0.5} \sin(2\pi(1 + 0.05)/(t + 0.05)).$$

As can be seen in Figure 4.2, the Doppler signal in the middle is expressed in all 7 IMFs with first IMFs expressing the higher frequency parts of the signal and the latter IMFs expressing the lower frequency sections. Moreover, no single IMF is ever purely signal or purely noise necessitating a local change point detection and signal cleaning. Running the change point detection algorithm in

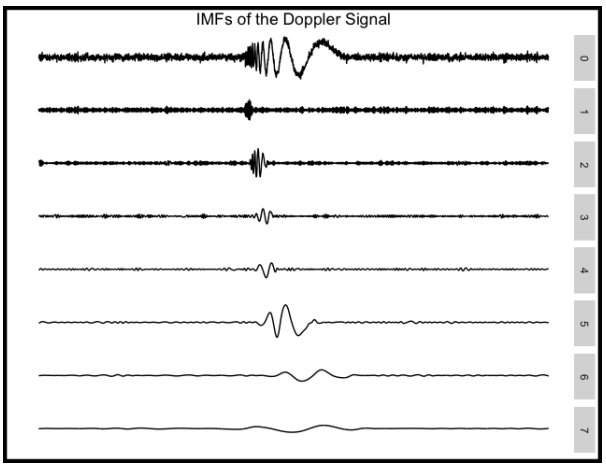


Figure 4.2: EEMD of the Local Doppler Signal. The IMF numbers are on the right with "IMF 0" referring to the original signal. In the EEMD, none of the IMFs are purely signal or noise necessitating a local signal cleaning procedure.

Figure 4.3 at an $\alpha = 0.05$ type I error level and $\gamma = 1$ identifies many locations at which a change in the signal was detected. While IMFs 1-3 have perfect identification of the Doppler signal, in IMFs 4-7, many spurious change points are detected that are not necessarily due to the Doppler signal. To remove these, the F-test cleaning step is performed. The resulting cleaned signal in Figures 4.4 and 4.5 illustrates how all of the change points outside of the duration of the Doppler signal were deemed nonsignificant by Holm-Bonferroni and set to zero. Not only does this provide a good estimation of the shape of the Doppler signal, matching the general sinusoidal shape and increasing frequency, but `LCDSC` provides a good estimate of when the Doppler signal starts, as the first nonzero point in IMF1 is at point 1010, only 1% of the way into the start of the Doppler signal.

4.3.2 Simulation 2: Doppler Signal-Comparison Study

To compare the performance of our algorithm, we extend our Doppler simulation from Simulation 1 and compare our performance against other EEMD signal cleaning techniques.

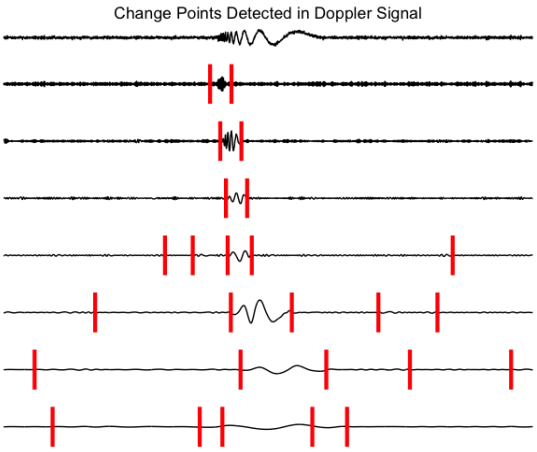


Figure 4.3: Change points that were detected in the Local Doppler Signal in Figure 4.2 when employing the normal likelihood ratio objective function and the Modified Bayesian Information Criterion over-fitting penalty.

These techniques come in two general varieties. Techniques 2-5 in Table 4.1 are based on identifying some subset of the IMFs as containing only noise and cleaning the signal by completely removing the noise IMFs. Many times, the correct number of IMFs to remove is determined subjectively by the experimenter though a trial-and-error process. To account for any possible variability in performance due to these judgements, we will come up with an upper-bound for the performance of each algorithm by computing the best possible set of IMFs for each of the algorithms in question.

As for the Wavelet Hard Thresholding (WHT) and Wavelet Interval Thresholding (WIT) cleaning techniques, these are based on performing a Wavelet-like thresholding on each of the IMFs (Kopsinis and Stephen, 2009). These compute the base noise level within each IMF and perform a hard or soft thresholding if the IMF lies within the expected noise band. While this method does not suffer from a subjective choice of IMF removal, it assumes that the true signal occurs throughout the entire duration of the signal, leading to a biased estimation of the base noise level.

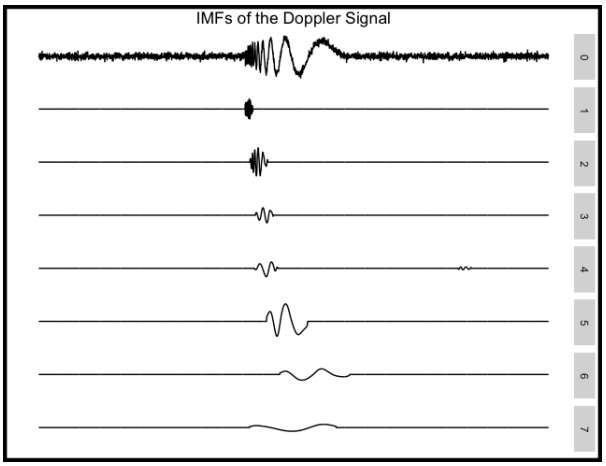


Figure 4.4: The IMFs in Figure 4.3 after each section that was identified by the change point detection algorithm was cleaned using the F-test/Hole-Bonferroni procedure with $\gamma = 1$. Notice how the basis functions are set to 1 when the signal is not present within the basis function.

The data model for the simulation will utilize the Local Doppler Model with the middle containing our desired signal but with the total signal length T at differing values:

$$X(t) = \begin{cases} R(t) & \text{if } t < \frac{2}{5}T, t > \frac{3}{5}T \\ S(t) + R(t) & \text{if } \frac{2}{5}T \leq t \leq \frac{3}{5}T. \end{cases}$$

T is tested at 1000, 2000, and 2500 time steps. $R(t)$ will again be Gaussian white noise but with the noise level varying from 0.2 to 0.5. The cleaned signal is then compared to the underlying Doppler signal and error computed in terms of Residual Sum of Squares (RSS) as this corresponds to the total power difference between the estimated and the cleaned signal,

$$RSS = \sum_{t=1}^{T} (X(t) - Cleaned(t))^2.$$

At each level of noise and signal length, 20 replicates of the simulation were performed.

The results in Figure 4.6 illustrate that across a wide scale of noise levels and sample sizes, the `LCDSC` performs well at local signal cleaning, uniformly outperforming other non-local signal cleaning techniques.

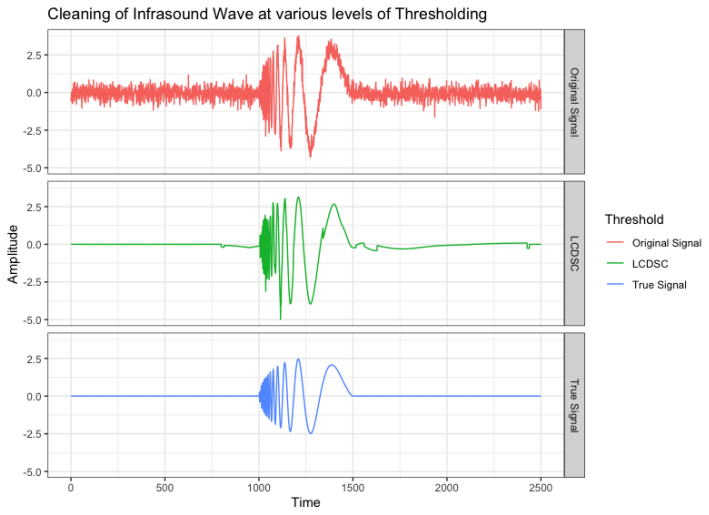


Figure 4.5: Comparison of the original signal with the cleaned signal. The recovers much of the original signal. It performs especially well at cleaning the signal to closely match the true start and end points.

Cleaning Method	Description
k-Highest	Removal all but the k-highest IMFs (Wu et al., 2019)
l-Lowest	Removal all but the k-lowest IMFs (Hotradat et al., 2019)
k-Highest & l-Lowest	Combination of k-Highest and l-Lowest (Huimin et al., 2017)
Power Set Cleaning	Perform a best subset selection over all possible subsets.
WHT	Wavelet Hard Thresholding each IMF (Kopsinis and Stephen, 2009)
WIT	Wavelet Interval Thresholding each IMF (Kopsinis and Stephen, 2009)
No Cleaning	No Signal cleaning

Table 4.1: List of Signal Cleaning Techniques

4.3.3 Simulation 3: Comparison Study- What if the signal is not local?

While the `LCDSC` is built for the problem of local signal detection and cleaning, it is important to determine its performance as the duration of true signal is increased or decreased. We can express how local our signal is in terms of a "locality Ratio":

$$\text{locality Ratio} = \frac{len(A)}{T} - 1.$$

$len(A)$ is the length of the interval A when the true signal is being expressed and T is the total length of the noisy signal. We vary the locality Ratio between 0 to 4, making the local signal cleaning problem increasingly local and favorable to `LCDSC`.

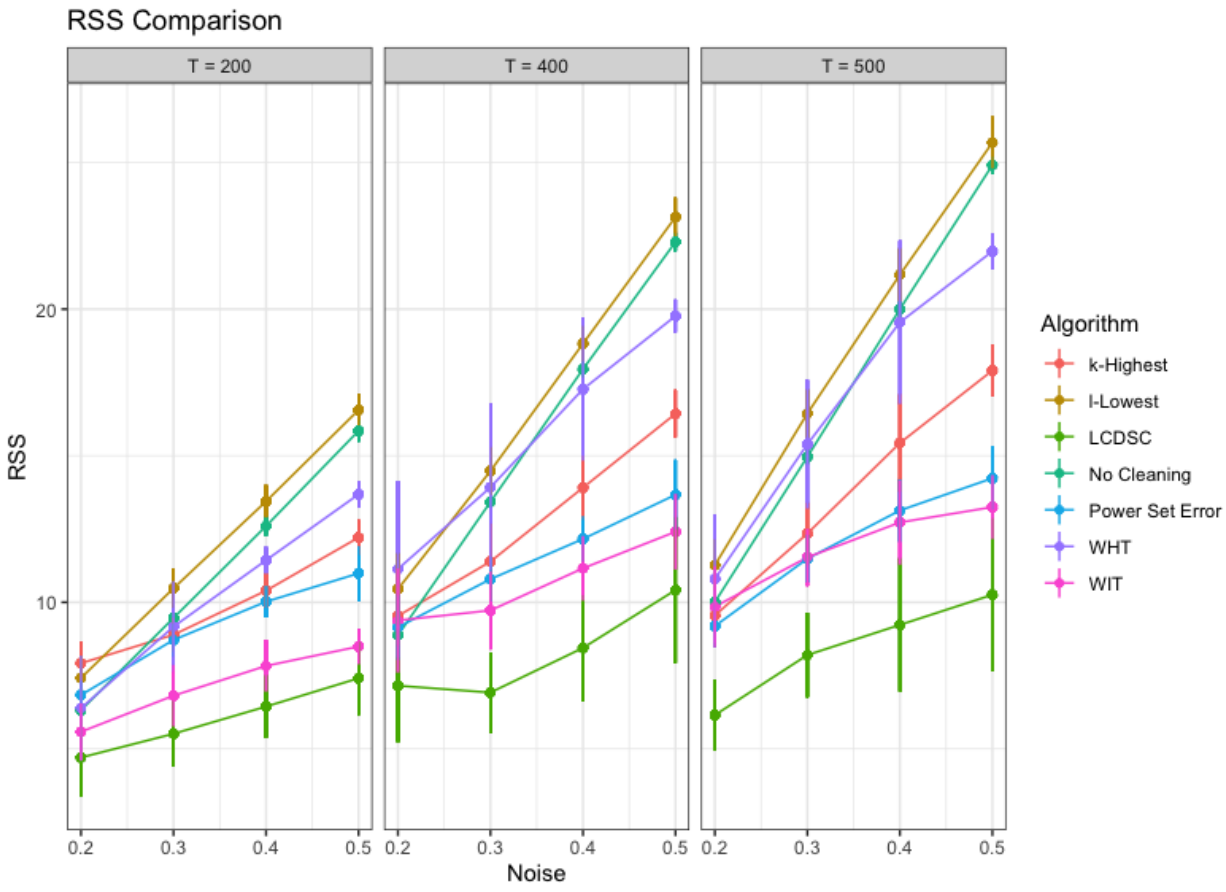


Figure 4.6: RSS Comparison of common cleaning methods vs `LCDSC`. The center point represents the mean RSS across 20 replicates and the bar represents one standard deviation from the center. From this, we observe that `LCDSC` performs better than competing signal cleaning methods, able to create the closest representation of the true signal.

Figure 4.7 illustrates that when the locality Ratio is at or below one, then `LCDSC` performs approximately the same as the best performing method such as k-Highest. However, once the locality ratio goes beyond one, `LCDSC` becomes the dominant signal cleaning technique followed by WIT. This gives us a rough guide for when to start considering a signal cleaning problem local or global. When the Noise Ratio is below one, it can be better to clean with global cleaning methods whereas local cleaning methods are better when the ratio is greater than one.

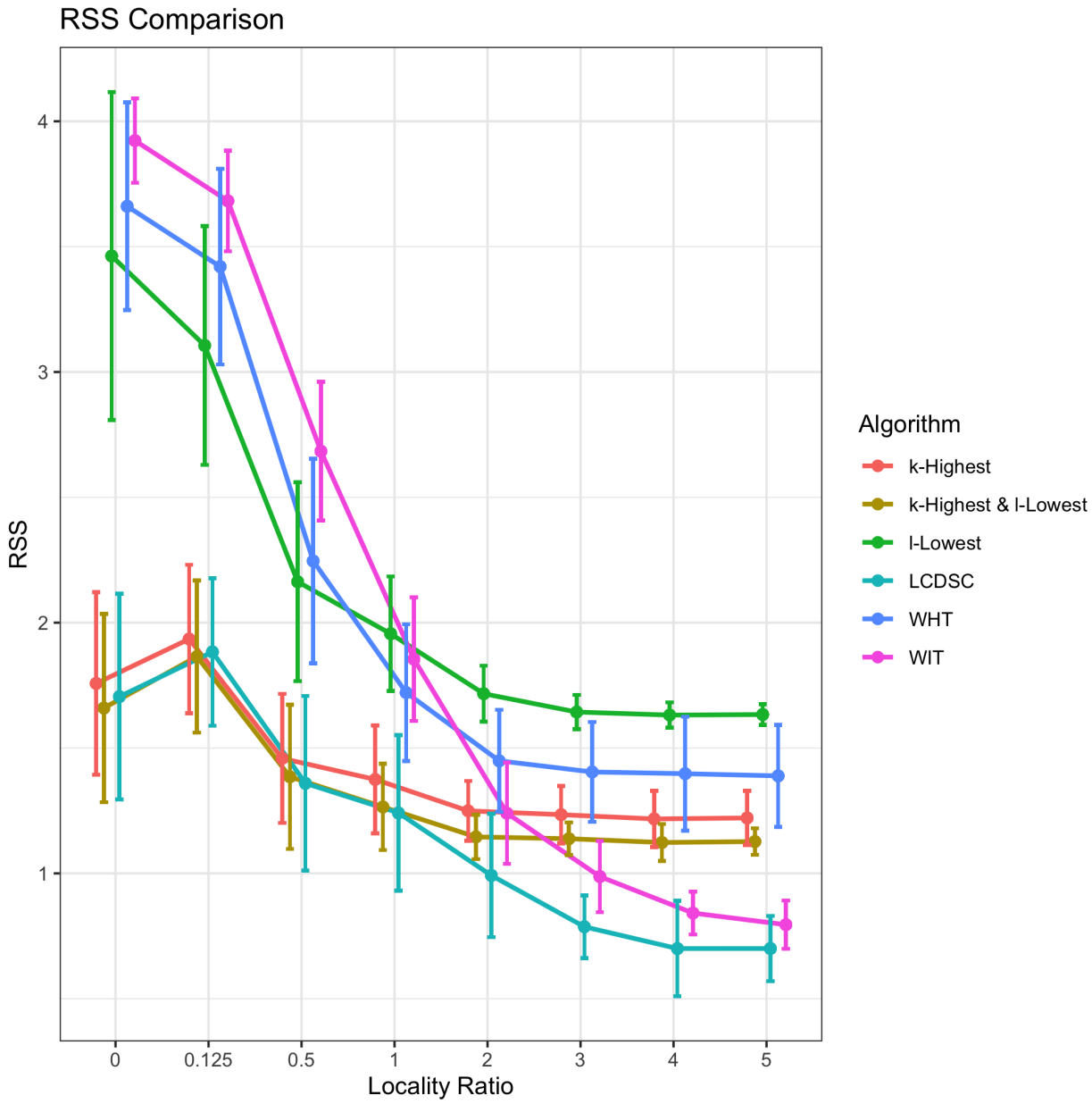


Figure 4.7: Changes in Residual Sum of Squares as the Locality Ratio is increased. When the noise ratio is low, the `LCDSC` performs slightly worse than k-Highest and k-Highest & L-lowest, but once the noise ratio increases above 1, the `LCDSC` becomes the best performing method. No cleaning was not plotted as it had a much higher error than all the others and Power Set was near equivalent to k-Highest & l-Lowest.

4.4 Application: Detection of Gliding events in Acoustic Explosions

On October 28, 2014 an Anteras rocket operated by Orbital Sciences Corporation exploded shortly after takeoff (Northon, 2015). The resulting explosion was powerful enough that acoustic shockwave arrivals were observed at stations over 2000km away from the launch site. At the time, 226 acoustic and atmospheric stations from the Transportable USArray network were located within range of the explosion, resulting in arrivals from the explosion being picked up by the array's infrasound sensors. Many of these arrivals exhibited characteristics of dispersive waves at the infrasound level (<20 Hz). This is of interest as dispersive waves were only recognized recently in the infrasound domain (Negraru and Herrin, 2009) and because the Anteras explosion was one of the largest demonstrations to date of the existence of infrasound dispersive waves (Vergoz, 2018). These dispersive waves are a result of the arrivals being reflected at different heights in the troposphere as well as being influenced by atmospheric conditions such as temperature and windspeed. This makes studying infrasound arrivals important tools in evaluating atmospheric density models (Vergoz, 2018).

Isolating these dispersive waves can be complicated due to the relatively short time periods when the explosion was detected as well as the complex weather and atmospheric factors affecting recording conditions at each sensor.

However, this problem is well suited for `LCDSC`. First, each infrasound is relatively quick (on the order of a few seconds within the 24-hour monitoring of the USArray sensors). Second, as seen in Figure 4.8, one of the canonical features of an infrasound dispersive wave is the presence of a "gliding" or steadily increasing frequency in the signal. This makes infrasound dispersive waves display gliding similar to a Doppler signal reversed, which the `LCDSC` has performed well at cleaning. Performing `LCDSC` on the signal from one of the acoustic stations, we do indeed observe in Figure 9 that `LCDSC` cleans the signal well especially compared to WIT which has made very little change to the signal due to the large period of noise throwing off the estimation WIT's baseline noise estimation.

Moreover, in Figure 10, by increasing the Threshold value, γ, we can clean the signal further and further until only the acoustic explosions are singled out. This occurs when gamma is around

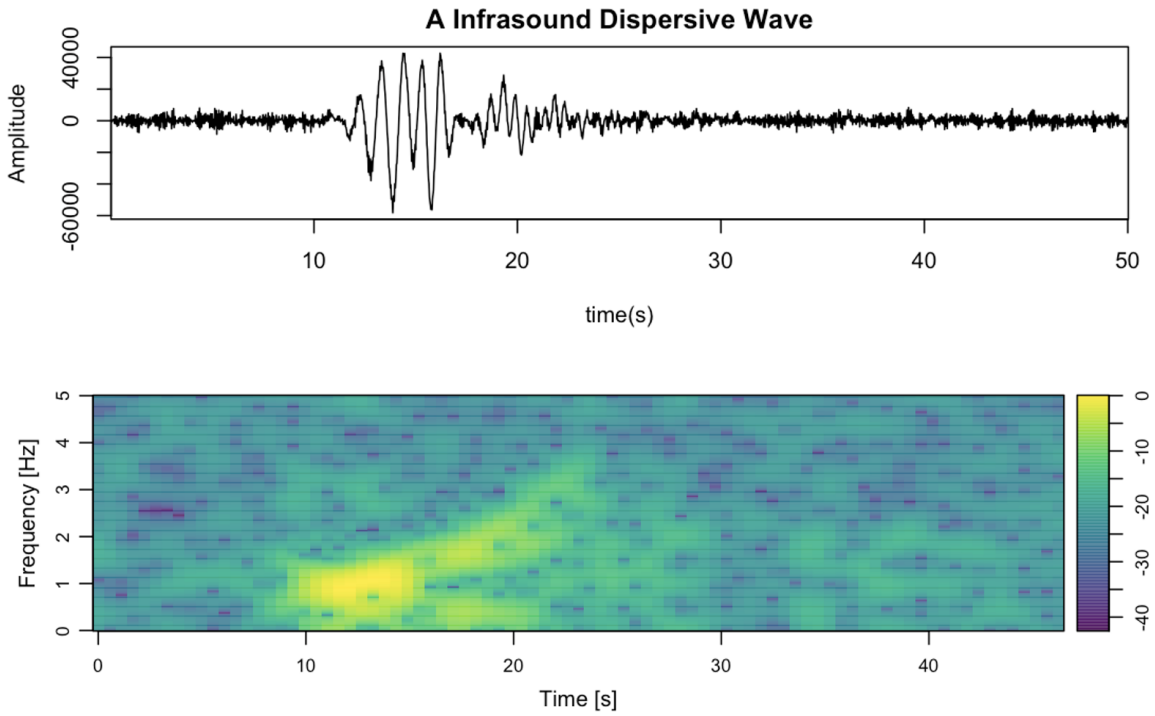


Figure 4.8: A canonical example of an Infrasound Dispersive wave. Note the increase in frequency over the duration of the signal in the Spectrogram plot.

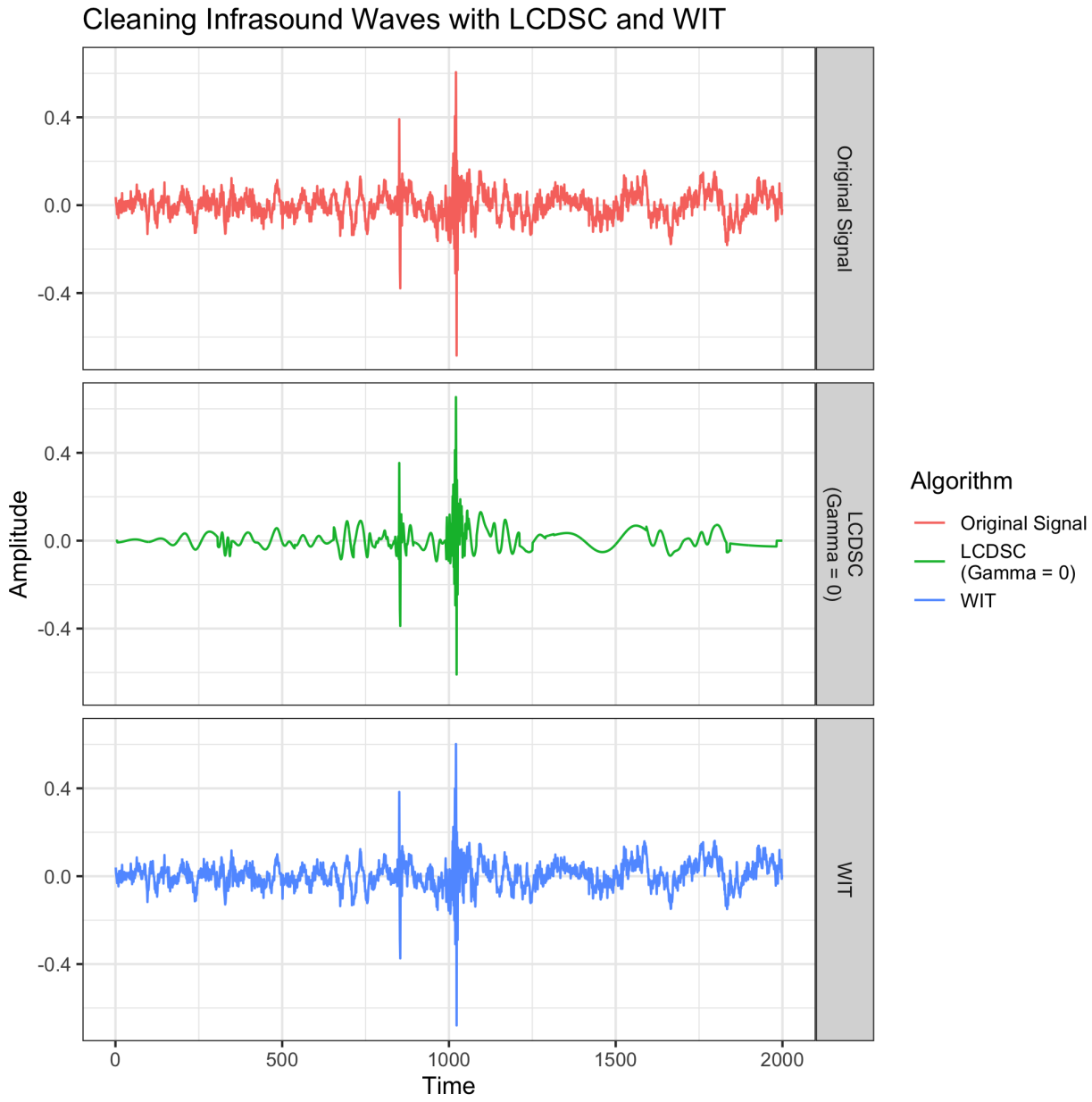


Figure 4.9: Comparison of uncleaned gliding signal with `LCDSC` cleaned signal and WIT cleaned signal. Note that `LCDSC` performs a better job at cleaning the signal than WIT, especially in helping to isolate the two spikes between 800-1000 that represent the infrasound dispersive wave.

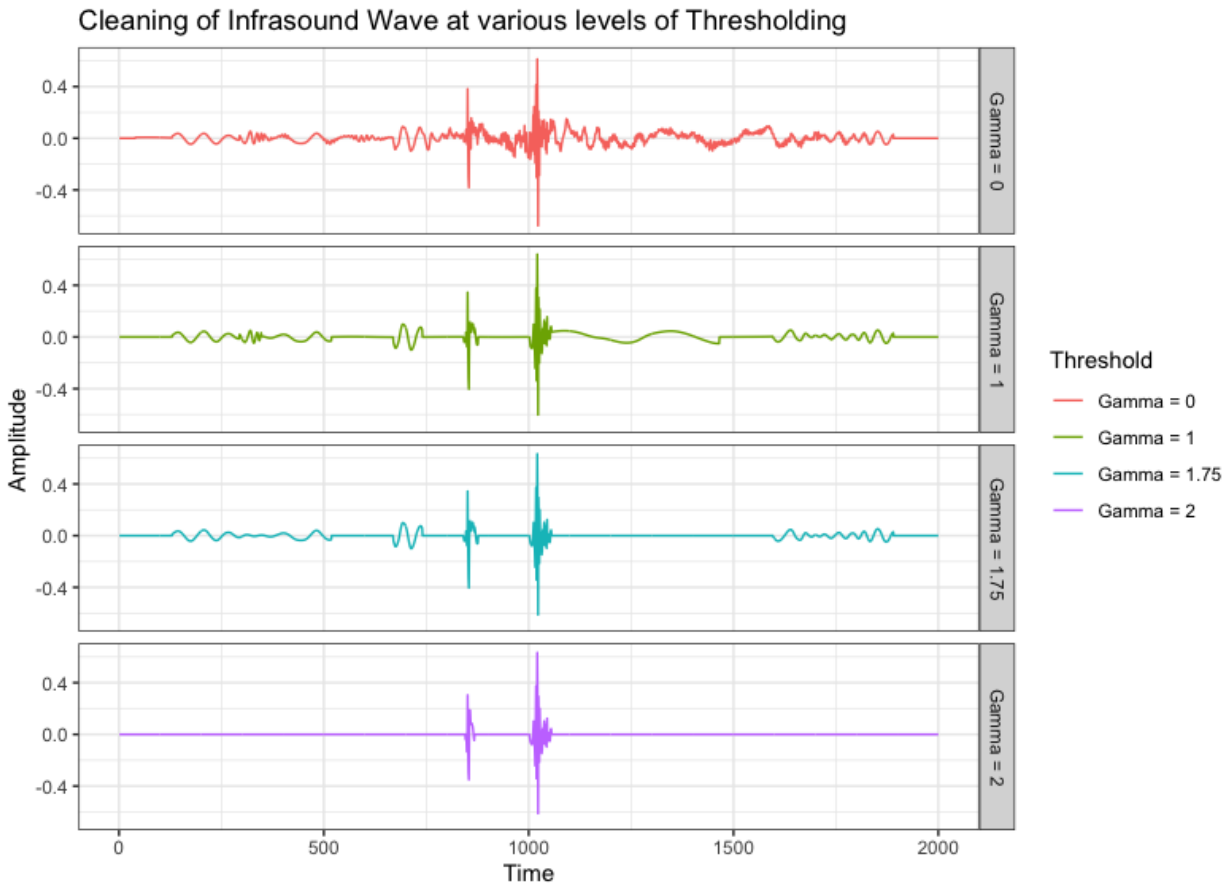


Figure 4.10: Comparison of `LCDSC` signal cleaning as γ is increased. As γ is increased, this results in a sparser and sparser signal cleaning, with $\gamma = 2$ most cleanly isolating the dispersive wave. This indicates that the dispersive wave can be identified in each IMF as a 2 fold increase in SNR compared to background noise.

2. This informs us that dispersive waves seem to lead to at least a 2 times increase in power in all of the IMFs.

4.5 Conclusion

Here we provided a demonstration of the utility of `LCDSC` for the problem of local change point detection and signal cleaning. While other EEMD signal cleaning algorithms can exhibit drawbacks when there are long periods of no signal, our `LCDSC` does not suffer from the same disadvantage. This makes it ideal for the cleaning of short-term signals such as acoustic shock

waves. We believe that the future development of EEMD signal decomposition will benefit greatly from the further development of methods based on local changes in basis functions.

Chapter 420pt

APPENDIX A: PATIENT CHARACTERISTICS

Table 4.2: Preadmission Patient Covariates

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%)

APPENDIX B: 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 4.3) for calculating the drug concentrations $D_{i,t,j}$ in the blood using the PK model.

Table 4.3: 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

APPENDIX C: 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 4.4.

Table 4.4: APPENDIX C: Binning of EA burden

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

APPENDIX D: SUMMARY OF NOTATION

Table 4.5: 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 E: METHODOLOGY

Let us describe how we applied the framework from Section 3.2 to analyze the EA data to obtain the results in Section 3.4. We divide the estimation pipeline into three stages (Figure 4.11):

1. In the *first stage*, 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 (Ge et al., 2021b; Zafar et al., 2021; Jing et al., 2016). We use the predictions to compute EA time series (Z_t^ω). As described in Section 3.4, $E_{\max}$ and E_{mean} are computed directly from Z_t^ω. Details are in the appendix in Section **??**.

2. In the *second stage* (Section 4.5), we fit a personalized pharmacokinetic/pharmacodynamic (PK/PD) model to each patient's response to ASM (Hill, 1909).

3. In the *third stage* (Section 4.5) 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, Parikh et al., 2018).

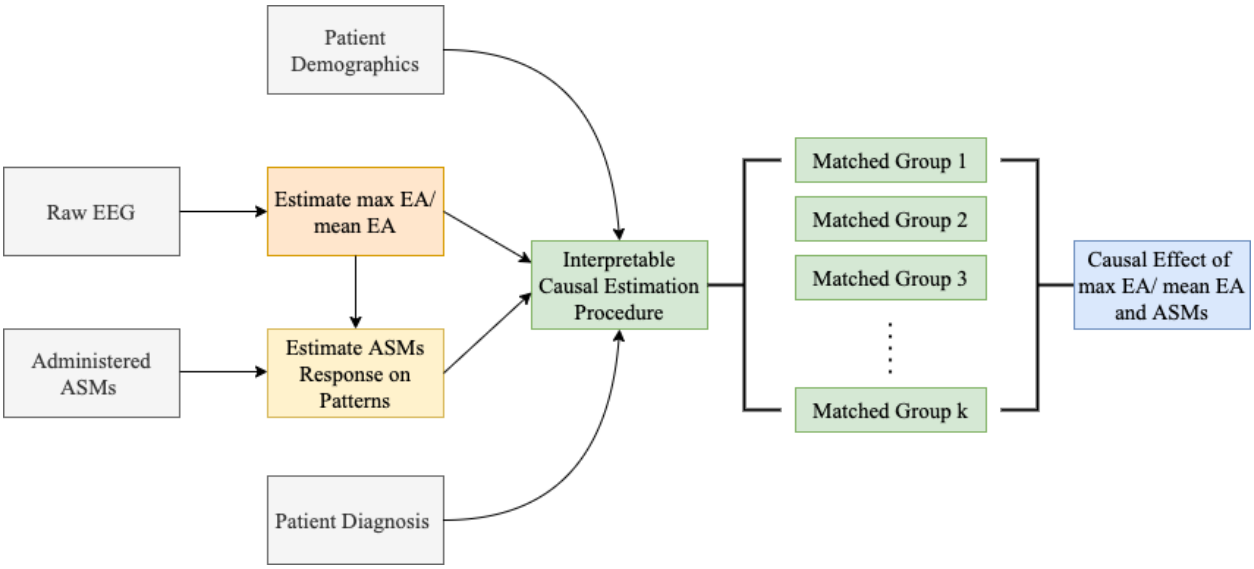


Figure 4.11: The overall analysis framework, consisting of three parts (indicated by different colors): EA burden computation, individual PK/PD modeling, and MALTS matching and effect estimation.

APPENDIX F: MECHANISTIC PHARMACOLOGICAL MODEL

As noted in section 3.3, 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 4.5.

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 (Hill, 1909) to estimate a short-term response to drugs:

$$\frac{dD_{i,t,j}}{dt} = -\frac{1}{\kappa_j} D_{i,t,j} + W_{i,t,j} \ , \tag{4.3}$$

$$Z_{i,t} = 1 - \sum_j \frac{D_{i,t,j}^{N_{i,j}}}{D_{i,t,j}^{N_{i,j}} + ED_{50,i,j}^{N_{i,j}}} \ . \tag{4.4}$$

Here κ_j is the average half-life of the drug (see Appendix 4.3 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 4.12).

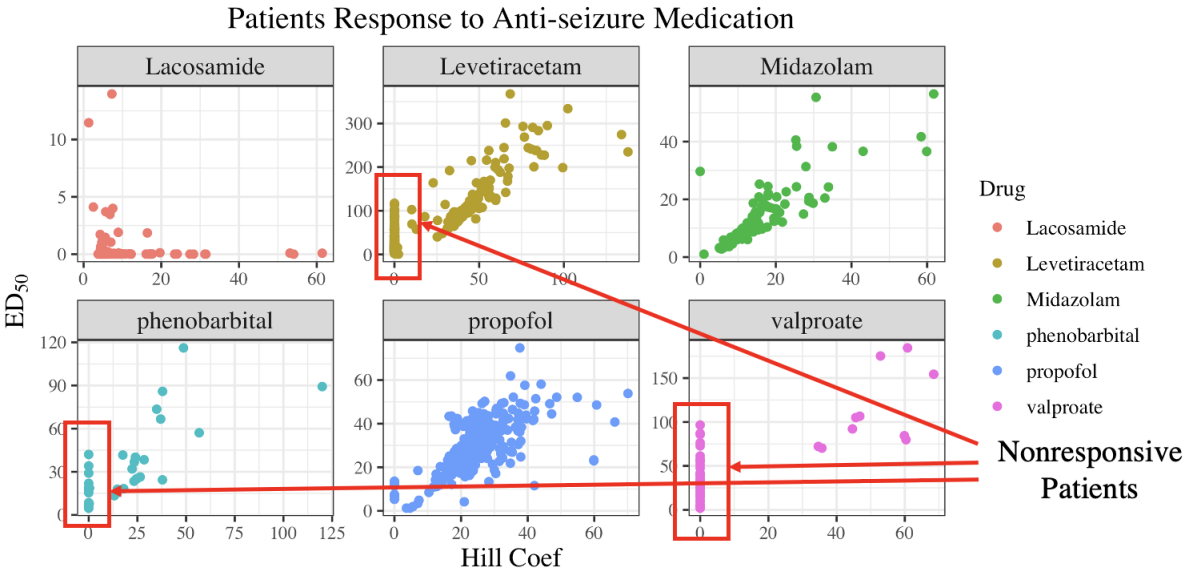


Figure 4.12: Hill coefficient vs. ED_{50} for the six drugs. Each point is a patient. The non-responsive patients with Hill coefficient of zero are highlighted.

APPENDIX G: INTERPRETABLE 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 invented by future Dr. Harsh Parikh 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 (Parikh et al., 2018, 2019).

The conventional objective function of MALTS, described in Parikh et al. (2018), 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 H: EXPERT LABELING OF EEG SIGNALS

The EEG signals of 1309 patients at Massachusetts General Hospital who met the inclusion criteria (described in Section 3.3) 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 (Benbadis, 2006) 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* (Jing et al., 2016) 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.

APPENDIX I: 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 (Huang et al., 2017); a Hidden Markov model was added to smooth the outputs (Ge et al., 2021a). By treating this as 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 (Hirsch et al., 2021). 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.

APPENDIX J: OPERATIONALIZING DENSENET

We used DenseNet with 7 blocks (Figure 4.13). 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 precedent of first order HMMs providing good smoothing for other EEG problems (Sun et al., 2017).

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 4.14) to reduce the chance of error since these patterns are intrinsically on a continuous spectrum.

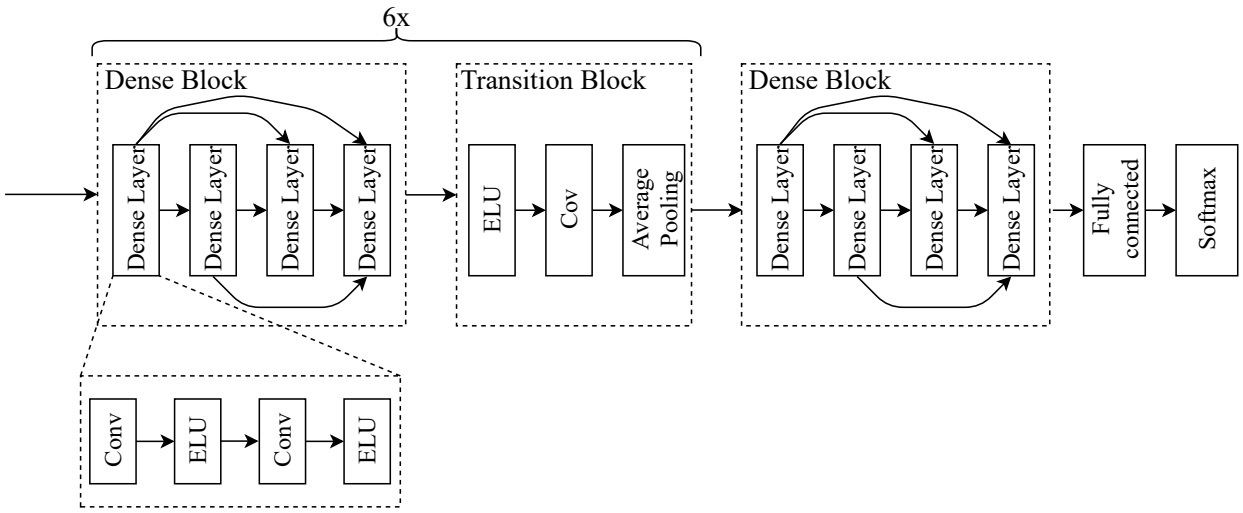


Figure 4.13: Structure of the DenseNet for automatic EA labeling.

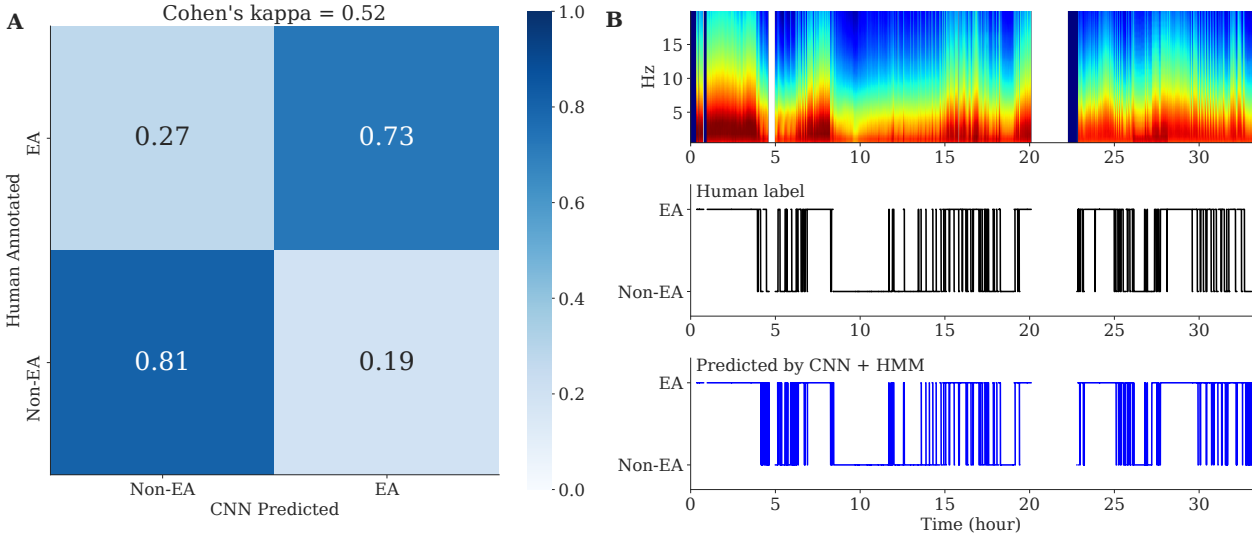


Figure 4.14: (A) Confusion matrix for the CNN prediction vs. human annotation, where each row represents the fraction of 2-second segments classified into EA (seizure/GPD/LPD/LRDA) or Non-EA (GRDA/other/artifact). The overall Cohen's kappa is 0.52. (B) The top panel shows the spectrogram of the EEG signal 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. [Credit to Dr. Haoqi Sun for the figure]

APPENDIX K: 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 4.15):

- 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 $\mathbb{E}[Y([\rho_2), 1.0]]$

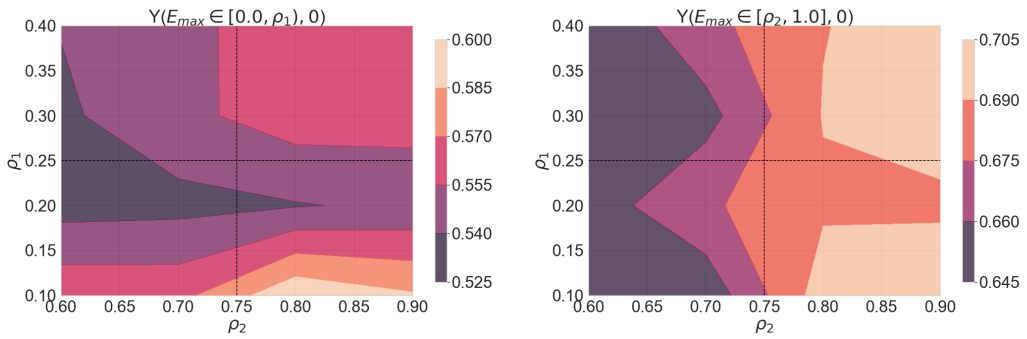


Figure 4.15: Sensitivity to quantization of EA burden into four levels. ρ_1 is the boundary between mild and moderate EA burden and ρ_2 is the boundary between severe and very severe EA burden. The contour plot shows estimated average potential outcomes – $Y([0, \rho_1), 0)$ and $Y([\rho_2, 1], 0)$ – 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. [Credit to future Dr. Harsh Parikh for the figure]

APPENDIX L: MISSINGNESS PATTERN

To check for possible selection bias, we compared the discharge mRS in patients with different missing conditions in Figure 4.16 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.

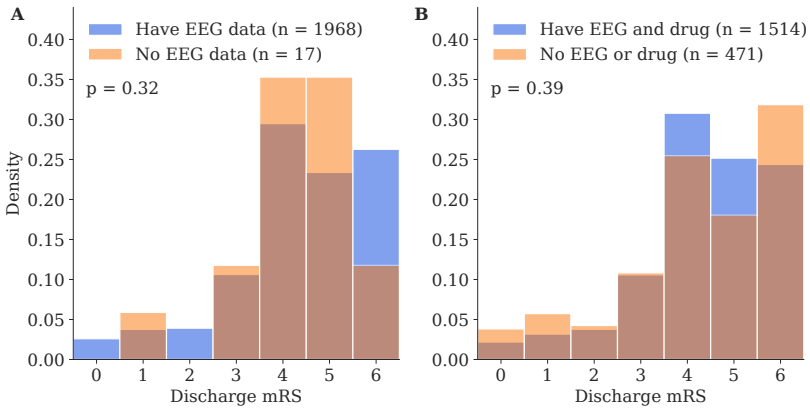


Figure 4.16: (A) The histogram of patients' discharge mRS (possible values are 0,1,2,3,4,5,6). The two subsets that are compared are patients who have EEG data (n = 1968) vs. patients who do not have EEG data (n = 17). To make the subsets comparison, the y-axis shows the density instead of the count. 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).

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.

APPENDIX M: 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.

Assumption 1): Control of pre-admission covariates and PD parameters

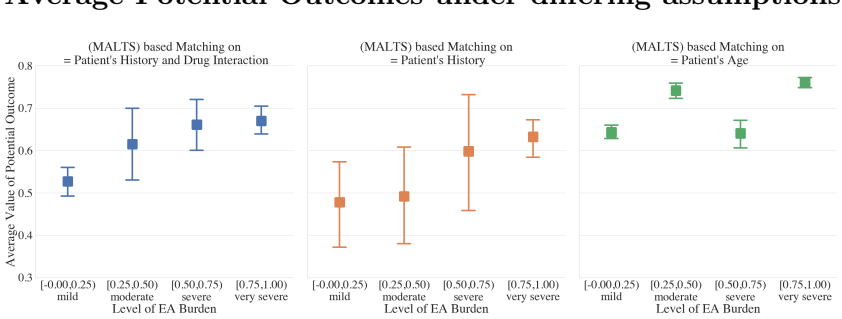


Figure 4.17: Estimated average potential outcome for different levels of $E_{\max}$ by matching on (left) all pre-admission covariates and PD parameters, (middle) all pre-admission covariates, and (right) only age of the patients.

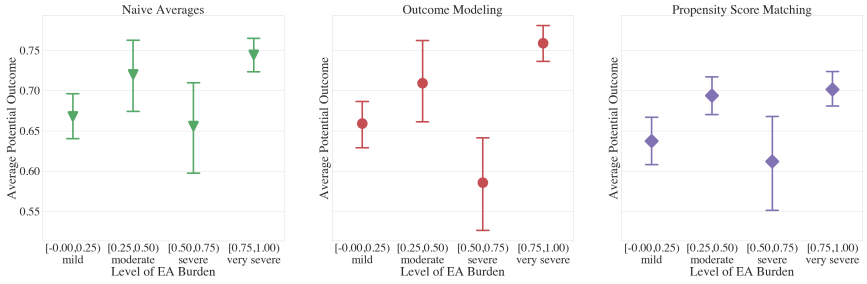


Figure 4.18: Estimated average potential outcomes computed using (left) Naive Average approach, (middle) Outcome modeling approach, and (right) Propensity Score matching.

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 4.17 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.

Assumption 2): Outcomes are a function of EA burden and 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 4.18. 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 4.18, 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 4.18, 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 4.17 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.

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 Blackwell (2014). We parameterize $q(\cdot)$ as a logarithmic function of e.

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

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

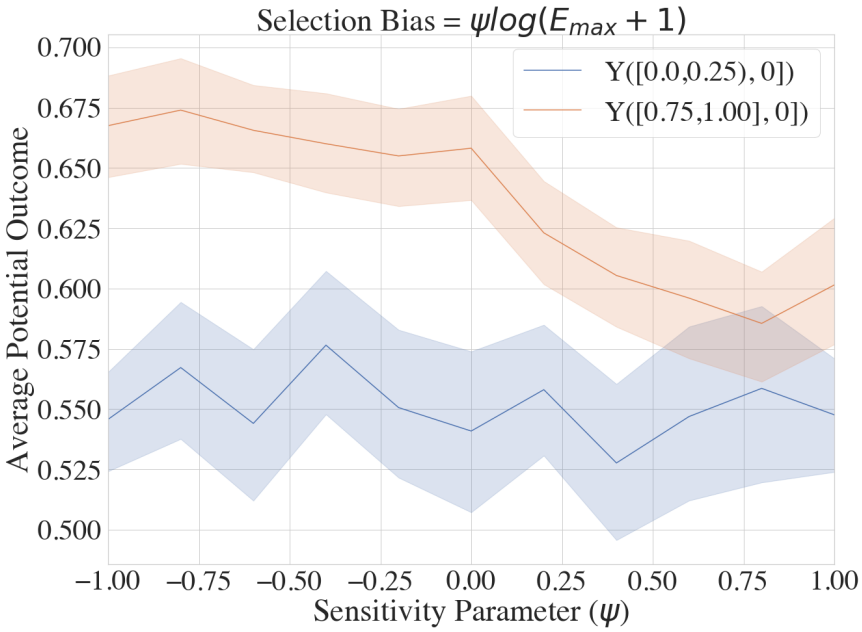


Figure 4.19: Sensitivity to unobserved confounding The results show that even at very high levels of selection bias, the effect of EA burden is not lost, indicating a degree of robustness in our results.

$E_{\max}$ 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.

BIBLIOGRAPHY

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Balakrishnan, S. and Wasserman, L. (2018). Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, 12(2):727 – 749.

Beals, R. (2004). *Analysis: An Introduction*. Cambridge University Press.

Benbadis, S. R. (2006). Introduction to sleep electroencephalography. *Sleep: A Comprehensive Handbook*, pages 989–1024.

Bernoulli, J. (1713). *Ars conjectandi*. Thurneysen Brothers.

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

Boggs, J. (2002). Seizures and organ failure. In *Seizures*, pages 71–83. Springer.

Burn, J. (1992). Reliability of the modified rankin scale. *Stroke*, 23(3):438–438.

Casella, G. and Berger, R. (2001). *Statistical Inference*. Duxbury Resource Center.

Chen, X. and Cui, B. (2016). Efficient modeling of fiber optic gyroscope drift using improved eemd and extreme learning machine. *Signal Processing*, 128.

Chen, X., Zhang, X., Zhou, J., and Zhou, K. (2019). Rolling bearings fault diagnosis based on tree heuristic feature selection and the dependent feature vector combined with rough sets. *Applied Sciences*, 9(6).

Chong DJ, H. L. (2005). Which eeg patterns warrant treatment in the critically ill? reviewing the evidence for treatment of periodic epileptiform discharges and related patterns. *J Clin Neurophysiol*, 22:79–91.

Cormier, J., Maciel, C. B., and Gilmore, E. J. (2017). Ictal-interictal continuum: when to worry about the continuous electroencephalography pattern. In *Seminars in Respiratory and Critical Care Medicine*, volume 38, pages 793–806. Thieme Medical Publishers.

de Campos, C. and Benavoli, A. (2011). Inference with multinomial data: why to weaken the prior strength. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2107–2112. AAAI Press. (acc.rate 30

De Marchis, G. M., Pugin, D., Meyers, E., Velasquez, A., Suwatcharangkoon, S., Park, S., Falo, M. C., Agarwal, S., Mayer, S., Schmidt, J. M., et al. (2016). Seizure burden in subarachnoid hemorrhage associated with functional and cognitive outcome. *Neurology*, 86(3):253–260.

Dempster, A. and Shafer, G. (1976). *A Mathematical Theory of Evidence*. Limited paperback editions. Princeton University Press.

Dempster, A. P. (1966). New Methods for Reasoning Towards Posterior Distributions Based on Sample Data. *The Annals of Mathematical Statistics*, 37(2):355 – 374.

Dempster, A. P. (1967). Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, 38(2):325 – 339.

Dempster, A. P. (1968). Upper and Lower Probabilities Generated by a Random Closed Interval. *The Annals of Mathematical Statistics*, 39(3):957 – 966.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, USA, 1st edition.

Fraser, D. A. S. (1968). *The Structure of Inference*. John Wiley & Sons.

Gaci, S. (2016). A new ensemble empirical mode decomposition (eemd) denoising method for seismic signals. *Energy Procedia*, 97:84–91.

Ganesan, S. L. and Hahn, C. D. (2019). Electrographic seizure burden and outcomes following pediatric status epilepticus. *Epilepsy & Behavior*, 101:106409.

Garoud, F., Lequeux, P., Bejjani, G., and Barvais, L. (2006). The influence of the dose on the time to peak effect of propofol. *European Journal of Anaesthesiology*.

Gaspard, N., Manganas, L., Rampal, N., Petroff, O. A., and Hirsch, L. J. (2013). Similarity of lateralized rhythmic delta activity to periodic lateralized epileptiform discharges in critically ill patients. *JAMA Neurol*, 70(10):1288–1295.

Ge, W., Jing, J., An, S., Herlopian, A., Ng, M., Struck, A. F., Appavu, B., Johnson, E. L., Osman, G., Haider, H. A., et al. (2021a). Deep active learning for interictal ictal injury continuum eeg patterns. *Journal of neuroscience methods*, 351:108966.

Ge, W., Jing, J., An, S., Herlopian, A., Ng, M., Struck, A. F., Appavu, B., Johnson, E. L., Osman, G., Haider, H. A., Karakis, I., Kim, J. A., Halford, J. J., Dhakar, M. B., Sarkis, R. A., Swisher, C. B., Schmitt, S., Lee, J. W., Tabaeizadeh, M., Rodriguez, A., Gaspard, N., Gilmore, E., Herman, S. T., Kaplan, P. W., Pathmanathan, J., Hong, S., Rosenthal, E. S., Zafar, S., Sun, J., and Brandon Westover, M. (2021b). Deep active learning for interictal ictal injury continuum EEG patterns. *Journal of Neuroscience Methods*, 351:108966.

Gelman, A. (2006). The boxer, the wrestler, and the coin flip. *The American Statistician*, 60(2):146–150.

Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica*, 19(2):491–544.

Hasegawa, R. B., Webster, D. W., and Small, D. S. (2019). Evaluating Missouri's handgun purchaser law: A bracketing method for addressing concerns about history interacting with group. *Epidemiology*, 30(3):371–379.

Hill, A. V. (1909). 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.

Hirsch, L. J., Fong, M. W., Leitinger, M., LaRoche, S. M., Beniczky, S., Abend, N. S., Lee, J. W., Wusthoff, C. J., Hahn, C. D., Westover, M. B., et al. (2021). American clinical neurophysiology society's standardized critical care eeg terminology: 2021 version. *Journal of Clinical Neurophysiology*, 38(1):1–29.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6(2), 65–70.*

Hotradat, M., Balasundaram, K., Masse, S., Nair, K., Nanthakumar, K., and Umapathy, K. (2019). Empirical mode decomposition based ecg features in classifying and tracking ventricular arrhythmias. *Computers in Biology and Medicine*, 112:103379.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269.

Huang, N. and Shen, S. (2014). *Hilbert-Huang transform and its applications.* World Scientific, 2nd edition.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N., Tung, C. C., and Liu, H. H. (1998). The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454:903 – 995.

Huang, N. E., Wu, M.-L. C., Long, S. R., Shen, S. S., Qu, W., Gloersen, P., and Fan, K. L. (2003). A confidence limit for the empirical mode decomposition and hilbert spectral analysis. *Procedings of the Royal Society Series A.*

Hughes, M. A., Glass, P. S., and Jacobs, J. R. (1992). Context-sensitive half-time in multicompartment: pharmacokinetic models for intravenous anesthetic drugs. *The Journal of the American Society of Anesthesiologists*, 76(3):334–341.

Huimin, Z., Meng, S., Wu, D., and Xinhua, Y. (2017). A new feature extraction method based on eemd and multi-scale fuzzy entropy for motor bearing. *Entropy*, 19(1).

Inclan, C. and Tiao, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923.

Jacob, P. E., Gong, R., Edlefsen, P. T., and Dempster, A. P. (2021). A gibbs sampler for a class of random convex polytopes. *Journal of the American Statistical Association*, 116(535):1181–1192.

Jing, J., Dauwels, J., Rakthanmanon, T., Keogh, E., Cash, S., and Westover, M. (2016). Rapid annotation of interictal epileptiform discharges via template matching under dynamic time warping. *Journal of neuroscience methods*, 274:179–190.

Johnson, E. L. and Kaplan, P. W. (2017). Population of the ictal-interictal zone: the significance of periodic and rhythmic activity. *Clinical Neurophysiology Practice*, 2:107–118.

Kim, J. A., Boyle, E. J., Wu, A. C., Cole, A. J., Staley, K. J., Zafar, S., Cash, S. S., and Westover, M. B. (2018). Epileptiform activity in traumatic brain injury predicts post-traumatic epilepsy. *Ann Neurol*, 83(4):858–862.

Kopsinis, Y. and Stephen, M. (2009). Development of emd-based denoising methods inspired by wavelet thresholding. *Signal Processing, IEEE Transactions on*, 57:1351 – 1362.

Lawrence, E., Liu, C., Wiel, S. V., and Zhang, J. (2009). A new method of multinomial inference using dempster-shafer theory. *Unpublished Manuscript*.

Lee, M. H., Kong, D.-S., Seol, H. J., Nam, D.-H., and Lee, J.-I. (2013). Risk of seizure and its clinical implication in the patients with cerebral metastasis from lung cancer. *Acta neurochirurgica*, 155(10):1833–1837.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag, New York, NY, USA, second edition.

Lei, Y. and Zuo, M. J. (2009). Fault diagnosis of rotating machinery using an improved HHT based on EEMD and sensitive IMFs. *Measurement Science and Technology*, 20(12):125701.

Li, T., Zhou, M., Guo, C., Luo, M., Wu, J., Pan, F., Tao, Q., and He, T. (2016). Forecasting crude oil price using eemd and rvm with adaptive pso-based kernels. *Energies*, 9(12).

Lin, J. J., Chou, C. C., Lan, S. Y., Hsiao, H. J., Wang, Y., Chan, O. W., Hsia, S. H., Wang, H. S., Lin, K. L., Group, C. S., et al. (2017). Therapeutic burst-suppression coma in pediatric febrile refractory status epilepticus. *Brain and Development*, 39(8):693–702.

Lin, X., Banks, D. L., Scott, D. W., Genest, C., Molenberghs, G., and Wang, J.-L. (2014). *Past Present and Future of Statistical Science*. CRC Press.

Liu, D., Yang, X., Wang, G., Ma, J., Liu, Y., Peng, C.-K., Zhang, J., and Fang, J. (2012). Hht based cardiopulmonary coupling analysis for sleep apnea detection. *Sleep Medicine*, 13.

Liu, G. & Luan, Y. (2015). An adaptive integrated algorithm for noninvasive fetal ecg separation and noise reduction based on ica-eemd-ws. *Medical & Biological Engineering & Computing*, 53(11):1113–1127.

Lozano, M., Fiz, J. A., and Jané, R. (2016). Performance evaluation of the hilbert–huang transform for respiratory sound analysis and its application to continuous adventitious sound characterization. *Signal Processing*, 120:99–116.

Lucke-Wold, B. P., Nguyen, L., Turner, R. C., Logsdon, A. F., Chen, Y.-W., Smith, K. E., Huber, J. D., Matsumoto, R., Rosen, C. L., Tucker, E. S., et al. (2015). Traumatic brain injury and epilepsy: underlying mechanisms leading to seizure. *Seizure*, 33:13–23.

Marchi, N. A., Novy, J., Faouzi, M., Stähli, C., Burnand, B., and Rossetti, A. O. (2015). Status epilepticus: impact of therapeutic coma on outcome. *Critical care medicine*, 43(5):1003–1009.

Martin, R., Zhang, J., and Liu, C. (2010). Dempster–Shafer Theory and Statistical Inference with Weak Beliefs. *Statistical Science*, 25(1):72 – 87.

Muhlhofer, W. G., Layfield, S., Lowenstein, D., Lin, C. P., Johnson, R. D., Saini, S., and Szaflarski, J. P. (2019). Duration of therapeutic coma and outcome of refractory status epilepticus. *Epilepsia*, 60(5):921–934.

Negraru, P. T. and Herrin, E. T. (2009). On Infrasound Waveguides and Dispersion. *Seismological Research Letters*, 80(4):565–571.

Ng, K., Tian, G., and Tang, M. (2011). *Dirichlet and Related Distributions: Theory, Methods and Applications*. Wiley Series in Probability and Statistics. Wiley.

Northon, K. (2015). Nasa statement regarding oct. 28 orbital sciences corp. launch mishap.

Oddo, M., Carrera, E., Claassen, J., Mayer, S. A., and Hirsch, L. J. (2009). Continuous electroencephalography in the medical intensive care unit. *Critical Care Medicine*, 37(6):2051–2056.

Osman, G. M., Araújo, D. F., and Maciel, C. B. (2018). Ictal interictal continuum patterns. *Current Treatment Options in Neurology*, 20(5):1–20.

Parikh, H., Rudin, C., and Volfovsky, A. (2018). Malts: Matching after learning to stretch.

Parikh, H., Rudin, C., and Volfovsky, A. (2019). An application of matching after learning to stretch (MALTS) to the ACIC 2018 causal inference challenge data. *Observational Studies*, 5:118–130.

Payne, E. T., Zhao, X. Y., Frndova, H., McBain, K., Sharma, R., Hutchison, J. S., and Hahn, C. D. (2014). Seizure burden is independently associated with short term outcome in critically ill children. *Brain*, 137(5):1429–1438.

Pearl, J. (1988). On probability intervals. *International Journal of Approximate Reasoning*, 2(3):211–216.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Rossetti, A. O., Hirsch, L. J., and Drislane, F. W. (2019). Nonconvulsive seizures and nonconvulsive status epilepticus in the neuro icu should or should not be treated aggressively: A debate. *Clinical Neurophysiology Practice*, 4:170–177.

Rossetti, A. O., Logroscino, G., and Bromfield, E. B. (2005). Refractory status epilepticus: effect of treatment aggressiveness on prognosis. *Archives of Neurology*, 62(11):1698–1702.

Rubinos, C., Reynolds, A. S., and Claassen, J. (2018). The ictal–interictal continuum: to treat or not to treat (and how)? *Neurocritical Care*, 29(1):3–8.

Rudin, W. (1953). *Principles of mathematical analysis*. McGraw-Hill Book Company, Inc., New York-Toronto-London.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464.

Sentz, B. and Ferson, S. (2002). Combination of Evidence in DempsterShafer Theory. Technical report, Sandia National Laboratory.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472.

Sun, H., Jia, J., Goparaju, B., Huang, G.-B., Sourina, O., Bianchi, M. T., and Westover, M. B. (2017). Large-scale automated sleep staging. *Sleep*, 40(10).

Tabaeizadeh, M., Nour, H. A., Shoukat, M., Sun, H., Jin, J., Javed, F., Kassa, S., Edhi, M., Bordbar, E., Gallagher, J., et al. (2020). Burden of epileptiform activity predicts discharge neurologic outcomes in severe acute ischemic stroke. *Neurocritical Care*, 32(3):697–706.

Tao, J. X., Qin, X., and Wang, Q. (2020). Ictal-interictal continuum: a review of recent advancements. *Acta Epileptologica*, 2(1):1–10.

Trinka, E., Cock, H., Hesdorffer, D., Rossetti, A. O., Scheffer, I. E., Shinnar, S., Shorvon, S., and Lowenstein, D. H. (2015). A definition and classification of status epilepticus–Report of the ILAE Task Force on Classification of Status Epilepticus. *Epilepsia*, 56(10):1515–1523.

Vergoz, Julien, e. a. (2018). "the antares explosion observed by the usarray: An unprecedented collection of infrasound phases recorded from the same event.". *Infrasound Monitoring for Atmospheric Studies*, page 349–386.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, T., Zhang, M., Yu, Q., and Zhang, H. (2012). Comparing the applications of emd and eemd on time–frequency analysis of seismic signal. *Journal of Applied Geophysics*, 83:29–34.

Wang, W.-c., Xu, D.-M., and Chen, X.-Y. (2015). Improving forecasting accuracy of annual runoff time series using arima based on eemd decomposition. *Water Resources Management*, 29:2655–2675.

Wang, X., Liu, C., Bi, F., Bi, X., and Shao, K. (2013). Fault diagnosis of diesel engine based on adaptive wavelet packets and eemd-fractal dimension. *Mechanical Systems and Signal Processing*, 41(1):581–597.

Wu, Y.-X., Wu, Q.-B., and Zhu, J.-Q. (2019). Improved eemd-based crude oil price forecasting using lstm networks. *Physica A: Statistical Mechanics and its Applications*, 516:114–124.

Wu, Z. and Huang, N. (2009). Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv. Data Sci. Adapt. Anal.*, 1:1–41.

Wu, Z. and Huang, N. (2014). Statistical significance test of intrinsic mode functions.

Xie, M.-g. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81(1):3–39.

Yager, R. and Liu, L. (2008). *Classic Works of the Dempster-Shafer Theory of Belief Functions*. Studies in Fuzziness and Soft Computing. Springer Berlin Heidelberg.

Zafar, S. F., Postma, E. N., Biswal, S., Boyle, E. J., Bechek, S., O'Connor, K., Shenoy, A., Kim, J., Shafi, M. S., Patel, A. B., et al. (2018). Effect of epileptiform abnormality burden on neurologic outcome and antiepileptic drug management after subarachnoid hemorrhage. *Clinical Neurophysiology*, 129(11):2219–2227.

Zafar, S. F., Rosenthal, E. S., Jing, J., Ge, W., Tabaeizadeh, M., Aboul Nour, H., Shoukat, M., Sun, H., Javed, F., Kassa, S., et al. (2021). Automated annotation of epileptiform burden and its association with outcomes. *Annals of neurology*, 90(2):300–311.

Zhang, J. and Liu, C. (2014). Dempster-shafer inference with weak beliefs. *Statistica Sinica*, 21.

Zhang, N. and Siegmund, D. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32.

Zheng, J., Pan, H., Yang, S., and Cheng, J. (2017). Adaptive parameterless empirical wavelet transform based time-frequency analysis method and its application to rotor rubbing fault diagnosis. *Signal Processing*, 130:305–314.