University of California

Los Angeles

Developing and Integrating Computer-Aided Diagnostic Tools into Clinical Medicine

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biomathematics

by

Wesley Thomas Kerr

2015

Abstract of the Dissertation

Developing and Integrating Computer-Aided Diagnostic Tools into Clinical Medicine

by

Wesley Thomas Kerr

Doctor of Philosophy in Biomathematics

University of California, Los Angeles, 2015

Professor Mark S. Cohen, Co-chair

Professor Henry Huang, Co-chair

The focus of this graduate thesis is the development and optimization of clinically applicable computer-aided diagnostic tools (CADTs) for seizure disorder. This thesis is comprised of two parts (1) development of unimodal and multimodal CADTs for seizure disorder and (2) a novel method for optimization of hyperparameters in machine learning models. The aims of CADTs are to address key challenges in the diagnosis and treatment of seizure disorder, including reducing the time to an accurate diagnosis, improving the sensitivity and specificity of diagnostic neuroimaging, and the understanding of the diagnostic value of interictal scalp electroencephalography (EEG). This could improve the long-term prognosis of patients with non-epileptic seizures (NES) and candidates for potentially curative resective surgery for epilepsy because treatment earlier in these patients' disease course has been shown to be more effective. Our novel method for optimizing hyperparameters has the potential to slightly improve the accuracy of machine learning models, while substantially increasing the interpretability of learned estimates and reducing computational cost. We define hyperparameters as variables that contribute to machine learning models but are not optimized jointly with parameters inherent to the model. When viewed as a whole, this body of work represents contributions both to the statistical development and application of machine learning to important clinical challenges.

The dissertation of Wesley Thomas Kerr is approved.

John M. Stern

Elliott Landaw

Marc A. Suchard

Henry Huang, Committee Co-chair

Mark S. Cohen, Committee Co-chair

University of California, Los Angeles

2015

Table of Contents

1 Introduction . . . 1

1.1 Public Health, Diagnostic and Treatment Challenges in Seizures . 1

1.2 Diagnostic Modalities Utilized in Seizures . . . 4

1.3 Significance & Impact of Developing Computer-Aided Diagnostic Tools (CADTs) for Seizures . . . 7

1.4 Statistical Challenges in Training & Validating Machine Learning Models . . . 11

1.5 Use of Electronic Health Records (EHRs) in Clinical Research . . 15

1.6 Summary of Content Herein . . . 17

2 Didactic Background Material . . . 18

2.1 Simple Linear Models . . . 18

2.2 Generalized Linear Models . . . 26

2.3 Machine Learning Statistics . . . 35

2.4 Machine Learning Classifiers . . . 48

2.4.1 Nearest Neighbor Classifiers and Norms . . . 48

2.4.2 Decision Trees . . . 51

2.4.3 Discriminant Analysis . . . 55

2.4.4 Support Vector Machines . . . 56

2.4.5 Neural Networks . . . 59

2.5 Experimental Design, Curse of Dimensionality & Feature Selection 62

2.5.1 Subsampling . . . 66

2.5.2 Filter methods . . . 67

2.5.3 Projection methods 68

2.5.4 Iterative methods 74

2.5.5 Regularization and Bayesian priors 75

2.5.6 Biological priors 80

2.6 Population Selection in the Development of CADTs 82

3 The Future of Medical Diagnostics: Large Digitized Databases 87

3.1 Abstract 87

3.2 Introduction 88

3.3 Mandates and Policies Driving the Change 90

3.4 Large Databases 91

3.5 Anticipated Challenges to Database Creation 95

3.6 Current Applications and Benefits of CAD 97

3.7 Benefits and Challenges of Databases in the Development of CAD Tools 102

3.8 The Future of Medical Diagnostics 105

3.9 Conclusions and Outlook 107

4 The utility of data-driven feature selection: Re: Chu *et al.* 2012109

4.1 Abstract 109

4.2 Comment 110

5 Balancing Clinical and Pathologic Relevence in the Machine Learning Diagnosis of Epilepsy 118

5.1 Abstract 118

5.2 Introduction 119

5.3 Methods 122

5.3.1 Dataset 122

5.3.2 Machine Learning Details 123

5.4 Results 123

5.4.1 Cross-Validation Accuracy Differences 123

5.4.2 Insight into Focality of the Epilepsies 124

5.5 Discussion 124

6 Poisson Noise Obscures Hypometabolic Lesions in PET 128

6.1 Abstract 128

6.2 Introduction 129

6.3 Materials and Methods 132

6.4 Results 134

6.5 Discussion 137

6.6 Supplementary Material 142

6.6.1 Regressions of Trends in Reconstructed Voxel Intensity . . 142

6.6.2 Reconstruction Algorithms 142

6.6.3 Variance Proof for the Filtered Back-Projection 143

6.6.4 Sample Size Calculation using Signal to Noise Ratios . . . 144

7 Accurate differentiation of epileptic and non-epileptic seizures through quantitative combination of findings in the clinical history145

7.1 Abstract 146

7.2 Introduction 147

7.3 Methods 148

7.3.1 Clinical Features 148

7.3.2 Statistical Methods 151

7.4 Results 155

7.5 Discussion 159

7.5.1 Diagnostic Value of Particular Findings 161

7.5.2 Developing an Outpatient Assessment Score 163

7.5.3 Limitations 165

7.5.4 Future Potential and Perspective 165

8 Automated Diagnosis of Epilepsy using EEG Power Spectrum 167

8.1 Abstract 167

8.2 Introduction 168

8.3 Patients and Methods 170

8.4 Results 171

8.5 Conclusion 173

9 Computer aided diagnosis and localization of lateralized temporal lobe epilepsy using interictal FDG-PET 176

9.1 Abstract 176

9.2 Introduction 177

9.3 Materials and Methods 183

9.3.1 Patient data 183

9.3.2 Computer aided diagnostic tool training and validation . . 184

9.3.3 Machine learning algorithmic details 185

9.3.4 Manual analysis of PET and MRI records 186

9.3.5 Combination of clinical information with computer aided diagnostic information 187

9.4 Results 189

9.5 Discussion 193

9.5.1 Clinical Impact 196

9.5.2 Combination of automated analysis with clinical wisdom . 197

9.5.3 Pathophysiological Insights 198

9.5.4 Limitations and future directions 202

9.6 Conclusion 204

10 Multimodal diagnosis of epilepsy using conditional dependence and multiple imputation 205

10.1 Abstract 205

10.2 Introduction 206

10.3 Methods 208

10.4 Results 212

10.5 Discussion 214

10.6 Conclusion 216

11 Parameter Selection in Mutual Information-Based Feature Selection in Automated Diagnosis of Multiple Epilepsies Using Scalp EEG 217

11.1 Abstract 217

11.2 Introduction 218

11.3 Methods 221

11.3.1 Patient & EEG Processing Information 221

11.3.2 Sampling, Feature Selection and Classification 222

11.4 Results 223

11.4.1 Variance with respect to Feature Number 223

11.4.2 Variance with respect to Quantal Level 223

11.5 Discussion 226

12 Hyperparameter Optimization with Random Field Theory without Nested Cross-Validation 229

12.1 Abstract 229

12.2 Introduction 230

12.3 Theoretical Framing of Random Field Theory Application 233

12.3.1 Chosen Examples 234

12.4 Analytical Exploration of 'Spatial' Dependence of Hyperparameter Choice 236

12.4.1 Regularization Hyperparameter, λ 237

12.4.2 Soft Margin Hyperparameter, C, in SVM 238

12.4.3 Feature Selection Hyperparameters 240

12.4.4 Summary of Analytical Exploration of Dependence 242

12.5 Empirical validation of "Spatial" Dependence in Null Data 242

12.5.1 Regularization Hyperparameter, λ 243

12.5.2 Soft Margin Hyperparameter, C, in SVM 243

12.5.3 Feature Selection Hyperparameters 244

12.5.4 Summary of Dependence in Null Datasets 244

12.6 Empirical Validation of "Spatial Dependence" in Epilepsy Data 245

12.6.1 Regularization Hyperparameter, λ 246

12.6.2 Soft Margin Hyperparameter, C, in SVM 246

12.6.3 Feature Selection Hyperparameters 247

12.6.4 Summary of Empirical Validation in Epilepsy Data 248

12.7 Comparison to Alternative Correction Methods 248

12.8 Verification of Asymptotic Approach of the Optimal Value of θ . . 249

12.9 Discussion 250

12.9.1 Recommendations for Determining Significance of Observed Maximum Cross-Validation Accuracy 252

12.9.2 Planning Sampling Schemes for Hyperparameters 253

12.9.3 Limitations and Future Directions 254

12.10Conclusion 255

13 Discussion & Conclusions 256

14 Conflicts of Interests 262

15 Abbreviations, Notation & Glossary 263

15.1 Abbreviations 263

15.2 Variable & Notation Definition 269

15.3 Glossary of Terminology 272

References 275

List of Figures

1.1 Flow chart of a statistical experiment. Dashed lines reflect optional steps. 12

1.2 3-Fold Cross-validation when θ is selected *a priori*, as compard to optimized. Therefore, no data is needed to learn θ. A separate model is learned for each fold, resulting in three overall models to aggregate later. 14

2.1 Flow chart of a statistical experiment. Dashed lines reflect optional steps. 19

2.2 Example of simulated input data (black) and a linear regression line (blue) fit of these data. Data were sampled from a linear line with Guassian noise. 20

2.3 Example of simulated input data (black) and a log-normal regression (blue) fit of these data. Data were sampled from an exponential line with Guassian noise in the transformed, exponential space. . . 27

2.4 Example of simulated input data (black) and a Poisson regression (blue) fit of these data. Data were sampled from a Poisson line with Guassian noise in the transformed, Poisson space. 29

2.5 Example of simulated input data (black) and a logistic regression (blue) fit of these data. Data were sampled from a logistic trend used to define the probability of success in a Bernoulli trial. Uniform noise was added to the Y dimension of the logistic input data to aid in visualization, but this noise was not included in modeling. 30

2.6 Illustration of how Newton's method is used to find a local critical point of the data. The initial guess is X_1 and the initial guess for the zero point is X_2. The value of the function is assessed at X_2 and iterated until the zero point is found. To find the critical point of a log-likelihood function, the zero of the derivative of the log-likelihood is found. 32

2.7 Qualitative illustration of the ideal and approximated penalty function for major categories of machine learning models. In this illustration, $y \in \{\pm 1\}$ and $\hat{y} \in \mathbb{R}$. 38

2.8 Splitting the data into three mutually exclusive groups to avoid overfitting and use data to estimate all important parts of the model and its performance. 39

2.9 Bias-Variance Tradeoff: As the number of training samples increases, both the mean and variance of the cross-validation accuracy increase monotonically. Mean is indicated by the black line, and standard deviation is illustrated by the grey error bars. Exact values and shape are illustrative, and do not represent real or simulated results. 40

2.10 3-Fold Cross-validation when θ is selected *a priori*, as compard to optimized. Therefore, no data is needed to learn θ. A separate model is learned for each fold, resulting in three overall models to aggregate later. 41

2.11 3 then 2 Fold Nested Cross-validation. The data is split into 3 mutually exclusive sets, and each one is successively used as the validation set. For each validation set, the remaining data is split into 2 mutually exclusive sets to estimate the optimum ψ and θ. As shown, the inner splits correspond to the higher level splits, but this is not necessarily the case. A separate model is fit for each fold. These models must then be aggregated in post-processing. . 42

2.12 An example receiver-operating curve which shows performance of a prediction with respect to different balances of sensitivity (True Positive Rate) compared to specificity (1-False Positive Rate). The black line illustrates a classifier with good performance, whereas the gray line indicates theoretical chance. 46

2.13 This illustrates the complexity of a nearest neighbor solution for two dimensional input data, where the closest $k = 5$ neighbors based on an Euclidean or L_2 norm were considered. The lighter color indicates the region in which validation data would be predicted to be from the blue or red class. 49

2.14 The distance between two points based on the L_1 and L_2. 50

2.15 The shape of "circles" calculated according to various norms. . . 52

2.16 An example decision tree. All exemplars begin in the largest circle, then are triaged towards the daughter circles based on maximizing the Gini impurity. Exemplars are predicted to be the most common class within the last daughter circle. Note that once a covariate is used, it can be reused lower down the tree. Depending on the stopping criterion, the depth of the tree is not necessarily uniform. 53

2.17 An example of an overfit decision tree. Black sections indicate where the model would predict class below, whereas white sections indicate class above. Simulated data (black dots) were distributed uniformly over the interval and randomly assigned to be class below or above, with a stable probability of 50%. 54

2.18 Examples of lines separating two classes based on Fisher Linear and Quadratic Discriminant Analysis. Simulated data is centered around a point with Gaussian noise added. The variance of the blue class is twice that of the red class. 56

2.19 Example of a soft-margin SVM solution for simulated Gaussian data. The black line indicates the SVM separating hyperplane. The red exemplar on the wrong side of the hyperplane is an example of a misclassified training exemplar that is a distance, ξ, from the maximum separating hyperplane. The green squares highlight the three support vectors that would define the hyperplane, if the misclassified exemplars were not present. The value of α_i is greater than 0 for all suport vectors, by definition. Lastly, the separating hyperplane is defined by Xw where w is perpendicular to a line that linearly projects the multivariate data onto a single decision dimension, β. 58

2.20 An illustration of the structure of a multilayer neural network. The input data, X, are combined linearly and pushed through a link function, g. In the case of a multilayer perceptron, this link is a logit function. Other neural networks use other link functions. These successive combinations of the input data are eventually used to predict the final output class. Through these successive combinations, high level interactions can be modelled, and a more nuanced penalty function can be used to approximate the ideal step function. 60

2.21 An example of the curse of dimensionality for one (black) versus two (red) dimensional data uniformly distributed over the interval $[0, 1]$. The colored interval and circle represent the size a hypersphere with matching radius in each dimensionality. 64

2.22 An illustration of the fraction of the total space taken by hyperspheres of various radii as the dimensionality of the input data increases. Observe that the fraction of the total space taken decreases more than exponentially, as seen by the slight positive inflection on these trend lines. 65

2.23 An example of the PCA (red) versus ICA (green) projection of input data. Observe that the first components of ICA and PCA are identical, whereas the second components vary. The second IC reflects the generative process of the data by identifying the next largest statistically independent component of the data, whereas the second PC is required to be orthogonal to the first. Figure reproduced from gael-varoquaux.info. 71

2.24 A) A two-dimensional manifold transformed into a three dimensional space. Distance along the manifold is meaningful, but distances based on three-dimensional coordinates have little to no meaning. B) Data sampled from this two-dimensional manifold. C) The data in B projected onto a two-dimensional space based on local-linear embedding. The black outlines in B and C reflect a neighborhood around a point in the original and projected spaces. Figure reproduced from Rowes & Saul [1]. 73

2.25 For linear solutions, there exist a set of solutions that result in the same log-likelihood (black line). The L_1 penalty chooses the solution with the minimum L_1, which also happens to be the solution with minimum L_0, in the vast majority of cases (b). In contrast, the L_2 and L_0 solutions do not coincide (a). 78

3.1 This figure illustrates the number of PubMed citations using each of the Mesh terms listed. Since 2002, the number of publications regarding computer-aided diagnostics has increased substantially. We are already seeing a commensurate increase in the number of publications regarding computerized medical record systems and electronic health records.[2] 89

3.2 Even before the ARRA in 2009, the number of physicians utilizing EHR systems was increasing. There are already a substantial percent of physicians using electronic records. Consequentially, it is relatively inexpensive to combine and mine these EHR systems for high quality clinical information. 92

3.3 The creation and utilization of EHR databases is complex; however, each of the steps in the data and implementation system are well defined. We expect that responsible researchers will be capable of tackling each of these steps to create unparalleled databases and develop high quality, clinically applicable CAD tools. 98

4.1 A reproduction of Chu *et al.*'s[3] figure 9F where the added shading indicates the 95% confidence interval for the no feature selection accuracy using the normal approximation of the binomial distribution. Accuracy using all voxelized features was no significantly higher than data-driven feature selection accuracy at the optimum C, C^*. At multiple non-optimum C values, the accuracy using data-driven feature selection was significantly higher than using all voxelized features. 115

5.1 CL1OCV accuracy of our computer-aided diagnostic tool using each control group. Error bars indicate standard error from the mean. Shading indicates performance of a naive classifier. PWN: Patients with non-epileptic seizures; NNP: Neurologically normal patients; L or R TLE: Left or Right Temporal Lobe Epilepsy. 124

5.2 Percent of the 47 ROIs that contributed to the best CL1OCV accuracy. Error bars indicate accuracies within the same significant random field theory cluster. 125

6.1 These circles illustrate examples in which each pixel intensity is initialized using a discrete uniform distribution with range of 1 to 100. An independent Poisson random variable with parameter equal to this intensity is then realized for each pixel. We then used the filtered back projection (FBP) and expectation maximization algorithm (EM) to reconstruct this circle based on its projection, as is done for PET images. 132

6.2 This figure illustrates the probability distribution of reconstructed voxel intensity for each of the reconstruction algorithms. For comparison, the right panel illustrates the original probability distribution before reconstruction. 135

6.3 This figure illustrates the magnitude of the reconstructed intensity bias of each of the algorithms. The line thickness represents the standard error for each point. This standard error is small due to the large sample size. The FBP is indicated by cyan and the EM is indicated by green. 136

6.4 This figure illustrates the magnitude of the signal to noise ratio of the reconstructed intensity each of the algorithms. Signal to noise ratio was calculated as the original intensity divided by the standard deviation of the biased reconstructed intensity. This corresponds to a hypothesized two-fold change in intensity. The line thickness represents the standard error for each point. This standard error is small due to the large sample size. The FBP is indicated by cyan, and the EM is indicated by green. 137

7.1 Odds ratio of all factors with a significant factors with at least 95% confidence based on multivariate logistic regression (MLR) with multiple imputation. The size of the colored bar indicates the magnitude of the odds ratio, and the color indicates its sign (Green for ES, Blue for NES). The transparent gray overlay indicates the 95% empirical confidence interval of chance. For prevalence, see Supplementary Table 1. All factors were indicators unless the units are specified otherwise. All starred units (*) were normalized across the whole dataset therefore bars reflect the odds ratio of an increase of one standard deviation. Because p-values were not independent across factors, no correction for multiple testing was applied. Abbreviations: number (#,n), seizure disorder (SzD), years old (y.o.), traumatic brain injury (TBI), with (w/), anti-seizure medication (ASM), seizure (Sz), movements (mvmts). 157

7.2 Leave-one-out cross-validation performance statistics of the data-driven L1 regularized logistic regression (L_1-LR) to differentiate between Non-Epileptic Seizures (NES) and Epileptic Seizures (ES). Accuracy, sensitivity and specificity (7.2A) describe the point on the receiver operating curve (ROC, 7.2B) that maximizes accuracy. Shading reflects empirical 95% confidence intervals of chance. Abbreviations: area under the ROC (AUC). 158

8.1 (A) directly compares the summary statistics of our computer-aided diagnostic (CAD) tool to the same statistics regarding conventional analysis of EEG. (B) assesses the likelihood ratios that can be achieved when our CAD is combined with conventional analysis. Error bars denote 95% CIs and are calculated without normal assumptions. Dashed lines indicate chance or 95% CIs of chance. All effects are significantly different from chance ($p < 0.001$) except when CAD is positive and manual analysis is negative. No comparative effects are significantly different. 172

8.2 Diagnostic yield of long-term video-EEG monitoring. Numbers indicate how many patients are in each class, and the size of the bar denotes percent of total, listed on the right side of the figure, that belong to each class. When the presurgical and intractable classes are combined, 6% of the patients have inconclusive results. NES, nonepileptic seizures. 174

9.1 CAD tool performance matches manual analysis. These figures indicate the accuracy, sensitivity and specificity of the LTLE (A), RTLE (B) and trinary (C) classifiers. The performance of our CAD tools matched that of MA and was superior to just using gender alone. The error bars indicate standard error of the mean performance for each measure. The translucent region indicates the performance of a naive classifier. *Indicates significant differences from the naive classifier with a confidence level of 95% or more. . 190

9.2 CAD tool is not redundant with manual analysis. The squared correlation of our CAD tools' results with those of MA of the iPET or sMRI from the same patients was below 50%. This indicates that while some information is shared, the majority of information provided by our CAD tools is not captured by MA. The correlation between MA of iPET and sMRI is similar in magnitude to the correlation of CAD with MA, therefore the CAD could potentially be seen as similar to another informative modality. *Indicates significant differences of the correlation from zero with a confidence level of 95% or more. 192

9.3 Automated combination of clinical information with automated analysis of iPET images. The automated combination of clinical information and/or MA with our analysis produced no significant change in performance for the LTLE (A), RTLE (B) or trinary (C) classifiers, relative to the CAD operating on automated values alone. The unshaded bars indicate the performance of similarly constructed CAD tools using clinical information or the results of MA alone. The shaded bars indicate the modified performance when information from NeuroQ is added. The horizontal line indicates the mean accuracy of each CAD tool without clinical information. The translucent region indicates the performance of a naive classifier. 194

9.4 Combination of clinical information and CAD results using likelihoods. Columns in this log plot above 1 indicate that the seizures are more likely to be epileptic whereas the columns below 1 indicate a non-epileptic etiology is more probable. (A) Illustrates the positive and negative likelihood ratio of each analysis method considered individually. (B,C) Illustrate the likelihood ratios of each possible outcome when two analysis methods are combined. (D) Indicates the likelihood ratios of each possible outcome when all analysis methods are combined. If all modalities agree, the likelihood non-significantly increases with the addition of each modality. However, if there is disagreement, the likelihood ratio is generally not significantly different from chance. The translucent bars indicate the 95% confidence interval for chance with the relevant sign (see Materials and Methods).The numbers above the translucent bars indicate the total number of patients with each outcome. The bars that go off the scale of the graph diverge toward zero or infinity because no patients of a certain class had that outcome. *Indicates significant differences of the correlation from zero with a confidence level of 95% or more. 195

10.1 Overall accuracy (A) and per-class sensitivity (B) of each classifier. Error bars reflect binomial theoretical standard error bars, with multiple imputation. Red shading reflects the 95% quantile bounds from permutation tests. Vector concatenation and conditional dependence are indicated by + and -, respectively. For conditional dependence, the order of modalities is read from left to right. Abbreviations: Clinical information (CI). 213

11.1 The cross validation accuracy of all four classifiers. The unsampled points are filled using Akima bivariate interpolation [4]. The bottom right corner is set to 0 due to lack of support. All values less than 40% are rounded up to 40% to maintain contrast. Without multiple testing correction, individual yellow to red points are significantly more accurate than a naive classifier whereas deep blue points are significantly worse. 224

11.2 Variance of cross validation accuracy with respect to number of input features. Thickness represents standard error. 225

11.3 Variance of cross validation accuracy with respect to number of quantal levels. Thickness represents standard error. 225

12.1 (Left) The correlation of cross-validataion accuracy by subject for the clinical dataset at one value of F_1 compared to F_2 where the distance between F_1 and F_2 is modified. (Right) The log PDF of maximum cross-validataion accuracy achieved over a range of F. . 247

List of Tables

3.1 A quick summary of notable databases of high quality information that have been developed and are being used for large scale studies. 93

5.1 The mRMR Rank of the top 5 ROIs based on the full data. Note that these rankings utilize the full dataset and therefore do not necessarily coincide with any individual training set, each of which are missing data from just one patient. The preceding L and R refer to left and right respectively. The lower case i, l, a, p, s, and m stand for inferior, lateral, anterior, posterior, superior, and medial respectively. The other abbreviations are for temporal cortex (Temp), thalamus (Thal), associative visual cortex (Ass Vis), and Sensorimotor cortex (SM). The temporal, parietal and frontal ROIs are all cortical ROIs. Colors indicate repeat ROIs. 125

7.1 Diagnostic subtype breakdown of patients included in our population. Lobe indicates location of seizure-onset zone. If seizure onset zones were in multiple lobes and/or had multiple seizure etiologies, patient was listed in each category. Abbreviations: Not Otherwise Specified (NOS), Lobe Epilepsy (LE), Frontal (F), Temporal (T), Parietal (P), Occipital (O), Non-Epileptic Seizures (NES), Epileptic Seizures (ES). 150

7.2 A complete list of the studied factors from each clinical note. For a detailed description of each factor, please refer to the online-only appendix. Abbreviations: number (#,n), anti-seizure medications (ASMs), medications (meds), review of systems (ROS), gastro-esophageal reflux disorder (GERD), transient ischemic attacks (TIA), neurofibromatosis type 1 (NF1), supplements (suppl). 152

7.3 The significant differences between patients with epileptic seizures (ES), non-epileptic seizures (NES) and inconclusive/mixed diagnosis based on video-electroencephalography. All factors are listed in percentages, unless otherwise specified. *Statistics of seizure duration, seizure frequency, and duration of seizure disorder were calculated in log space, but are displayed in linear values (i.e. displayed value is $10^{[mean[\log_{10} data]]}$). All unlisted factors exhibited no statistically significant differences between ES and NES with greater than 95% confidence (false discovery rate, α=0.05). Dashes in missing data indicate a variable assumes to be biased (see Methods). Abbreviations: number (#), missing data (M), seizure disorder (SzD), standard error (SE), seizures (Sz), anti-seizure medications (ASMs), review of systems (RoS), postictal (PI), deep tendon reflexes (DTRs), comorbidities (comorbid). 159

7.4 See legend for Table 7.3. 160

9.1 This table reflects the clinical information known before the application of the CAD tool. All times are listed in years (y) unless otherwise specified by days (d), weeks (w), or months (m). Manual analysis of all patients' iPET and sMRI were not done, therefore we list the number with available manual results. *, §, or ¶ indicate that the value for NES vs. LTLE, NES vs. RTLE, or LTLE vs. RTLE, respectively, is statistically significant from both the LTLE and RTLE groups with at least 95% confidence using a two-sample z-test of proportions or MannWhitney U test, where appropriate. No other differences are statistically significant ($p > 0.10$). 188

9.2 This table illustrates the top six informative and non-redundant regions of interest (ROIs) that may contribute to each of the CAD tools, as determined by the minimum redundancy-maximum relevancy criteria (mRMR)[5, 6]. The illustrated rank order of features was calculated based on the full dataset and does not necessarily match the rank list of any individual training sample. The leading L or R indicates left or right. The lowercase letters indicate inferior (i), lateral (l), median (m), anterior (a), and posterior (p). The lagging C signifies cortex. Note that the LTLE vs. NES and trinary classifiers include information from 42 and 30 ROIs, respectively. To better understand the benefit of mRMR, this list can be directly compared to the list of ROIs ranked by t-statistics in Table A1 in Appendix. 191

10.1 Summary of the most prevalent features in each diagnostic group, prior to multiple imputation. Abbreviations: Temporal Lobe Epilepsy (TLE), Other Focal Epilepsy (OFE), Generalized-onset epilepsy (Gen), Unspecified-onset Epileptic Seizures (UES), Non-Epileptic Seizures (NES), Magnetic Resonance Imaging (MRI), fluoro-deoxyglucose Positron Emission Tomography (PET). 208

Acknowledgments

This work was supported by the University of California, Los Angeles (UCLA)-California Institute of Technology Medical Scientist Training Program (NIH T32 GM08042), the UCLA Systems and Integrative Biology Training Grant (NIH T32 GM008185), the UCLA Neuroimaging Training Program (NIH T90 DA022768 & R90 DA023422), the National Institutes of Drug and Alcohol Abuse (NIH R33 DA 026109 to MSC), the UCLA Department of Biomathematics, Psychiatry & Biobehavioral Sciences, and Mathematics, the UCLA Graduate Division's Dissertation Year Fellowship, the UCLA Brain Research Institute's Travel Award, the Carol M. Newton Fund, research grants to JE (NS03310 & NS080181), the UCLA Applied Math Research Experience for Undergraduates (NSF DMS-1035536 to AB), the Hearst Foundation (to GEO), and the William M Keck Foundation.

This graduate thesis includes reproductions of published material with multiple contributors. The relative contributions and acknowledgements of co-authors is listed at the beginning of each relevant chapter.

Like all modern research, this work could not have been possible without the support of many collaborators and research assistants. First of all, we would like to thank our patients for consenting for the use of their records in research, and for coming to UCLA for their care.

This work was heavily based upon the good records and clinical care provided by the UCLA Seizure Disorder Center. In particular, we thank Drs. John Stern and Jerome Engel, Jr. for their indispensible guidance in the clinical treatment of seizures, their expert knowledge of the key diagnostic modalities, their help with experimental design and framing our work for a clinical audience. We thank Dr. Noriko Salamon for her help with experimental design and understanding of the neuroimaging data. Our ability to get high quality information from such a large patient sample was due to the high quality databasing and record keeping of the head EEG technologist, Kirk Shattuck. We also thank Dr. Patricia Walshaw for sharing her knowledge of the psychogenic seizure population, and helping determine key factors in the clinical work. Lastly, we also thank the epilepsy and neurophysiology fellows in the UCLA Epilepsy Monitoring Unit for teaching Wesley about the practice of epilepsy, and how these methods and data would be used. These fellows include, but are not limited to, Eileen Reed, Chrystal Ried, Shennan Weiss, Brian Moseley, and Raphael Lopez.

We also thank the group within the Ahmanson Translational Imaging Center headed by Dr. Daniel H.S. Silverman. Specifically, we thank Cheri Geist for organizational support for the Fontiers in Neurology manuscript[7]. Stefan Nguyen, Regina Ahn and Navya Reddy organized and processed the FDG-PET images through the NeuroQ software. That paper could not have been completed so quickly and efficiently without their assistance.

The members of the Laboratory of Integrative Technology, led by Mark S. Cohen, also provided invaluable support through discussions and other activities. Ariana Anderson provided statistical guidance throughout this work, and appeared on many associated publications. Pamela Douglas helped design experiments and think through many of the feature selection and machine learning methods involved in this work. Hongjing Xia helped identify mRMR as a good feature selection method, and assisted with the understanding of scalp EEG data. Edward Lau and Andrew Cho were critical to the parallelization of the computationally intensive portions of this work, and provided manuscript drafting and computational support throughout this work.

In addition to the people that provided support for research, we thank the volunteers that assisted with the transfer and organization of the EEG records, whose names are listed on the title pages. Without their hours of work, we could not have built such a large and comprehensive database of scalp EEG records from patients with seizure disorder.

We also thank the other researchers in the Center for Health Sciences B8-169 suite of offices from the Carrie Bearden, Jamie Feusner and Susan Bookheimer labs, including Tessa Harrison, Rachel Jonas, Teena Moody, Laurel Martin-Harris, Jesse Brown, Kevin Japardi, Steffie Thomson, Kevin Terashima, Brian Renner, Caroline Montojo, Carolyn Chow, Leila Kushan and Nicole Enrique. Each of them provided informal guidance, mental and psychological support throughout this work.

While these researchers and research assistants provided invaluable support, we cannot overlook the support provided by loved ones including Wesley's wife, Jane Kim Kerr; and his parents, Sharon and Brad Kerr. Even though they did not understand the math or statistics involved, they had continual and unwavering confidence in our ability to find solutions and work through the difficult portions, even if the solutions we found were not those that we thought that we would find.

Vita

2005–2006 Research Assistant with Charles Abrams & Tami Bach. Department of Hematology & Oncology, University of Pennsylvania.

2006–2007 Research Assistant with Seema Bhatnagar. Department of Psychology, Children's Hospital of Philadephia.

2007 Research Assistant with Jonathan Huppert. Department of Psychology, University of Pennsylvania

2008 Research Assistant with Michael Klein. Department of Chemistry, University of Pennsylvania.

2008 Nurse's Assistant. Department of Neurosurgery, University of Pennsylvania.

2007–2009 Research Assistant with Geoffrey K. Aguirre. Department of Neurology, University of Pennsylvania.

2009 B.A. (Biological Basis of Behavior & Biological Mathematics), University of Pennsylvania.

2009–Present M.D.-Ph.D. Student, UCLA-California Institute of Technology Medical Scientist Training Program.

2009 Rotating Graduate Student with Mark S Cohen, Psychiatry & Biobehavioral Sciences Department, UCLA.

2010 Rotating Graduate Student with Kenneth Lange, Biomathematics Department, UCLA.

2010–2015 Graduate Student with Mark S Cohen, Biomathematics Department, UCLA.

2012 M.S. (Biomathematics), UCLA.

2013 Teaching Assistant, Mathematics Department, UCLA. Taught one section of Math 3C (Probability for Life Sciences) under direction of Professor David Weisbart.

2014 Teaching Preceptorship, Biomathematics Department, UCLA. Taught as part of BiomathM230 (Computed Tomography: Theory & Applications) under direction of Professor Henry Huang.

2015 Research Assistant with Ariana Anderson, Psychiatry & Biobehavioral Sciences Department, UCLA.

2015 Research Assistant with Alex Bui & Cyrus Raji, Radiology Department, UCLA.

Publications and Presentations

Peer-Reviewed Publications

1. **Kerr WT**, Hwang ES, Raman KR, Barritt SE, Patel AB, Le JM, Hori JM, Davis EC, Braesch CT, Janio EA, Lau EP, Cho AY, Anderson A, Silverman DHS, Salamon N, Engel J, Jr., Stern JM, Cohen MS. Multimodal diagnosis of epilepsy using conditional dependence and multiple imputation. in: 4th International Workshop Pattern Recognition in Neuroimaging. (Tuebingen, Germany: Conference Publishing Services). 2014.

2. **Kerr WT**, Douglas PK, Anderson A, Cohen MS. The utility of data-driven feature selection: Re: Chu et al. NeuroImage 84:1107-1110, 2014.

3. Anderson A, Douglas PK, **Kerr WT**, Haynes VS, Yuille AL, Xie J, Wu YN, Brown JA, Cohen MS. Non-negative matrix factorization of multimodal MRI, fMRI and phenotypic data reveals differential changes in default mode subnetworks in ADHD. NeuroImage [epub ahead of print], 2013.

4. Douglas PK, Pisani M, Reid R, Head A, Lau E,Mirakhor E, Bramen J, Gordon B, Anderson A, **Kerr WT**, Cheong C, Cohen MS. Method for simultaneous fMRI/EEG data collection during a focused attention suggestion for differential thermal sensation. J. Vis. Exp. (), e3298, doi:10.3791/3298, 2013.

5. LaMoyne R, **Kerr WT**, Zanjani K, Mastroianni T. Implementation of an iPod wireless accelerometer application using machine learning to classify dispartity of hemiplegic and healthy patellar tendon reflex pair. JMIHI. (accepted 2013).

6. **Kerr WT**, Nguyen ST, Cho AY, Lau EP, Silverman DH, Douglas PK, Reddy NM, Anderson A, Bramen J, Salamon N, Stern JM, Cohen MS. Computer aided diagnosis and localization of lateralized temporal lobe epilepsy using interictal FDG-PET. Frontiers in Neurology. 4:31. 2013.

7. **Kerr WT**, Cho AY, Anderson A, Douglas PK, Nguyen SY, Reddy NM, Lau EP, Hwang E, Raman K, Trefler A, Silverman DH, Cohen MS. Balancing clinical and pathologic relevance in the machine learning diagnosis of epilepsy. in: 3rd International Workshop Pattern Recognition in Neuroimaging. (Philadelphia: Conference Publishing Services). 2013.

8. Douglas PK, Lau EP, Anderson A, **Kerr WT**, Head A, Wollner MA, Moyer D, Durnhofer M, Li W, Bramen J, Cohen MS. Single trial decoding of belief

decision making from EEG and fMRI data using ICA features. Frontiers in Human Neuroscience. (in press 2013).

9. **Kerr WT**, Anderson A, Lau EP, Cho AY, Xia H, Bramen J, Douglas PK, Braun ES, Stern JM, Cohen MS. Automated diagnosis of epilepsy using EEG power spectrum. Epilepsia. 53(11):e189-e192. 2012.

10. **Kerr WT** & Lau EP. Poisson noise obscures hypometabolic lesions in PET. Yale J Biology & Medicine. 85:541-549. 2012.

11. **Kerr WT**, Lau EP, Owens GE, Trefler A. The future of medical diagnostics: large digitized databases. Yale J Biology & Medicine. 85:363-377. 2012.

12. **Kerr WT**, Anderson A, Xia H, Braun ES, Lau EP, Cho AY, Cohen MS. Parameter selection in mutual information-based feature selection in automated diagnosis of multiple epilepsies using scalp EEG. in: 2nd International Workshop Pattern Recognition in Neuroimaging. (London: Conference Publishing Services). 2012.

13. Drucker DM, **Kerr WT**, Aguirre GK. Distinguishing conjoint and independent neural tuning for stimulus features with fMRI adaptation. J Neurophysiol. 101(6):3310-24. 2009.

14. Bach TL, **Kerr WT**, Wang Y, Bauman EM, Kine P, Whiteman EL, Morgan RS, Williamson EK, Ostap EM, Burkhardt JK, Koretzky GA, Birnbaum MJ, Abrams CS. PI3K regulates pleckstrin-2 in T-cell cytoskeletal reorganization. Blood. 109(3):1147-55, 2007.

15. Grissom N, **Kerr W**, Bhatnagar S. Struggling behavior during restraint is regulated by stress experience. Behav Brain Res. 191(2):219-226. 2008.

16. Bach TL, Chen QM, **Kerr WT**, Wang Y, Lian L, Choi JK, WU D, Kazanietz MG, Koretzky GA, Zigmond S, Abrams CS. Phospholipase cbeta is

critical for T cell chemotaxis. J Immunol. 178(4):2223-7, 2007.

Presentations

1. Aguirre G, **Kerr W**, Drucker D. Cortical representation of texture and scale studied with fMRI. J Vision. 9(8): 899-899. 2009.

2. **Kerr WT**, Anderson A, Lau EP, Cho AY, Xia H, Bramen J, Douglas PK, Braun ES, Stern JM, Cohen MS. Automated diagnosis of epilepsy using EEG power spectrum. Society for Neuroscience. New Orleans, USA. October 2012.

3. **Kerr WT**, Cohen MS. Pattern analysis in the diagnosis of epilepsy. American Society of Neuroradiology Symposium. San Diego, USA. May 2013.

4. **Kerr WT**, Trefler A, Raman KR, Hwang ES, Salamon N, Cohen MS. Predicting when epilepsy is observable using MRI and FDG-PET. Society for Neuroscience. San Deigo, USA. November 2013.

5. **Kerr WT**, Trefler A, Raman KR, Hwang ES, Salamon N, Cohen MS. Predicting when MRI and FDG-PET will exhibit epilepsy-related findings. American Epilepsy Society. Washington, District of Columbia, USA. December 2013.

6. Patel AB, Barritt SE, **Kerr WT**, Torres-Barba D, Trefler A, Raman KR, Hwang ES, Le JM, Hori JM, Davis EC, Braesch CT, Janio EA, Stern JM, Salamon N, Cohen MS. Computer-aided diagnosis of epilepsy using clinical information. Center for Undergraduate Research Posters on the Hill. Washington, District of Columbia, USA. March 2014.

7. **Kerr WT**, Cho AY, Nguyen ST, Reddy NM, Silverman DHS, Salamon N, Stern JM, Cohen MS. Interictal metabolic alterations in patients with psychogenic non-epileptic seizures. Organization for Human Brain Mapping. Hamburg, Germany. 2014.

8. **Kerr WT**, Janio EA, Breasch CT, Hori JM, Le JM, Raman KR, Patel AB, Barritt SE, Hwang ES, Davis EC, Torres-Barba D, Salamon N, Engel J, Jr., Stern JM, Cohen MS. Differentiating epileptic from non-epileptic seizures through patterns of comorbidities and pharmacologic management. American Epilepsy Society. Seattle, Washingon, USA. December 2014.

Awards & Honors

1. International Workshop in Pattern Recognition in Neuroimaging Best Student Paper 2012

2. UCLA Brain Research Institute & Semel Institute for Neuroscience & Human Behavior 2012 Graduate Student Travel Award in Recognition of Excellence in Neuroscience Research

3. UCLA University Fellowship 2011-2012 & 2012-2013

4. Beverly Hills Iranian-American Doctors (BHIAD) Research Merit Award 2011

5. Southern California Lambda Medical Association Hagan-Schedier Medical Student Grant Award 20110 & 2011

6. International Workshop in Pattern Recognition in Neuroimaging Best Student Paper 2013

CHAPTER 1

Introduction

1.1 Public Health, Diagnostic and Treatment Challenges in Seizures

We aim to address important public health problems in the long process of diagnosing and treating patients with seizures. Ten percent of the US population will experience at least one isolated seizure in their lifetime [8, 9]. Every year, one hundred thousand Americans are diagnosed with epilepsy, defined by a chronic predisposition for seizures [10, 11]. Anti-seizure medications (ASMs) effectively treat two-thirds of these patients [12, 13, 14, 15, 16]. Failing an ASM is defined by a less than 50% reduction in seizure frequency. After failing two or more ASMs, the probability of seizure freedom on medical treatment is low [15]. After failing appropriate doses of at least two ASMs appropriate to their seizures, patients are triaged to tertiary care centers for epilepsy to identify if there are other medical, surgical or technological treatments for their seizures [17]. Even if they have not failed two ASMs, patients with atypical seizures also may be referred for differential diagnosis of their episodes so that the most effective treatment can be identified [18, 17].

To effectively characterize and localize the seizure-onset zone, patients may be admitted for simultaneous video-electroencephalography (vEEG) [19, 20, 21]. Neuroimaging including structural and diffusion magnetic resonance imaging (MRI) and deoxyflouroglucose positron emission tomography (PET) can supplement this diagnostic process [17, 22, 23]. Patients with a single seizure focus may be eligible for surgical treatment that results in seizure reduction in two thirds of patients, depending on the etiology of their seizures [24, 25]. Surgery is more effective when a structural and/or metabolic lesion is visible using neuroimaging, and is consistent with patients' semiology and ictal EEG [22, 23, 24]. In particular, it can be challenging to discriminate bilateral from unilateral temporal lobe epilepsy [26, 27, 28]. Due to the functional role of the mesial temporal lobe in memory formation, bilateral resection of the temporal lobe results in irreversible deficits in new memory formation [29, 30, 31]. Therefore, only patients with unilateral temporal lobe epilepsy are surgical candidates. At UCLA, we have shown that resective surgery is more effective earlier in the patient's disease course in both adults and children [32, 33, 34]. Unfortunately, the average time to resective surgery is 18.8 years [35].

Of patients admitted for vEEG, one-third experience non-epileptic seizures (NES) [8]. Throughout this manuscript, we will refer to these events as seizures even though they are not caused by abnormally synchronous, epileptic neural activity [18, 36]. This terminology respects the patients' experience of these events as seizures. We find the alternate terminology of "events" or "attacks" unnecessarily vague or suggestive of an external source. These terms also do not reflect the difficulty in discriminating between NES and ES without vEEG.

NES can be split into two subtypes: psychogenic and physiologic [18]. Physiologic NES are seizure-like signs and symptoms caused by organic dysfunction including, but not limited to, complex migraines, syncope, transient ischemic attacks, polypharmacy, confusion episodes in dementia [18]. Psychogenic NES are understood as a conversion disorder, in which patients translate psychological challenges into physical symptoms [18]. At our center, 90% of NES are psychogenic [37]. Both subtypes of NES are not due to epileptic neural activity, therefore the primary mechanism of action of ASMs will not treat the seizures. Instead, one should focus on treating the underlying psychiatric challenges through cognitive behavioral therapy and/or medications [38]. Similar to the surgical patients, the long-term seizure outcome for patients with psychogenic NES is better if the disorder is diagnosed earlier [39, 40]. However, this long and complex diagnostic process makes a quick diagnosis and triage challenging, at best. The average time between first seizure and the diagnosis of NES is 9.2 years [41, 42].

These unfortunate statistics have a large impact on the cost of care and quality of life for patients with seizures. Because patients with seizures are treated, most frequently, as if they have epilepsy until proven otherwise, the cost of care for all medication resistant patients is similar. Without resective surgery, the lifetime cost of medication resistant seizures is US$100,000. In the case of NES, cost could be reduced drastically because these patients clearly would not need ASMs, but they do require targeted treatment of the cause of their seizures [18, 38, 43]. For patients with focal seizures, surgery for epilepsy is effective treatment and cost effective, due to decreasing the need for ASMs and reducing, if not eliminating, seizures [44, 32, 45, 46]. For patients with ES that are not surgical candidates, different medications tend to be effective for focal versus generalized onset seizures, and certain medications are effective for particular epilepsy syndromes [17]. Due to lost ability to work and other factors, the annual economic cost of epilepsy to the US is $34 billion [47, 48, 49, 50]. In addition to these economic factors, recent research has shown a decreased quality of life for patients with seizures [51, 52], and in particular patients with psychogenic seizures [53, 54], using almost any measure of quality of life. By more effectively and efficiently identifying the cause of the seizures, we can target therapies better. By understanding and identifying the common comorbidities in each subtype of seizures, we also can target resources and interventions towards treating and managing the non-seizure factors that contribute to these quality of life statistics.

1.2 Diagnostic Modalities Utilized in Seizures

Numerous diagnostic modalities are used to measure the clinical, structural, metabolic and electrographic findings that are associated with seizures, to characterize subtypes and to determine surgical candidacy. We will review these diagnostic modalities in the context of diagnosing and treating a patient with medication resistant seizure disorder.

After the patient experiences their first seizure, most patients seek medical attention through an emergency department or outpatient facility. If they don't seek attention after their first seizure, patients certainly seek help after their second unprovoked seizure. During this encounter, a physician assesses the history of seizures, medical history, psychiatric history, social history, family history and conducts a physical and, potentially, a neurological exam. All of these data help the physician and patient determine the cause and treatment for the seizures [17]. These data typically are recorded in free form text written by the physician. These notes are made for the purpose of communication with other health care providers and, frequently, insurance companies. However, patient-physician teams are not known to create reliable, reproducible data: different physicians ask and record different data from the same patient, and the same patient will answer in different ways depending on how the question is asked and their relationship with the physician [55]. This results in missing data and inconsistent records. This complicates the modeling of these data. These inconsistencies will be present when implementing or applying the knowledge gleaned from modeling.

The main challenge in utilizing clinical data is the extraction of meaningful data from the free text. This free text can be coded as binary presence or absence, categorical data, counts, or continuous quantities. The complexity of the data is too large for individual readers to code every detail in the note. Instead, we use our knowledge from treating these patients, in combination with past literature, to identify factors with a higher prior probability that they will help make a meaningful distinction between populations of patients.

If the physician determines that more data is needed or could be helpful in treating the patient, they will order a routine outpatient scalp EEG [17]. This outpatient EEG hopes to capture epileptic activity during the 20-minute recording, either in the form of a seizure or interictal epileptiform discharges. This occurs in 50% of patients on their first recordings, and is highly sensitive [56, 57]. When these findings are not present, the EEG is inconclusive. After three or more outpatient EEGs, 90% of patients with ES have exhibited epileptic activity. The goal of EEGs is to identify the seizure focus, if it exists. Different medication is effective for different seizure foci and for generalized versus focal seizures.

These assessments allow for identification of effective treatment in two thirds of patients with seizures [12, 13, 14, 15, 16]. A patient is considered to have failed a medication if they are not seizure free on a therapeutic dose or they experienced detrimental side effects while on the medication. Patients that fail two or more appropriately chosen ASMs are, by definition, medication resistant. Medication resistant patients should be referred to tertiary care centers for epilepsy so that the appropriate treatment plan can be made [17].

Tertiary care centers frequently utilize more extensive monitoring and more modern technology. These modern technologies include x-ray computed tomography (CT), MRI, PET and magnetoencephalography (MEG) [17]. These modalities provide a unique view into the pathologic process. CT is sensitive to skull fractures and acute bleeds, but is insensitive to the soft tissue changes that are expected in seizure disorder. MRI can provide contrast between grey and white matter, and thereby help visualize epileptogenic malformations of cortical development including but not limited to focal cortical dysplasia, heterotopias, polymicrogyria, and hippocampal sclerosis [58]. Diffusion tract imaging (DTI), a subtype of MRI, can provide an even more detailed assessment of white matter tracts [27]. FDG-PET provides a complementary picture by focusing on measuring glucose metabolism in cortical areas [22, 23]. Between seizures, a focus of hypometabolism can indicate the seizure onset zone [59, 60]. During a seizure, a focus of hypermetabolism can indicate regions involved in the seizure network, through FDG-PET or SPECT imaging. The challenge to all of these imaging modalities is aligning the complex cortical structure across patients so that intensity values can be compared. Additionally, each patient's seizure network is different subtly, which makes it difficult to make generalizable comparisons across patients, even if the patient's general seizure onset zone matches.

The gold standard diagnostic method for seizure disorder is a long-term vEEG [10, 18, 61, 36]. The limitation in the above modalities is that most of them rely on observing secondary aspects of the seizure network. VEEG has the benefit of visualizing the behavior and the electrophysiologic signs of the seizures simultaneously. This gives an educated observer a unique ability to determine if the behavior can be explained by the neural signal recorded by the EEG. When psychogenic seizures are part of the differential, this direct pairing of behavior with neural activity allows for an experienced observer to diagnose the patient definitively. This pairing also is useful to distinguish between potential seizure onset zones. If a single resectable focus can be identified that does not also hold critical functionality for normal function (i.e. language), then the patient could benefit from surgery.

If the patient is a potential candidate for surgery, additional more invasive diagnostic modalities are available, like intracranial grids, depth electrodes and intraoperative electrocorticography. These modalities help differentiate between similar seizure onset zones for the sake of surgical planning, especially when previous modalities disagreed or conflicted. Because our focus is on developing diagnostics, these invasive methods are outside the scope of this work.

1.3 Significance & Impact of Developing Computer-Aided Diagnostic Tools (CADTs) for Seizures

The goal of computer aided diagnostic tools (CADTs) is to supplement, not replace, clinical expertise and reasoning. If designed as such, CADTs can evaluate data differently than human experts [62]. Humans are exceptional at detecting complex trends in small numbers of features with a strong signal to noise ratio. Computers are exceptional at detecting subtle, noisy trends across hundreds or thousands of features. Our goal for CADTs is to translate complex data best read by an automated algorithm into simple outputs that expert human observers can integrate into their clinical decision making process. In this section, we discuss the many junctures at which CADTs could make an impact in clinical care of seizures.

The earliest point for intervention is during or after the first, or subsequent, outpatient assessment for seizures. Conventional assessment of seizures relies on a description of the seizure events. These descriptions, combined with the knowledge of where functions are localized in the brain, can help form an idea about where the seizures come from. Unfortunately, witnesses and patients are notoriously unreliable in the description of the seizures and neural networks are complex [63]. The specific networks involved in the seizure, and the propagation patterns frequently vary across patients, even if the seizure-onset zone is the same. Therefore, while these descriptions provide some evidence to localize the seizures, more data is needed to provide a complete and definitive assessment of the seizures.

CADTs can help at this stage by quantifying the relative value of each reported factor and integrating multiple historical factors into a single, objective likelihood score for NES. When the algorithm indicates NES, this can be used to triage patients at risk for NES towards tertiary care centers to rule out ES and positively diagnose NES. The objectivity of this score allows the clinician to maintain their therapeutic relationship with the patient because they do not need to challenge the patient's idea that their seizures are due to epilepsy. As well as we try to explain what the cause of NES is, due to the stigma around mental disorders means that there is a potential for the patient to feel betrayed when the clinician brings up the possibility of NES [64, 65]. By using an objective score to indicate NES, the patient-physician alliance can be maintained.

Note that we propose that a CADT-based score could be used to triage towards tertiary care centers, instead of used to diagnose the patient. This is based on the knowledge that the specificity of any CADT will be less than 100%, just as the specificity of the clinical assessment is imperfect because of the limitation of the quality of outpatient interview data. Therefore, it is necessary to refer these patients to tertiary care for a more detailed assessment of their seizures by an epilepsy specialist, potentially including EEG and other imaging modalities.

If and when scalp EEG is deemed helpful or necessary to help diagnose and localize the seizures, the interpretation of these data relies on the observation of interictal epileptiform discharges or overt seizure activity during the recording, as discussed above [57]. Seizure and spike detection protocols, ultimately, aim to replace neurologists by being able to identify the activity that neurologists use to understand the seizure onset and propagation [66, 67, 68, 69, 70]. Seizure prediction uses similar methods to seizure and spike detection to give patients advance warning that their seizures will occur soon [71]. Recent developments in seizure prediction utilize intracranial EEG to achieve sufficient performance to be applicable to patients. Patients that require intracranial monitoring have already had their seizure onset zone localized enough to allow for placement of recording electrodes. Seizure detection, prediction and intracranial monitoring address important issues in the management of established ES, instead of in the diagnosis of ES.

In addition to the extensive work that has been done in automated seizure and spike detection, as well as seizure prediction, CADTs have the potential to identify diagnostic information in the resting state, interictal periods [72, 73, 8, 74]. This addresses a fundamental limitation of scalp EEG, for a scalp EEG recording to be diagnostic; some interpretable activity must be present. If we can identify reliable resting state EEG changes, then we could improve the diagnostic yield of every scalp EEG assessment ordered. This could have an impact similar to the CADT based on the clinical information: we can triage patients at risk for NES towards tertiary care faster, in addition to improving the localization and characterization of epileptic seizures to facilitate medical and surgical management.

If the patient fails to respond to an ASM, an MRI is indicated to further characterize the seizures [17, 75]. MRIs have the unique ability to identify tumors, cortical dysplasia, heterotopias, other cortical and subcortical malformations, as well as findings thought to be secondary to repeated uncontrolled seizures, like hippocampal sclerosis. These findings are apparent to an experienced radiologist. However, quantitative morphometry has shown repeatedly that there exist subtle changes that are not appreciated by visual inspection [76, 77, 26, 28, 78]. Focal cortical thinning or thickening, sometimes accompanied by increase in signal intensity, can indicate the seizure-onset zone. These findings help identify the epileptogenic zone: the area where, if resected, the patient would be seizure free. However, there exist radiologic changes outside the seizure-onset zone that were not appreciated until quantitative morphometric methods were developed [26, 79]. Unfortunately, patients that have more pathologic changes outside of their primary seizure focus have a worse prognostic outcome after resective surgery for epilepsy.

CADTs have the potential to leverage these findings into an objective predictive score for epilepsy in general, or even discriminate between subtypes of epilepsy. A number of CADTs have been developed to identify the epileptogenic zone, and have shown that a more complete resection of the identified area led to improved seizure control [80]. While these methods seem promising, none has been validated enough to begin a randomized control trial. This may be because the current method of determining the site of resection is through a multispecialty discussion of all the available evidence, instead of relying on a single information modality to define the resected area. Alternatively, CADTs can aim to assist in this diagnostic process by helping to identify the general subtype of seizures, either epileptic or non-epileptic. The CADTs we developed below specifically address the question of lateralization. Differentiation between left and right temporal lobe epilepsy (TLE) is critical for pre-surgical planning because patients with bilateral TLE are not candidates for resective surgery due to irreversible memory loss, as we learned with patient HM [29, 7, 81, 31]. Patients with unilateral TLE are surgical candidates because of the capability for a single hippocampus to hold and generate memories. The recent advent of responsive neurostimulation (RNS) has given hope to patients with bilateral TLE [82], but these types of decisions later in the diagnostic process are outside the scope of this work. CADTs that aim to diagnose, as compared to identifying the region to resect, contribute to the diagnostic process instead of aim to replace it. Therefore, there may be more support for the integration of the latter CADTs into the clinical diagnostic and pre-surgical process.

MRIs are ordered prior to tertiary care and within tertiary care, therefore there is potential for their use in the early diagnosis of disease. However, MRIs are more expensive than scalp EEG and clinical interviews, therefore current protocols only indicate their use after failure of at least one ASM [17]. During assessment at a tertiary care center, MRI also is used to subtype epileptic seizures, even if the patient is not a surgical candidate. Therefore, development of diagnostic CADTs outside the pre-surgical process can still have an impact on clinical decision-making.

The last pivotal piece of the diagnostic and pre-surgical assessment for epilepsy is FDG-PET imaging. When clinical evidence, scalp EEG and MRI are discordant, PET imaging can help further localize the epileptogenic region [83, 84, 85]. The goals of algorithms based on PET are similar to those based on MRI [7]. Some methods seek to identify a region to resect, and others seek to assist in the diagnostic process [72, 26, 27, 73, 28, 74]. Because FDG-PET scans are such a late part of the diagnostic assessment and are only done at some tertiary care centers, the clinical impact of these tools may be reduced.

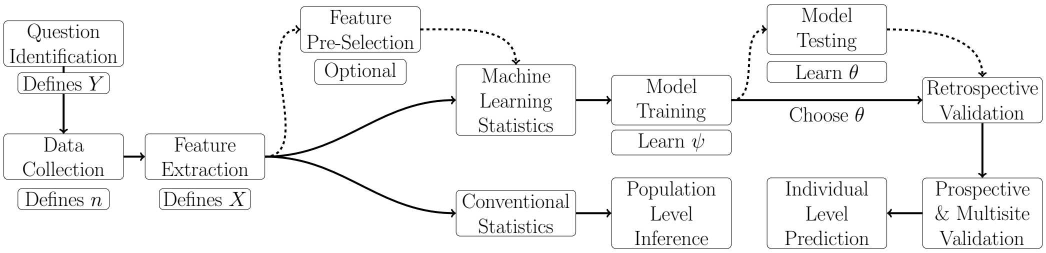
Lastly, we recognize that MEG is another useful diagnostic modality used to diagnose seizure disorder. The information gleaned from MEG is similar and complementary to EEG. At our center, it is acquired after the other modalities if more information is needed. Because of the relatively low impact of MEG in the pre-surgical assessment, we do not address the development of CADTs based on MEG in this work.

1.4 Statistical Challenges in Training & Validating Machine Learning Models

The rate limiting steps in the development of CADTs are statistical, in our opinion. Figure 1.1 illustrates the general process to train and validate machine-learning models. For our purposes, we define machine learning as statistical models to predict binary or categorical outcomes. The statistical challenges include insufficient sample size, a huge number and complexity of potentially diagnostic features, difficulties in finding the meaningful trends in this complex data through structured algorithms and inefficiencies in using data to train these algorithms.

One of the central challenges in developing these CADTs is the collection of sufficient high quality and clinically relevant data to train and validate these models. One limitation to many neuroimaging studies is that they are underpowered vastly [86, 87, 88]. Using conventional models, a model is underdetermined if the number of independent features to be studied, p, outnumbering the number of sta-

Figure 1.1: Flow chart of a statistical experiment. Dashed lines reflect optional steps.



tistically independent data points, n. Using some clever statistical optimization criteria, we can find unique and stable solutions when $n << p$ (see Didactic Background Material, Feature Selection). These criteria, however, do not obviate the curse of dimensionality (CoD): as the p increases, the average distance between the data points increases supra-exponentially in the measurement space. The only way to negate the effect of the CoD is to collect databases that increase supra-exponentially. In the modern age of electronic health records (EHRs) and novel advancements in data management and storage by giant technology companies like Google and Facebook, this may be possible.

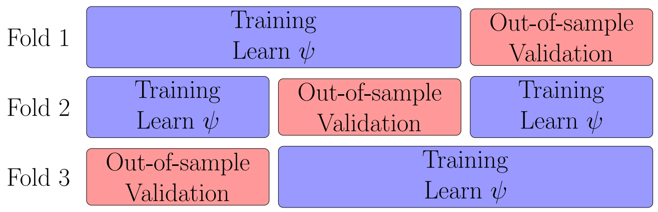
However, we must ensure that the information collected from these patients is meaningful. While it would be tempting to collect all possible data from each data point, we remember that if these features simply add noise, then we are decreasing our ability to generalize from a training data point to validation data because these data are farther apart. The best method for focusing this search is by using biological and clinical prior knowledge in combination with novel statistical methods [3]. For epilepsy, this means initially utilizing the data that is collected as part of the diagnostic and pre-surgical assessment. Clinicians have determined that these information modalities hold valuable information, so it is logical to use rigorous statistical methods to verify and build upon that knowledge.

Even if we can collect huge quantities of data, we must make assumptions about the structure of the diagnostic information in these features (see Didactic Background Material, Machine Learning Statistics). In lay language: we can only find diagnostic information where we look for it. If we assume that the diagnostic information is captured by linear changes in the input data, then we can only find linear changes. Currently, the best method for integrating interaction and non-linear information is a neural network model. These models, however, can be too flexible and thereby require huge amounts of information. If limited information is available, the neural network has a tendency to overfit the data by using its flexibility to capture trends in the training data that improve training performance, but do not translate to the validation data. Therefore, if there is known structure in the diagnostic information, imposing that structure could improve the ability for the model to generalize to validation data. Again, we only can see diagnostic information in data if we look in the appropriate way.

When limited data is available, it is important to make full utilization of these data. We define the types of machine learning algorithms based on how they estimate the optimum value of the parameters inherent to the model, which we will designate ψ. However, numerous models or protocols include hyperparameters that are not optimized jointly with ψ, which we will designate θ. Conventional methods to optimize both ψ and θ is to split the data into three groups: training for ψ, testing for θ and validation data to assess the generalizability of the learned ψ and θ. In order to minimize the variance of the estimate the generalizability of the model well, the size of the validation set needs to be maximized. However, in order to learn a good model, the size of the training and test sets also must be maximized. Limiting the size of the training set biases the generalization performance to be worse. This is called the bias-variance tradeoff (see Didactic Background Material, Machine Learning Statistics).

A popular method to reduce the effect of the bias-variance tradeoff is nested cross-validation. In cross-validation, all of the data is used in the validation set, just not at the same time (see Figure 2.10). In ten-fold cross-validation, the data is split into ten mutually exclusive subsets. Nine of the subsets are used for training and testing, whereas the last subset is used for validation. Subsequently, the identity of the validation subset is permuted such that each set is the validation subset once and only once. In nested ten-fold cross-validation, the nine subsets are split into ten mutually exclusive sub-subsets. Nine of these sub-subsets are used to train ψ for a range of θ. The choice of θ that performs the best on the last sub-subset is used without modification on the validation subset. Consequently, there are ten different sub-sub-models that are applied to the same validation subset. Pooling the result through averaging or pooling scores can estimate the overall generalization performance while maintaining the out-of-sample nature of both the testing and the validation data.

Figure 1.2: 3-Fold Cross-validation when θ is selected *a priori*, as compard to optimized. Therefore, no data is needed to learn θ. A separate model is learned for each fold, resulting in three overall models to aggregate later.



If we are mostly interested in assessing the generalization performance, we could perform a complete cross-validation within each nested fold, but this has the propensity to overfit the data. To be clear, a complete cross-validation means that instead of having sub-sub-models, we choose the single θ that maximizes performance on the nine subsets, and re-train ψ based on that value of θ. This maintains the out-of-sample nature of the validation data, and thereby maintains the validity of our estimates of generalizability. Unfortunately, this reduces the interpretability of the ψ and θ because they were "peaked." Although there are

10-fold less pairs of ψ and θ, the variance of ψ and θ across cross-validation folds can be larger. This increase in variance reduces our ability to estimate the effect of each parameter and hyperparameter within ψ and θ accurately.

In part 2 of this work, we propose and empirically validate a method to utilize more of the available data for training ψ and θ, without reducing the size of the validation set. This is accomplished by reducing the need for nested cross-validation. Our method also explicitly addresses the variability in θ across cross-validation folds and thereby improves our ability to interpret the sensitivity of our model to our choice of θ.

1.5 Use of Electronic Health Records (EHRs) in Clinical Research

One of the ways we addressed these statistical challenges was through effective utilization of electronic health records (EHRs). As described in detail in our manuscript, the recent mandate for EHRs has great potential impact on the development of computer-aided diagnostic tools (CADTs) and clinical research [89]. However, there are certain considerations that must be discussed to understand the limitations of this work.

The benefit of EHRs is that they provide large amounts of detailed patient information in the form that is used to treat patients from an unselected population. One of the limitations to databases of recruited patients is that we cannot assess if the patients that volunteer for additional research tests truly reflect the general patient population. When unselected populations are used, the diversity of the studied population exactly matches the population that is being treated at the institutions that contributed to the EHR.

Another benefit of EHRs is that the populations available for study can be orders of magnitude larger than specially recruited for a specific research project.

One of the challenges of building large databases is financing the acquisition of data. If data were acquired as part of patients' clinical treatment, then researchers do not need to spend precious research dollars on data acquisition and subject recruitment. Instead, research funds can be spent on organizing, studying the data and developing useful models of the data. Given the recent literature showing that many studies are underpowered and therefore difficult to reproduce, the increase in sample size addresses a key limitation present in the literature [86, 87, 88].

However, these volumes of data are clinical quality, not research quality. The quality of the data matches the quality used during routine treatment of patients. Research quality data may be more consistent and clean than clinical quality data, but the algorithms, tools and models developed using research quality data may or may not apply to clinical quality data. Consequentially, to advocate for CADTs developed using research quality data, one must advocate both for the data acquisition protocol the developers used and for the utility of their CADT. If a CADT has been developed using clinical quality data, then clinicians can better assess how the CADT would perform on the clinician's data.

In this work, we make judicious use of the UCLA Seizure Disorder Center's EHR including all records from patients admitted for vEEG monitoring. As part of their clinical care, each patient underwent continuous scalp or intracranial vEEG and a subset of the following diagnostic procedures: CT, structural and/or diffusion MRI, FDG-PET, MEG, and/or SPECT. The goal of monitoring was to correlate ictal behavior with electrographic changes indicative of hypersynchronous epileptic neural activity. This monitoring is the gold standard for determining if seizures are epileptic and, if they are, localizing the seizure-onset zone. UCLA admits 140 patients per year for this monitoring. Information from each modality is saved for a variable amount of time. EEG data and clinical reports are archived back to 2000, resulting in 2,100 unique patients. In comparison, other studies of ictal semiology or the risk factors for epilepsy and NES rely on data from between

15 and 100 patients. Neuroimaging records are saved back to 2006, resulting in up to 1,260 patient records. Because not all patients underwent all imaging, this resulted in 800 unique patients with diagnostic neuroimaging. In comparison, the AD neuroimaging project is spending millions of dollars to acquire research quality data from 1,000 subjects. Previous seizure studies have included no more than 150 highly selected patients. Therefore, the UCLA database has unprecedented statistical power to study seizures. However, by including all patients admitted to vEEG, the database has increased heterogeneity than these previous studies. This heterogeneity also gives us the opportunity to study more rare and less appreciated subtypes of epilepsy.

1.6 Summary of Content Herein

This graduate thesis is organized in the following manner. In chapter 2, we start with didactic background material to establish notation and the statistical perspective we take on the problem. Chapters 3 through 6 are reprints of peer-reviewed manuscripts that review and address basic challenges related to the foundation of the main work. Chapters 7 through 10 describe our published and pre-published manuscripts describing the development of single modality and multimodality CADTs for seizure disorder. In particular, we address the development of a CADT for three major information modalities in epilepsy: clinical information, scalp EEG, and FDG-PET. In chapters 11 and 12, we discuss the foundational problem of hyperparameter training and interpreation with novel statistical perspective. Subsequently, we conclude by addressing how this work fits into the greater context of the literature and dicuss the necessary follow up steps that must be taken to before CADTs are implemented in epilepsy clinics. Additionally, we discuss follow up studies regarding our random field theory based method for optimizing hyperparameters.

CHAPTER 2

Didactic Background Material

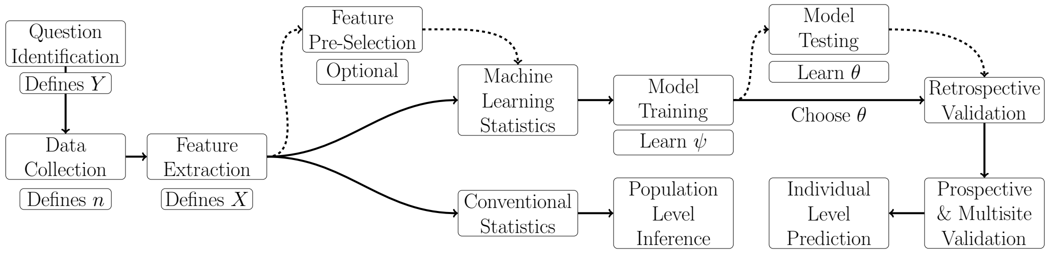
To provide a strong theoretical basis for the novel work described in this manuscript, and to define notation, the following sections provide a more didactic background in generalized linear statistical modeling, machine learning statistics, experimental design and the development of clinically applicable computer-aided diagnostic tools (CADTs). This also illustrates the perspective from which we tackle the important clinical and statistical challenges inherent to this work.

The goal of statistical modeling is to determine if there is a relationship between the input data, X, and the outcome variable of interest, Y. In both conventional and machine learning statistics, it is difficult to measure causality, but we can determine if there is some relationship. This has great power to understand predictive factors, as well as answer important questions in biology, medicine and other fields. It is our opinion that every responsible scientist should have a basic understanding of statistics. The following section aims to cover some of these basic principles, as well as some more advanced concepts. For a general overview, we illustrate the complete process necessary to study data using statistics (Figure 2.1).

2.1 Simple Linear Models

Given generic input and outcome data, one should think of the simplest relationship first; namely, is the relationship between X and Y linear? We note that

Figure 2.1: Flow chart of a statistical experiment. Dashed lines reflect optional steps.



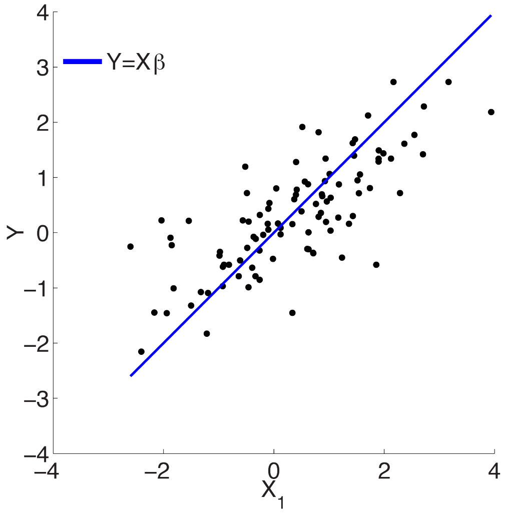
this is equivalent to asking, "Are X and Y correlated?" This is because correlation implies a linear relationship. To test this question, we consider the following relationship (Figure 2.2):

$$Y = X\beta + \epsilon \text{ such that } \epsilon \sim N(0, I_{n\times n}\sigma^2) \tag{2.1}$$

where β is a vector of linear weights of the input data, X, including an intercept term, $I_{n\times n}$ is the n by n identity matrix and ϵ is a vector of the error of the model on each exemplar. Exemplars are assumed to be independent, and identically normally distributed across exemplars with variance of σ^2. The assumption of identical distribution suggests that the distribution of each element of the error, ϵ_i, has the same mean and variance. Using this expression, we ask if X and Y have a linear relationship by finding the best β and estimating how likely this β and predictive performance would be achieved if there was no linear relationship.

In this case of assuming linearity, we define the "best" β as the β that minimizes the sum of the squared error, $\sum_{i=1}^{n} e^2$, where the English e reflects the observed error as compared to the theoretical error, ϵ. We do this because we are interested in minimizing the distance between the estimated outcomes, $\hat{Y}$, to the observed outcomes, Y. The error can be positive or negative, whereas distances are positive, therefore simply summing e is ineffective. (Actually, $\sum_{i=1}^{n} e$ is guaranteed to be zero, which will become clear below.) One could propose to

Figure 2.2: Example of simulated input data (black) and a linear regression line (blue) fit of these data. Data were sampled from a linear line with Guassian noise.



optimize $\sum_{i=1}^{n}|e|$, but the kink of the absolute value function at zero leads to more difficult (but not impossible) optimizations. In addition, for reasons that we don't show here, minimizing the least squared error is identical to the maximum likelihood estimate. This gaurantees that our estimates of β are unbiased: $E(\hat{\beta}) = \beta$. Therefore, we choose to minimize the sum of squared error.

Performing this optimization relies on the simple principles of finding a critical point in a function and checking (or assuming) that it is a minimum. We note that the only critical point in $\sum_{i=1}^{n}\epsilon_i^2$ is at the point $\epsilon = 0$, which is a minimum. A critical point is defined as a point for which the derivative of the function is zero. If the second derivative is negative at the critical point, then the critical point is a local maximum. Conversely, if the second derivative is positive, the critical point is a local minimum. In this case, we seek to find the β that minimizes:

$$\begin{aligned}
\min_{\beta} \sum_{i=1}^{n} e^2 =& \sum \left(Y - X\beta\right)^T \left(Y - X\beta\right) && (2.2)\\
=& Y^TY - Y^TX\beta - X^TY\beta + X^TX\beta^2 && (2.3)\\
& \text{Taking the derivative with respect to } \beta \\
0 =& -Y^TX - X^TY + 2X^TX\hat{\beta} && (2.4)\\
& \text{Recognizing that } Y^TX \text{ and } X^TY \text{ are scalars and} \\
& \text{are therefore equal} \\
2X^TX\hat{\beta} =& 2X^TY && (2.5)\\
\hat{\beta} =& (X^TX)^{-1}X^TY. && (2.6)
\end{aligned}$$

Note that we use linear algebra to simplify notation. One important assumption of linear modeling is that X^TX is invertible. This occurs when the row rank of X is greater than or equal to the number of columns in X. In non-math speak, this means that there are more independent samples of data than there are factors that you would like to study. (In fact, you would like the row rank of X to be much greater than the number of columns in X, for reasons that will become clear later.)

To answer our question if there is a linear relationship between X and Y, then we must test if β is significantly different from zero. First, we address if any elements of $\hat{\beta}$, $\hat{\beta}_j$ are significantly different from zero. Given our $\hat{\beta}$ is normally distributed, this relies on estimating the standard error of $\hat{\beta}$. When performed on each element of β individually, this test is called a Wald test. If we substitute $\hat{\beta}$ into the expression for $I_{n\times n}\sigma^2 = E(\epsilon\epsilon^T)$, then relatively simple linear algebra shows us that:

$$
\begin{aligned}
Var(\hat{\beta}) =& E\left[(X^TX)^{-1}X^TY\left((X^TX)^{-1}X^TY\right)^T\right] && (2.7)\\
=& E\left[(X^TX)^{-1}X^T(X\beta+\epsilon)(X\beta+\epsilon)^TX(X^TX)^{-1}\right] && (2.8)\\
& \text{Factoring out the constant portions} && \\
=& E\left[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}\right] && (2.9)\\
=& E\left[(X^TX)^{-1}X^T\sigma^2I_{n\times n}X(X^TX)^{-1}\right] && (2.10)\\
=& \sigma^2E\left[(X^TX)^{-1}X^TX(X^TX)^{-1}\right] && (2.11)\\
=& \sigma^2(X^TX)^{-1}. && (2.12)
\end{aligned}
$$

However, we must now recognize that σ^2 is *estimated* from the data, and not known *a priori*. Because this is a simple linear model, we know that the β are normally distributed. If, for some reason, we knew *a priori* the covariance of β, Σ_β, then we can use normal statistics to compare $\hat{\beta}$ to the null hypothesis that there is no relation between the input data X and the outcome variable, Y as follows:

$$\hat{\beta}_j - 0 \sim N(0, \Sigma_{\beta,jj}) \quad (2.13)$$

where the subscript, jj, reflects the j^{th} diagonal entry of the covariance matrix. If, however, we do not know the covariance then we need to estimate the covariance, $\hat{\Sigma}_\beta$, from the data. When we use an estimated covariance instead of a known covariance, we must take into account the uncertainty in this estimate. We do this by using a t distribution, as follows:

$$\frac{\hat{\beta}_j - 0}{\sqrt{\hat{\Sigma}_{\beta,jj}}} \sim t_\nu \quad (2.14)$$

where ν is the number of degrees of freedom in the data. As $\nu \to \infty$, the t distribution becomes a normal distribution. Intuitively, degrees of freedom are the number of independent data points, after estimating the parameters inherent to the model. In the case of a conventional univariate t test, ν is typically $(n-1)$

where n is the number of data points because the mean has been estimated. In a linear model studying one input factor, ν typically is $(n-2)$ because β consists of an estimated intercept and slope. In general, we express the number of estimated parameters as p, which is the length of β that is required to be less than the row rank of X. This can be proven by deriving an unbiased expression for $\hat{\sigma}^2$. We do this as follows, with a judicious use of linear algebra. To do so, we define $Var_S(e) = E(e^Te)$ as the sum of the squared deviation of the entries of e from the mean of e.

$$Var_S(e) = Var_S(Y - X\beta) = Var_S(Y - \hat{Y}) \tag{2.15}$$

Recognize that $(X^TX)^{-1}X^TY = HY = \hat{Y}$ where H is idempotent.

$$= Var_S(Y - HY) = Var_S\left[(I-H)Y\right] \tag{2.16}$$

$$= Var_S\left[(I-H)(X\beta+\epsilon)\right] \tag{2.17}$$

Factoring out the constant portion

$$= Var_S\left[(I-H)\epsilon\right] = E\left[((I-H)\epsilon)^T)(I-H)\epsilon\right] \tag{2.18}$$

$$= E\left[\epsilon^T(I-H)^T(I-H)\epsilon\right] \tag{2.19}$$

Simplifying due to symmetry and idempotency of $(I-H)$

$$= E\left[\epsilon^T(I-H)\epsilon\right] \tag{2.20}$$

The expression within the expectation is a scalar, so we

can apply the trace operator:

$$= E\left[tr\left(\epsilon^T(I-H)\epsilon\right)\right] \tag{2.21}$$

The trace operator is invariant to cyclic permutation of the arguments.

$$= E\left[tr\left((I-H)\epsilon\epsilon^T\right)\right] \tag{2.22}$$

Reversing the order of the operators,

$$= tr\left[E\left((I-H)\epsilon\epsilon^T\right)\right] \tag{2.23}$$

Using that $I-H$ is constant and $E(\epsilon\epsilon^T)$ is scalar

$$= tr\left[I-H\right]E(\epsilon\epsilon^T) = tr\left[I-H\right]Var(\epsilon) \tag{2.24}$$

By assumption of the linear model

$$=\sigma^2 tr\left[I-H\right]=\sigma^2\left[tr(I)-tr(H)\right]=\sigma^2(n-p) \tag{2.25}$$

Rearranging this expression

$$\sigma^2 =\frac{SSE(e)}{n-p}. \tag{2.26}$$

Variance typically is calculated as the squared deviation of the data from the mean, over the number of independent data points. Overall, this shows that even though the number of data points is n, the number of independent data points is $n-p$ after estimating β.

Returning to our original question, we want to ask if any individual $\hat{\beta}_j$ is significantly different from zero. We define something as statistically significant if the probability of our observed result, p, is less than a particular cutoff, α. This α is the false positive rate: the probability we would conclude that an effect exists even if it does not. Conventionally, α is set to 5%. Mathematically, we write this as:

$$P\left(|t_\nu|>\left|\frac{\hat{\beta}-0}{\left(\hat{\Sigma}_{\beta,jj}\right)^{1/2}}\right|\middle|\nu=n-p\right)\overset{?}{<}0.05 \tag{2.27}$$

In addition to asking if individual $\hat{\beta}_j$, we also can ask how significant the aggregate model is from chance. To do this, we take a slightly different, but equivalent, perspective than above. We define the sum of squared error (SSE) as the quantity we minimized before, $\sum e^2$. In addition, we define the variance accounted for by the model, abbreviated MSE, as $MSE = Var_S(Y) - SSE$. We then consider the ratio of variance accounted for by the model to the unmodeled variance:

$$\frac{MSE}{SSE}=\hat{F}\sim F_{\nu_1,\nu_2} \tag{2.28}$$

As noted above, this ratio follows an F distribution where ν_1 is the number of parameters in the model, and ν_2 is the degrees of freedom in the error. The theoretical basis for the F distribution is based upon the following knowledge. Each ϵ is normally distributed. The sum of the square of ν independent and identically distributed normally variables is, by definition, a χ^2_ν distribution with ν degrees of freedom. The ratio of two χ^2_ν distributed variables with varying degrees of freedom is, by definition, an F distribution with ν_1 and ν_2 degrees of freedom, where ν_1 is the degrees of freedom of the numerator and ν_2 is the degrees of freedom of the denominator.

For our test of significance, we test if the probability of observing this $\hat{F}$ is less than 5%, given there is no linear relationship between X and Y. If this occurs, then we can reject the null hypothesis that there is no linear relationship. To remove this double negative statement, this means that there is a linear relationship between X and Y. We note that if $\nu_2 = 1$, then the F distribution is equivalent to the t-distribution. Additionally, we note that a subset of the model can be tested in this way, by comparing the variance accounted for by a subset of factors, compared to the SSE of the full model. We provide two examples of this.

First, consider if there are two information sources that contribute to X. Let's split X into two block matrices reflecting this: $X = [X_1, X_2]$. We can ask if there is a linear relatioship between X_1 and Y, controlling for X_2. This is done by calculating the F statistic between the portion of the sum of squared variation in Y modeled by the β corresponding to information source 1, MSE_1 to the residual error, SSE. This is done as follows:

$$Var_S(Y) = MSE_1 + MSE_2 + SSE \tag{2.29}$$

$$\frac{MSE_1}{SSE} \sim F_{\nu_1,\nu_e} \tag{2.30}$$

where ν_1 is the number of entries of β corresponding to information source 1, and $\nu_e = n - p$, where p is the length of β.

In our second example, consider that you want to know if a particular input factor, X_j, has a particular non-linear relationship with Y. Suppose that you want to ask if there is a quadratic relationship between X_j and Y. This requires estimating two separate β_j, one corresponding to the linear relationship, $\beta_{j,1}$, and the next corresponding to the quadratic relatiosnhip, $\beta_{j,2}$. Using the same structure as above, one could test if there is a significant quadratic relatisonhip between X_j and Y by testing the F statistic corresponding to the ratio of the sum of squared variation modeled by $\beta_{j,1}$ and $\beta_{j,2}$ to the residual error. Note that this is different from the hierarchical test of if a second order term in X_j produces a significant decrease in the residual error. This latter test can be done through a t-test of if $\hat{beta}_{j,2}$ is significantly different from the null hypothesis, $\beta_{j,2} = 0$.

Our software of choice for implementing these simple linear models is the lm function in R because of its ease of use, clear treatment of the error and zero cost of the software. In addition, a good number of regression diagnostics come with the lm function in R, so that one can check the assumptions of the model, including if ϵ are independent and identically normally distributed.

2.2 Generalized Linear Models

The generalization of linear models can be written succinctly in terms of math, but the challenge occurs in effectively estimating the parameters in these models. A general linear model supposes that

$$E(Y) =g^{-1}\left(X\beta\right) \tag{2.31}$$

$$Var(Y) =Var\left[g^{-1}\left(X\beta\right)\right] \tag{2.32}$$

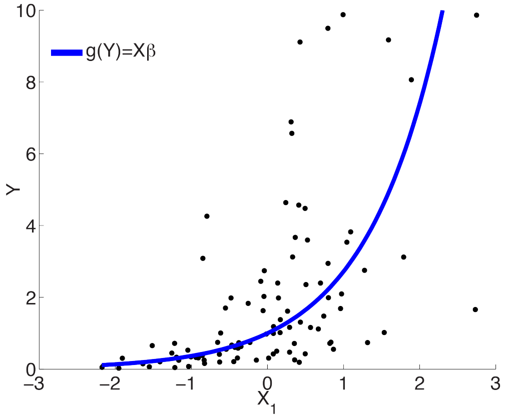
where g is any specified function or probability statement. We specify g based on prior knowledge of structure of Y. If g is a probabilistic statement, then the structure of the error is implicitly defined.

First, we will consider relatively trivial generalized linear models and when to apply them. In simple linear models, we implicitly assumed that Y is normally distributed by assuming that ϵ was normally distributed. Suppose instead that Y is log-normally distributed. For example, seizure duration appears to be exponentially distributed with most seizures being short, and some rare seizures being very long (i.e. status epilepticus). To model this, we transform Y such that the error in our model will be normally distributed variable, and see if X is correlated with that transformed value. Mathematically, this is written as (Figure 2.3):

$$E(Y) = e^{X\beta} \text{ or } \log Y = X\beta. \tag{2.33}$$

$$Var(Y) = Var(e^{X\beta}) \sim I_{n\times n}\sigma^2 \tag{2.34}$$

Figure 2.3: Example of simulated input data (black) and a log-normal regression (blue) fit of these data. Data were sampled from an exponential line with Guassian noise in the transformed, exponential space.



Because the transformed variable is normally distributed, the same optimization procedure and software that was used for simple linear models can be used in these settings. In this case, we used an exponentially distributed variable as an example. In fact, if there exists a transform $g(Y)$ such that Y is not normally distributed but $g(Y)$ is well defined and the error is normally distributed, then the simple linear model optimization tools can be used to study the relationship between $g(Y)$ and X.

Less trivial generalized linear models rely on the knowledge or assumption that there is a known distribution of Y that is not normal and, frequently, $E(Y)$ provides information about the $Var(Y)$. The most frequent generalized linear models of this type are Poisson regression for count outcome data and logistic regression for binary outcome data. Due to our focus on binary classification, we will cover Poisson regression first and in less detail.

If the outcome data, Y, is a count of independent events that occur over a given time frame, where events cannot occur simultaneously and the presence of an each event is independent, then one should use a Poisson regression. In seizure disorder, a clear example of this is in the modeling of seizure frequency, which is equivalent to modeling the count of seizures over a given time period. Two seizures clearly cannot occur simultaneously in the same patient. However, seizures may not be independent in time, especially if a patient has a tendency to have clusters of seizures. In that case, the number or frequency of seizure clusters could be a Poisson variable.

In a Poisson regression, we assume that the relationship between X and Y take the following form (Figure 2.4):

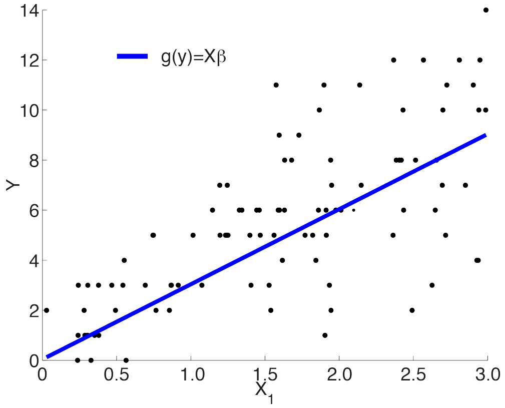
$$P(y_i = y) = \frac{\lambda^y e^{-\lambda}}{y!} \text{ such that } \lambda = X\beta. \tag{2.35}$$

To find the best β to maximize the posterior probability of the data, we seek to maximize the likelihood that the observed data would occur, given our model. This concept of maximum likelihood can be written as:

$$\max_\beta P(Y = \hat{Y}|X, \beta) = \max_\beta \prod_{i=1}^{n} P(y = \hat{y}_i|X, \beta) \tag{2.36}$$

This product can be done because we assume each exemplar is independent, and the probability associated with independent variables multiplies. Next, we recognize that differentiation of multiplied variables is messy, compared to differentiation of summed variables, therefore we maximize the log-likelihood of the

Figure 2.4: Example of simulated input data (black) and a Poisson regression (blue) fit of these data. Data were sampled from a Poisson line with Guassian noise in the transformed, Poisson space.



observed data, given the model, resulting in the following expression:

$$\max_{\beta} \sum_{i=1}^{n} \log P(y = \hat{y}_i | X, \beta) = \max_{\beta} \sum_{i=1}^{m} -\log(y!) + y_i \log(X\beta) + X\beta \qquad (2.37)$$

Looking at this expression of the log-likelihood, it is apparent that differentiation with respect to β will not lead to a simple analytical formula for β. Therefore, we must use other iterative or approximating optimization methods to estimate $\hat{\beta}$. Although not always used, clever optimization schemes could utilize the knowledge that the variance of a Poisson variable is equal to its expectation. Therefore, the standard error of $\hat{\beta}$, estimated from the error in prediction, can provide further information to better estimate $\hat{\beta}$.

In practice, this assumption of equal expectation and variance may not be accurate. This can occur if any other source of variation is present than simply observing the Poisson variable. Common sources of additional variance include measurement noise, missing counts, and approximating counts to the nearest 5th, 10th, or 100th integer. In this case, one can use an 'over-dispersed' Poisson model or a more complex model outside the scope of this background material [90, 91, 92].

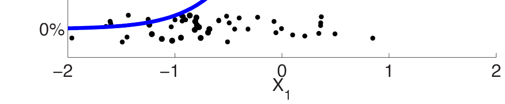
The most common form of over-dispersed Poisson is a negative binomial model, where the shape of the distribution matches the Poisson, but the variance assumptions are less strong. The challenge to utilizing a negative binomial model is that one must choose *a priori* the number of failed trials allowed (see Random Field Theory section). Despite these assumptions and challenges, utilization of Poisson or negative binomial regression allows for accurate and statistically rigorous modeling of count data.

Logistic regression is prevalent in modeling of binary outcome data because it models outcomes as Bernoulli random variables. The key insight of logistic regression is to translate binary data into a continuous predicted outcome so that conventional statistics can be used. This transformation of the simple linear model is accomplished through the logit transform (Figure 2.5):

$$logit(Y) = X\beta \text{ or, equivalently, } \pi = \frac{e^{X\beta}}{1+e^{X\beta}} \tag{2.38}$$

where π is the vector of the probabilities, π_i, such that $P(y_i = 1|\beta, X) = \pi_i$.

Figure 2.5: Example of simulated input data (black) and a logistic regression (blue) fit of these data. Data were sampled from a logistic trend used to define the probability of success in a Bernoulli trial. Uniform noise was added to the Y dimension of the logistic input data to aid in visualization, but this noise was not included in modeling.



When multiple trials are aggregated, Bernoulli random variables become binomial random variables. Therefore, the likelihood function for logistic regression is:

$$L(Y|X,\beta) \propto \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{2.39}$$

This is a binomial distribution, without the leading combinatorial term, which is why we used the "proportional to" notation. We can omit the combinational term without loss of rigor because the derivative of this term, with respect to β is zero. To differentiate the likelihood more easily, we optimize this using the log of the likelihood:

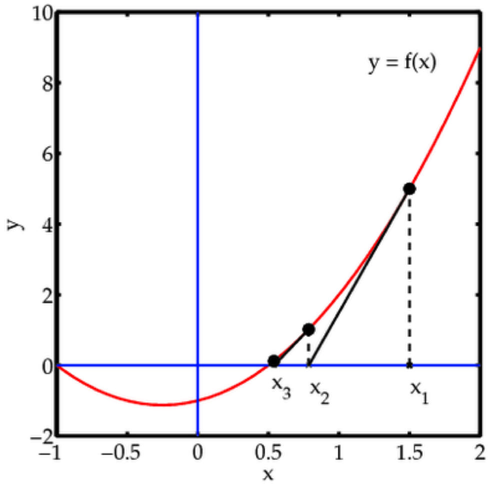
$$\ell(Y|X,\beta) \propto \sum_{i=1}^{n} y_i \log \pi_i + (1-y_i)\log(1-\pi_i) \tag{2.40}$$

To find the critical point of this function with respect to β, we first write the log-likelihood in terms of β then differentiate, as follows:

$$\begin{aligned}
\ell(Y|X,\beta) &= \sum_{i=1}^{n} y_i \log\left[\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right] + (1-y_i)\log\left[\frac{1}{1+e^{X_i\beta}}\right] && (2.41)\\
&= \sum_{i=1}^{n} y_i \log e^{X_i\beta} - \log\left[1+e^{X_i\beta}\right] && (2.42)\\
\frac{\partial \ell(Y|X,\beta)}{\partial\beta} = 0 &= \sum_{i=1}^{n} y_i X_i - \frac{e^{X_i\beta}}{1+e^{X_i\beta}} X_i && (2.43)\\
0 &= \sum_{i=1}^{n} \left[y_i - \pi_i\right] X_i && (2.44)
\end{aligned}$$

Unfortunately, we can recognize this as the transcendental equation that does not have an analytical solution. Therefore, we estimate $\hat{\beta}$ using Newton's method (Figure 2.6). This method guarantees that we find a local critical point, but does not guarantee that this local critical point also is a global critical point. Therefore, it is important to assess the stability of the Newton-Raphson solution with respect to initial guesses at $\hat{\beta}$. We denote the intial guess as $\beta^{(0)}$. Newton's method supposes that a function can be approximated effectively based on a second order

Figure 2.6: Illustration of how Newton's method is used to find a local critical point of the data. The initial guess is X_1 and the initial guess for the zero point is X_2. The value of the function is assessed at X_2 and iterated until the zero point is found. To find the critical point of a log-likelihood function, the zero of the derivative of the log-likelihood is found.



Taylor expansion:

$$f(\beta) \approx f(\beta^{(0)}) + \frac{1}{1!}\left(\beta - \beta^{(0)}\right)\left.\frac{\partial f(\beta)}{\partial \beta}\right|_{\beta=\beta^{(0)}} + \frac{1}{2!}\left(\beta - \beta^{(0)}\right)^2 \left.\frac{\partial^2 f(\beta)}{\partial \beta^2}\right|_{\beta=\beta^{(0)}} \tag{2.45}$$

where the $\left.\frac{\partial^k f(\beta)}{\partial \beta^k}\right|_{\beta=\beta^{(0)}}$ notation refers to the k^{th} derivative of f calculated at β_0. We differentiate this expression with respect to β, we get:

$$0 = f'(\beta^{(0)}) + \frac{1}{2}f''(\beta^{(0)})2\left(\beta^{(1)} - \beta^{(0)}\right) \tag{2.46}$$

$$\beta^{(1)} = \beta^{(0)} - \frac{f'(\beta^{(0)})}{f''(\beta^{(0)})} \tag{2.47}$$

where primes are used as short hand for derivatives with respect to β. Note that β is a vector, so we can write this statement equivalently using linear algebra:

$$\beta^{(1)} = \beta^{(0)} - H^{-1}(\beta^{(0)}) \bigtriangledown f(\beta^{(0)}) \tag{2.48}$$

where $H(\beta)$ is the Hessian of β and $\bigtriangledown f(\beta)$ is the gradient of f with respect to β. We already wrote an expression for $\bigtriangledown \ell(Y|X,\beta)$ above:

$$\bigtriangledown \ell(Y|X,\beta) = \sum_{i=1}^{n} \left[y_i - \pi_i\right] X_i. \tag{2.49}$$

What remains is to differentiate this expression again to yield the Hessian:

$$H(\beta) = \frac{\partial}{\partial \beta}\sum_{i=1}^{n} \left[y_i - \pi_i\right] X_i \tag{2.50}$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \beta}\frac{e^{X_i\beta}}{1+e^{X_i\beta}} X_i \tag{2.51}$$

$$= \sum_{i=1}^{n} \frac{e^{X_i\beta}X_i(1+e^{X\beta}) - e^{X\beta}e^{X\beta}X}{\left(1+e^{X_i\beta}\right)^2} X_i^T \tag{2.52}$$

$$= \sum_{i=1}^{n} \frac{e^{X_i\beta}}{1+e^{X_i\beta}}\left[\frac{1}{1+e^{X\beta}}\right] X_i X_i^T \tag{2.53}$$

$$= \sum_{i=1}^{n} \pi_i(1-\pi_i) X_i X_i^T. \tag{2.54}$$

If we provide an initial guess of $\beta^{(0)}$, this iterative optimization will converge to a particular $\hat{\beta}$. Note that a guess of $\beta^{(0)} = \overrightarrow{0}$ is not a good initial guess because it is difficult to disentangle each individual element of β, each of which provide little effect on the final solution. One example of feasible initial values for $\beta^{(0)}$ start by randomly selecting each $\beta_j^{(0)}$ from independent standard Gaussian distributions.

This iterative optimization accurately estimates $\hat{\beta}$ but does not necessarily estimate the standard error of β. The standard error of β is estimated by assuming that the linearity assumption inherent to Newton's optimization holds. Consequentially, β are distributed normally with standard deviation defined by the Hessian of β. This means that for binary or logical input data, the variation of β_j is:

$$Var(\beta) = (X^T W X)^{-1} \tag{2.55}$$

for some weighting matrix, W, of the data. From (2.53), we know that

$$X^T W X = \sum_{i=1}^{n} \pi_i (1 - \pi_i) X_i X_i^T. \tag{2.56}$$

As a reminder, to test if the relationship between the outcome and input data is significant with 95% confidence, we test if

$$P\left(|\hat{\beta}_j| < \beta_j \,\middle|\, E(\beta_j) = 0, Var(\beta_j) = (X^T W X)^{-1}_{j,j} \right) < 0.05 \tag{2.57}$$

Using words, this equation asks if the probability of observing a $\hat{\beta}_j$ as or more extreme to the $\hat{\beta}_j$ we observed is less than 5% given the null distribution that $E(\beta) = 0$ and the variance is $(X^T W X)^{-1}$.

In addition to understanding how logistic regression models are trained, it is important to understand the interpretation of the resulting model. Firstly, if the number of exemplars in each class ($y_i = 0$ or $y_i = 1$) is equal and each of the input variables have zero mean, then the intercept, β_0 should be zero, reflecting that 50% (inverse logit of 0 is 50%) of the outcome data came from exemplars with $y_i = 1$. For β_j that correspond to the linear effect of input data, the $\hat{\beta}_j$ is a log odds ratio. An odds ratio of 2 suggests that the oods of the outcome is $y_i = 1$ is twice the probability that the outcome is $y_i = 0$ (66% vs 33%).

To combine these multiple odds ratios, we assume that each variable is conditionally independent. Mathematically, we can express this as:

$$P(y_i = 1|X, \beta) = \prod_{j=1}^{m} P(y_i = 1|X_j, \beta_j). \tag{2.58}$$

When we extend this to combine multiple odds ratios (OR), we get:

$$OR(y_i) = \frac{P(y_i = 1|X, \beta)}{P(y_i = 0|X, \beta)} = \frac{\prod_{j=1}^{m} P(y_i = 1|X_j, \beta_j)}{\prod_{j=1}^{m} P(y_i = 0|X_j, \beta_j)}. \tag{2.59}$$

The calculation of this can be simplified by taking the logarithm of that product:

$$\log OR(y_i) = \sum_{j=1}^{m} \log \frac{P(y_i = 1|X_j, \beta_j)}{P(y_i = 0|X_j, \beta_j)} = X_i\beta. \tag{2.60}$$

On finite precision machines, sums take much less computational power than products, therefore this simplification can be useful practically without loss of interpretability. When creating a receiver operating curve (ROC), one can use $OR(y_i)$ or $\log OR(y_i)$ to determine the trade-off between sensitivity and specificity (see below).

To implement these generalized linear models, there are numerous software packages. Our software package of choice is the standard lm and glm functions in R. This is because of its efficient implementation, good documentation (see `http://www.ats.ucla.edu/stat/r/`), clear treatment of standard error of all estimates, and good support for goodness of fit regression diagnostics.

2.3 Machine Learning Statistics

This section covers a unified theory of machine learning and gives examples of key machine learning classifiers that are necessary to understand the state of the field. By no means is this a comprehensive list and treatment of each classifier, but one will be able to understand the theory behind novel classifiers based on those presented here.

In this setting, we define machine learning as the development of automated methods to predict binary (or categorical) outcome data ($y \in \{\pm 1\}$ or $y \in \{0, 1\}$). In these models, we aim to maximize the predicted accuracy or surrogates of the accuracy. We can give each solution a quantitative score by combining a confusion matrix with a risk matrix. A binary risk matrix is defined as:

$$R = \begin{bmatrix} r_{0,0} & r_{0,1} \\ r_{1,0} & r_{1,1} \end{bmatrix} \tag{2.61}$$

We note that risk matrices are sometimes referred to as penalty functions. The first index of $r_{k,l}$ represents the known class whereas the second index refers to the class predicted by the model. Typically, $r_{0,0}$ and $r_{1,1}$ are zero, because they represent correctly classified exemplars. The off diagonals, $r_{0,1}$ and $r_{1,0}$, are non-negative and represent the penalty desired for either false positive ($r_{0,1}$) or false negative ($r_{1,0}$) classifications. The relative magnitude of $r_{0,1}$ and $r_{1,0}$ define our desired balance of maximizing sensitivity ($r_{1,0} > r_{0,1}$) versus maximizing specificity ($r_{0,1} > r_{1,0}$). When $r_{0,1} = r_{1,0}$, we are optimizing the overall accuracy of the model. We will return to this concept of the balance of sensitivity and specificity when we discuss receiver operating curves below.

The binary confusion matrix is defined as:

$$C = \begin{bmatrix} n_{0,0} & n_{0,1} \\ n_{1,0} & n_{1,1} \end{bmatrix} \tag{2.62}$$

where the indices are defined as they are in the risk matrix below. The n reflects the number of exemplars that has been classified in each way.

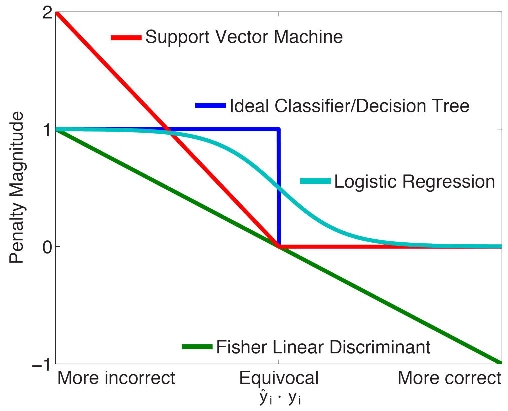
We calculate the weighted accuracy by the following element-wise sum:

$$\sum_{l,k} r_{l,k} n_{l,k}. \tag{2.63}$$

The challenge to optimizing these models is that this cost or penalty inherently is based on step functions that place exemplars into one item of the confusion matrix, or another. This step function is both non-continuous and non-differentiable (see Figure 2.7). This step function is the desired penalty that one would like to apply: a cost of unity for misclassifying a single training exemplar, and zero cost for correctly classifying that exemplar. The key to developing an efficient and effective machine-learning algorithm is to approximate this step function with an alternate penalty function. The key difference between each of the classifiers is how each approximates the step function (Figure 2.7).

The simplest approximation is taken by Decision Trees. In decision trees, we assume that the meaningful information in the input data also can be harnessed through step functions. Therefore, we sequentially choose input features that maximize the predictive accuracy in the training set. The next level of complexity is accomplished by approximating the step function by a linear function, as is done in Fisher Linear and Quadratic Discriminant Analysis (LDA and QDA). This linear assumption is sensitive to both easy to classify and hard to classify outliers. The soft-margin support vector machine (SVM) assumption uses the concept of a margin (see SVM below) to reduce sensitivity to easy-to-classify outliers. The next level of complexity is logistic regression, which assumes that the probability of class membership changes linearly in logit space (see Logisic Regression in Generalized Linear Models section above). This perspective has the benefit of being based firmly in statistical rigor through the Bernoulli distribution. Lastly, we do not assume the shape of the penalty function. In the canonical neural network model (not shown in Figure 2.7), we construct hidden layers of data, which are linear combinations of input data modeled through a logistic function. These hidden layer(s) of data are combined linearly through more logistic regression functions. The consequence of this deep model is that we can more closely approximate the step function with a series of logistic functions. These approximations allow for mathematical optimization of our chosen objective function (see below), but do not necessarily translate to improved accuracy of the generalized model. Prior to reviewing each of these models, we cover overarching themes of machine learning statistics that apply to each of these models.

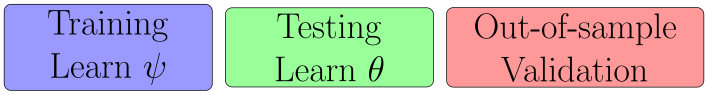
Figure 2.7: Qualitative illustration of the ideal and approximated penalty function for major categories of machine learning models. In this illustration, $y \in \{\pm 1\}$ and $\hat{y} \in \mathbb{R}$.



The general method to develop any machine-learning model is to use data to train the various aspects of the models. In ideal situations, one has enough data to split the available data into two or three subgroups: training group for parameters (ψ), testing group for hyperparameters (θ) and validation group for out-of-sample performance (Figure 2.8). The core of a machine learning model is the structure and parameters, ψ, that are optimized with respect to some objective function. Most machine learning models require the selection of hyperparameters, θ, that are critical to the structure of the model, but traditionally are fixed and are not optimized jointly with ψ. An optional testing set, generally smaller than the training set, can be used to optimize θ. Without this optimization, traditionally θ are considered fixed. Lastly, the learned ψ and θ are applied, without change, to the validation group to assess the generalizability of the model to 'unseen' data.

Because the ultimate goal of developing machine learning models is to apply them to data that the outcome we are trying to predict is unknown, the performance of the model on this dataset is critical.

Figure 2.8: Splitting the data into three mutually exclusive groups to avoid overfitting and use data to estimate all important parts of the model and its performance.



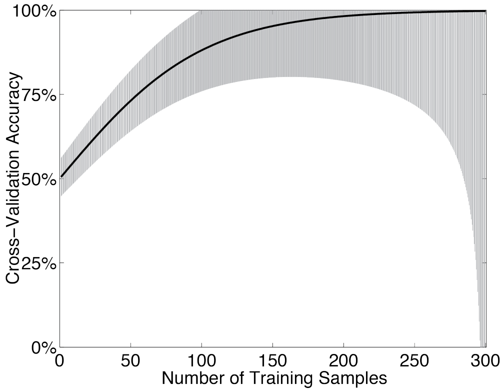
To estimate the performance of the learned model accurately, one would suggest that the validation set would be as large as possible, but in creating these sets, we must consider the bias-variance trade-off [93]. This trade-off states that as the validation set grows, the variance in the estimate of performance decreases. In terms of the overall accuracy, the binomial distribution suggests that the standard error of the estimate of variance of a proportion is defined by (Figure 2.9):

$$SE(\pi = Accuracy) = \sqrt{\frac{\pi(1-\pi)}{n}} \tag{2.64}$$

where n is the number of samples in the validation set. We define "overall" accuracy as the accuracy considering all exemplars equally. (We will address later if this binomial assumption is valid.) However, the bias-variance trade-off states that when one increases the size of the validation set, the size of the training and/or test sets must decrease, given a constant overall sample size. This decrease in size of the training and/or test sets results in reduction of the generalization performance, likely due to a decreased ability to estimate the predictive value of the input data. Thereby, we have the trade-off where an increase in the validation set allows us to estimate generalization performance better, but the same increase in the size of the validation set decreases the generalization performance.

When large or unlimited datasets are available, this trade-off is a non-issue, but in realistic situations, it is critical to use all the available data. One method

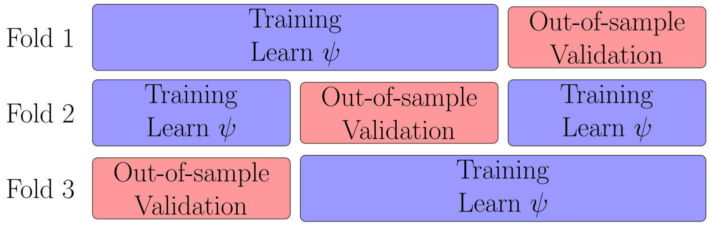
Figure 2.9: Bias-Variance Tradeoff: As the number of training samples increases, both the mean and variance of the cross-validation accuracy increase monotonically. Mean is indicated by the black line, and standard deviation is illustrated by the grey error bars. Exact values and shape are illustrative, and do not represent real or simulated results.



to accomplish this is through cross-validation (Figure 2.10). In cross-validation, the entire dataset is split into the two groups and θ is chosen *a priori*. After the model has been trained and validated, the data in the validation set is permuted and the process is repeated such that each data point is in the validation set once and only once. The generalization results across all validation sets are pooled. In this way, all of the data is used to make a large validation set, and the size of the training group is maximized. If one data point is left out of training and testing, as in leave-one-out cross-validation, then the size of the training and testing sets are maximized while the whole dataset is used for assessing generalization performance.

When there are hyperparameters, θ, to optimize, a testing group is required, because one must use data to estimate the optimum θ. Throughout this work, we define θ as parameters that cannot or are not estimated simultaneously with the parameters intrinsic to the model, ψ. If one uses the same data to optimize ψ and θ by choosing the θ-ψ pair that results in the best training performance, then

Figure 2.10: 3-Fold Cross-validation when θ is selected *a priori*, as compard to optimized. Therefore, no data is needed to learn θ. A separate model is learned for each fold, resulting in three overall models to aggregate later.



the model frequently is over-fit. By using separate data to optimize ψ (training) and θ (testing), the propensity for overfitting is reduced because the testing set is statistically independent, at least theoretically, from the training set.

This testing set typically is chosen through nested cross-validation (Figure 2.11). Just as data is left out for the validation set to estimate the generalization accuracy, data can be left out from the training group to estimate the optimal θ. In double leave-one-out nested cross-validation, one exemplar is used for validation and a separate exemplar is used for testing. For each θ in the search space (range of possible θ), ψ should be optimized using the training data. Those learned ψ are applied to that testing exemplar. The θ-ψ pair that best predicts the testing exemplar, by a chosen criterion, is then applied without change to the validation to assess the generalization performance.

With our knowledge of the bias-variance trade-off, we recognize that this progressive splitting further biases the model towards worse generalization performance and contributes to another source of variance in the estimate of generalization performance. However, this is the least biased and most principled method for estimating both θ and ψ. Addressing this critical and basic challenge is one of the pillars of this graduate thesis (see chapters 11 & 12).

Cross-validation both is intensive computationally and can complicate the in-

Figure 2.11: 3 then 2 Fold Nested Cross-validation. The data is split into 3 mutually exclusive sets, and each one is successively used as the validation set. For each validation set, the remaining data is split into 2 mutually exclusive sets to estimate the optimum ψ and θ. As shown, the inner splits correspond to the higher level splits, but this is not necessarily the case. A separate model is fit for each fold. These models must then be aggregated in post-processing.

Fold 1-1	Training Learn ψ	Testing Learn θ	Out-of-sample Validation
Fold 1-2	Testing Learn θ	Training Learn ψ	Out-of-sample Validation
Fold 2-1	Training Learn ψ	Out-of-sample Validation	Testing Learn θ
Fold 2-2	Testing Learn θ	Out-of-sample Validation	Training Learn θ
Fold 3-1	Out-of-sample Validation	Training Learn ψ	Testing Learn θ
Fold 3-2	Out-of-sample Validation	Testing Learn θ	Training Learn ψ

terpretation of $\hat{\psi}$ and $\hat{\theta}$. Because multiple models need to be trained, the computational cost increases linearly with the number of validation folds. Additionally, each training and test set result in a different model. In some cases, averaging across validation folds results in models similar to when all the data was used for training and testing (i.e. β in logistic regression). In most cases, however, averaging either cannot be done or does not reflect what a larger model would look like (i.e. thresholds and variable order in decision trees). Therefore, interpretation of models across folds can be extremely difficult.

These problems can be reduced by utilizing k-fold cross-validation, where the data is split into k validation groups. This results in only k different models to train and interpret (or k^2 when performing nested cross-validation).

With the exception of logistic regression, one additional challenge to many machine learning models is the lack of good statistical theory to predict the sensitivity of the solution to changes in chosen or estimated parameters, or design choices. In some cases, this results in an inability to estimate the probability distribution of summary statistics under the null hypothesis that X has no predictive relationship with Y. In machine learning literature, the "null" hypothesis frequently is referred to as "chance." A chance classifier is defined by a predictive algorithm based on an X matrix that has no predictive relationship with Y. In other words, a chance classifier randomly guesses Y without consideration of X. When we seek to test statistical significance, we are asking what the probability of our observed results are if the classifier was a chance classifier. If we do not have a clear probability distribution for a summary statistic, the key to estimating significance is determining a quantifiable value that represents the aspect of the solution that we seek to test. As long as this quantifiable value is defined for each solution, its significance can be assessed.

There are two subtly different methods to establish empirical probability distributions for arbitrarily defined values: permutation testing and bootstrapping.

In bootstrapping, artificial datasets are created by artificially resampling the data with replacement, then retraining the model on each of these artificial datasets. Each artificial dataset is the same size as the original dataset. The variation around the observed value due to resampling allows us to estimate the 95% confidence interval around the value. However, the assumptions critical to bootstrapping frequently do not hold for many machine learning models, therefore bootstrapping is not used frequently.

In permutation testing, we seek to form an empirical probability distribution of any summary statistic achieved by a chance classifier. This empirical probability distribution can estimate the p-value of the observed value. If the observed value has less than 5% probability, then we consider the observed value statistically significant. This is achieved by randomly permuting the class information without replacement. The input data is left unchanged even though the class label is changed, thereby breaking the association between data and class. The correlation structure within the data, however, is maintained. The significance of the observed value with real data can be estimated by counting the number of times the null data achieved the observed value, or one more extreme.

An alternate method of permutation testing is to create artificial null data, which we define subsequently. The class information is kept the same, but fake input data is generated according to the null hypothesis. For binary input data, when the probability of $x_i = 1$ is not related to the input class, π_i is set to be 50%. For continuous input data, typically one assumes the data is Gaussian with a mean of 0 and a variance of 1. The actual value of mean and variance are irrelevant, except for normalization relative to other features and, depending on the objective function, relative magnitude to the regularization term (see below for explanation of regularization).

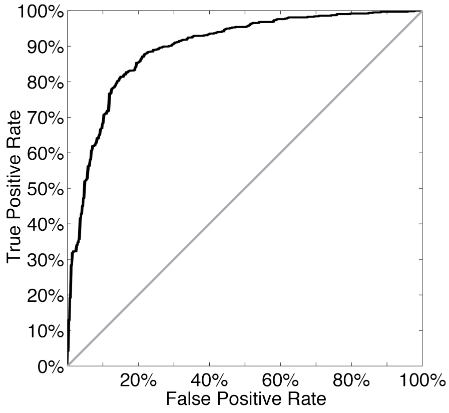
The difference between the two types of permutation testing is the latent structure within the input data. When the class information is permuted, but the input data is unchanged, there is latent structure in the data. When artificial data is created, this latent structure also is removed. One reason for removing this structure is that even when the class information is permuted, there is a quantifiable similarity between the original and permuted class labels. As that similarity increases, the similarity of the "null" data and the original data increases. This breaks the assumption of permutation testing that there is no relationship between the class labels and the input data. (This may be one reason why the distributions of predicted accuracies estimated by permutation tests are not exactly binomial.) However, when there is correlation within the input data, the correlation can result in numerical instabilities and other unpredictable effects on the value of interest. Therefore, there is no best choice between the two types of permutation testing.

Even though an false positive rate, α, of 0.05 can be estimated for any summary statistic with just 20 permutations in each permutation scheme, many more need to be conducted to be confident in the estimated cutoff. While the mean or median value is stable when 20 independent permutations, the ordinal statistics are unstable relatively. This is clear when one considers the distribution of the maximum value of 20 permutations. Due to heavy tailed distributions, the probability of huge values is small, but finite, thereby potentially huge variations in the maximum. Therefore, rules of thumb are that to estimate the α cutoff of 5% or 1%, 10,000 and 50,000 independent permutations must be done, respectively.

Permutation tests, as described above, can be used to determine significance for any summary statistic. While we motivated these tests above based on a desire to test the significance of parameters in the model, ψ and θ, permutation tests are used most commonly to estimate the significance of the performance of the overall model. When one is interested in applications, accurately estimating the significance of performance statistics is critical to assessing the applicability of the model.

Performance statistics include cross-validation accuracy, sensitivity, specificity and area under the receiver operating unit curve (AUC, see Figure 2.12). Up to a certain degree of error, one can assume that accuracy is binomially distributed by assuming the probability of predicting the class of each exemplar is equal. An astute reader would notice that binomially distributed variables are counts, whereas accuracies are percents. Accuracy can be transferred trivially to a count of the number of accurately predicted exemplars by multiplying by the number of validation exemplars. The binomial assumption is useful when performing power calculations and for pilot studies where we seek to avoid the computational cost of at least 10,000 permutations. However, other work has shown that this assumption is an approximation [94].

Figure 2.12: An example receiver-operating curve which shows performance of a prediction with respect to different balances of sensitivity (True Positive Rate) compared to specificity (1-False Positive Rate). The black line illustrates a classifier with good performance, whereas the gray line indicates theoretical chance.



The inexact nature of this assumption becomes clear when we explore situations where the number of exemplars in each class is uneven, or we seek to estimate probability distributions for the other performance statistics. Consider the extreme example where 90% of exemplars are class $y_i = 0$. A naive classifier would predict every validation case came from class zero and thereby acheive 90% accuracy. If the distribution of the validation data matches the training data, then the accuracy of these naive classifiers would be binomially distributed around 90%. However, the probability of success on each exemplar is not equal: it is more likely for exemplars that are class zero to be classified correctly. Next, consider a less extreme example where leave-one-out cross validation is used on a dataset with 150 exemplars, 75 of which are from each class. Within each training set, the classes are mismatched slightly so if the model assumes the class distribution in the training set is the same as the validation set, then the overall accuracy would be 74/149 or 49.7%. If, however, the model takes the naive perspective and classifies the validation data as the most common class in the training set, then the overall accuracy would be 0%. In order to tell the difference between these two options, empirical tests of chance provide invaluable evidence for what "chance" means.

Theoretically, the balance between the naive and chance classifiers is due to the amount of overfitting done by the model in the training set. If the model overfits the data, then it can be more confident in the class of the validation exemplars. If one class is more common, this confident classifier could classify everything as the most common class, irrespective of its input data, resulting in a fully naive classifier. If the data is not as overfit, then the decision boundary could be softer, where a larger fraction of the input space would result in classification of validation data as the less common class. Because validation data would lie, randomly, on one side or the other of this decision boundary, the classifier would act more like chance classifier. Because the degree of overfitting cannot be predicted theoretically, it is difficult to knowing when and the degree to which a classifier will overfit the data.

Therefore, when possible, permutation tests are the preferred method of determining the significance of any summary statistic. In addition to the benefits discussed above, permutation tests also allow us to estimate the null distribution of any summary statistic that we can calculate from the results. As long as the summary statistic is defined and consistently quantitative, it can be calculated both in the original data and the permutated datasets. Suppose, for instance, we desire to report the area under the receiver operating curve (AUC). Estimating the probability distribution of this can be difficult [95]. Alternatively, suppose we want to report the number of predictive factors with non-zero log odds ratios in a L_1 regularized logistic regression model so that we can comment on the complexity of the problem at hand (see Feature Selection section in Didactic Background). There is not good theory that could be used to determine a null probability value for that. In these situations, permutation testing provides a simple and statistically valid method for determining significance and interpreting values, even though they are expensive computationally.

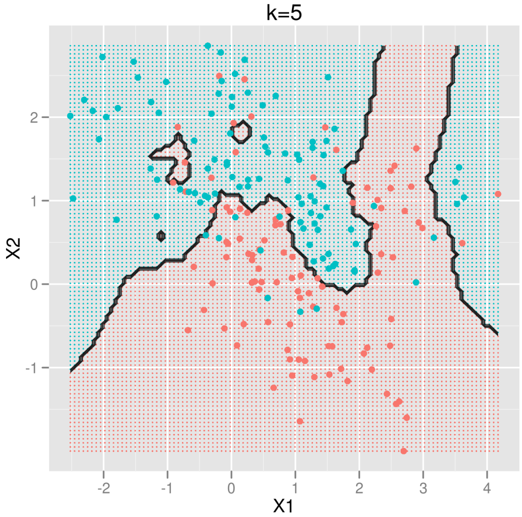
2.4 Machine Learning Classifiers

We will review the key classifiers discussed briefly above. We cover both the basic theory behind each algorithm and selected quantitative results whose significance can be estimated empirically for these models without strong statistical theory to suggest a distribution.

2.4.1 Nearest Neighbor Classifiers and Norms

The classifiers with the simplest and most generalizable design concept are nearest neighbor classifiers. This relies on the generalizable insight that data points that are closer in X are therefore more likely to have similar output class. A classical nearest neighbor algorithm simply classifies the validation data as the same class as the closest exemplar within the training data. This can result in an extremely complicated, and binary, decision boundaries (see Figure 2.13). A decision boundary is defined as a line in the input space where if a validation exemplar lies on one side, it is classified as $y = 1$, and if it lies on the other side, it is classified as $y = 0$. Highly flexible and binary decision boundaries have a greater propensity to overfit the data. If we want to make a softer boundary that takes into account more training data at each point, we can use a k-nearest neighbor classifier, where we consider the class of the k nearest training exemplars. The fraction of these training classifiers that were $y = 1$ is used to estimate the probability the validation exemplar is $y = 1$.

Figure 2.13: This illustrates the complexity of a nearest neighbor solution for two dimensional input data, where the closest $k = 5$ neighbors based on an Euclidean or L_2 norm were considered. The lighter color indicates the region in which validation data would be predicted to be from the blue or red class.



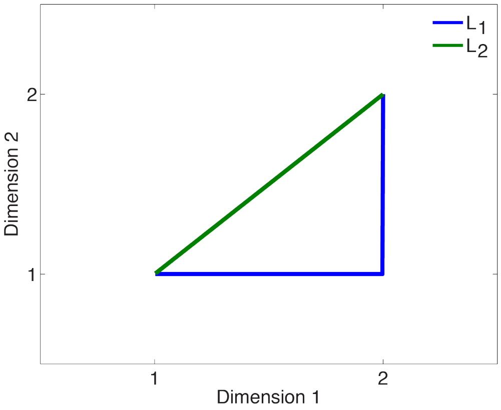
An important remaining question regarding nearest neighbor classifiers is 'how do we define close?' Typically, we assume that all input data are equally important and linearly scaled; and the orientation of vectors within the X space does not hold information, then the L_2 norm is appropriate. The L_2 norm is the most common distance norm, and is defined by:

$$\|X_{\text{training}} - X_{\text{validation}}\|_2 = \sqrt{\sum_{j=1}^{m} \left(x_{\text{training},j} - x_{\text{validation},j}\right)^2}. \qquad (2.65)$$

Many people use this norm without considering the strong assumptions listed above. Potentially, this is because it is difficult to make an argument for any other distance function without making an explicit local model (see below for model descriptions). However, both of the main assumptions can be modified easily. First, if we do not want to assume that data is spherical in the input space (orientation doesn't matter), then we can use an L_1 norm, where distance is defined by the sum of the distance along each input dimension (Figure 2.14), as follows:

$$\|X_{\text{training}} - X_{\text{validation}}\|_1 = \sum_{j=1}^{m} \left|x_{\text{training},j} - x_{\text{validation},j}\right|. \qquad (2.66)$$

Figure 2.14: The distance between two points based on the L_1 and L_2.



Next, we can assume that not all features should be equally weighted. If we say that only the maximum difference along an individual input feature matters, we can use L_∞ norm:

$$\|X_{\text{training}} - X_{\text{validation}}\|_\infty = \max_{m} \left|x_{\text{training},j} - x_{\text{validation},j}\right|. \qquad (2.67)$$

Lastly, if we want to count the number of features where the input data differs, by any amount, we can use the L_0 norm:

$$\|X_{\text{training}} - X_{\text{validation}}\|_0 = \sum_{j=1}^{m} \left|x_{\text{training},j} - x_{\text{validation},j}\right|^0 . \tag{2.68}$$

Any non-negative real number, r, can be used in the exponent, resulting in a continuous variation in the "circle" equidistant from a center (Figure 2.15). These unconventional r values balance the strength of the various norms by being intermediate between them, and are defined by the following formula:

$$\|X_{\text{training}} - X_{\text{validation}}\|_r = \left[\sum_{j=1}^{m} \left(x_{\text{training},j} - x_{\text{validation},j}\right)^r\right]^{\frac{1}{r}} . \tag{2.69}$$

If instead we want to use a mix of these classifiers, we can use a mixed norm, defined by

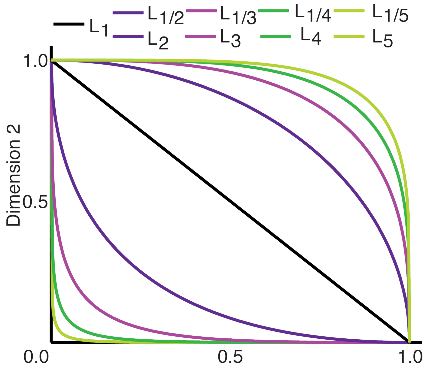
$$\begin{aligned}\|X_{\text{training}} - X_{\text{validation}}\|_{r,c} =&\gamma\|X_{\text{training}} - X_{\text{validation}}\|_r \\ &+ (1-\gamma)\|X_{\text{training}} - X_{\text{validation}}\|_c\end{aligned} \tag{2.70}$$

where $\gamma \in [0, 1]$, and r and c are in $\mathbb{R}_{\geq 0}$. While there are a great variety of norms, it is difficult to know which norm will result in the most generalizable solution prior to applying it to the validation data. This is there is not clear statistical theory behind the nearest neighbor classifier that would suggest which norm is the best for a given type or structure of input data. Therefore, we prefer classifiers with a clear objective function that has some theory behind it.

2.4.2 Decision Trees

The classifiers that best match the step function are decision trees (Figure 2.7). The simplest version of a decision tree is called an alternating decision tree (ADT). These trees are built by making successive binary decisions to separate the output classes. At each decision point, called nodes, we search through each of the available input data and test each threshold to maximize the purity of the daughter

Figure 2.15: The shape of "circles" calculated according to various norms.



leaves, as quantified by their Gini impurity. Formally, this search is defined as:

$$\begin{aligned}\max_{j,k} G(j,k) = \max_{j,k} & P(Y=-1|X_j<k)\cdot[1-P(Y=-1|X_j<k)]+ \\ & P(Y=1|X_j\geq k)\cdot[1-P(Y=1|X_j\geq k)] \qquad (2.71)\end{aligned}$$

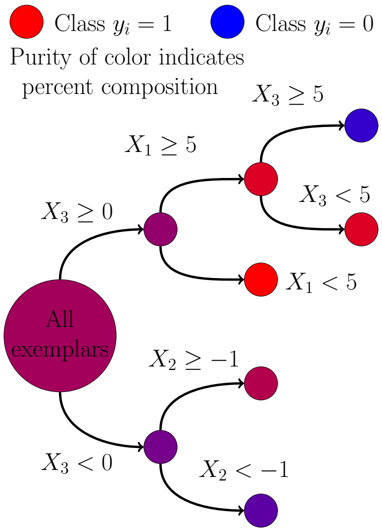
where Y is a vector of exemplars that were present at the node (see Figure 2.16). For each daughter leaf, the successive maximization and split are repeated until one of the following conditions is met:

- The daughter node consists of exemplars of only one type.
- The daughter node consists of just one exemplar.
- A predefined stopping point is reached.

There are a large variety of predefined stopping criterions including but not limited to a minimum size of daughter node; a maximum number of binary decisions; not allowing the same X_j to be used at multiple levels of the tree; or a minimum Gini impurity achieved for the best X_j and k. Alternatively, pruning methods can be applied after a dense tree is trained. These criterions hope to reduce the propensity for decision to overfit the data.

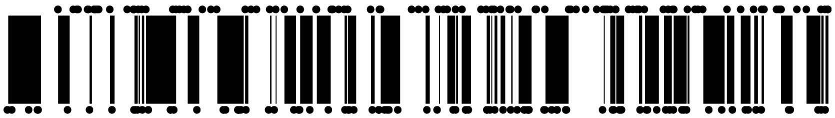
The challenge in using decision trees is that their propensity to overfit the data and the strong assumption that binary decisions can capture the full complexity

Figure 2.16: An example decision tree. All exemplars begin in the largest circle, then are triaged towards the daughter circles based on maximizing the Gini impurity. Exemplars are predicted to be the most common class within the last daughter circle. Note that once a covariate is used, it can be reused lower down the tree. Depending on the stopping criterion, the depth of the tree is not necessarily uniform.



of the data. Both of these challenges can be shown through a simple example. Consider a single input feature where a linear increase in x logistically increases the probability that $y = 1$. For any finite sample and no predefined conditions, successive binary decisions can split the range of x into sections that are predicted to be all $y = 1$ and other sections predicted to be all $y = -1$ (see Figure 2.17). The decision points will be more dense in around the region where $P(y_i = 1|x_i = \gamma) = 50\%$, but they will give the false impression that the certainty of the predicted class is equally accurate in this intermediate region as it is along the outer ranges of x.

Figure 2.17: An example of an overfit decision tree. Black sections indicate where the model would predict class below, whereas white sections indicate class above. Simulated data (black dots) were distributed uniformly over the interval and randomly assigned to be class below or above, with a stable probability of 50%.



One method to overcome this particular limitation of decision trees is addressed partially by random forests. As is connoted by calling the algorithm a 'forest,' this method uses a combination of many decision trees to make a probabilistic decision. If all trees were constructed with the same input data and training exemplars, each tree would be identical. Randomly sampling the input features and training exemplars with replacement generates variation across trees. The overall sample size and dimensionality of the data are kept constant, but by placing multiple identical points at certain locations and not sampling others results in changes to the Gini impurity, thereby creating varying trees. When the forest is used to classify an exemplar, a probabilistic decision is generating by dividing the number of decisions that the classifier was $y_i = 1$ divided by the total number of trees in the forest. However, when limited training exemplars and input features is available, the variation across trees is limited.

2.4.3 Discriminant Analysis

Instead of assuming that all information in the features can be captured through binary decisions, one can relax that assumption slightly, as is done in Fisher Linear or Quadratic Discriminant Analsyis (LDA or QDA). Instead of using binary decisions, LDA and QDA use a linear changes in X to produce linear changes in the probability that $y = 1$. The LDA algorithm officially is trained by maximizing the ratio of the predicted within-class variance to the between-class variance (Figure 2.18). Formally, this is done through either of the following equivalent maximizations:

$$\min_{\beta} -\beta^T \Sigma_B \beta \text{ s.t. } \beta^T \Sigma_W \beta = 1 \text{ or } \max_{\beta} \frac{\beta^T \Sigma_B \beta}{\beta^T \Sigma_W \beta}. \tag{2.72}$$

where Σ_W is the within class variance and Σ_B is the between class variance. The solution to both of these maximizations is:

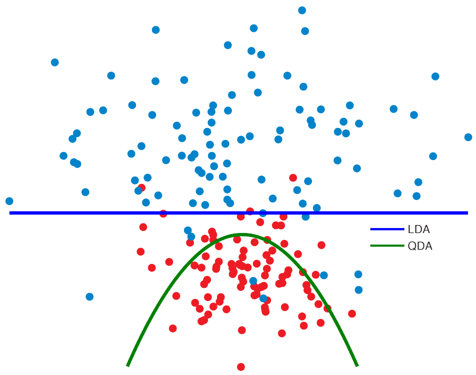
$$\beta = \left[E(X_i|y_i = 1) - E(X_i|y_i = 0)\right] \Sigma_W^{-1} = \left[\mu_1 - \mu_0\right] \Sigma_W^{-1} \tag{2.73}$$

However, this phrasing of the algorithm is unnecessarily opaque. LDA is equivalent mathematically to k-means clustering when k is 2. In k-means clustering, the multivariate means of all the exemplars where $y = 1$ and $y = 0$ are taken separately. The probability that an out-of-training-sample exemplar is defined as:

$$P(y_i = 1|X) = \frac{\frac{<y_i, E(X_i|y_i=1)>}{Var(X_i|y_i=1)}}{\frac{<y_i, E(X_i|y_i=1)>}{Var(X_i|y_i=1)} + \frac{<y_i, E(X_i|y_i=0)>}{Var(X_i|y_i=0)}}. \tag{2.74}$$

In LDA, one assumes that $Var(X_i|y_i = 1) = Var(X_k|y_k = 0)$. If all exemplars are classified as the most likely exemplar (or a threshold probability is defined), this results in a linear decision boundary (a hyperplane $w \in \mathbb{R}^m$) that is perpendicular to the $\hat{\beta}$ learned above. In QDA, we estimate the within-class variance separately,

Figure 2.18: Examples of lines separating two classes based on Fisher Linear and Quadratic Discriminant Analysis. Simulated data is centered around a point with Gaussian noise added. The variance of the blue class is twice that of the red class.



resulting in a quadratic decision boundary that curves towards the class with the smaller variance (Figure 2.18).

There are a few key limitations to LDA and QDA. The most apparent limitation is the assumption of linearity in X. Both LDA and QDA would, therefore, be insensitive to quadratic (or higher order) trends in individual X_j or nonlinear interactions between dimensions in X_j and X_k for $j \neq k$. Additionally, all training exemplars contribute equally to the final solution, including potential easy to classify outliers.

2.4.4 Support Vector Machines

In contrast, the main innovation in a Support Vector Machine (SVM) classifier, relative to LDA and QDA, is unequal weights on training exemplars, although other innovations also are present [96]. Support vectors are the hardest to classify exemplars, in that their input data lie closest to the other class. The exemplars that are easier to classify are ignored, for sake of determining the maximally separating hyperplane. This makes SVM less sensitive to outliers, if they are easy to classify. The line that is used to separate the two classes is based on maximizing the margin between the exemplars from each class (Figure 2.19). A margin is defined by the distance between the closest two exemplars along the line that is perpendicular to the separating hyperplane, w. Because less data is used to determine this hyperplane, we must apply regularization so that there is one unique maximum to the objective function. A hard-margin SVM is trained by optimizing the following Lagrangian objective function:

$$L(X, Y|\psi = \{w, \alpha\}) = \frac{\|w\|_2^2}{2} - \sum_{i=1}^{n} \alpha_i \left[y_i X_i w - 1\right] \tag{2.75}$$

where $y_i \in \{\pm 1\}$ and α denotes the support vectors: $\alpha_i = 0$ if and only if the exemplar is not a support vector. In comparison to the β that we learned in other machine learning applications, w is the hyperplane perpendicular to β.

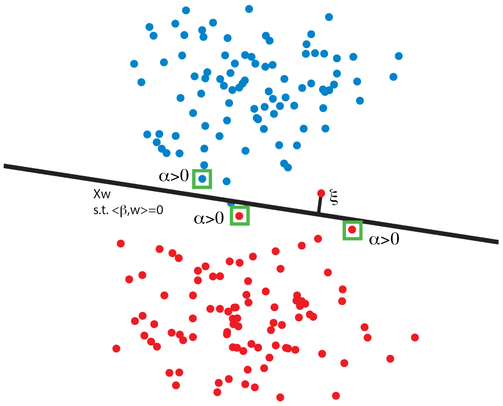
The last innovation of SVM is the use of a kernel. The objective function above finds a linear separating hyperplane in the original input space. Alternatively, we can find a linear separating hyperplane in the kernel space. The kernel space is a higher dimensional space that can be used to train non-linear separating planes.

The SVM objective listed above reflects a "hard-margin" SVM, where there exists a linearly separating hyperplane in which none of the training data is misclassified. In the original X space, this hyperplane may not exist. In a sufficiently higher dimensional kernel space, one can guarantee that such a hyperplane exists. If we do not want to use that type of kernel space, we can use a "soft-margin" SVM, where misclassified training exemplars penalize the Lagrangian objective function linearly with the distance from the separating hyperplane, as follows:

$$\begin{aligned} L(X, Y|\psi = \{w, \alpha, \xi, r\}, \theta = \{C\}) = & \frac{\|w\|_2^2}{2} + C\|\xi\|_1 \\ & - \sum_{i=1}^{n} \alpha_i \left[y_i X_i w - 1 + \xi_i\right] + \sum_{i=1}^{n} r_i \xi_i \end{aligned} \tag{2.76}$$

where ξ_i is the distance of training exemplar i to the correct side of the separating hyperplane, and r is another Lagrange parameter indicating the support

Figure 2.19: Example of a soft-margin SVM solution for simulated Gaussian data. The black line indicates the SVM separating hyperplane. The red exemplar on the wrong side of the hyperplane is an example of a misclassified training exemplar that is a distance, ξ, from the maximum separating hyperplane. The green squares highlight the three support vectors that would define the hyperplane, if the misclassified exemplars were not present. The value of α_i is greater than 0 for all suport vectors, by definition. Lastly, the separating hyperplane is defined by Xw where w is perpendicular to a line that linearly projects the multivariate data onto a single decision dimension, β.



vectors. In this case, support vectors are the points that define the separating hyperplane plus all misclassified training exemplars. This guarantees that a unique solution exists for any X and kernel. However, we note that as the penalty for misclassification decreases ($C \to 0$), SVM approaches L_2 regularized LDA.

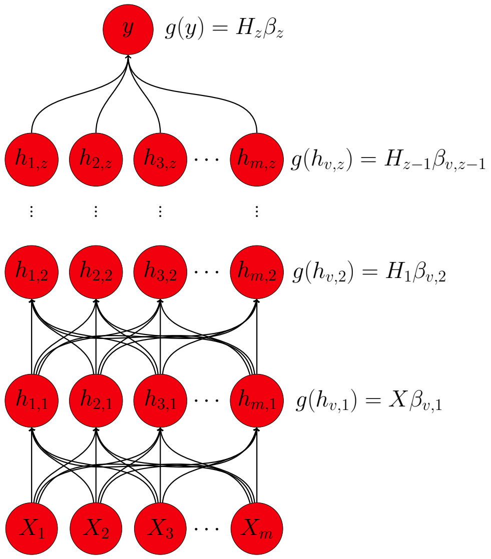
SVM models are very popular in neuroimaging because of their demonstrated applicability, their resistance to outliers and stability in high dimensional spaces despite limited exemplars. As discussed in detail below, regularization increases our ability to find stable solutions in underdetermined systems where the number of exemplars, n, is far less than the number of input features, m. All of these models induce an assumed structure of the loss function and the shape of the predictive information held within the input data.

2.4.5 Neural Networks

The most flexible model, a neural network, places very little assumptions on the data. While there are multiple structures of neural networks, the most common is a multilayer perceptron (see Figure 2.20). A multilayer perceptron makes multiple and successive linear combinations of the original input data, which are placed through a logistic function after each linear combination. After multiple layers of these combinations, any loss function can be approximated. Additionally, any structure of predicted information, including quadratic and exponential relationships between particular X_j and Y, can be utilized fully. Because of this flexibility, the performance of a deep neural network is at least equivalent to any the performance of any other machine learning model, when infinite data is available.

The benefit of neural networks is also the challenge. Training the many modifiable parameters, $\psi = \{\beta_{\nu,z}\}$, and hyperparemeters, $\theta = \{\nu, z\} \in \mathbb{Z}^2_{\geq 0}$, within the neural network requires a large amount of data and computational power. When limited data is available, neural networks tend to overfit the data, because the

Figure 2.20: An illustration of the structure of a multilayer neural network. The input data, X, are combined linearly and pushed through a link function, g. In the case of a multilayer perceptron, this link is a logit function. Other neural networks use other link functions. These successive combinations of the input data are eventually used to predict the final output class. Through these successive combinations, high level interactions can be modelled, and a more nuanced penalty function can be used to approximate the ideal step function.


y
$g(y) = H_z\beta_z$
$h_{1,z}$ $h_{2,z}$ $h_{3,z}$ $\cdots$ $h_{m,z}$
$g(h_{v,z}) = H_{z-1}\beta_{v,z-1}$
$h_{1,2}$ $h_{2,2}$ $h_{3,2}$ $\cdots$ $h_{m,2}$
$g(h_{v,2}) = H_1\beta_{v,2}$
$h_{1,1}$ $h_{2,1}$ $h_{3,1}$ $\cdots$ $h_{m,1}$
$g(h_{v,1}) = X\beta_{v,1}$
X_1 X_2 X_3 $\cdots$ X_m

large number of parameters allows each exemplar to completely define the region around it, just as the deep alternating decision tree did earlier. In that way, the performance of the model on the training data can be much greater than the performance on the validation data. Because we are more concerned with the ability of the learned solution to generalize to unseen data, this propensity to overfit is a major limitation.

Now that we have reviewed the major categories of classifiers (but by no means, not all classifiers), a logical question is how one goes about choosing a classifier for a given application. Based on the insight above about overfitting, we should choose a classifier that best matches the known or hypothesized structure of the input data. In lay terms: "keep it simple, stupid." In more nuanced language: unless you know there exists a more complicated structure, don't assume that it exists. The simplest and easiest to interpret models are from logistic regression because of the wide applicability to multiple types of input data, clear distributions of error, and rigorous statistical motivation.

If the input data is known to be from an unbounded Gaussian distribution, then it is feasible to assume linearity, as is done by LDA or QDA. If, in addition, we know that the data has a problem with outliers, an SVM could result in more stable and generalizable solutions. However, due to how SVM maximizes the margin of the hyperplane, SVM may not be the best algorithm for binary data: since all data lie in two locations, the margin is the same no matter which exemplars are the support vectors.

Lastly, if the previous classifiers are not effective, a large amount of training data is available and/or a complex interaction structure in the input data is expected, then a neural network classifier could result in the best performance. The challenge is that in deep neural network classifiers, the number of parameters is prohibitively hard to interpret. This limits our ability to learn about the underlying system, even if we can produce highly accurate class predictions.

Even though this may seem like a comprehensive review of machine learning algorithms, we have left out key categories that are not pertinent to the work below. These omitted topics include, but are not limited to, unsupervised and semi-supervised algorithms. A supervised algorithm is defined by using known class information to train the algorithm.

While there are a plethora of good software to implement each individual model, we recommend using the Java-based software package *Weka*. The main benefit to *Weka* is its implementation of many diverse classification methods within the same software, and the ability to modify the risk matrix easily. This facilitates the comparison of results across different models. Additionally, *Weka* is relatively easy to use in a graphical user interface (GUI), and can be implemented non-trivially on the command line. The major limitations to *Weka* are inefficient memory usage, especially for large scale models, and lack of good documentation for the command line version of the software.

2.5 Experimental Design, Curse of Dimensionality & Feature Selection

The design of experiments for the sake of machine learning is both similar and different from experiments that plan on utilizing conventional statistics. A key point is that, even though machine-learning statistics has great power, they are still limited by the same need for a sufficient amount of high quality data to estimate the predictive ability of the input data.

To illustrate this, we will illustrate this by drawing a parallel between underdetermined simple linear models and generalized linear models. When Y is Gaussian, the solution to simple linear regression is:

$$\beta = (X^T X)^{-1} X^T Y, \tag{2.77}$$

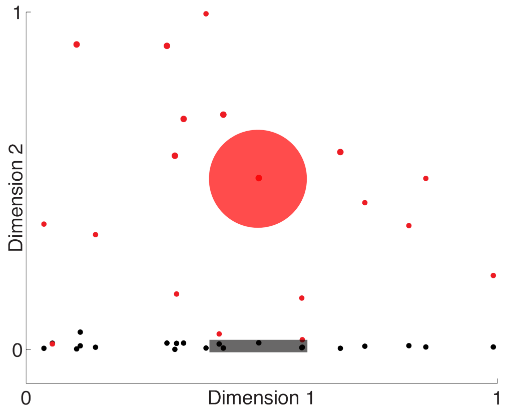
where β is the multivariate slope, and the other variables are defined as they have been throughout this manuscript. We note that if $X^T X$ is not invertible, the system is underdetermined: there is insufficient data to independently estimate all elements of β. When these methods are applied using computers with finite precision, an $X^T X$ that is close to an uninvertible $X^T X$ also may provide inaccurate or noisy estimates of β. In this latter case, $X^T X$ is said to have a poor condition number. Non-invertibility and poor condition numbers occur when there is insufficient or close to insufficient row rank. This occurs when there are fewer training exemplars, n, than features, m. Alternatively, if two or more X variables are collinear or approximately collinear then the inverse is similarly undefined because at least two columns are linear combinations of each other. This is just one, of many, examples of the types of relationships that cause poor condition number. In a generalized linear model, we model Y through a transformation function $g(Y)$, like the logistic function. While the solution to a generalized linear model is not always based on the inverse of $X^T W X$. In the case of logistic regression, Newton's algorithm is based on the Hessian of X, which depends on the inverse of $X^T W X$ (see Logistic Regression Section above). Therefore, if simple linear regression is underdetermined, there is a possibility that logistic regression also may be underdetermined.

To avoid these issues, we should aim to sample the full range of X from as many independent patients as possible. This minimizes the potential for collinear variables. Additionally, by sampling from a wide distribution of patients, we also maximize our ability to estimate the predictive ability of each feature.

However, it is sometimes unavoidable to have collinear variables or insufficient data ($n << m$). To overcome these, we can apply penalties and optimization methods to result in stable, unique solutions. While these methods are not unique to machine learning, they are applied more widely in machine learning settings. Before we describe regularization penalties and feature selection methods to overcome these limitations in the data, we must describe the curse of dimensionality (CoD).

The curse of dimensionality (CoD) is an unavoidable fact of sampling data [97]. No matter what method of feature selection or regularization we apply, the curse of dimensionality applies. The basic premise of the curse is that as the dimensionality of X increases (more features are measured), the average distance between exemplars increases more than linearly (Figure 2.21). When we make predictions, each method is based on applying knowledge from similar exemplars in X to the validation data. If the distance between exemplars increases, then the data are less similar and, consequentially, less can be learned from the "nearby" exemplars, subject to a few assumptions.

Figure 2.21: An example of the curse of dimensionality for one (black) versus two (red) dimensional data uniformly distributed over the interval $[0, 1]$. The colored interval and circle represent the size a hypersphere with matching radius in each dimensionality.

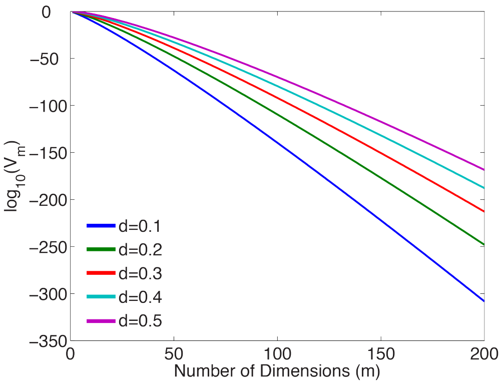


The proof of this is simple relatively. Let's assume that data is uniformly distributed from 0 to 1 along each X dimension, and that the distribution in each dimension is independent. If we choose an arbitrary data point and draw a circle of radius d around it, the area of that circle is the fraction of the space that is a given distance from the data point. The formula for the volume of a Euclidean hypersphere with radius d and dimensionality m is:

$$V_m(d) = \frac{\pi^{m/2}}{\Gamma\left(\frac{m}{2}+1\right)} d^m. \tag{2.78}$$

Due to the gamma function, it is difficult to gain intuition about the behavior of the volume from the formula alone. This becomes more apparent when we plot $V_m(d)$ as a function of m (see Figure 2.22). From this figure, it is clear that the volume decreases more than exponentially as d increases. The reason for the supraexponential increase is the combination of an exponential function with a factorial. While we do not show it here, this proof is not restricted to Euclidean distance or uniformly distributed X. As long as each dimension of X is independent, the result would be the same, but the integrals would be (much) harder.

Figure 2.22: An illustration of the fraction of the total space taken by hyperspheres of various radii as the dimensionality of the input data increases. Observe that the fraction of the total space taken decreases more than exponentially, as seen by the slight positive inflection on these trend lines.



The impact of the CoD is that if each X dimension was independent, the only way to combat this supraexponential increase in distance between the points is to increase the number of data points supraexponentially, to achieve the same sampling density as would be achieved in a lower dimensional space.

If, however, the data point lie in a lower dimensional subspace or manifold within the high dimensional space, then the CoD does not hold. The premise behind feature selection methods is to find this lower dimensional space, or limit the search region to a smaller section of the full X space. For the sake of this work, we split feature selection into the following categories: (1) subsampling, (2) filter methods, (3) projection methods, (4) iterative methods, (5) regularization and Bayesian priors and (6) biological priors. We note that this is, by no means, a comprehensive review of feature selection methods, which is a field unto itself [98]. After reviewing key examples of each of these methods, we will discuss briefly a couple key challenges in the implementation of feature selection methods, namely, the optimization of hyperparameters involved in each method and the need for careful cross-validation.

2.5.1 Subsampling

Feature selection through subsampling was discussed when we discussed random forests. In random forests, the exemplars and the features are sampled with replacement such that the artificial datasets match the size of the original dataset. When making the artificial datasets, we can choose the dimensionality of X in each. Even if we choose X to be of the same dimensionality as the original datset, the double sampling that occurs due to sampling with replacement causes the functional dimensionality to decrease. If each, or a large fraction of, original input features held predictive information, then limiting the dimensionality within each tree would improve the tree's ability to model that predictive information. If, however, a sparse subset of X held predictive information, and a limited size tree was used such that at least part of that subset was not included in each tree frequently, then the majority of the trees would be modeling noise, resulting in an ineffective prediction.

2.5.2 Filter methods

One of the most popular methods of feature selection is filter feature selection. In this case, we set a quantifiable criterion that ranks or orders the features. We then choose to include the top F of these features in the subsequent machine learning classifier training. The remaining features are left out. While this decreases the dimensionality of the dataset that is seen by the classifier, it does not obviate the CoD. If the filter feature selection is applied within each cross-validation fold (and it should be), then the identity of the F features selected will not be consistent. To determine the predictive ability of any feature, one must consider both how frequently it was selected, and how much it was utilized when it was selected. Just because a feature was ranked highly by the filter does not mean that the machine learning classifier, in fact, used that feature to predict the outcome. Additionally, if we choose a filter feature selection that ignores the dependency and inter-relatedness between the features, then the subsequent classifier will not have access to this dependency or potentially informative interaction, which we discuss more below.

The simplest and most popular filter is the ANOVA or mass univariate t-statistic filter where we calculate the absolute value of the t statistic for each input feature. We then include features with the highest magnitude of t, or choose a univariate significance cut-off that a feature must pass to be included. The difference between these two criterions is practical, so we view them as duals of each other (i.e. functionally equivalent). We note that this filter is different from stepwise regression (see iterative methods below). While this filter is simple and easy to apply, it also ignores the potentially rich correlation structure within the input data and biases towards features that are discriminative on their own. Therefore, after this type of selection, one should utilize simpler linear models of X, like a decision tree, LDA, QDA or logistic regression, where each variable is considered independent. The t statistic is linear inherently; therefore it will be less sensitive to non-linear information. If a neural network is applied after this mass univariate t-statistic filter, then it will look for interaction effects that were ignored in the initial filter. While it may be feasible to suggest that we are only interested in interaction effects when main effects also are present, the most discriminative information may be held within higher level interactions or when sources of noise are controlled for, as we discuss in Kerr *et al.* [99].

An example of a filter feature selection that explicitly incorporates the dependency within X is the minimum redundancy, maximum relevancy toolbox (mRMR) [5, 6]. This is applied through either forward or backward selection. We describe the forward selection algorithm. In mRMR, we rank features based on their mutual information with the predicted class labels, with a penalty for sharing information with higher ranked features, as such:

$$Score_{X^{(r+1)}} = \max_j MI(Y, X_j) - \frac{1}{r} \sum_{X_k \in X^{(r)}} MI(X_k, X_j) \tag{2.79}$$

Where $MI(A, B)$ is the mutual information between A and B, and $X^{(r)}$ is the set including the top r ranked features. The function of this criterion is to maximize the independence within the modeled set, with the hope that the non-redundant information in the lower ranked features holds no predictive information. If we apply this type of filter, then assumptions of independence between the modeled features are more likely to be true. Additionally, the non-linearity of the mutual information allows for us to consider non-linear relationships between Y and X_j, but when $X_j \in \mathbb{R}$ then we must quantize or smooth X_j prior to calculation of the mutual information. This leads to another free hyperparameter that can be optimized.

2.5.3 Projection methods

The next version of feature selection is projection methods, where we search for a lower dimensional projection of the data that holds the predictive information.

The general form of a linear projection is:

$$Z = XA \tag{2.80}$$

Where Z is an n by p matrix, X is our original n by m matrix, and A is a m by p projection matrix. The difference between the many linear projection methods is their criterion for finding A.

The canonical projection method is principle component analysis (PCA) or singular value decomposition (SVD), which are identical for our purposes. PCA seeks to find a set of orthogonal axes within the space that capture the maximum variance within the higher dimensional set, irrespective of the source of that variance. This is equivalent to finding the eigenvalues and eigenvectors of the dataset, defined by:

$$D = U^T XV \tag{2.81}$$

where D is a diagonal matrix of eigenvalues, and U and V are unitary n by p and m by p matrices, respectively. When using this as feature selection, we choose the F eigenvectors with the F highest magnitude eigenvalues for inclusion into subsequent modeling. This assumes that all other dimensions of variation are noise. It is important to note that scaling of individual X_j will result in larger eigenvalues along that axis, so all X_j must be scaled similarly prior to PCA.

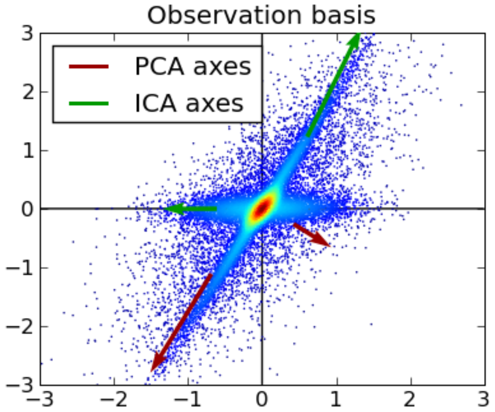
While the choice of F can be arbitrary, in some settings there is solid statistical theory to suggest an effective choice. Suppose that there exists a lower dimensional space with meaningful information, and all subsequent dimensions are noise. Because we rank the eigenvalues, the eigenvalues from noise dimensions should grow steadily as the rank increases. However, when an inflection point is reached where the magnitude of the eigenvalue is suddenly higher, then that suggests that there is another source of variation that the eigenvector is capturing. If this alternate source is not noise, then it could be signal. Therefore, if there is an inflection point in the curve of the magnitude of eigenvalues, then F could be selected to include all principle components with rank higher than that inflection point.

The challenge in PCA is that because of the quadratic loss function inherent in maximizing the variance, almost every X_j has a non-zero projection onto each PCA. Therefore interpreting the principle components (PCs) can be difficult. That has led to the development of more sparse methods of linear projection, where sparse refers to the number of X_j that contributes to each projected component.

In neuroimaging, the most popular projection method outside PCA is independent component analysis (ICA). Instead of finding mathematically orthogonal components as PCA does, ICA finds statistically independent components that maximize the variance captured in the data [100]. This can result in non-orthogonal components that are oriented along the major axes of variation in the data (see Figure 2.23). Although there are a number of definitions for statistical independence in this setting, the fast ICA toolbox (`http://research.ics.aalto.fi/ica/fastica/`) is the easiest to implement. The effect of using statistical independence as compared to mathematical orthogonality is that a larger number of X_j have a zero-weighted projection onto each independent component (IC). In neuroimaging, this has resulted in our ability to identify that particular ICs correspond to particular networks of anatomical regions. When a high number of ICs are fit, biological prior beliefs of the shape and structure of ICs that correspond to "real" signal can be used to exclude or filter out "noise" ICs [101, 102, 103, 100, 104, 105, 106].

The selection of how many components to include, F, is even more challenging in ICA than it was in PCA. It is trivial to show that PCA is nested: when a smaller F is selected, it causes no change to the highest ranked components. ICA is not nested: selecting a smaller F results in changes to each of the components.

Figure 2.23: An example of the PCA (red) versus ICA (green) projection of input data. Observe that the first components of ICA and PCA are identical, whereas the second components vary. The second IC reflects the generative process of the data by identifying the next largest statistically independent component of the data, whereas the second PC is required to be orthogonal to the first. Figure reproduced from gael-varoquaux.info.



Additionally, the rank of ICs is not stable for the same F, but the identity of the ICs are. Therefore, F needs to be selected prior to ICA and the argument of the inflection point in variance cannot be easily applied.

If your input data allows it, even more sparsity can be applied through the utilization of non-negative matrix factorization (NNMF). Similar to ICA and PCA, NNMF finds a linear projection from the original X space to a lower dimensional space, with the strong prior assumption that noise results in zero or near zero values, and signal results in positive values in X. It is required that all elements of X are non-negative. Similar to ICA, there are multiple algorithms to implement NNMF that are based on optimizing subtly different criterion that are the outside the scope of this work. It is sufficient to note that, similar to ICA, NNMF is not nested, and can provide an even high degree of sparsity on how many X_j contribute to each component. NNMF has shown to be highly effective in natural language processing (NLP) applications [107], as well as in some neuroimaging applications [108].

The last two projection methods that we will discuss are non-linear projection methods called local-linear embedding (LLE) [1] and multidimensional scaling (MDS). The premise behind these methods is that there exists a manifold within the X space that may or may not be captured by a linear projection. This presumes that it is the distance along this manifold that is predictive, instead of the distance traveled in the original space.

In local-linear embedding (LLE), [1] we calculate the pairwise distance between all data with respect to some norm. We then place exemplars on a new, lower dimensional space such that these pairwise distances are maximally preserved, according to the following loss function:

$$\min_{W} \sum_{i=1}^{n} \| X_i - \sum_{j} W_{ij} X_j \|^2 \tag{2.82}$$

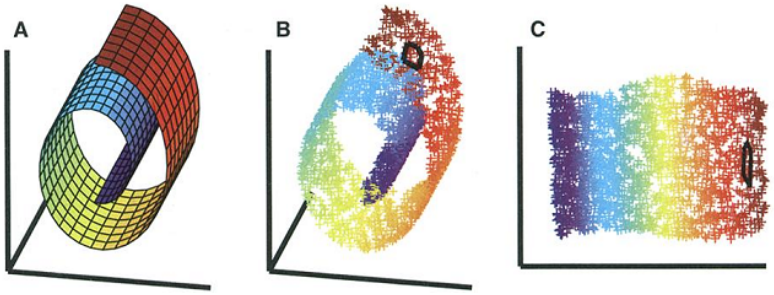
where j is the set of neighbors or local points, and W is a reconstruction matrix to predict X_i from its neighbors X_j. After learning W for set X, the final step in LLE is to map the higher dimensional X onto a low dimensional manifold, Z, such that the following loss function is minimized:

$$\min_{Z} \sum_{i=1}^{n} \| Z_i - \sum_{j} W_{ij} Z_j \|^2 \tag{2.83}$$

where W is fixed and Z are allowed to vary. This allows us to plot the higher dimensional data on a lower dimensional subspace, according to the learned manifold.

This assumes that the relationship between data is linear within the manifold. If we want to relax this assumption to suggest that within a certain distance, all non-linear spaces can be approximated by linear functions, we can implement a variation of LLE that only seeks to maintain the pairwise distances for local data (with some criterion). This perspective has the unique ability to model complex manifolds within the original space, like U and S shapes (see Figure 2.24), that would be otherwise unappreciated.

Figure 2.24: A) A two-dimensional manifold transformed into a three dimensional space. Distance along the manifold is meaningful, but distances based on three--dimensional coordinates have little to no meaning. B) Data sampled from this two-dimensional manifold. C) The data in B projected onto a two-dimensional space based on local-linear embedding. The black outlines in B and C reflect a neighborhood around a point in the original and projected spaces. Figure reproduced from Rowes & Saul [1].



Similar to LLE, MDS seeks to find a new set of input data that maintain the pairwise distance between points. (MDS is built into the statistics toolbox of MATLAB.) In MDS, one can choose the desired norm that should be used to calculate distance in the new space. If we seek to apply a particular norm during our machine-learning model (see below), then this guarantees that the input data obeys the assumptions behind that norm. However, this is because all information that is not captured by this norm is filtered out. Further, in non-metric MDS, we don't seek to match the continuous value of distance between points. Instead, we seek to maintain the rank of distances across points. Therefore, if the norm that we assumed in the original space was incorrect, but was correct up to the rank of the points, the relationship between points in the non-metric MDS would not be affected.

The limitation of both of these manifold learning methods is that the relationship between the Y and the original input data X is broken. If we seek to learn about the discriminative power of a given X_j or group of X_j, it is difficult to back-project from solutions in the manifold space onto the original X. Therefore, similar to the neural network classifier above: manifold learning can result in improved class predictions, but at the cost of the interpretability of the model.

2.5.4 Iterative methods

The next category of feature selection methods is iterative methods. Just as in the above methods, iterative feature selection is not incorporated into the likelihood or objective function fit by the classifier, but it uses the result of classifier learning to help select which features to include. The underlying assumption behind iterative methods is that there exists a subset of features that, when included in the model, result in the best performance. Therefore, we must search the combinatorial space of subsets of features to find this optimal model.

The simplest version of iterative methods is forward selection. Forward selection is initiated by fitting a univariate predictive model for each X_j and choosing the X_j that maximizes the performance of the model. Subsequently, models are built with this high ranked feature, plus each individual X_j that was not already included. The subset $X^{(2)}$ that results in the best performance is saved and the process is iterated until either (1) no improvement in performance is reached or (2) all features have been ranked. Note that both the number of features in the model, F, and the criterion that models were trying to maximize can be selected by the user.

Alternatively, models can be built using backward selection. In this case, a full model using all the data is trained. Then, models excluding one X_j of are trained. We again select the smaller model with the best performance and repeat this process until (1) no change in the performance is reached or (2) the model includes no X_j.

In some cases, backwards selection can be accomplished without fitting smaller candidate models. If we choose to model data using logistic regression, features can be ranked according to the significance of their odds ratio. Similarly, in SVM recursive feature elimination (SVM-RFE), features with zero or close-to-zero weights can be excluded [109, 110]. Both of these criteria allow for iterative exclusion of groups of X_j, in addition to just singular X_j. However, even though it is tempting to suggest that non-zero weighted features are related to the outcome class, this is not guaranteed. These features either provide predictive information, or control for important sources of noise that allow for more predictive information to be gleaned from other variables (see Haufe *et al.* [111] for more discussion). Additionally, when using a regularized model (see below), the risk of false-negative findings is high.

We note that each of these iterative methods do not fully search the combinatorially large space of all subsets of X. Instead, they provide heuristics for searching a reasonable subset of this space. There exist stochastic methods to more effectively search a larger number of the subsets, but those methods are outside the scope of our work.

2.5.5 Regularization and Bayesian priors

The most statistically appealing method of feature selection is the application or regularization or Bayesian priors about the nature of informative features. The difference between the terms "regularization" and "Bayesian priors" is semantic, and not based on differences in theory, as will become clear below. In contrast to all of the above methods, regularization and Bayesian priors are integrated into the likelihood and objection functions that we optimize using the training data.

By integrating feature selection into the function we optimize, we can study the sensitivity of our fitted model to these choices better, by considering the derivative of the solution with respect to the regularization term, and/or modeling the distribution of the solution with the regularization term. Determining significance and the effect of these terms is a novel and active area of research called post-inferential statistics.

First, we discuss regularization terms. Regularization terms apply the prior hypothesis that a small subset of X is informative, and the other terms are just noise. Therefore, most of the values of the β vector that weights the X should be zero or close to zero (i.e. β are sparse). We impose this hypothesis by saying, mathematically, that we prefer solutions that this is the case. Regularization can be applied to any machine learning classifier that optimizes a likelihood or objective function by optimizing the sum of the regularization penalty and the likelihood or objective function:

$$\ell_R(Y|X, \psi, \theta) = \lambda\|\beta\| + \ell(Y|X, \psi, \theta). \tag{2.84}$$

In this way, non-sparse solutions that produce high likelihood are penalized, resulting in a more sparse solution.

A consequence of applying this prior hypothesis is that stable, unique solutions can be achieved despite underdetermined data. If $n << m$, then many machine learning models discussed above do not have unique solutions. If a regularization term is applied, then the requirement of invertibility of the Hessian ($X^T X$) is not required because the solution is not based on inverting the Hessian. (Frequently, the solution relies on the inversion of the Hessian plus a term derived from the regularization penalty.)

Observant readers will remember that the original formulation of SVM includes an L_2 regularization term. This is included because when we consider just the support vectors when learning the maximally separating hyperplane, we are working essentially in a system where $n << m$. Therefore regularization is needed to guarantee the uniqueness of the solution.

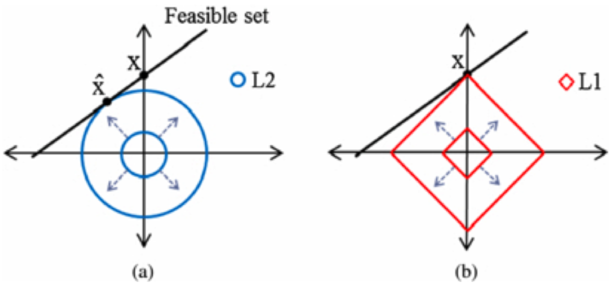
However, regularization has a tendency for false negative results. Consider an X_j and X_k that both have unique and shared predictive information. If the magnitude of the unique information in X_j or X_k is hard to distinguish from noise, then a regularized solution could include X_j but not X_k, or vice-versa. This is because most of the predictive information is in X_j. The addition of weight of X_k, given that X_j is included, may not result in enough improvement of the likelihood or objective function. Therefore, regularized solutions will identify sets of features that are predictive, but interpreting a weight of zero or near zero is difficult (see Haufe *et al.* [111] for why we say sets of features, as compared to individual features).

The easiest regularization penalty to optimize is the L_2 norm, because of its differentiability. While L_2 regularization is effective in a wide variety of applications, L_2 regularized solutions do not tend to place zero-weight on any X_j. Instead, small values are allowed.

Suppose that $\beta_j = 0.01$. In an L_2 norm, this would contribute a penalty of roughly $(\beta_j)^2 = 0.0001$. Therefore, small weights are no different than zero weights in terms of the final solution. If all features contribute to the end solution, then it is difficult to interpret which features are the most discriminative, especially when utilizing a kernel space in an SVM [111].

A solution to this problem of non-zero weights is to use an L_1 norm. A challenge behind an L_1 norm is that it is undifferentiable at zero. Because we seek for many values to be uniquely zero, this can be a large optimization problem. The functional effect of an L_1 norm is that it makes the w_j that were small, but non-zero, actually have a weight of zero. This results in more interpretable solutions because we can assess our performance as if we only sampled the features with non-zero weight. In fact, clever proofs from the compressed sensing literature have shown us that L_1 solutions are identical to L_0 solutions (see Figure 2.25). L_0 solutions seek to minimize the number of features with non-zero weight and ignore the actual weight assigned. This has a clear interpretation: L_1 and L_0 regularization results in the solution that includes the minimum number of X_j to achieve the given performance. Therefore, when performing a pilot analysis of a large number of features, with the hope of identifying a small set of highly important features, it may be attractive to apply L_1 or L_0 regularization even if it is not necessary for the uniqueness of the solution.

Figure 2.25: For linear solutions, there exist a set of solutions that result in the same log-likelihood (black line). The L_1 penalty chooses the solution with the minimum L_1, which also happens to be the solution with minimum L_0, in the vast majority of cases (b). In contrast, the L_2 and L_0 solutions do not coincide (a).



An important limitation to L_1 regularization is that the estimates it generates are biased towards zero. In other words, L_1 will underestimate the effect of the X_j that it gives non-zero weight, relative to when infinite data is available.

The last example of regularization terms we will discuss is a total variation (TV)-L_1 norm hybrid. While it is trivial to suggest that norms can be mixed through weighted sums, or norms other than the L_1 and L_2 can be applied, the TV penalty let's us apply a prior about the structure of X_j. If X_j are pixels or voxels (volumetric pixels) from an image, then a reasonable prior hypothesis is that nearby X_j could provide similar information. Therefore, the weight of X_j should not change drastically cross the image. This is achieved by applying the following TV penalty in one dimension:

$$\|\bigtriangledown\beta\|_1 = \sum_{j=2}^{m} |\beta_j - \beta_{j-1}| \,. \tag{2.85}$$

Both L_1 and L_2 regularization of logistic regression and SVM have been implemented within the LibLinear toolbox, available for C and MATLAB (`http://www.csie.ntu.edu.tw/~cjlin/liblinear/`, [112]). The TV penalty has been implemented within *scikitlearn*, a Python toolbox for machine learning (`https://github.com/nilearn/nilearn/pull/219`) [113, 114].

Lastly, more general Bayesian priors can be applied to likelihood functions. Suppose we apply the given prior on β: $P(\beta)$. Using Bayes formula, this is integrated into the posterior probability in the following way:

$$P(Y|X, \psi, \theta) = \frac{P(X|Y, \psi, \theta)P(Y)}{P(\psi, \theta)} \tag{2.86}$$

where β is an element of ψ, the set of parameters optimized using the objective function. As a reminder, θ are the set of hyperparameters that are not estimated jointly with the objective function. Since optimizing sums is easier than optimizing products, we can log-trans form this expression to generate log-posterior probability:

$$\log P(Y|X, \psi, \theta) = \log P(X|Y, \psi, \theta) + \log P(Y) - \log P(\psi, \theta). \tag{2.87}$$

The subtraction of $\log P(\psi, \theta)$ is eerily similar to the addition of a penalty term to the objective functions listed above, when we recognize $\log P(X|Y, \psi, \theta)$ as the log-likelihood. Therefore, the regularization terms discussed above can be expressed in terms of prior hypothesis on ψ and θ.

One example of a Bayesian feature selection prior is a spike and slab prior. For each β_j, we place a large spike of probability mass at $\beta_j = 0$. The rest of the probability mass we spread uniformly (or with a slow decay constant) on the rest of the possible range of β_j. Consequentially, most β_j will have zero weight, and some β_j will have non-zero weight.

The challenge to the application of Bayesian methods is their computational intensity. Even though there is a strong statistical basis for their construction, only a limited number of posterior probabilities have analytical solutions to their optimization. The most popular method to find the optimum of Bayesian expressions is through Markov chain Monte Carlo (MCMC) simulations, generally using Metropolis Hastings sampling (i.e. OpenBUGS software). For even moderately sized problems, these MCMC simulations can be prohibitively slow to converge. Novel hardware and software is being developed to address these limitations.

2.5.6 Biological priors

The last feature selection method is important to discuss, even if it has no mathematical or statistical motivation. If we have prior biological knowledge of the system that we are studying, applying that knowledge to assist in feature selection can result in improvements in performance [3]. For example, if we are studying the difference between left and right temporal lobe epilepsy, we can hypothesize that the strongest signal is going to in the temporal lobes, and/or the difference between the temporal lobes. (Note that based on current research, including our own, [7, 81] this is not necessarily a good hypothesis.) This hypothesis can be applied by considering only the X that consist of features from those regions.

We highlight that some biological hypotheses can be applied within a statistical framework. In particular, we know that human brains are organized into similar functional regions; therefore the contribution of nearby regions should be similar. This hypothesis matches the assumptions behind TV regularization, therefore TV may be effective for neuroimaging data (as in Dubois *et al.* [114]).

As another example, the structural changes due to dementia may be focused in particular regions, but they also involve wide regions of the brain. Therefore, when predicting class information about dementia using brain information, an L_2 prior that considers small but dispersed contributions may be more appropriate. In contrast, some epilepsies are due to highly focal lesions. When predicting class information about these, an L_1 prior may be more effective in eliminating the large amount of irrelevant information while focusing on a small number of informative features.

A common theme through each of these feature selection methods is that they require the selection of seemingly arbitrary hyperparameters, θ. With the exception of Bayesian priors, these θ are not explicitly written into the likelihood or objective function. Therefore, it is difficult to balance the choice of θ with the goodness of fit of the model, or understand the sensitivity of the model to changes in θ, which otherwise could be done by differentiating the likelihood or objective function with respect to θ. The selection of these θ and understanding the sensitivity of the performance to these choices is discussed in chapters 11 & 12 in detail.

The other important caveat to feature selection methods that we must re-emphasize is that they are partial solutions to the curse of dimensionality, but they are no means full solutions in all situations. The addition of free parameters and hyperparameters also increases the propensity to overfit the data, especially when limited data is available. Therefore, feature selection must be applied within each cross-validation fold so that we can ensure that the accuracy of the estimates of how well the models generalize to out-of-sample data are preserved. If a general feature selection is done prior to the cross-validation split, then information from the validation data can bleed into the training solution. Therefore, unless it is impossible to split feature selection from cross-validation, we advise that it should be done. This is the case even for unsupervised feature selection methods.

2.6 Population Selection in the Development of CADTs

The key to developing clinically applicable CADTs is the identification and detailed description of the diagnostic challenge, the clinical context, the patient population, and the data available for study. Even as mathematicians and statisticians, it is critical to have a close collaboration with clinicians that are familiar with the practical challenges inherent in the desired application.

The first and most important part of a CADT is to define clearly the diagnostic challenge and the appropriate control population. One example of this is in the early diagnosis of Alzheimer's dementia (AD). Once clinical signs and symptoms of dementia are apparent, the pathologic challenges responsible for the dementia may be irreversible. If patients at high risk for AD could be identified prior to onset of clinical signs, then interventions could be developed prior to the onset of permanent damage. Therefore, CADTs for AD focus on the prediction of later development of clinical signs of dementia in normal or high-risk populations. Based on this insight, diagnostic studies focus on discriminating age-matched controls from AD, stable mild cognitive impairment (MCI) and progressive MCI. Patients with MCI have early signs of dementia, but have not yet progressed to overt AD. The discrimination between controls and AD serves as a proof of concept that the classification scheme is effective. The discrimination between progressive and stable MCI reflects the desired application of the CADT. The Holy Grail would be to identify patients at high risk for AD prior to the development of clinical signs. This Holy Grail study requires long-term follow up of a large population of health patients, with the knowledge that some would develop AD.

The diagnostic challenge in seizures is different from dementia because physicians are seeking to determine the etiology of seizures that occur prior to assessment. Therefore, comparison of patients with epilepsy to seizure-naive controls does not address the diagnostic challenge at hand because clinicians would never consider epilepsy as a potential diagnosis in patients who have never experienced a seizure-like event. In our opinion, there are two potential control populations. If we seek to diagnose epilepsy as early as possible, we should compare patients with epilepsy to patients who experience an isolated or provoked seizure but do not progress to experience repeated, unprovoked seizures, which is the definition of epilepsy. If we seek to diagnose the etiology of repeated seizures, these patients with isolated seizures are irrelevant. Instead, we should compare to patients that experience NES (see chapter 1). In contrast to the patients with isolated seizures, patients with NES are treated as if they have epilepsy until they are diagnosed. Consequentially, their exposures to ASMs and other iatrogenic exposures match patients with epilepsy [8, 81]. Therefore, when developing CADTs for epilepsy, we believe that the first, critical stage is to differentiate epilepsy from NES.

This mirrors the clinical question of interest because both populations have a prior history of seizures, but complicates the interpretation of the learned models. Any observed differences reflect the difference between patients with NES and ES, as compared to just the difference attributable to epilepsy. Therefore, without other corresponding evidence, changes could be due to ES, NES or a combination of both conditions.

For example, in the chapter 9, [7] we find that metabolic alterations in sensorimotor cortex, which we interpret as a novel finding suggestive of extratemporal changes due to temporal lobe epilepsy (TLE). This was based on the supposition that patients with NES have no focal lesions visible by FDG-PET. Recent research, however, found reductions in cortical thickness in sensorimotor cortex in patients with psychogenic non-epileptic seizures compared to healthy controls [115]. This reduction in cortical thickness would likely be accompanied by a concomitant reduction in glucose metabolism. Therefore, we should have interpreted our change in sensorimotor cortex as at least in part due to metabolic changes in NES.

This reasoning can be generalized beyond Alzheimer's, psychogenic seizures or epilepsy. We argue that the proper control for a novel CADT is the differential diagnosis (DDx) for the signs or symptoms of interest. In clinical medicine, the term "symptoms" reflect observations made or reported by the patient, witnesses or their caregivers. The term "signs" reflect objective observations made by a health care provider. Lastly, a DDx reflects all of the potential conditions or syndromes that can result in the sign or symptom. These lists are well defined in the medical literature and, frequently, are long because they include all possibilities, no matter how unlikely it is. Note that this requires us to recognize signs or symptoms to address, as compared to conditions, the latter of which is the conventional organization of research and medicine.

A relevant example of a differential diagnosis is the DDx of episodic, self-resolving loss of awareness, responsiveness or consciousness is as follows:

1. Epilepsy
    (a) Generalized-onset seizures
    (b) Focal dyscognitive seizures
2. Psychogenic non-epileptic seizures
3. Physiologic non-epileptic seizures
    (a) Complex migraines
    (b) Confusion episodes in dementia
    (c) Syncope
    (d) Narcolepsy or cataplexy
    (e) Hyperventiliation syndrome
    (f) Movement disorders

4. Provoked causes
    (a) Stroke
    (b) Temporary Ischemic Attacks (TIA)
    (c) Myocardial infarction (MI)
    (d) Polypharmacy
    (e) Medication or substance effect
5. Malingering
6. Munchausen's Disease

The sheer number of conditions on a differential diagnostic list is long, requiring sufficient patients in each group to be able to characterize the trends of the input data within the group. The sheer number of patients required for this can be prohibitive, so initial CADTs can and should address the most salient facets of this differential. For example, much of our work focuses on answering the following question: who will benefit from treatment with ASMs? This simplifies the differential into two groups, at the cost of a certain level of detail.

Alternatively, one can utilize data from an unselected population of patients that present with episodic, self-resolving loss of awareness, responsiveness or consciousness. Some groups of patients will not be represented well in this dataset because they are rare, compared to the other groups. This naturally will prioritize research towards the most prevalent groups. If, however, one seeks to study a more rare group, then the minimum size of the database will be defined by how many exemplars of the rare group of interest are expected to be present, for a given size.

For example, if we seek to study psychogenic NES, we know from previous literature that psychogenic NES comprises 30% of patients admitted to vEEG.

If we want a minimum sample of 100 patients with psychogenic NES, then we would need to collect data from at least 334 patients admitted to vEEG. If we seek to study malingering or Munchausen's Disease, we recognize that roughly 5 of every 300 patients with NES are malingering (estimate from UCLA Seizure Disorder Center data). If we want at least 100 patients who are malingering or have Munchausen's Disease in our dataset, then we need at least 20,000 patients admitted to vEEG. While these sample size calculations are important for experimental design, they also are important concerns about implementation. If we expect to apply our CADT to hundreds of patients, then it may be feasible to ignore the possibility of malingering or Munchausen's Disease in favor of identifying psychogenic NES. If, however, we propose that our CADT should be applied nationwide to the hundreds of thousands of patients per year that present with episodic loss of awareness, then a non-negligible number of these patients are malingering or have Munchausen's Disease. Therefore, it would be inappropriate for us to not attempt to describe that population.

Therefore, when seeking to develop a clinically relevant CADT, one must consider the specific clinical question that the CADT seeks to assist with. For diagnostics, this question frequently refers to a differential diagnosis of signs or symptoms present in a patient. Prior to implementation, the CADT should be trained on an appropriate amount of data to study the population of interest.

CHAPTER 3

The Future of Medical Diagnostics: Large Digitized Databases

This chapter is a reproduction of our work published in the Yale Journal of Biology & Medicine.[89] This work includes contributions from Edward P. Lau, Gwen E. Owens and Aaron Trefler. WTK initiated, organized and wrote the majority of this article. EPL created the third figure and contributed significantly to sections regarding computational efficiency, the structure of databases and the content strategy problem. GEO created the first and second figures, as well as the first table, and contributed significantly to sections regarding patient and physician attitudes toward databases and computer-aided diagnostics. AT contributed by conducting substantial literature review to support the ideas expressed.

3.1 Abstract

The electronic health record mandate within the American Recovery and Reinvestment Act of 2009 will have far reaching impacts on medicine. In this article, we provide an in-depth analysis of how this mandate is expected to stimulate the production of large scale digitized databases of patient information. There is evidence to suggest that millions of patients and the NIH will fully support the mining of such databases to better understand the process of diagnosing patients. This data mining likely will reaffirm and quantify known risk factors for many diagnoses. This quantification may be leveraged to further develop computer-

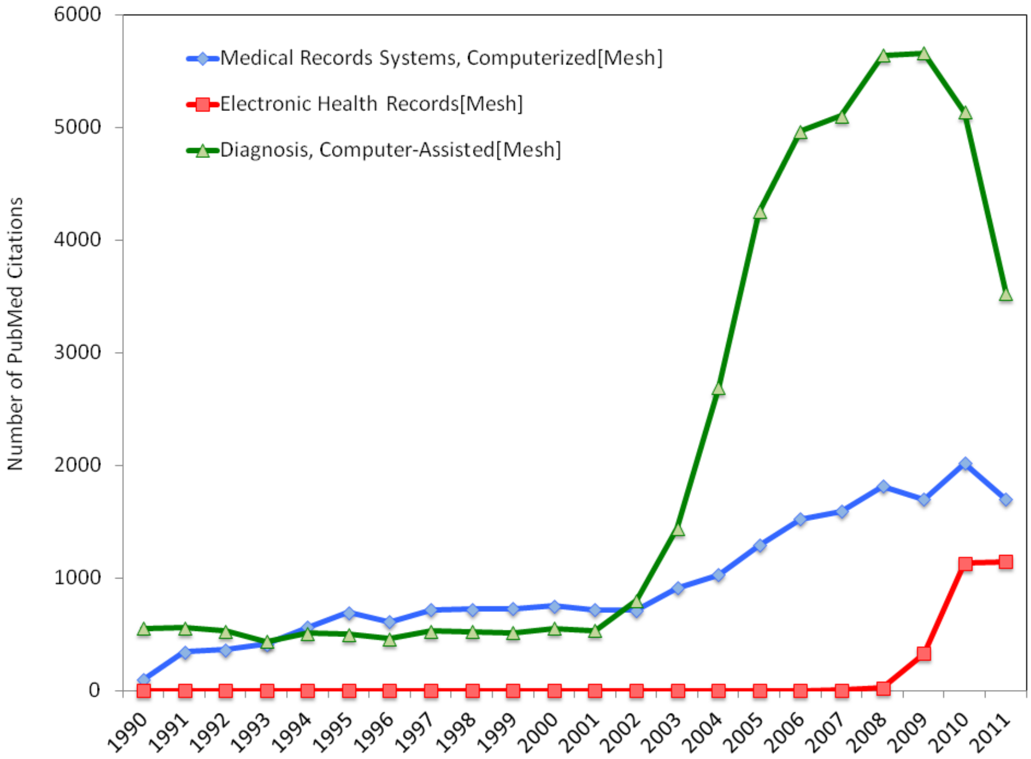
aided diagnostic tools that weigh risk factors and provide decision support for health care providers. We expect that creation of these databases will stimulate the development of computer-aided diagnostic support tools that will become an integral part of modern medicine.

3.2 Introduction

The impact of nationwide implementation of electronic health record (EHR) systems will change the daily practice of medicine as we know it. With medical records in their current state, it is extremely difficult to efficiently collate records and mine clinical information to understand trends in and differences between various patient populations. This limits the size of patient groups and thereby reduces the statistical power of many research protocols [62]. The EHR mandate will stimulate institutions to digitize their records in common formats that are amenable to collating data into large databases. These databases with records from potentially millions of patients can then be processed using sophisticated data mining techniques. There are numerous regulatory, practical and computational challenges to creating and maintaining these databases that will need to be appropriately addressed. Many groups are already compiling large databases of high quality patient information with great success [116, 117, 118, 119, 120, 121, 122, 123, 124]. Based on its previous efforts, we expect the National Institute of Health (NIH) to fully support researchers that seek to tackle the challenges of creating EHR-based databases that include clinical notes and other data points such as laboratory results and radiological images. Such databases will be invaluable to the development of computer aided diagnostic (CAD) tools that, we believe, will be responsible for many advances in the efficiency and quality of patient care[62]. CAD tools are automated programs that provide synthesized diagnostic information to providers that are not otherwise readily accessible. The rate of development of

CAD tools and the mining of medical record systems has increased markedly since 2002 (Figure 3.1) and we expect the development of large EHR-based databases will only stimulate this activity further [2][2][2][2]. In this article, we provide an in-depth analysis of the effect of the EHR mandate on the development of databases that could be mined to create high quality CAD tools. Further, we illustrate how computer aided diagnostics can be integrated efficiently into daily medical practice.

Figure 3.1: This figure illustrates the number of PubMed citations using each of the Mesh terms listed. Since 2002, the number of publications regarding computer-aided diagnostics has increased substantially. We are already seeing a commensurate increase in the number of publications regarding computerized medical record systems and electronic health records.[2]



3.3 Mandates and Policies Driving the Change

Although the growth of large digitized databases is stimulated by numerous sources, there are two key policy decisions that have the potential to dramatically speed this growth and change medical diagnostics as we know it. These key policies are the final NIH statement on sharing research data in 2003 and the EHR mandate in the American Recovery and Reinvestment Act of 2009 (ARRA) [125, 126][125, 126][125, 126][125, 126]. The seed for developing large open databases of medical information began was planted initially by the NIH statement on sharing research data. In 2003, the NIH mandated that investigators submitting an NIH application seeking $500,000 or more in direct costs in a single year are expected to include a plan for data sharing [126][126][126][126]. A large portion of academic medicine research is funded through grants of this type and therefore the amount of high quality information about patients in the public domain is growing rapidly. This may be one reason why interest in computerized medical record systems increased in 2003 (Figure 3.1). Unfortunately, the NIH has identified that this policy has not led to the degree of data sharing they anticipated, as evidenced by NOT-DA-11-021 entitled "Expansion of sharing and standardization of NIH-funded human brain imaging data" [127]. The focus of this request for information (RFI) was to identify the barriers to creating an open-access database for brain imaging data, including medically relevant images. This RFI implies that the NIH likely would support efforts to establish large open, digitized databases that include patient information.

Those who designed the ARRA presumably recognized the potential of digitized medicine and decided to support its development. In the ARRA 20 billion dollars was provided to establish EHRs for all patients treated in the US [125]. Healthcare providers that do not establish an EHR system after 2014 will be subject to fines. This was intended to further stimulate the trend of increased utilization of EHR systems (Figure 3.2). As stated in the bill, the reasons for this mandate include reduction of medical errors, health disparities, inefficiency, inappropriate care and duplicative care. Further, the ARRA EHR mandate has and is meant to improve coordination, the delivery of patient-centered medical care, public health activities, early detection, disease prevention, disease management and outcomes [128, 125]. To facilitate these advances, the knowledge about and methods for bioinformatics must be applied to millions of EHRs to develop automated, computer aided diagnostic (CAD) tools. For example, one efficient way to avoid inappropriate care is for an automated program to produce an alert when a health care provider attempts to provide questionable service. The development of such CAD tools is not trivial; however large, high-quality, open EHR databases will greatly decrease development costs and accelerate testing. Below we discuss why it is our firm belief that these databases will make the implementation of computer-aided diagnostics virtually inevitable.

3.4 Large Databases

There are a growing number of these large databases populated with clinically relevant information from patients suffering from a diverse range of medical conditions, some already including detailed multimodal information from hundreds to millions of patients. Here we will briefly review the General Practice Research Database (GPRD), the Alzheimers Disease Neuroimaging Initiative (ADNI), the Personal Genome Project (PGP), the European Database on Epilepsy (EDE) and the Australian EEG Database. These and other databases are summarized in Table 3.1.

The GPRD includes quality text based records from over 11 million patients primarily from the UK but also includes patients from Germany, France and the US [118, 129]. The database is used primarily by pharmacoepidemiologists

Figure 3.2: Even before the ARRA in 2009, the number of physicians utilizing EHR systems was increasing. There are already a substantial percent of physicians using electronic records. Consequentially, it is relatively inexpensive to combine and mine these EHR systems for high quality clinical information.

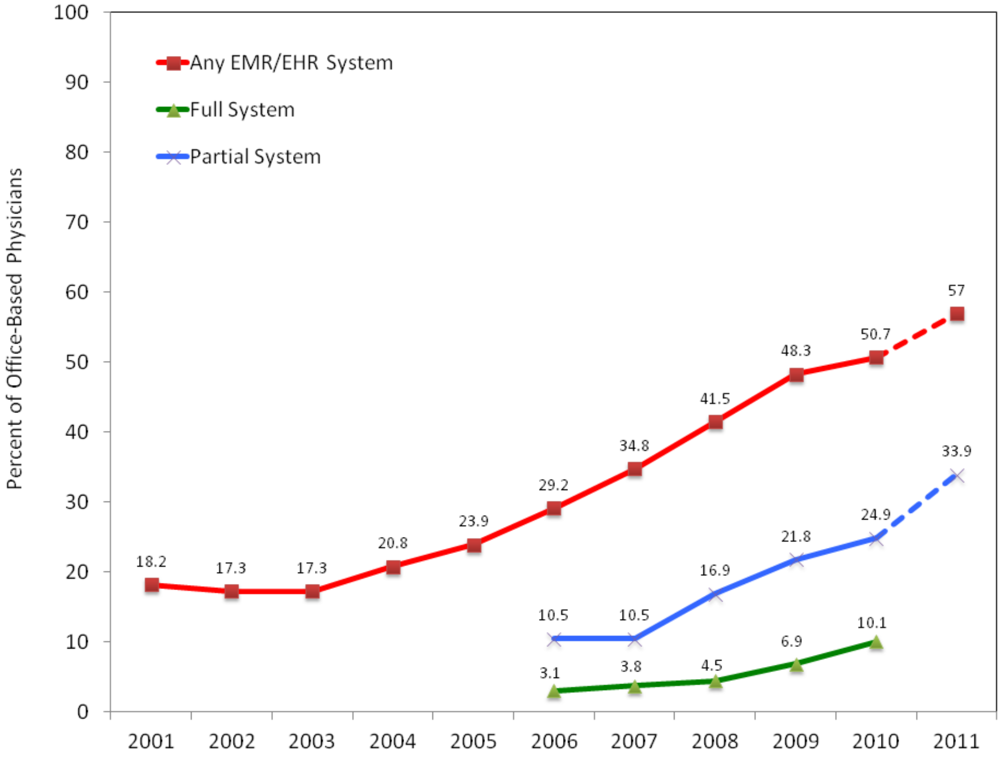


Table 3.1: A quick summary of notable databases of high quality information that have been developed and are being used for large scale studies.

Database	Information Contained	Funding Source(s)	Access	Website
ADHD-200	776 resting-state fMRI and anatomical datasets along and accompanying phenotypic information from 8 imaging sites; 285 of which are from children and adolescents with ADHD aged 7-21	NIH	Research community	fcon_1000.projects.nitrc.org/indi/adhd200/index.html
Alzheimer's Disease Neuroimaging Initiative (ADNI)	Information on 200 control patients, 400 patients with mild cognitive impairment, and 200 with Alzheimer's disease	NIH	Public access	www.adni-info.org/
Australian EEG Database	18,500 EEG records from a regional public hospital	Hunter Medical Research Insitute and the University of Newcastle Research Management Committee	User access required (administrator, analyst, researcher, student)	aed.newcastle.edu.au:9080/AED/login.jsp
Clinical Trials	Registry and results of >100,000 clinical trials	NIH	Public access	clinicaltrials.gov/
Epilepsiae European Database on Epilepsy	Long-term recordings of 275 patients	European Union	Research community	www.epilepsiae.eu/
Healthfinder	Encyclopedia of health topics	Department of Health and Human Services	Public access	healthfinder.gov/
Kaiser Permanente National Research Database	Clinical information on >30 million members of the Kaiser Foundation Health Plan	Kaiser Foundation Research Institute	Kaiser Permanente researchers and collaborating non-KP researchers	www.dor.kaiser.org/external/research/topics/Medical_Informatics/
National Patient Care Database (NPCD)	Veterans Health Administration Medical Dataset	U.S. Department of Veterans Affairs	Research community	www.virec.research.va.gov/DataSourcesName/NPCD/NPCD.htm
Personal Genome Project (PGP)	1,677+ deep sequenced genomes. Goal is 100,000 genomes.	NIH and private donors	Open consent	www.personalgenomes.org/
PubMed	Article titles and abstracts	NIH	Public access	www.ncbi.nlm.nih.gov/pubmed/

though other researchers are mining this database actively to create automated tools that extract, at base, the diagnostic conclusions reported in each note [130, 131]. Although the recall and precision of these tools was good–86% and 74% respectively in one study [131]these tools are constantly improving. We expect the increasing size of this and other databases will further stimulate high quality research in this field and result in highly efficient and effective data extraction tools. This conclusion is supported by the fact that over 73 scholarly publications utilized the GPRD in the first three quarters of 2011 alone [129]. This database, however, is limited to the text of the clinical notes.

Other databases go further by providing complex data regarding large cohorts of patients. The ADNI database contains data fields that track the rate of change of cognition, brain structure and function from 800 patients, including 200 with Alzheimers disease (AD) and 400 with mild cognitive impairment (MCI) [124]. Researchers are planning to add 550 more patients to this cohort in ADNI2 [117]. The current ADNI database includes full neuroimaging data from all of these patients in the hope that this data can be used to discover the early warning signs for AD. ADNI has been used already to develop machine learning (ML) tools to discriminate between AD and normal aging . Another database compiled by the PGP currently has 1,677 patients and researchers plan to expand this to include nearly complete genomic sequences from 100,000 volunteers using open-consent [121]. Researchers involved in the PGP anticipate that this sequence information will be used to understand risk profiles for many heritable diseases [119]. Other similarly large databases of complex data already exist; the EDE contains long-term EEG recordings from 275 patients with epilepsy [116, 123] and the Australian EEG Database holds basic notes and full EEG results from over 20,000 patients [120, 122]. These databases have been used to develop sophisticated seizure prediction and detection tools. Here at UCLA we are compiling a database of clinical notes, scalp EEG, MRI, PET and CT records from over 2,000 patients admitted for video-EEG monitoring.

The existence of these databases containing detailed clinically relevant information from large patient cohorts confirms that the international research establishment and the NIH are extremely excited about and supportive of large clinical databases. This suggests that as the EHR mandate simplifies collation of patient data, the limiting factor in generating large databases of thousands to millions of patient records will be for organizations to work through the practical hurdles of consenting patients and making data available for efficient searching and processing.

3.5 Anticipated Challenges to Database Creation

Our conclusion that large clinical databases will continue to expand is based on key assumptions that important regulatory and computational hurdles will be overcome. These challenges include, but are not limited to: (1) patient consent, (2) IRB approval and (3) consistent improvements in processing these large datasets. We believe the probability that these potential problems will be solved is high.

Forming open databases requires that patients consent to the sharing of pertinent parts of their medical records. In the development of the Personal Genome Project (PGP), Church et al. established open-consent so that all de-identified records can be shared freely [121]. Patients in EHR databases would likely utilize an identical open-consent process. We have personal experience analyzing datasets that require consenting adult patients admitted for video-EEG monitoring for epilepsy as well as pediatric epilepsy patients undergoing assessment for resective neurosurgery at UCLA. After we explained that consent would have no impact on their care, every patient admitted for these reasons (716/716) consented to their records being used for research. Weisman et al. reported that 91% of respondents would be willing to share their records for health research and that most would be more comfortable with an opt-in system [132]. Other surveys of patients report a consent rate of approximately 40% for providing de-identified information to researchers [133, 134]. Even after consenting, patients are relatively uninformed about the safeguards and risks to sharing their health information [135]. A more detailed and careful explanation of these procedures and the potential impact of the research may result in an increased consent rate. Any national patient database is likely to face pushback from a public already concerned about invasions of privacy by corporations and the government, therefore we suspect consent rates would be lower than what we have experienced. Additionally, the rate of consent is likely to decline, in part, due to media coverage of previous unethical practices in research. A prime example is the novel, The Immortal Life of Henrietta Lacks, published in 2010 by Rebecca Skloot, that recounts how, due to lack of proper regulation in 1951, Ms. Lacks cells were immortalized without her consent and were used widely for important advances in medical research[136]. We expect that patients and regulators sensitive to the concept of information about their care being stored indefinitely for research use may not consent on the basis of this and other salient examples.

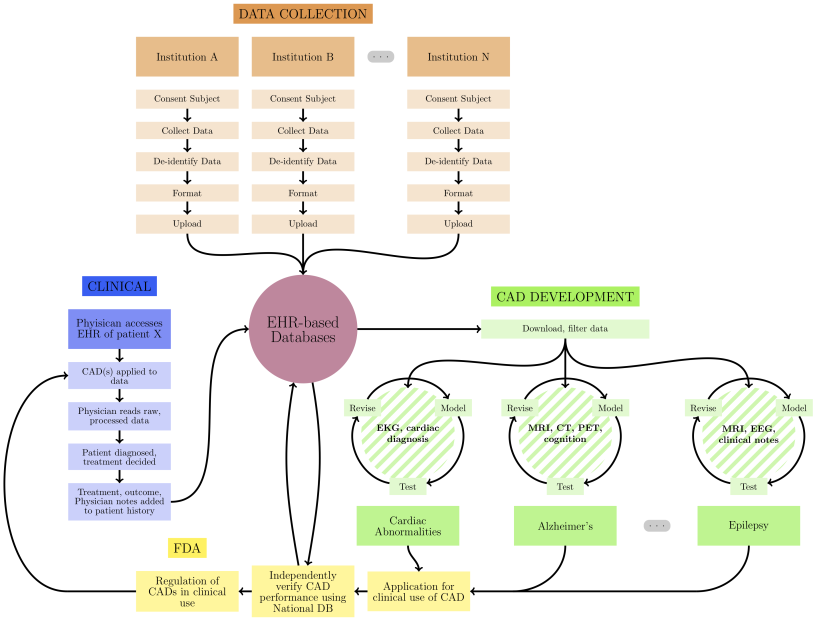
The key regulatory challenge to the creation of such large databases, however, is the complex multicenter IRB approval process. The most important concern that current IRBs have expressed is whether the data stream includes adequate de-identification of all records before they are released for research use, as illustrated in Figure 6.3. This would likely require each contributing institution to develop a reliable and consistent method of de-identifying all records. For written records, this includes removing all protected patient information (PPI), as defined by HIPAA regulations and the Helsinki Declaration [137, 138]. In order to do this effectively, numerous safeguards must be put in place. For example, if a nation-wide database is updated in real time malicious individuals could potentially re-identify individual patients by their treatment location, date and basic information as to what care they received. One solution to minimize these risks, suggested by Malin et al., is to granulize dates and treatment locations to ensure that the potential re-identification rate of patients remains well below 0.20%[135]. This granulation may also allow for inclusion of patients older than 89 years, the maximum reportable age under HIPAA regulations [137]. Although specific dates and locations are important, especially to the Center for Disease Control (CDC), simply generalizing days to months and towns to counties is required to maintain patient privacy. When dealing with more complex records as in neuroimaging, all centers would be required to be proactive in using the most up to date software for de-identification including, but not limited to, the removal of the bone and skin structure of the face that can be used to recreate an image of the patients face and thereby identify the patient. Automated software to do these complex steps has already been made publically available by the Laboratory of Neuroimaging (LONI) at UCLA [139]. Due to the unprecedented quality and applicability of these large databases, we are confident that responsible researchers will work to identify and address these regulatory hurdles.

Lastly, the computational burden of utilizing such large databases is not trivial. The question is not if mining this database is possible: it is when. Moores law has accurately predicted the biennial doubling of computer processing power [140] and, though this rate is showing signs of slowing, growth still is exponential [141]. Current ML methods have been effectively applied to the ADNI database of 800 patients [3, 142, 143] and as well as the GPRD of almost 12 million patients from the UK [129]. This suggests that if adequate computational technology does not already exist to effectively mine US-based EHR databases, it will be available soon.

3.6 Current Applications and Benefits of CAD

The application of CAD to patient data is not a novel idea. Numerous CAD tools have been demonstrated to be extremely useful to clinical medicine but few have been approved for routine clinical use [144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 62, 156, 157, 158]. In general, these tools attempt to predict the outcome of more expensive or practically infeasible gold standard diagnostic assessments. Humans are capable of weighing at most 30 factors at once

Figure 3.3: The creation and utilization of EHR databases is complex; however, each of the steps in the data and implementation system are well defined. We expect that responsible researchers will be capable of tackling each of these steps to create unparalleled databases and develop high quality, clinically applicable CAD tools.



using only semi-quantitative modeling [159]. The key exception to this is visual processing in which the visual heuristic reliably removes noise from images to readily detect the underlying patterns [160]. This exquisite pattern detection, however, is limited by our inability to detect relationships separated widely in space or time or whose patterns evolve out of changes in randomness. Further, human performance is highly variable due to the effects of expertise, fatigue and simply human variation [161]. Computational analysis, on the other hand, can integrate complex, objective modeling of thousands to millions of factors to reliably predict the outcome of interest [162]. During validation, the performance of a CAD tool is described in detail to understand its strengths and weaknesses. Unlike manual analysis, given a similar population of test samples, a CAD tool can be expected to perform exactly as it did during validation. In some cases, the constantly updating algorithms inherent in human decision-making may result in deviation from the previously studied ideal. It is not certain that this deviation always results in improved sensitivity and specificity. The cost of expert analysis of clinical information also is increasing continually. Effective implementation of automated screening tools has the potential to not only increase the predictive value of clinical information but also to decrease the amount of time a provider needs to spend analyzing records. This allows them to review more records per day and thereby reduce the cost per patient so that the effective public health impact of each provider is increased [163]. This will complement the numerous potential benefits quoted above. Here we review the success of implemented CAD tools and highly promising new tools that have demonstrated the potential for wider application. In particular, CAD tools have been applied to aid in the diagnosis of three extremely prevalent maladies in the US: heart disease, lung cancer and Alzheimers disease (AD).

The most widely recognized CAD tool in clinical medicine is built into electrocardiogram (EKG) currently available software and reads EKG records and reports any detected abnormalities. These algorithms are responsible for the life saving decisions made daily by automated electronic defibrillators (AEDs). The diagnosis of more complex cardiac abnormalities is an extremely active area of research [144, 164, 145, 147, 148, 165, 150, 151, 152, 153, 166, 155, 157, 158]. In one recent example, a CAD tool differentiated between normal beats, left and right bundle block (LBBB and RBBB), and atrial and ventricular premature contraction (AVP, PVC) with over 89% accuracy, sensitivity, specificity and positive predictive value [158]. This and other automated algorithms detect subtle changes in the shape of each beat and variations in the spectral decomposition of each beat over an entire EKG recording that often includes thousands of beats. As a result of this accuracy, conventional EKG readouts in both hospitals and clinics frequently include the results of this entirely automated analysis. When taught to read EKGs, providers are instructed that the automated algorithm is largely correct but to better understand the complex features of the waveforms, providers must double check the algorithm using their knowledge of the clinical context. This CAD tool was the first to be widely applied because, in part, EKG analysis is simplified by the presence of the characteristically large amplitude QRS wave that can be used to align each beat. Other modalities do not necessarily have features that are as amenable to modeling.

One example of overcoming this lack of clear features is the semi-automated analysis of thoracic x-ray computed tomography (CT) images to detect malignant lung cancer nodules. This tool segments the CT into bone, soft tissue and lung tissue then detects nodules that are unexpectedly radiolucent and assesses the volume of the solid component of non-calcified nodules [156]. This method effectively detected 96% of all cancerous nodules with a sensitivity of 95.9% and a specificity of 80.2% [156]. Even though this tool is not part of routine care, Wang et al. demonstrated that when radiologists interpret the CTs after the CAD tool they do not significantly increase the amount of cancer nodules detected [156]. In fact, they only increase the number of false positive nodules, indicating that the CAD tool is operating on meaningful features of the nodules that are not reliably observable even by trained radiologists. This suggests that in some cases, computer aided diagnostics can reduce the number of images that radiologists have to read individually while maintaining the same high quality of patient care.

The success of CAD tools in Alzheimers disease (AD) shows exactly how automated tools can utilize features not observable by trained radiologists by reliably discriminating AD from normal aging and other dementias. Because of its unique neuropathology, AD requires focused treatment that has not been proven to be effective for other dementias [167]. The gold standard diagnostic tool for AD is cerebral biopsy or autopsy sample staining of amyloid plaques and neurofibrillary tangles [167]. The clear drawback of autopsy samples is that they cannot be used to guide treatment and cerebral biopsy is extremely invasive. An alternative diagnostic is critical for reliably distinguishing between the two classes of patients at a stage that treatment is effective. In 2008, Kloppel et al. demonstrated how a support vector machine (SVM)-based CAD tool performed similarly to six trained radiologists when comparing AD to normal aging and fronto-temporal lobar dementia (FTLD) using structural magnetic resonance imaging (MRI) alone [168]. Numerous other applications of ML on other datasets all have achieved similar accuracies ranging from 85 to 95% [142, 169, 170, 143]. All of these tools do not require expertise to read; therefore they can be applied both at large research institutions and in smaller settings as long as the requisite technology is available. These tools, with appropriate validation using large databases, could indicate which patients would benefit most from targeted treatment and therefore substantially reduce morbidity.

These cases are exemplary; however, many other attempts to develop CAD tools have had more limited success. In particular, the automated analysis of physicians notes has proven particularly difficult. In a publication in 2011 using a total of 826 notes, the best precision and recall in the test set were 89% and 82%, respectively [171]. These values are extremely encouraging when considering a similar study in 2008 that attempted to measure the health related quality of life in 669 notes and achieved only 76% and 78% positive and negative agreement between the automated algorithm and the gold standard [172]. When viewing these accuracies in terms of the potential of applying these tools to patients, these accuracies are far from adequate. Physicians can quickly scan these notes and immediately understand the findings within them and therefore these CAD tools would not improve upon the standard of care if used to summarize the note. Nevertheless, note summaries are useful in an academic setting. It is possible that these tools can be used to interpret thousands of notes quickly and without using any physician time. Even though more than 10 percent of the interpretations are inaccurate, the findings of the CAD tool could be used in a research setting to estimate the risk of other outcomes in these patients including their risk for cardiovascular disease and even death.

3.7 Benefits and Challenges of Databases in the Development of CAD Tools

The establishment of databases that are made possible by the EHR mandate has enormous potential for the development of CAD tools. A telling quotation from Rob Kass, an expert in Bayesian statistics, reads: “the accumulation of data makes open minded observers converge on the truth and come to agreement” [173]. In this setting, the accumulation of a gigantic body of clinical data in the form of EHR databases will be invaluable for the description of numerous clinical syndromes and disease. If these databases are unbiased, high quality samples of patients from the general population, there will be no better dataset with which to apply bioinformatics methods to understand the epidemiology, co-morbidities, clinical presentation and numerous other features of most syndromes and diseases. In addition to quantifying what is known already, these large databases can facilitate the development of new hypotheses regarding neurobiological and genetic underpinnings of these conditions through machine learning approaches [174][174][174][174]. One of the constant factors that limit many clinical and research studies is the steep cost of obtaining high quality data that can be used to develop and test continually updated hypotheses. EHR databases would drastically reduce this cost and thereby allow more funds to be dedicated to the development of models that better elucidate the biology underlying each condition.

In addition to facilitating more applicable and statistically powerful modeling, increased sample size also results in increased machine learning performance. In theory, as sample size increases, the amount of detected signal grows, resulting in an accuracy that is a sigmoid function of sample size. Each feature would therefore have a maximum discriminatory yield that can only be achieved with sufficiently large training sample size. Using the ADNI database, Cho et al. confirmed this theoretical result by demonstrating that the accuracy of all tested discriminations increased monotonically with the number of training subjects[175]. Therefore, in order to develop the most accurate and therefore applicable CAD tool, one must train it on as large a representative sample size as can be obtained. As noted by van Ginneken et al. [62], if one CAD tool is already FDA approved, securing adequate funding to prove a new tool performs better is a major hurdle. Large EHR databases would lower this barrier and foster innovation that will benefit patient care. If even ten percent of US patients consented to the addition of their records to databases, millions of cases would be available. It is important to note, however, that the accuracy of a tool developed on an infinite sample is not 100 percent. Instead, it is limited by the ability of the model to understand trends in the data and the discriminatory power of the features used in the model. This discriminatory power, and thereby CAD tool performance, is based on a few key assumptions about the databases.

The most important assumption is that the gold standard reported in the database is definitive. At best, supervised machine learning can only recreate the performance of the gold standard. If, for example, clinicians consistently misdiagnose bipolar disorder as depression, then any database would confuse the two disorders and replicate this misdiagnosis. Thereby, any CAD tool can only be as good as the experts used to train it. This suggests than when training CAD tools, the developers should limit the training and validation sets to clear examples of each condition to minimize but not eliminate this bias. This limitation also leaves space for research groups to develop improved gold standards or clinical procedures that could outperform the CAD tool. Thereby, we expect that CAD tools cannot replace the great tradition of continual improvement of clinical medicine through research or the advice of the national and international experts that study and treat specific conditions.

Another key assumption is that the training sample is an unbiased representation of the population in which the CAD tool will be applied. Correction of this bias is critically important because a supervised CAD tool is only as applicable as its training and validation set is unbiased. We expect that these databases will endow modern statistical methods the power needed to identify, quantify and control for possible sources of bias that have not been appreciated in smaller databases[176]. In many clinical research protocols, it is common practice to ignore this assumption because the practical cost of obtaining a truly unbiased sample is prohibitive. For example, it is often the case that patients recruited at large academic medical centers have more severe disease than other centers. This assumption of an unbiased sample is justified because, in most cases, there is little evidence that the pathophysiology underlying disease in research subjects or patients with severe disease differs from the full population. Because of their size, EHR based databases would be expected to include patients that would not ordinarily be recruited into research studies. Research based on these databases would then be more representative of the affected population than current research methods.

Current experimental design methods produce high quality clinical information that minimizes noise in the sampled data. As the number of patients increases, so does the number of independent health care providers and institutions that collect data associated with each patient. This in turn substantially increases the number of possible sources of uninformative noise that must be adequately controlled. Controlling for some of these sources of noise is simply a statistical exercise but others require more complex biostatistical modeling. One particularly egregious source of noise is if providers at particular institutions do not write clinical notes that fully represent the patients symptoms and the providers findings. No matter how effective CAD tools become, providers will always need to speak to patients, ask the right questions and provide consistent, high quality care. Patients are not trained, unbiased observers. Patients frequently omit pertinent details regarding their complaints unless they trust the provider and the provider asks the right question in the right way. On the scale of the entire database, detecting low quality or biased information is difficult because it requires testing if the data from each institution varies significantly from the trends seen in the rest of the dataset. These differences, however, could reflect unique characteristics of the patient population being treated at that institution. The development of reliable techniques to identify and control for these sources of noise will be critical to the effective mining of the EHR databases.

3.8 The Future of Medical Diagnostics

The key hurdle to deploying CAD tools in academic and clinical medicine is the efficient implementation of these tools into software already utilized by clinicians.

As stated by van Ginneken et al., the requirements of a CAD are that it has sufficient performance, no increase in physician time, seamless workflow integration, regulatory approval and cost efficiency [62]. We have already discussed how the sheer size of the EHR database will substantially improve the performance and applicability of CAD tools. The improvements that were the basis for the ARRA EHR mandatewhich we believe will be implemented using computer-aided diagnosticsprovide clear evidence for the issue of cost effectiveness. Each of the improvements from the reduction of duplicative or inappropriate care to the increase in early detection, will decrease the cost of health care nationwide [125][125][125]. Given these benefits and improved performance, it would only be a matter of time before these tools would be given regulatory approval. The only facet of CAD implementation left would be efficient implementation that does not increase physician time. This is a content strategy problem

Before seeing a patient, many providers scan the patient note for information such as the primary complaint given to the intake nurse, if available, and the patients history. A CAD tool could provide a formatted summary of such notes, making it more accessible. Reviewing other test data is also routine. A CAD tool that pre-reads radiological images could simply display the predicted result as part of the image header. Radiologists could then see and interpret the results of the CAD tool as well as confirm these results and provide additional details in their subsequent clinical note. Outputs similar to these could be provided at the top or bottom of reports for EEGs, metabolic panels and other medical procedures. Regardless, physicians should have access to the raw data so that they can delve deeper if they desire more detailed information [62].

During a patient visit, the CAD tool could help remind the physician of key issues to cover that are related to previous clinical notes to address patterns that the computer notices but the physician may have overlooked. The Agile Diagnosis software is already exploring how best to design this type of tool [177].

After the visit, the tool could then operate on the aggregate information from this patient and provide recommendations and warnings about medications and treatments. The inclusion of citations that verify the evidence-based efficacy of the recommended medications and warnings is simple and requires very little space and processing power though frequent updating may be necessary.

Although, the CAD reminders would likely be ignored by experienced providers, their constant presence could serve as a quality assurance measure. As discussed by Dr. Brian Goldman, M.D., at his TED talk, all providers make mistakes[178]. These CAD based reminders have the potential to improve upon the rate at which these mistakes are made and important details are missed. The most impactful benefits of CAD, however, are not in improving the care given by experienced providers that rarely make mistakes or miss details. Instead, these CAD tools will help inexperienced providers, those with limited medical training or special expertise or experienced practitioners who lack current expertise to provide basic health care information to underserved populations. In this way, the development of CAD tools could reduce the magnitude of health disparities both inside the US and worldwide.

3.9 Conclusions and Outlook

The EHR mandate will likely have widespread beneficial impacts on health care. In particular, we expect that the creation of large scale digitized databases of multimodal patient information is imminent. Based on previous actions of the NIH, we expect it to substantially support the development of these databases that will be unprecedented in both their size and quality. Such databases will be mined using principled bioinformatics methods that have already been actively developed on a smaller scale. In addition to other potential impacts, these databases will substantially speed up the development of quality, applicable CAD tools by providing an unprecedented amount of high quality data at low cost upon which models can be built. We believe that these tools will be responsible for many of the improvements quoted in the motivation for passing ARRA including the reduction of medical errors, inefficiency, inappropriate care and duplicative care while improving coordination, early detection, disease prevention, disease management and, most importantly, outcomes [125].

The development of widespread CAD tools validated on large representative databases has the potential to change the face of diagnostic medicine. There are already numerous examples of CAD tools that have the potential to be readily applied to extremely prevalent, high profile maladies. The major limiting factor is the validation of these methods on large databases that showcase their full potential. The development, validation and implementation of these tools, however, will not occur overnight. Important regulatory, computational and scientific advances must be achieved to ensure patient privacy and the efficacy of these automated methods. The problem of mining large databases also introduces numerous statistical problems that must be carefully understood and controlled.

The goal of these methods is not to replace providers but to assist them in delivering consistent, high quality care. We must continue to respect the science and art of clinical medicine. Providers will always be needed to interact with patients, collect trained observations and interpret the underlying context of symptoms and findings. In addition, providers will have the unique ability to understand the applicability of computer-aided diagnostics to each patient. Thereby, we believe that bioinformatics and machine learning will likely support high quality providers in their pursuit of continual improvements in the efficiency, consistency and efficacy of patient care.

CHAPTER 4

The utility of data-driven feature selection: Re: Chu *et al.* 2012

This is a reproduction of our work published in Neuroimage.[99] This work was a collaboration with Pamela K. Douglas, Ariana Anderson and Mark S. Cohen. PKD is a co-first author on this manuscript. Wesley identified the manuscript and brought his comments to the attention of the rest of the Laboratory of Neuroimaging Technology. He organized the article, collaboration and was responsible for the feature selection discussion. PKD was responsible for the section discussing the optimization of the C parameter and the overall direction and tone of the article. AA created the figure and assisted with editing the whole manuscript. MSC encouraged WTK and PKD to pursue the manuscript, and helped direct the tone of the manuscript, in addition to providing significant editing.

This is a comment on the following publication:[3]

Chu C, Hsu AL, Chou KH, Bandettini P, Lin C, ADNI Initiative. "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images." NeuroImage. 2011;60(1):59-70.

4.1 Abstract

The recent Chu *et al.* [3] manuscript discusses two key findings regarding feature selection (FS): (1) data driven FS was no better than using whole brain voxel data

and (2) *a priori* biological knowledge was effective to guide FS. Use of FS is highly relevant in neuroimaging-based machine learning, as the number of attributes can greatly exceed the number of exemplars. We strongly endorse their demonstration of both of these findings, and we provide additional important practical and theoretical arguments as to why, in their case, the data-driven FS methods they implemented did not result in improved accuracy. Further, we emphasize that the data-driven FS methods they tested performed approximately as well as the all-voxel case. We discuss why a sparse model may be favored over a complex one with similar performance. We caution readers that the findings in the Chu *et al.*, report should not be generalized to all data-driven FS methods.

4.2 Comment

Recently, Chu *et al.* [3] assessed how feature selection (FS) affected classification accuracy on a series of two class problems using grey matter voxels features. FS techniques are categorized typically as filter based, embedded, or wrapper based methods [179]. Within the neuroimaging community, data-driven FS (DD-FS) methods have been used commonly because they are generally effective: univariate t-test filtering (e.g. [180, 181]) and wrapper-based SVM recursive feature elimination (RFE) approaches (established in [110]; effective in [182, 183, 184]).

Chu *et al.* [3] presented a principled analysis that compared the performance of these two DD-FS approaches with voxelized features from a region of interest (ROI) based on a biological hypothesis, t-test in combination within an ROI constraint, and in absence of any first stage FS. Their analysis revealed that the DD-FS methods tested were unable to outperform simply using all ∼300,000 voxel features for discrimination, similar to results published by [**?**] who tested a series of FS methods. While Chu *et al.*[3] clearly discusses that these results are data specific, their findings nonetheless highlight the essential importance for further analysis of FS methods in neuroimaging applications where the data is both noisy and vast. We emphasize that their findings that DD-FS did not improve accuracies should be limited to a certain class of FS methods, for a limited set of parameter choices and kernels. The sensitivity of SVM accuracy to DD-FS methods with respect to changing kernels was discussed by [185], so we focus on the specific DD-FS methods implemented by Chu *et al.*[3]. We caution readers that their results should not be generalized to other DD-FS methods.

We first discuss the two DD-FS methods that were tested, and point out certain theoretical constraints that are common across both techniques. These limitations are well established in the machine learning (ML) literature, and have been discussed by the primary author of the fundamental RFE manuscript [109]. Both t-test filtering and RFE favor selection of features that maximize accuracy individually, assuming that these will provide the highest discrimination accuracy when used in aggregate [110]. Consider however, examples where multiple features provide largely redundant, yet highly diagnostic, information (i.e., spatially adjacent neuroimaging voxels), while other features with lower margins and t statistics hold unique information [186]. Within this framework, the redundant features will be retained, while the features that provide unique information that could improve performance will be discarded. Both of these factors contribute to a decrease in classification accuracy, rather than an increase, as discussed for neuroimaging data by [187, **?**, 188].

Further, features that are not themselves diagnostic, but which control for nuisance factors (e.g. age-associated atrophy [26]) are expected to have extremely low univariate $|t|$ values and reduced margins. To determine the utility of each feature in RFE, the multivariate separability vector, w, is projected onto each feature-dimension to get a univariate margin, w_j. In RFE, features with the smallest univariate margin, $\|w_j\|$, are excluded iteratively until the desired number of features is achieved. We expect that the margin of nuisance-controlling factors would be greater than noise but smaller than the margin of the diagnostic feature. In this case, the smallest margin and $|t|$ statistic features would be excluded before the diagnostic features by these DD-FS methods because the stopping criterion used by Chu *et al.*[3] was the *number* of selected input features. The stopping criterion is defined by the criteria used to determine exactly how many features are included in the final model. If one had used the observed training or testing accuracy (as in backward or forward selection) or an arbitrary fixed criterion for $\|w_j\|$ or $|t|$ to determine the stopping criterion, we would expect that these nuisance features may be included in the final model learned using RFE, but not using t-statistic filtering.

In contrast, the least-squares (ℓ_2) regularization in SVM, itself a multivariate DD-FS method, likely includes these nuisance factors: in regularization, features are selected based on the degree to which they maximize classification accuracy instead of reducing the number of input features using an indirect proxy for classification accuracy. The RFE model is mathematically equivalent to the ℓ_2 SVM model in which the smalleest SVM margins are set to be identically zero instead of their small estimated value. Similarly, t statistics assumes that the margins of low $|t|$-statistic features should be zero. This assumption is identical to the sparsity assumption of an ℓ_1 regularized SVM. However, ℓ_1 SVMs only outperform ℓ_2 SVMs when the underlying solution itself is sparse [189]. By extension, we believe that RFE and t statistic filtering will only outperform ℓ_2 SVM if the best diagnostic model is sparse.

As shown by Chu *et al.*[3], RFE and t-statistics did not improve performance, suggesting that these assumptions of non-redundancy and sparsity may have been violated. These shortcomings suggest that, while t statistics and RFE have practical value, they are not general panaceas.

The limited efficacy of RFE, or univariate t statistics, does not predict that alternate unsupervised DD-FS algorithms will, or will not, outperform regularization. Independent and Principal Component Analysis (ICA and PCA), for example, can both in effect project multiple linearly correlated, or redundant, features onto reduced number of features [103, 100, ?, 190]. In contrast to RFE and t-statistics, these methods that combine highly correlated and, frequently, spatially continuous voxels into regional features improve generalization substantially (e.g. [191, 192, 193, 194, 195]). Both ICA and PCA can control for the variation in highly diagnostic independent or principle components due to nuisance factors. Other DD-FS methods such as information criteria [5, 6], genetic algorithms [196], and Markov Chain Monte Carlo methods [197] select a single representative of each set of redundant diagnostic features. This perspective on DD-FS does not modify the original input features; instead it aims to more efficiently select the minimum subset of non-redundant features that maximizes performance. Numerous other DD-FS approaches employ clever algorithms that overcome some of the limitations of RFE and t statistics (i.e [198, 199, 200, 201, 202, 203, 204, ?, 205]); for review see [206]. Therefore, we emphasize again that the findings for RFE and t statistics should not be generalized to all DD-FS methods.

As a second practical point, we consider the conclusion that DD-FS performed worse than the feature selection inherent to SVM. We direct attention to figure 9F of the original manuscript, which shows how accuracy changes with decreasing values of the SVM regularization parameter, C, as a function of the DD-FS method employed for the largest sample size. We remind the reader of the original soft margin SVM formulation presented famously by [96] that presents the Lagrange functional for the two-class problem as:

$$L(w, b, R, \xi) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(x_i \cdot w + b) - 1 + \xi_i \right] - \sum_{i=1}^{n} r_i \xi_i + C \sum_{i=1}^{n} \xi_i \quad (4.1)$$

where n, i, w, b, Y, X, α, r, and ξ are the total number of exemplars, exemplar index, margin, intercept, output class vector, input data matrix, support vector

Lagrange parameter, soft margin Lagrange parameter and soft margin misclassification penalty, respectively. The linear decision function in the feature space takes the form:

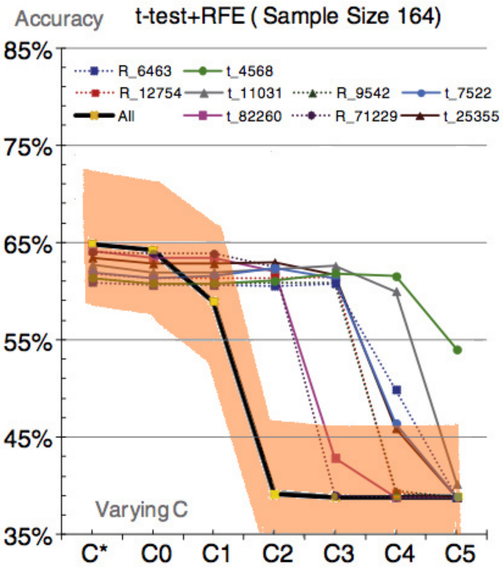
$$I(z) = \text{sign}\left(\sum_{\text{Support Vectors}} \alpha_i x_i \cdot z + b\right) \tag{4.2}$$

Where z is the hyperplane perpendicular to w. If $\alpha_i = 0$, then the corresponding sample was classified correctly and are irrelevant to the final solution. If $\alpha_i = C$, then the sample was misclassified, and if $0 < \alpha_i < C$, then the sample is located on the margin. If $\alpha_i > 0$, the sample is called a support vector. When solving for very large values of C, the problem tends towards the hard margin solution that can be solved using quadratic programming. With smaller C, the soft margin functional can be optimized through its dual formulation with quadratic programming.

Within their analysis, Chu *et al.*[3] assessed their accuracy with several parameter choices of C without cross-validation. The global optimum accuracy was obtained in absence of FS. However, we would like to emphasize that even for the optimum C case (indicated by C*), the performance of the other FS algorithms were all within the 95% confidence interval of the no FS (Figure 4.1). For moderate to small choices of C, FS methods systematically outperformed no FS, and were overall less sensitive and more robust to the choice of C. As discussed by Chu *et al.*[3], the selection of this C is computationally intense therefore it is frequently simply selected *a priori*.

While we agree that DD-FS does not always improve classification accuracy, it can help elucidate the pathology or physiology of the system under study, and can reduce the sensitivity of performance to tuning parameters when applied to the data in a principled manner. Overall, a parsimonious model made possible by DD-FS allows models to be more transparent, and thereby more useful for neuroscientific interpretation [**?**]. This sparsity can be implemented through separate FS methods, or within the SVM itself. While ℓ_2 regularization already applies a

Figure 4.1: A reproduction of Chu *et al.*'s[3] figure 9F where the added shading indicates the 95% confidence interval for the no feature selection accuracy using the normal approximation of the binomial distribution. Accuracy using all voxelized features was no significantly higher than data-driven feature selection accuracy at the optimum C, C^*. At multiple non-optimum C values, the accuracy using data-driven feature selection was significantly higher than using all voxelized features.



degree of sparsity [96], ℓ_0 regularization imposes a stricter penalty and has been used to interpret dynamic causal modeling features [207].

In the ML literature, it is common to evaluate methods primarily, or solely, on their classification accuracy. For typical cases, this is entirely appropriate: the goal is to classify, and not to explain. In investigative research, however, the needs are broader and more nuanced. In our own work, we use ML to aid in our understanding of brain function and dysfunction. We have shown previously that in some cases high classification accuracy can be obtained either from nuisance factors [208], or from factors in the data, such as demographics, unrelated to neuroimaging [209]. While these have the potential to generate clinically meaningful accuracies, they provide limited neurophysiological insight. If, on the other hand, the feature space is selected to project onto well defined, neurally-oriented subspaces, it is possible to jointly achieve excellent accuracy and explanatory power to aid in neuroscientific discovery. For example, independent components identified from functional MRI data frequently identify the default mode network [210] and have been used for classification [193] as well as the generation of meaningful feature dictionaries [208, 211]. Although these dictionary elements would vary across subjects and scans, we and others have shown that they are consistent enough to have an identifiable manifestation, an assumption underlying group-ICA methods [212, 213]. Therefore, these methods accomplish the tasks of feature 'identification' and 'selection' simultaneously.

The goal of feature selection is to minimize the number of estimated parameters in the final machine-learning model to improve performance and generalizability. The concept of balancing the empirical performance of the model to the data with the number of estimated parameters is well established in conventional statistics. For generalized linear models, the pervasive F test explicitly divides the explained variance of a model by the number of estimated parameters in the model to calculate the mean squared error. Additionally, the reference F distribution for determining significance is wider for models with more estimated parameters. Similarly, the Akaike and Bayesian information criteria (AIC and BIC) explicitly penalize the observed log likelihood of models using a function of the number of estimated parameters. While these criteria cannot formally be applied directly to cross-validation accuracy, our perspective is that the concept behind these criteria is applicable to machine learning models. Based on that idea, machine-learning models that achieve similar accuracy by operating on a selected set of features are preferred in investigative research over machine-learning models that are saturated with input features. We recognize that, unlike the likelihood or explained variance, cross-validation accuracies do not monotonically increase with the number of estimated parameters. We believe that DD-FS methods, in some situations, can be used effectively to accomplish this dual goal of model simplicity and high empirical cross-validation accuracy.

Despite the shortcomings of the methods tested mentioned herein, we also find it interesting that removal of a vast number of potentially irrelevant features with FS did not offer improvement, despite the theoretical caveats we detail above. It is possible that this lack of improvement is informative in and of itself. We suggest that pre/post FS accuracy should be reported more often, as these results may be helpful in conceptualizing how feature interactions are related to information representation in neural systems.

Because of this improvement in interpretability, we emphasize that FS methods are valuable beyond improving classification accuracy; just as a picture is a thousand words, an interpretable model is oftentimes immensely more valuable than a marginally superior yet uninstructive classification tool.

CHAPTER 5

Balancing Clinical and Pathologic Relevence in the Machine Learning Diagnosis of Epilepsy

This chapter is a reproduction of our work that appeared in the proceedings of the International Workshop on Pattern Recognition in Neuroimaging.[?] This work was a collaboration with Andrew Y. Cho, Ariana Anderson, Pamela K. Douglas, Edward P. Lau, Eric S. Hwang, Kaavya R. Raman, Aaron Trefler, Stefan T. Nguyen, Navya M. Reddy, Daniel H. Silverman and Mark S. Cohen. Wesley organized the manuscript and the data, wrote the manuscript, wrote the necessary code and performed the majority of the analysis. AYC and EPL assisted with parallel processing that was necessary to complete the analysis in a timely manner. AA and PKD assisted with direction and discussion. ESH, KRR and AT assisted in identifying eligible candidates for the study and assisted in editing the manuscript. STN and NMR processed the PET images through NeuroQ and assisted in editing the manuscript. DHS supervised STN and NMR and assisted in direction and phrasing of the manuscript. MSC secured funding for the study, assisted with direction and editing.

5.1 Abstract

The application of machine learning to epilepsy can be used both to develop clinically useful computer-aided diagnostic tools, and to reveal pathologically relevant insights into the disease. Such studies most frequently use neurologically normal

patients as the control group to maximize the pathologic insight yielded from the model. This practice yields potentially inflated accuracy because the groups are quite dissimilar. A few manuscripts, however, opt to mimic the clinical comparison of epilepsy to non-epileptic seizures, an approach we believe to be more clinically realistic. In this manuscript, we describe the relative merits of each control group. We demonstrate that in our clinical quality FDG-PET database the performance achieved was similar using each control group. Based on these results, we find that the choice of control group likely does not hinder the reported performance. We argue that clinically applicable computer-aided diagnostic tools for epilepsy must directly address the clinical challenge of distinguishing patients with epilepsy from those with non-epileptic seizures.

5.2 Introduction

Machine learning (ML) has proven clinically useful in many aspects of the diagnosis and pathologic characterization of epilepsy. For decades, seizure detection and prediction algorithms have been applied to scalp and intracranial electroencephalography (EEG) to characterize each patients seizures [214, 215]. The seizure detection algorithms are integrated regularly into the clinical software for EEG review. More recently, ML has been applied to scalp EEG, intracranial EEG, structural and diffusion MRI and FDG-PET to diagnose epilepsy compared to both patients with non-epileptic seizures (PWN) and seizure-naive, neurologically normal patients (NNPs) [8, 7, 26, 28, 27, 216, 217]. These applications helped to provide meaningful pathological insight into the features of epilepsy. They likely also have the potential to become computer-aided diagnostic tools (CADTs) that can be applied readily in clinical medicine. The choice of which "control" group to compare patients with epilepsy (PWE) against is not uniform. In this manuscript, we provide evidence to suggest that that the optimal control group depends upon whether an ML tool is intended to elucidate facets of the neurophysiology of epilepsy, or to estimate the reliability of the tool to diagnose patients in a clinical setting.

Non-epileptic seizures are primarily psychiatric events. In greater than 90% of cases, these seizure-like attacks are a symptom of conversion disorder in which patients psychological challenges present as physical symptoms [218, 219]. These attacks tend to mimic seizures that the patient has seen or heard about; they are therefore difficult to distinguish from epileptic seizures. The diagnosis of non-epileptic seizures is based on ruling out all organic causes for the attacks; therefore these patients are frequently exposed to antiepileptic medication and other neurologic interventions prior to definitive diagnosis. There are no neuropathologic changes that predispose these patients to have attacks and the electrophysiological mechanism for the attacks is dissimilar from the mechanism for epileptic seizures. Therefore, antiepileptic medication and other neurologic interventions will not control these attacks, though their side effect profile remains the same. Although a small minority of PWN also suffer from epileptic seizures, it is important to distinguish PWN from PWE so that they can receive the treatment appropriate to the underlying cause of their seizures.

The pathologic benefit of comparing NNPs to PWE is clear. Using our knowledge of normal physiology, this comparison has been used to describe how the complexity of EEG recording during seizure (ictus) is less than that of interictal EEG [**?**, 215]; how epileptic lesions are associated with focal cortical atrophy and hyperintensity on MRI [58]; and how focal interictal hypometabolism in FDG-PET can indicate the seizure onset zone [22]. These differences are thought to be macro-scale features that confirm our understanding of epilepsy on a neural level.

The decreased complexity in EEG reflects hypersynchronized activity of neurons in the epileptic network, coupled with inhibited activity in the surrounding tissue [220]. Focal cortical atrophy and MRI signal hyperintensity are MRI-based signs of increased cell death of inhibitory cells and the ensuing gliosis [221]. The cause for focal hypometabolism is defines less clearly, but it has been shown that neurons within the epileptic network have altered metabolic activity [222]. Cell death and gliosis also may play a role in interictal hypometabolism [222]. While these findings help us to understand the neuropathologic features of epilepsy, there are, unfortunately, important caveats to their interpretation.

Patients with seizures are exposed to environmental factors that may artificially increase physicians ability to discriminate NNPs from PWE. The mechanism of action of many antiepileptic drugs (AEDs) is to decrease the synchronicity and excitability of neural networks [220], thereby potentially increasing the baseline complexity of EEG so that the contrast with seizure activity is enhanced. Similarly, some AEDs have psychiatric side effects that appear similar to the psychiatric co-morbidities of PWN [223, 224]. In contrast with NNPs, PWN frequently are treated with these AEDs before their seizures are determined to be non-epileptic, therefore the use of PWN as a control group more accurately controls for the potential effect of AEDs.

Even though the diagnoses of non-epileptic seizures and epilepsy are distinct, many of their risk factors are shared. PWNs model their seizures after those they have seen or heard about before, therefore the relationship with family history is difficult to describe [225, 226]. Similarly, both PWN and PWE are associated with traumatic brain injury (TBI), albeit PWN are more associated with mild TBI [218, 225]. The presence of psychiatric comorbidities increases the suspicion for non-epileptic seizures, but epilepsy also has been shown to be associated with significant psychiatric challenges, potentially as a side effect of AEDs, or to the loss of independence and the stigma associated with the disease [227]. Therefore, in order to assess reliably if ML models can detect signs of the underlying pathology associated with epilepsy, it is useful to compare PWE to PWN.

Lastly, and potentially most importantly, the comparison to PWN mirrors the clinical question at hand. Physicians only question if a patient has epilepsy if they present with seizure-like events. They do not, as suggested by the comparison with NNPs, consider epilepsy in all patients they encounter.

As a result of these concerns, we seek to study in our FDG-PET database which control group results in a more accurate diagnosis of lateralized temporal lobe epilepsy (TLE). This allows us to measure the potential effect of the confounding factors discussed above. We also inspect if there were detectable and/or interpretable differences between PWN and NNP.

5.3 Methods

5.3.1 Dataset

All of the 105 patients with seizures included in our analysis were admitted to the University of California, Los Angeles (UCLA) Seizure Disorder Centers video-EEG Epilepsy Monitoring Unit between 2005 and 2012. Each patients diagnosis was based on a consensus panel review of their clinical history, physical and neurological exam, neuropsychiatric testing, video-EEG, interictal FDG-PET, ictal FDG-PET, structural and diffusion MRI and/or CT scan. This multimodal assessment is the gold standard for epilepsy diagnosis, and for localization of the epileptic focus. The patients included in this analysis were chosen because they had an FDG-PET; had no history of penetrative neurotrauma, including neurosurgery; were determined by consensus diagnosis to have a single, lateralized epileptogenic focus and had no suspicion of mixed non-epileptic and epileptic seizure disorder. These patients were diagnosed with either left temporal lobe epilepsy (LTLE, n=39), right temporal lobe epilepsy (RTLE, n=34) or non-epileptic seizures (PWN, n=32). PET images were determined to be interictal by clinical findings and concurrent scalp EEG. Neurologically normal, seizure naive patients (NNP, n=30) were scanned for other reasons on the same clinical scanners and were age matched to the PWN. Details of PET acquisition and feature extraction using NeuroQ (Syntermed, CA) are described in [7].

5.3.2 Machine Learning Details

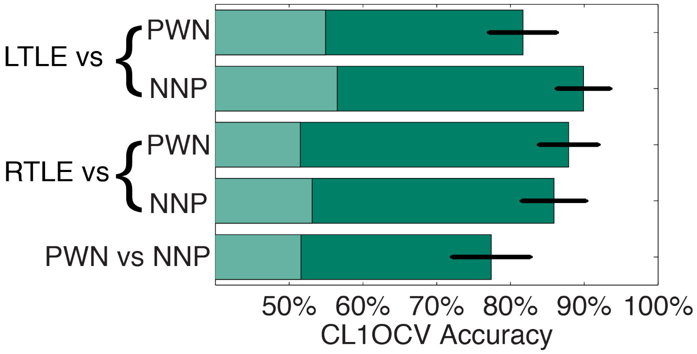
The Multilayer Perceptron was implemented to discriminate between either PWN or NNP versus RTLE or LTLE with default parameters in Weka [228] using the protocol described in [7] where the clinical implications of the PWN versus TLE discrimination are discussed. The minimum redundancy-maximum relevancy (mRMR) toolbox for MATLAB [6, 5] was used to generate a ranked list of the 47 ROI metabolisms (features) based on the training set. We used a random field theory correction (RFTC) to correct for the bias in selecting the maximum cyclical leave-one-out cross validation (CL1OCV) accuracy after testing multiple numbers of ROIs that contributed to the model [7]. Weka [228] was used to implement CL1OCV of a cost-sensitive MLP that was weighted to maximize balanced accuracy, defined by the mean of sensitivity and specificity.

5.4 Results

5.4.1 Cross-Validation Accuracy Differences

Using either control group, we diagnosed lateralized TLE effectively with greater than 81% CL1OCV accuracy (RFTC z-test of proportions versus naive classifier, z>5.8, p< 10^{-8}; Figure 5.1). There was no significant difference between the CL1OCV accuracies when NNP or PWN was used when diagnosing RTLE or LTLE (two sample z-test of proportions, $|z|$ <1.5, p >0.16). We discriminated between PWN and NNP with 77% CL1OCV accuracy which was significantly better than chance (RFTC z-test of proportions versus naive classifier, z=4.9, p< 10^{-5}).

Figure 5.1: CL1OCV accuracy of our computer-aided diagnostic tool using each control group. Error bars indicate standard error from the mean. Shading indicates performance of a naive classifier. PWN: Patients with non-epileptic seizures; NNP: Neurologically normal patients; L or R TLE: Left or Right Temporal Lobe Epilepsy.



5.4.2 Insight into Focality of the Epilepsies

Figure 5.2 illustrates the number of features that produced the random field theory corrected CL1OCV accuracy. Using the PWN as a control, the RTLE comparison required fewer ROIs than the LTLE comparison. Using the NNPs as a control, the same trend was seen, albeit with different mRMR feature rankings (Table 5.1).

5.5 Discussion

Even though there are substantial differences in the resting state neural metabolism of PWNs and NNPs, the choice of control group did not substantively affect our ability to diagnose PWEs, nor did it provide different pathologic insight into the difference between LTLE and RTLE. This suggests that comparing PWE to NNPs did not artificially increase our discriminative ability, contrary to our hypothesis. Our reliable discrimination between PWNs and NNPs, and the difference in feature rankings however, indicate that the multilayer perceptron may harness sepa-

Figure 5.2: Percent of the 47 ROIs that contributed to the best CL1OCV accuracy. Error bars indicate accuracies within the same significant random field theory cluster.

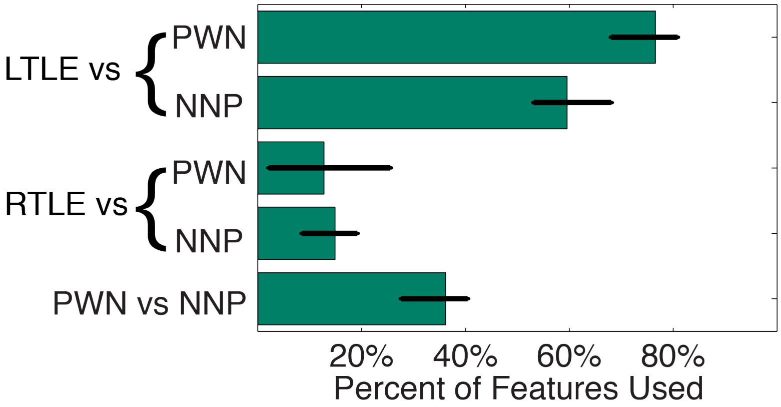


Table 5.1: The mRMR Rank of the top 5 ROIs based on the full data. Note that these rankings utilize the full dataset and therefore do not necessarily coincide with any individual training set, each of which are missing data from just one patient. The preceding L and R refer to left and right respectively. The lower case i, l, a, p, s, and m stand for inferior, lateral, anterior, posterior, superior, and medial respectively. The other abbreviations are for temporal cortex (Temp), thalamus (Thal), associative visual cortex (Ass Vis), and Sensorimotor cortex (SM). The temporal, parietal and frontal ROIs are all cortical ROIs. Colors indicate repeat ROIs.

ROI Rank	LTLE vs PWN	LTLE vs NNP	RTLE vs PWN	RTLE vs NNP
1	Midbrain	R Ass Vis	R ila Temp	R ila Temp
2	L ilp Temp	R pm Temp	L SM	R ilp Temp
3	R ilp Temp	L i Frontal	R ilp Temp	L Lentiform
4	L Ass Vis	R s Parietal	L sl Temp	L SM
5	L Broca’s	L SM	R Thal	R sl Temp

rate pathologic findings depending on the control group used. Interested readers are directed to [7] for an in depth description of this clinical and pathologic insight.

This suggests that there may be no control group that a priori will result in higher performance. The comparison to NNPs might describe better the combination of pathologic insults that results from, and/or causes, epilepsy. In our case, the lack of temporal ROIs in the top 5 rankings for LTLE suggests that the most salient pathologic consequences, and/or initiating factors, may lie outside the epileptogenic focus whereas the opposite may be true for RTLE. In contrast, the comparison to PWN demonstrates directly how the algorithm would perform in the clinic. In addition to utilizing the neurometabolic changes associated with epilepsy, this model may also harness the neurometabolic changes associated with PWN. Therefore, all of the observed differences cannot be attributed directly to epilepsy. As discussed above, depression was associated with hippocampal volume loss [229, 230]. Therefore, conversion disorder may also have characteristic FDG-PET findings.

These results reveal the challenge of developing a CADT to diagnose patients effectively. Just detecting epilepsy is not enough; we must also discriminate it reliably from disorders whose presentation is similar. The ideal CADT for epilepsy would effectively rule out transient ischemic attacks, confusion episodes, syncope, drug abuse and other disorders on the differential diagnosis for seizures (for full differential diagnosis see [231]). This presents a clear challenge: effectively recruiting and scanning enough patients with each of these disorders is prohibitively expensive in both time and money. Therefore, when planning experiments, we believe that one must choose the control group(s) that reflects the desired balance of clinical relevance to pathologic relevance.

There are a few limitations in the generalizability of these findings to the diagnosis of epilepsy and other disorders. In patients who present with their first seizure, the clinical question is not merely if the seizures are epileptic or non-epileptic: the patient also needs to know if they are at risk for future seizures. This clinical comparison may be better served by the contrast between NNPs and PWE. In addition, PWN are more frequently misdiagnosed as frontal lobe or generalized epilepsy instead of TLE. Therefore, there are also some caveats to using PWN as the control group for TLE diagnosis.

The challenge of identifying the proper control group to train CADTs is not unique to epilepsy. For example, many CADTs for Alzheimers disease frequently are controlled both by NNPs and patients with mild cognitive impairment [169, 175]. Few studies, however, consider the full differential diagnosis for dementia, including Parkinsons dementia, fronto-temporal lobe dementia, Lewy-body disease, and other dementias. Similarly, much work has been done in discriminating patients with schizophrenia from NNPs even though antipsychotic medication is associated with substantial neurologic changes [232].

We believe that there may be two divergent goals for machine learning in clinical populations: the pathologic description of disorders and the development of clinically applicable tools. Therefore, when describing the underlying pathophysiology of disease, the goal of machine learning is not necessarily to optimize classification accuracy. It is instead to pose a biologically plausible model that reflects trends seen in the data. This is related, but potentially separate, from the ultimate goal of using machine learning to maximize the clinical utility of computer-aided diagnostic tools. We argue here that to maximize clinical applicability, one must mimic the clinical question at hand by carefully selecting the control group.

CHAPTER 6

Poisson Noise Obscures Hypometabolic Lesions in PET

This chapter is a reproduction of our work published in the Yale Journal of Biology & Medicine. [233] This was a collaboration with Edward P Lau, who provided computation and editing support.

6.1 Abstract

The technology of flouro-deoxyglucose positron emission tomography (PET) has drastically increased our ability to visualize the metabolic process of numerous neurological diseases. The relationship between the methodological noise sources inherent to PET technology and the resulting noise in the reconstructed image is complex. In this study, we use Monte Carlo simulations to examine the effect of Poisson noise in the PET signal on the noise in reconstructed space for two pervasive reconstruction algorithms: the historical filtered back-projection (FBP) and the more modern expectation maximization (EM). We confirm previous observations that the image reconstructed with the FBP biases all intensity values towards the mean, likely due to spatial spreading of high intensity voxels. However, we demonstrate that in both algorithms the variance from high intensity voxels spreads to low intensity voxels and obliterates their signal to noise ratio. This finding has profound impacts on the clinical interpretation of hypometabolic lesions. Our results suggest that PET is relatively insensitive when it comes to

detecting and quantifying changes in hypometabolic tissue. Further, the images reconstructed with EM visually match the original images more closely, but more detailed analysis reveals as much as 40% decrease in the signal to noise ratio for high intensity voxels relative to the FBP. This suggests that even though the apparent spatial resolution of EM outperforms FBP, the signal to noise ratio of the intensity of each voxel may be higher in the FBP. Therefore, EM may be most appropriate for manual visualization of pathology but FBP should be used when analyzing quantitative markers of the PET signal. This suggestion that different reconstruction algorithms should be used for quantification and visualization represents a major paradigm shift in the analysis and interpretation of PET images.

6.2 Introduction

Positron emission tomographic (PET) images play a major role in the treatment and management of a growing number of maladies. In most cases, the interpretation of these images relies on the detection of high intensity lesions by quantifying the relative distribution of a radioactively decaying tracer. This tracer is most commonly fluoro-deoxyglucose (FDG) which allows PET to quantify the relative glucose metabolism in tissues. Hypermetabolic lesions indicate the presence of inflammation, malignancy and/or major functional changes. The observation of these changes has been critical to the characterization and clinical management of central nervous system cancers, paraneoplastic syndrome, Huntingtons and, when scanned during ictus, epilepsy [22, 234]. In some cases, PET is used to guide resective curative neurosurgery [22, 235].

The detection of hypometabolic lesions is equally clinically relevant. PET has been effectively used to characterize Parkinsons disease, Alzheimers disease, interictal epilepsy, cortical dysplasia, tuberosclerosis, and even mood disorders. In these cases, the hypometabolic lesions indicate functional abnormalities or portend the location of future atrophic lesions [236, 237, 238, 239, 240, 241, 242, 22, 235, 243, 234, 222]. In cortical dysplasia, tuberosclerosis and central nervous system infections multiple structural abnormalities frequently exist but only a small subset of these lesions generates epileptic seizures. The co-localization these structural abnormalities visualized in MRI with hypometabolic lesions observed using PET can be effectively used to determine which of these structural abnormalities is generating the seizures [22]. When these co-localized lesions are resected, 86% of patients achieve favorable outcomes compared to 30-76% without co-localization [244, 245, 246].

PET, however, may be biased against the detection of these hypometabolic lesions. The technology of PET relies on the emission of positrons from radioactively decaying isotopes. The number of positrons that are emitted from each volumetric pixel, or voxel, is Poisson distributed. In a Poisson distribution, the variance of a sample is equal to its mean. Consequently, the variance of positron count increases as overall signal increases. Due to the fact that these images are reconstructed based on projections, this noise could potentially spread to nearby voxels [247, 248]. For hypermetabolic lesions, this would result in the lesion dominating the signal in the reconstructed images. Unfortunately, the same signal interaction can allow surrounding normal tissue to mask hypometabolic tissue.

This potential bias against the detection of hypometabolic lesions may seem to be an issue of resolution. Modern reconstruction techniques like ordered subset expectation maximization (OSEM) substantially increase the resolution of reconstructed images relative to the canonical filtered back-projection (FBP) [249, 250]. One of the major hurdles to resolution in FBP is the streaking caused by high intensity voxels. In X-ray computed tomographic (CT) imaging this streaking is regularly caused by bone artifacts. It is also present, albeit to a lesser degree, in PET [251, 252]. The OSEM algorithm substantially decreases the effect of these streaks and thereby increases image resolution [253, 254]. This improvement is visually apparent even to the untrained observer and has resulted in the pervasive adoption of the OSEM algorithm for CT and PET reconstruction.

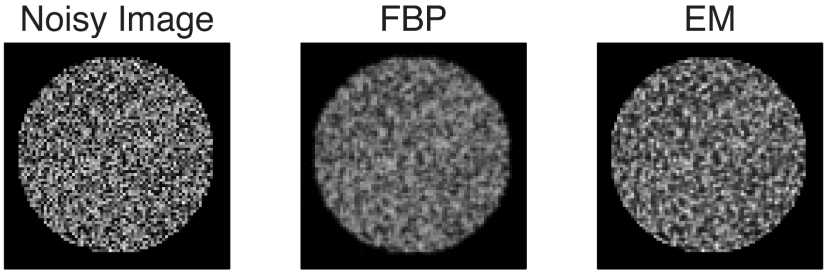
These improvements, however, only focus on decreasing the bias in signal intensity caused by surrounding tissues and ignore effects of noise. A simple theoretical proof illustrates that maximum variance of voxel intensity in image space provides an upper bound for the maximum reconstructed voxel intensity variance for FBP (see Supplementary Material) [249]. There is no analogous proof for the iterative EM algorithm, much less the OSEM algorithm. Therefore even though the OSEM algorithm decreases bias, it has the potential to increase variance and thereby decrease signal to noise ratio. This potential challenge has been largely ignored because, as humans, we are exquisitely capable of detecting changes in the mean but relatively weak at detecting changes in spread. The development of PET and CT reconstruction has focused on the generation of visually interpretable images; therefore previous literature has focused almost exclusively on trends of the mean. As more quantifiable markers of PET are developed, we believe that an in-depth treatment of the variance is critically important to achieving accurate and clinically relevant measurements.

In this paper, we use Monte Carlo simulations to characterize the statistical properties of the variance in both EM and FBP reconstructed images. We demonstrate that in both algorithms, the Poisson noise from hypermetabolic voxels obliterates the signal to noise ratio for hypometabolic lesions, resulting in a bias against the detection of hypometabolic lesions. The understanding of this effect has a profound impact on the interpretation of hypometabolic lesions on PET images.

6.3 Materials and Methods

In this Monte Carlo simulation, 10 million exemplars of reconstructed Poisson noise were measured from pixels with integer initial intensity from 1 to 100. Even though these are two dimensional (2D) images, the concepts are readily generalizable to three dimensions. Figure 6.1 illustrates one example of a noisy image and the two reconstructions with a common intensity scale. We ignore the effects of attenuation, randoms, scatter, deadtime, detector normalization, scan length, decay, interpolation and the specific reconstruction filter because their inclusion does not influence our conclusions. These factors either uniformly increase the variance of the reconstructed intensities or exaggerate the contribution of hypermetabolic voxels to the total variance of the image.

Figure 6.1: These circles illustrate examples in which each pixel intensity is initialized using a discrete uniform distribution with range of 1 to 100. An independent Poisson random variable with parameter equal to this intensity is then realized for each pixel. We then used the filtered back projection (FBP) and expectation maximization algorithm (EM) to reconstruct this circle based on its projection, as is done for PET images.



Images were sequentially realized until each intensity value had been reconstructed at least 100 thousand times. This took 2,614 realizations and approximately 7 cpu-days. Each simulation image was 80 voxels by 80 voxels with circle of radius 35 voxels centered on the 40th voxel in each dimension. This corresponds roughly to PET images of an average human brain with $(2mm)^3$ voxels. Each voxel within the circle was given an initial intensity from a discrete uniform distribution ranging from 1 to 100. All voxels outside the circle had intensity 0. Poisson noise with parameter equal to the initial intensity of each voxel was then added. Radon projections were used to simulate the actual data collected by sensors for integer angles from 0 to 179 degrees of this noisy image. By realizing many independent images in this way, edge effects and the effect of particular configurations were minimized.

The regular shape and voxel intensities were chosen to improve the interpretability of our results. This simplification resulted in a deeper understanding of the forces generating our results below. The results can be easily generalized to the interpretation of a diverse set of hypometabolic lesions on cranial PET. The diversity of lesion location and type is prohibitively large to address in a single publication.

Images were reconstructed from the simulated sensor data using the filtered back-projection (FBP) and expectation maximization (EM) algorithms. For the FBP reconstruction, the ramp filter and linear interpolation were used and the image was padded with zeros up to 126 voxels by 126 voxels. This reconstruction exactly mimics the canonical implementation of the algorithm. For the iterative EM reconstruction, the initial image had uniform intensity 1. The canonical full form formula was used for the updates of the EM. The A matrix was formed by calculating the explicit point spread function for all integer angles from 0 to 179 degrees (see Supplementary Information for algorithmic details). The pervasive OSEM algorithm is a subset of the EM algorithm that substantially decreases the computational load of reconstruction therefore all results shown for the EM algorithm generalize to OSEM.

Due to the high spatial frequency in the focus of the image, 300 iterations were used for each EM reconstruction. The image did not appear qualitatively different after 20 iterations. The magnitude of variance was also observed to decrease asymptotically with iteration number (data not shown). The choice of 300 was made to maximize the potential for high spatial frequency noise that may better match the underlying data.

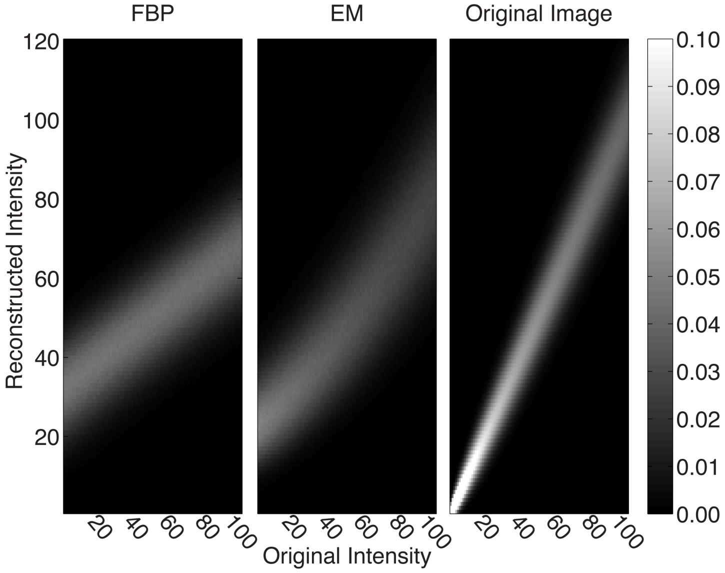
All simulations were conducted in MATLAB 7.14 (Mathworks) and all statistical analysis was conducted in R (see Supplemental Material). Signal to noise ratio was calculated as the ratio of original intensity to the standard deviation of the reconstructed intensity. This is equivalent to a hypothesized two fold change in original intensity.

6.4 Results

A detailed statistical analysis of the reconstructed images reveals important trends. Figure 6.2 illustrates the probability density of the reconstructed values with respect to their initial intensity. In this figure, all densities above 0.1 are rounded down to 0.1 to facilitate comparisons between the distributions. In the right panel the Poisson nature of the original image is evident: the spread increases linearly with respect to the original intensity. The probability densities of the reconstructed intensities are markedly different from that of the original image. For both algorithms, the variance is much more homogenous and more extreme values shrink towards the mean. In order to formalize these observations, we fit statistical models to these trends. All intervals below reflect 95% confidence intervals.

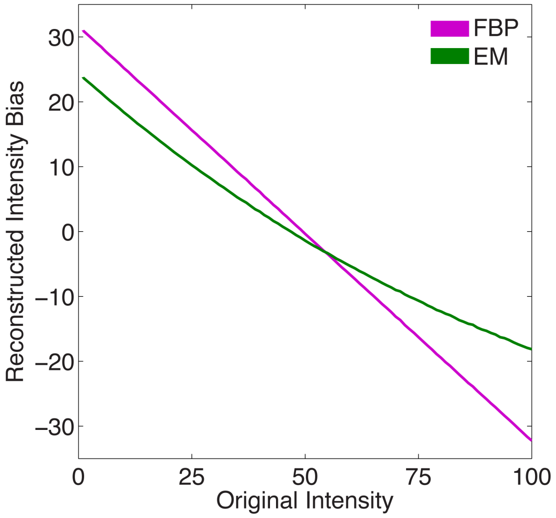
First, we address the observation that all intensities shrink back towards the mean, albeit less so for EM than FBP (Figure **??**).This regression back to the mean appears to be linear for FBP reconstructed voxel intensities (FBP-RVI) and quadratic for EM reconstructed voxel intensities (EM-RVI). This quadratic trend results in fitting high intensity voxels more closely compared to low intensity

Figure 6.2: This figure illustrates the probability distribution of reconstructed voxel intensity for each of the reconstruction algorithms. For comparison, the right panel illustrates the original probability distribution before reconstruction.



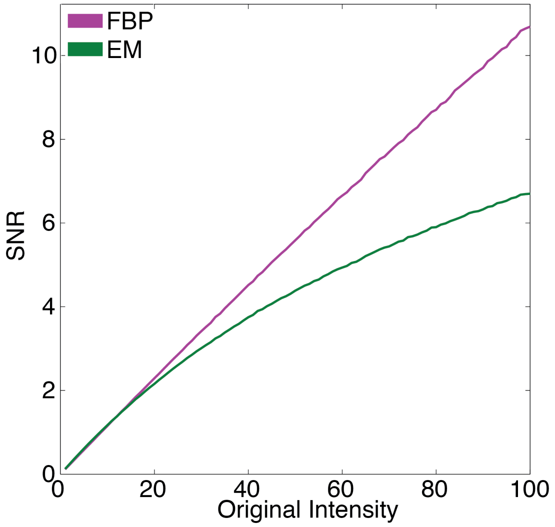
voxels. We fit statistical models to quantify and compare these trends across reconstruction algorithms. The FBP reconstructed voxel intensities (FBP-RVI) regressed back to this mean linearly with slope of -0.64 and intercept of 31 *units* (-0.0642 to -0.0634 and 31.56 to 31.63). The EM reconstructed voxel intensities (EM-RVI) regressed quadratically back to the mean with acceleration of 0.0018 $units^{-1}$ (0.00176 to 0.00179). After controlling for this quadratic term, the EM-RVI had a 7 units smaller intercept and a slope of 0.038 closer to zero than the FBP-RVI (-7.44 to -7.37 and 0.037 to 0.040). The F statistic of this composite model was 3.7 million with 4 and 195 degrees of freedom, resulting in a model-wide p value of less than 10-16. There was no evidence that the residuals deviated from the assumption of independent identically distributed Gaussians. Even though the EM algorithm converges quadratically to the maximum likelihood solution [253, 254], calculating more iterations does not significantly change any of these fitted parameters.

Figure 6.3: This figure illustrates the magnitude of the reconstructed intensity bias of each of the algorithms. The line thickness represents the standard error for each point. This standard error is small due to the large sample size. The FBP is indicated by cyan and the EM is indicated by green.



The focus of this report is the signal to noise ratio of reconstructed voxels. Controlling for the biases addressed above, the signal to noise ratio to detect a hypothesized two fold change in intensity was substantially reduced for EM reconstructions compared to FBP reconstructions (Figure **??**). The maximum signal to noise ratio for FBP-RVI was 60% larger that of the maximum for EM-RVI. For both algorithms, this original intensity dependent increase in the signal to noise ratio with respect to original intensity reflects similar trends in variance as seen in the bias. The FBP-RVI variance increases linearly with intercept of 73 $units^2$ and slope of 0.013 $units$ (73.36 to 73.82 and 0.009 to 0.017). The EM-RVI variance increased quadratically with acceleration of 0.004 (0.0040 to 0.0044). After controlling for this quadratic term, the EM-RVI variance had a 13 $units^2$ smaller intercept and a 1 $unit$ larger slope (-13.6 to -12.8 and 1.04 to 1.08). This means that the EM performs slightly better for extremely low intensity voxels but variance in EM-RVI and FBP-RVI also quickly increases as original intensity increases. The F statistic of this composite model was 280,000 with 4 and 195 degrees of freedom, resulting in a model-wide p value of less than 10^{-16}. There was no evidence that the residuals deviated from the assumption of independent identically distributed Gaussians.

Figure 6.4: This figure illustrates the magnitude of the signal to noise ratio of the reconstructed intensity each of the algorithms. Signal to noise ratio was calculated as the original intensity divided by the standard deviation of the biased reconstructed intensity. This corresponds to a hypothesized two-fold change in intensity. The line thickness represents the standard error for each point. This standard error is small due to the large sample size. The FBP is indicated by cyan, and the EM is indicated by green.



6.5 Discussion

These striking results have a profound impact on the interpretation of PET images using quantitative and visual measures. We demonstrated in our simulations that PET is insensitive to all but large scale changes in hypometabolic regions. Therefore, we caution against the interpretation of hypometabolic lesions when reading PET images both visually and quantitatively. However, we confirm that EM improves the spatial resolution of reconstructed images by decreasing the bias introduced by nearby voxels when compared to the FBP but we also illustrate that this bias correction results in a substantial decrease in the signal to noise ratio. Consequentially, even though EM reconstructed images are more consistent with our knowledge of the underlying biological structures, this increased spatial resolution comes at the cost of decreased statistical power of quantitative measures of the signal.

When interpreting PET images, our results suggest that one should focus on regions that are normally hypermetabolic relative to the surrounding tissue and caution against interpretation of changes in hypometabolism. For example, this is particularly important when interpreting images from patients with tuberosclerosis for identification of epileptic focus. The tubers that characterize this disease can be small and distributed throughout the brain. Lee & Salamon suggests that hypometabolic lesions corresponding with structural abnormalities are candidates for epileptic foci [22]. If a structural lesion is in a hypometabolic region, however, our results suggest that there is very little power to detect metabolic abnormalities. This has the potential to increase the false negative rate for foci detection and thereby lead to patients with multifocal epilepsy being diagnosed with focal epilepsy. This misclassification can lead to patients undergoing focal surgical treatment that fails to control their seizures [244, 245, 246].

However, this does not hinder the ability of PET to recognize changes in relatively hypermetabolic tissue. It is important when reading PET images for one to consider the expected metabolism in the region of interest. If the expected metabolism is high, then most observed changes are interpretable and clinically relevant. Conversely, if the expected metabolism is low, then one should recognize that only comparatively large changes in metabolism are interpretable.

These findings also provide further motivation for the development of focused radioactive PET tracers to improve sensitivity [255]. Focused tracers target particular receptors or tissue types. For example, in Parkinsons disease, there is increased neural death in the substantia nigra, resulting in decreased metabolism [256]. As we have shown, the power to detect these subtle, highly localized hypometabolic lesions is limited with FDG-PET. Our results suggest that the radioactive serotonin analog, 18F-DOPA, that differentially localizes to the substantia nigra in normal tissue has increased signal to noise ratio [257]. If the relative localization of this tracer is reduced, this may provide early diagnostic or more detailed prognostic information for the patient [258]. From a research perspective, this early detection could result in the development of novel pharmaceutical intervention that could slow the progression of disease. This also suggests that PET experiments will have higher signal to noise ratios if they are designed such that they focus on changes in tissue that is the target of the tracer.

The implication of these findings is particularly salient for quantitative PET analysis that has the potential to capture more subtle or distributed trends in metabolism. Conventional analysis of PET segments the brain into focused regions of interest then averages the reconstructed metabolic rate across the entire region [259, 260]. Although it is tempting to suggest that this averaging improves the signal to noise ratio with respect to the factors we have modeled, this is, unfortunately, not the case. Instead, the linearity of the noise spreading suggests that the variance from hypermetabolic voxels spreads across the entire projection and is not corrected by reconstruction algorithms. This suggests that the noise across a local hypometabolic region is correlated. Because of this correlation, the average then estimates the value of the signal plus the noise instead of separating the two. Therefore, our results suggest that PET is systematically insensitive to the detection of changes in hypometabolic tissue even when averaged over lower resolution regions of interest.

Our guidance to bias against interpretation of changes in hypometabolic tissue, however, is most generalizable to high resolution changes in metabolism. If these changes are widely distributed over tissue, then the number of hypermetabolic voxels that contribute to each projection decreases. Consequentially, the noise is dominated only by the voxel with highest metabolism within the larger region. This is especially relevant to current analysis of epileptic foci because, due to the low spatial resolution of surgical procedures, only large magnitude, low resolution changes are clinically meaningful.

The substantial decrease in signal to noise ratio caused by the EM reconstruction suggests that while it vastly outperforms the FBP in manual interpretability and spatial resolution, EM may be not be appropriate for quantitative analysis of the PET signal. Based on our results, studies based on EM reconstructed images need 2.5 more patients or images than studies based on FBP reconstructed images to achieve the same signal to noise ratio (see Supplementary Material). Computation time for both reconstructions is relatively inexpensive, therefore our results suggest that both methods should be employed. The EM reconstruction should be used for visual interpretation and the FBP reconstruction should be used for statistical inferences. This guidance, however, is balanced by the fact that FBP reconstructed images are more biased than EM reconstructed images. Using the mean squared error, which incorporates both bias and variance, the signal to noise ratio for the EM remains less than the FBP for the majority of voxel intensities.

One could naturally suggest acquiring multiple PET images from the same patient to better quantify the noise distribution, but this practice is limited by expense. Statistically, one can expect that collecting multiple samples will increase the signal to noise ratio by a multiplicative factor of the square root of the number of samples. For hypometabolic voxels, however, the signal to noise ratio is so low that hundreds PET images would be insufficient to reveal relatively large changes. Each PET, however, has substantial cost in physician, scanner and patient time and resources. Simply splitting each scanning session into smaller time windows also does not solve the problem because the spatial resolution is a function of the total number of positrons observed [251, 252].

The knowledge that the noise in reconstructed space is likely heteroschedastic can be incorporated into statistical models of the signal in two theoretically equivalent ways. From a frequentist perspective, this can be done by relaxing the assumption that the residuals of the model are identically distributed. Instead, the variance of the residuals can be modeled as a linear or quadratic function of signal strength. By modeling this source of noise, therefore removing its contribution to the standard error of the model, we expect that the fit of the model would increase [261, 262, 263]. From a Bayesian perspective, one could introduce a prior that linearly or quadratically deweights the contribution of hypometabolic regions. This deweighting may also help ill posed models like those used in machine learning reduce their propensity to over fit the data by incorporating additional knowledge. As shown recently by Chu et al., this incorporation of additional biological and physical information may result in improved predictive performance .

These simulated results can be extended to address the signal to noise ratio in specific regions of interest. In particular, this approach of simulating the Poisson noise can be used to determine the sensitivity of FDG-PET to detect differences in numerous regions of interest in the brain. This could be used to give a more detailed explanation of the power of PET to describe high resolution metabolic changes. This could lead to an improved interpretability of smaller magnitude changes that indicate subtle phenomena. In particular, these subtle changes could be used in the aging population to predict which patients will progress to AD, as is currently being actively addressed using genotypic and MRI-based measures [264].

6.6 Supplementary Material

6.6.1 Regressions of Trends in Reconstructed Voxel Intensity

All regressions were performed by directly comparing the results for the EM to the FBP by incorporating a logical indicator variable for the EM, δ_{EM}, into the regressions. The full equation of the quadratic model fit is:

$$Y = \beta_{0,FBP} + \beta_{1,FBP} I + \delta_{EM} \left(\beta_{0,EM} + \beta_{1,EM} I + \beta_{2,EM} I^2 \right) + \epsilon \qquad (6.1)$$

where the ϵ vector is the Gaussian stochastic error term, the I vector indicates original intensity, the Y vector reflects the bias or variance and the β vector reflects the fitted non-linear parameters. The model was fit in this way because $\beta_{0,EM}$ and $\beta_{1,EM}$ reflect the change in the model attributable to the EM algorithm. The quadratic term was not included for the FBP because, when fit, its coefficient was not significantly different from zero (p>0.4).

6.6.2 Reconstruction Algorithms

The Radon projection, $m(t, \theta)$, for angles from 0 to 179 degrees of this noisy image was defined by the line integral over the line $l(t, \theta)$ for $t = x \cos\theta + y \sin\theta$ for the image, $I(x, y)$:

$$m(t, \theta) = \int_{l(t,\theta)} I(x, y) ds \qquad (6.2)$$

The filtered back projection (FBP) reconstruction calculates the reconstructed image,$J(x, y)$, based on the convolution, $*$, of the projection with the ramp filter,$g(t)$, using the formula below. In this case, $\Delta\theta$ is 1 because the angles of projections are in integer steps.

$$J(x, y) = \sum_{\theta=0}^{179} m(t, \theta) * g(t) \Delta\theta \qquad (6.3)$$

In the EM reconstruction, the initial reconstructed image had uniform intensity 1. The following formula was used for the iterative updates of the EM, where the A matrix was calculated as the point spread functions of individual voxels of intensity 1 in each position of the image. The superscript indicates the iteration index. The sums with two indices indicate double sums.

$$J(x,y)^{(n+1)} = J(x,y)^{(n)} \frac{\sum_{t,\theta} \left(\frac{m(t,\theta)A(x,y,t,\theta)}{\sum_{x,y} A(x,y,t,\theta)J(x,y)^{(n)}} \right)}{\sum_{t,\theta} A(x,y,t,\theta)} \tag{6.4}$$

6.6.3 Variance Proof for the Filtered Back-Projection

In order to derive the relation between the noise in the image space and the reconstructed noise, we use slightly different notation than we used above for the FBP. Let Δt, $\Delta\theta$ and N_θ denote the step size in pixels and angles and the number of angles sampled, respectively. Define the projection as $m_{ij} = m(t = i\Delta t, \theta = j\Delta\theta)$ such that $t = x\cos j\Delta\theta + y \sin j\Delta\theta$. Further, let $\sigma^2_{\max} \geq Var\left[m(t,\theta)\right]$ for all t and θ. The discrete FBP is then:

$$J(x,y) = \sum_{i,j} m_{ij} g(t - i\Delta t)\Delta t \Delta\theta. \tag{6.5}$$

Consider then the variance of these reconstructed values and recognizing the filter as a linear operator:

$$\begin{aligned} Var\left[J(x,y)\right] =& Var\left(\sum_{i,j} m_{ij} g(t - i\Delta t)\Delta t\Delta\theta\right) && (6.6) \\ =& \sum_{i,j} Var(m_{ij}) g(t - i\Delta t)^2 \Delta t^2 \Delta\theta^2 && (6.7) \\ \leq& \sum_{i,j} \sigma^2_{\max} g(t - i\Delta t)^2\Delta t^2\Delta\theta^2 = \sigma^2_{\max}\Delta t\Delta\theta^2 \sum_{i,j} g(t - i\Delta t)^2 \Delta t && (6.8) \end{aligned}$$

Using Parsevals theorem, applying the Nyquist frequency cutoff for the ramp filter and recognizing that total degrees are sampled:

$$Var\left[J(x,y)\right] \leq \frac{\sigma^2_{\max}\Delta t\Delta\theta^2 N_\theta}{12\Delta t^3} = \frac{\pi^2}{12}\frac{\sigma^2_{\max}}{N_\theta \Delta t^2}. \tag{6.9}$$

This formula provides a reasonable upper bound for the variance in the reconstructed space. Both algorithms performed significant better than this upper bound.

6.6.4 Sample Size Calculation using Signal to Noise Ratios

We calculate the relative sample size required to achieve an equivalent effective signal to noise ratio when using an acquisition or processing stream with different statistical power. Let SNR_{EM}, SNR_{FBP}, n_{EM}, and n_{FBP} be the signal to noise ratio of EM and FBP and the sample size of EM and FBP, respectively. Because standard error is proportional to the square root of sample size, the following equivalence can be assessed:

$$SNR_{EM}\sqrt{n_{EM}} = SNR_{FBP}\sqrt{n_{FBP}}. \tag{6.10}$$

This equation can trivially be rearranged to show that, to achieve the same effective signal to noise ratio, the ratio of the sample sizes must be equal to the square of the ratio of the signal to noise ratios. Alternatively, this is equivalent to the ratio of the variances. This can be written in functional form as:

$$\frac{n_{EM}}{n_{FBP}} = \left(\frac{SNR_{FBP}}{SNR_{EM}}\right)^2 = \frac{Var_{FBP}}{Var_{EM}} \tag{6.11}$$

CHAPTER 7

Accurate differentiation of epileptic and non-epileptic seizures through quantitative combination of findings in the clinical history

This chapter describes work for future publication. This work was a collaboration with Chelsea T. Braesch, Emily A. Janio, Justine M. Le, Jessica M. Hori, Akash B. Patel, Norma L. Gallardo, Janar Bauirjan, Andrea M. Chau, Sarah E. Barritt, Eric S. Hwang, Emily C. Davis, Andrew Y. Cho, Joe Gordon, David Torres-Barba, Jerome Engel, Jr., Mark S. Cohen and John M. Stern. Wesley organized the collaboration, downloaded and curated the database, performed the statistical analysis, and wrote the majority of the manuscript. CTB, EAJ, JML, JMH, ABP, NLG, JB, AMC, SEB, ESH, and ECD annotated clinical notes, contributed to which factors would be studied and contributed to the interpretation and discussion of results. AYC assisted with parallelizing the code and helping make the computational intensive portions of this work happen. JG and DT-B provided pilot analysis of results that helped contribute to which factors would be included. JE provided substantial support for interpretation and framing of results. MSC assisted with organization, direction, experimental design and framing. JMS assisted with continual guidance throughout all stages of the study and manuscript.

7.1 Abstract

Objective: Early and accurate differentiation of patients with epileptic (ES) and non-epileptic seizures (NES) is critical to establish effective treatments, improve quality of life, and reduce the cost of intractable seizures. Hundreds of clinical measures have been shown to differentiate these populations, yet the diagnosis remains challenging, and often is questioned by non-epileptologists. We evaluated the real-world clinical applicability of a computer-aided diagnostic model that combined the diagnostic value of a large set of measures reported in routine outpatient clinical reports for a large population of 1,126 patients with intractable seizure disorder. Methods: We included all consecutive patients (634 ES, 314 NES, 178 mixed/inconclusive) admitted to our adult video-electroencephography epilepsy-monitoring unit between January 2006 and April 2014. We recorded 91 potentially diagnostic measures included in the first outpatient neurological report evaluating the seizures written at our center. We estimated an objective diagnostic score for ES versus NES by combining multiple imputation of missing data with multivariate and regularized logistic regression. Results: Of the 91 studied measures, 43 were independent predictors of ES or NES (empirical Wald, $p<0.05$). When we balanced the number of questions we needed to ask against the overall accuracy of the model, combining 31 findings resulted in an area under the receiver-operating curve (AUC) of 90.1% ($p<0.001$). At the point of highest overall accuracy (85%, $p<0.001$), the sensitivity and specificity were 92% ($p>0.6$) and 70% ($p<0.05$), respectively. Significance: This quantitative analysis expands our understanding of the role historical findings to estimate the likelihood of a NES diagnosis. No individual clinical finding was pathognomonic for NES. Instead, a combination of multiple findings provides a more reliable and complete assessment of each patient. This retrospective analysis provides the foundation for a prospective validation of an objective score to identify patients who are more likely to have a diagnosis of NES, so that they are triaged more rapidly for defini-

tive diagnosis and appropriate treatment. Earlier diagnosis is expected to reduce cost and iatrogenic harm, and to improve quality of life and long-term prognosis.

7.2 Introduction

Prior to an accurate diagnosis of non-epileptic seizures (NES), patients with NES are treated erroneously for epileptic seizures, resulting in increased cost, minimal seizure control, and unnecessary exposure to medications with risks of adverse effects. While epileptic seizures (ES) and non-epileptic seizures (NES) may be behaviorally similar, they are generated by different mechanisms, so treatment for ES generally does not affect NES.[18, 265] These treatments may include emergency department visits with the risk of iatrogenic harm from urgently treating presumed epileptic seizures with intravenous medications and sometimes intubation.[10] [56, 266] On average, there is a 9.2 year delay between seizure-onset and diagnosis of NES,[42] and long-term seizure burden is reduced when NES by early diagnosis.[39, 40] Patients with NES comprise 10% of patients with seizures seen by neurologists.[8, 9] With better understanding of the outpatient presentation of NES, earlier diagnosis may be possible and this would expedite more appropriate care.

Our group, and others, have shown that NES can be differentiated from certain types of ES prior to or during video-electroencephalography (vEEG) monitoring using focused psychiatric assessments,[267] video recordings of seizures,[268, 269] interictal scalp EEG,[8] and interictal flouro-deoxyglucose positron emission tomography (FDG-PET).[7] Numerous publications have reported the diagnostic value of particular clinical findings, or groups of findings, particularly in psychogenic NES,[18, 270, 271, 272, 63, 273] and a few risk scores for NES have been validated on smaller, selected patient populations.[274, 272, 63] Some smaller studies suggest that individual findings including sexual abuse, ictal hip thrusting, and lack of ictal incontinence may be almost perfect predictors of psychogenic NES. Follow-up studies, however, frequently fail to replicate these findings. Additionally, limited work has reliably identified physiologic NES either by ruling out ES or by identifying NES positively, prior to ineffective treatment with ASMs. Thus, previous work leaves an open question: can the plethora of findings associated with seizures be combined to improve the diagnostic accuracy for NES?

Once we establish how findings can be combined, how can we apply that learning in a busy outpatient clinic? While prior studies suggested important psychiatric and pathophysiologic information about the mechanism for NES, a limitation of all of these assessments is that they require more and higher quality data to be collected, and for patients to be selected from a population based upon their subjective risk for NES. In contrast, we focus on diagnostic findings that were reported during an outpatient clinical interview by patients that were admitted ultimately for vEEG monitoring. Our approach aims to maximize the potential applicability of our results because it does not require neurologists to change how they practice medicine: it leverages the massive amounts of data that are available at all epilepsy treatment centers. Studies on more selected populations regarding a limited number of diagnostic findings have been shown to have a remarkably reliable ability to detect NES.[270, 269, 272, 63] Our work aims to replicate and extend these findings on a large, unselected population, while developing a method to combine potentially diagnostic findings to improve our accuracy.

7.3 Methods

7.3.1 Clinical Features

For this study we included 1,126 patients who had been admitted consecutively to the UCLA adult vEEG monitoring unit between January 2006 and April 2014. This sample included 634 with a vEEG diagnosis of ES, 314 with NES, and 178 with either mixed ES and NES or inconclusive monitoring results. Inconclusive results were those who had an insufficient number of typical seizures during monitoring for any reason. Patients were considered to have mixed ES and NES if there was any evidence or suspicion of both seizure etiologies. For instance, if a patient had two seizure types, and one of which was proven with vEEG to be NES, but there was suspicion that the second type could be epileptic even though none were observed on vEEG, the result was considered mixed or inconclusive. We excluded such patients because logistic regression requires mutually exclusive diagnostic classes (ES or NES, not both). Additionally, the diagnostic label for patients with suspicion of mixed etiology and inconclusive monitoring was more uncertain, relative to the gold standard of a fully conclusive vEEG monitoring admission. Methods for integrating data with uncertain diagnostic labels into a predictive algorithm, called semi-supervised learning, are not yet well established. While there are many subtypes of ES and NES, we modeled these diagnostic groups as large and homogenous groups (see Table 7.1 for breakdown) with the understanding that future work can address the ability of predictive models to predict sub-groups within ES and NES that, likely, are very different from each other. In particular, that includes the differentiation of psychogenic from physiologic NES. All forms of NES were included in our analysis to have the results pertain to a real-world situation that includes both and does not require the clinician to exclude physiologic NES when considering whether the seizures are epileptic or non-epileptic. All patients consented for the use of their records in research, and the UCLA Institutional Review Board approved this study. This work is consistent with the Declaration of Helsinki.

To determine the presence or absence of diagnostic findings, WTK and trained undergraduate researchers (CTB, EAJ, JML, JMH, ABP, NLG, JB, AMC) manually annotated each patients first sufficiently detailed report of an outpatient neurology clinical visit at UCLA. There were 258 unique authors of the reports

Table 7.1: Diagnostic subtype breakdown of patients included in our population. Lobe indicates location of seizure-onset zone. If seizure onset zones were in multiple lobes and/or had multiple seizure etiologies, patient was listed in each category. Abbreviations: Not Otherwise Specified (NOS), Lobe Epilepsy (LE), Frontal (F), Temporal (T), Parietal (P), Occipital (O), Non-Epileptic Seizures (NES), Epileptic Seizures (ES).

Diagnostic Subtype	Count	% Total	% of Diagnostic Class
NES	314	28%	
Psychogenic NES	285	25%	91%
Physiologic NES	32	3%	10%
Inconclusive	178	16%	
Insuff. Typical Events	131	12%	74%
Mixed NES + ES	47	4%	26%
ES	634	56%	
Generalized-Onset	33	3%	5%
FLE	115	10%	18%
TLE	343	30%	54%
PLE	31	3%	5%
OLE	19	2%	3%
Focal-Onset NOS	22	2%	3%
NOS	105	9%	17%
Other ES	23	2%	4%
Total	1,126		

with 114 authors writing at least two reports. The most common author was JMS (143 notes). The mean and median number of reports written per author was five and one, respectively. All report authors were blind to this analysis. A report was considered sufficiently detailed if the seizure history was longer than 5 sentences and included a description of the patients seizure semiology. Annotation took approximately 15 minutes per patient for the selected findings. The considered measures were selected based on (1) previous literature reports examined the finding (see online-only appendix), (2) neurologists in the group mentioning the finding in a large number of reports or (3) pilot analysis suggested that a difference existed. This resulted in the inclusion of five categories of findings: demographics, comorbidities, remote history & physical exam findings, peri-ictal characteristics, and pharmacologic history. All findings were recorded as binary presence/absence, unless a natural quantification existed (i.e. age, comorbidity count or seizure frequency). Unclear history was assumed, conservatively, to indicate that the finding was not present. Comprehensive annotation of each finding mentioned in every report was used for training and took 90 minutes per patient, and all undergraduate researchers were trained by annotating at least 10 notes comprehensively prior to examining notes for the selected findings alone. The undergraduate researchers were blinded to the vEEG diagnosis when annotating each outpatient report. To assess inter-rater reliability, 21 patients were coded twice by different undergraduates and a random selection of 10 patients from each undergraduate also were annotated by WTK. For binary features, Cohens kappa was averaged across all logical features. For continuous features, paired correlation was averaged across all continuous features reported (no missing data imputation). The complete list of the 91 considered findings appears in Table 7.2. For a detailed description of how all of these findings were defined and how missing data were modeled, see the online-only appendix.

7.3.2 Statistical Methods

In these real-world clinical records, not all findings were discussed in every note, resulting in missing data. No report discussed all included findings. The findings with missing entries were split into two groups: findings that would be missing completely at random (MCAR), and findings whose absence holds information–

Table 7.2: A complete list of the studied factors from each clinical note. For a detailed description of each factor, please refer to the online-only appendix. Abbreviations: number (#,n), anti-seizure medications (ASMs), medications (meds), review of systems (ROS), gastro-esophageal reflux disorder (GERD), transient ischemic attacks (TIA), neurofibromatosis type 1 (NF1), supplements (suppl).

Demographics	**Comorbidities**	**Peri-Ictal Factors**
Age (years)	# medical comorbidities (n)	Median seizure duration (s)
Gender	# psychiatric comorbidities (n)	# Seizure Types (count)
Family history of seizures	Catastrophic illness	Trigger: sleep deprivation
	Metastatic neoplasia	Trigger: stress
Physical & Historical Exam	Non-metastatic neoplasia	Trigger: loud noises
Duration of seizure disorder (years)	Neurodegenerative disease	Trigger: Menses
Age of onset (<14, <19, >54,	Cardiovascular disease	Auras
year between 19 and 54)	Hypertension (HTN)	Aura: Headache
Seizure frequency (per month)	HTN encephalopathy	Aura: Metallic taste
Remote history of seizures	Atrial fibrillation	Aura: Anxiety/fear
Precipitating event	Diabetes mellitus (DM)	Post-ictal confusion/fatigue
Neuroinfection	Stroke	Sudden onset of seizure
Neurotoxin	TIA	Seizure directly from sleep
Febrile seizures	Developmental delay	Amnesia
Traumatic brain injury (TBI)	NF1	Aphasia
TBI with immediate effects	Hydrocephalus	Dialeptic seizures
TBI with prolonged effects	Psychosomatic disorders	Automatisms
Injury during seizure	GERD/Gastric Ulcers	Lip smacking
Sexual abuse	Migraines	Oral trauma
Physical abuse	Asthma	Tonic/clonic movements
Significant social challenges	Hypothyroidism	Hip thrusting
Substance abuse	Chronic pain	Head movements
Current substance use	Major depressive disorder	Myoclonus
Current smoking	Anxiety disorders	Freezing
Obesity		Secondary generalization
Employment/student status	**Pharmacologic History**	# limbs moving (count)
Premature birth	# current ASMs (n)	Gaze deviation
# complaints on RoS (n)	# failed ASMs (n)	Eye closure
Memory complaints on RoS	# psychiatric meds (n)	Hallucinations
Coordination complaints on RoS	# current meds (n)	Metallic taste
Muscle twitching on RoS	# non-medical suppl (n)	Cry/Scream
Deep tendon reflexes (score+)		Anxiety-like symptoms
		Incontinence

termed “biased” findings (see online appendix for list). For example, if a report did not discuss migraines, it is most likely that the patient does not have migraines or migraines were not an active problem. Therefore, we filled in the missing data with the most likely result. This will result in an underestimate of the diagnostic power of the biased finding. In contrast, an MCAR finding, i.e. seizure frequency, would be relevant for all patients but was not mentioned or quantified. The values of the MCAR finding entries were multiply-imputed 20 times using all other findings in a multivariable logistic, linear, ordinal or categorical regression based on the inherent distribution of the MCAR finding. In brief, single imputation uses the inherent co-linearity between findings to estimate the value of each missing entry. Multiple imputation estimates the value of the missing entry, plus statistical noise based on the goodness of fit of the imputation model. The 20 different realizations of the statistical noise allow the subsequent modeling to place the appropriate amount of confidence in the imputed value. Each realization was run for 10,000 iterations (7 hours per realization, Rhat<1.1). If the MCAR assumption is true, multiple imputation will yield in an unbiased estimate of the diagnostic power of the MCAR findings.[275, 276] All summary statistics were combined into a single aggregate statistic with respect to variance both within and between imputated datasets. Mixed and inconclusive patients were treated as if their diagnosis was MCAR and contributed to the multiple imputation because the relationship between the studied findings was expected to be preserved.

Our predictive analysis was conducted in two phases to answer two similar but separate questions. We used multivariate logistic regression (MLR) within MATLAB (Mathworks, MA) to ask which studied findings were independent predictors of NES, and how predictive is each finding, controlling for all others. However, to implement an objective score of the likelihood of NES, we would require patients and physicians to quantify each significant finding. When we study 91 different findings, this is infeasible. Therefore, we use L1-regularized logistic regression (L_1-LR) to ask: what is the smallest number of questions we can ask to achieve a similar predictive performance? Briefly, L1-regularization applies a penalty to the log-likelihood of the model for every finding that is estimated to have non-zero predictive power. Functionally, this allows us to misclassify more training examples, if it means that we can ask fewer questions. Although to our knowledge L1-regularization has not been applied in the setting of seizures, it is well established in machine learning statistics. As of this writing, the specific software we use to implement the L_1-LR model, Liblinear, has 688 citations. [112]

We built separate MLR and L_1-LR models on each imputed dataset independently using leave-one-out cross-validation. Each of these models predicts a probability of ES versus NES for each patient and imputed dataset. We logit-transformed, and averaged, the predicted probability across the imputed datasets. The transformed average gives more weight to imputed datasets that indicated better confidence in the prediction of ES versus NES. We trained and assessed the accuracy, sensitivity and specificity of our differentiation of patients with a conclusive vEEG diagnosis only (n=950). To estimate the clinical impact of our work, we also examine the accuracy of our algorithm based on each patients reason for vEEG admission.

To ensure the accuracy of our estimates of significance, we estimated the null probability distribution of all summary statistics empirically based on at least 10,000 permutation tests, in which the diagnostic labels were shuffled randomly without replacement. All stages of analysis (imputation and cross-validation) were conducted on each of the at least 10,000 permuted datasets and the quantiles of the observed summary statistic were used to determine significance. This ensures that any bias, non-normality, or overfitting that occurred in our original datasets also could occur in our permuted datasets. Therefore, our p-values should be more accurate than the assumption of Gaussian or binomial statistics.

To interrogate the diagnostic power of each finding independent of all other data, and in combination with other features, univariate diagnostic power was quantified using univariate t or hypergeometric statistics. For univariate statistics, MCAR findings were not filled in, but biased factors were. Binary findings were compared using Fisher exact tests. Continuous findings were compared using heteroschedastic t-statistics on original or, if there was theory suggesting log-normal behavior, log-transformed data. We split the age of onset into 3 indicators and a continuous range: onset before puberty (age 14), onset before age 19, onset after age 54. Age of onsets between 19 and 54 were modeled by their exact value. Due to the co-linearity of duration of seizure disorder, age of onset and current age, current age was excluded from the predictive models. Multivariate diagnostic power was quantified by averaging the log-odds ratio in the predictive models across all cross-validation folds and imputed datasets. This assesses the diagnostic power of each finding, when linearly controlling for every other examined finding. Due to the non-normality and dependence structure of L1-regularized log-odds ratios and to maintain consistency across models, empirical null distributions were calculated for each measure in both models using the permutations described above.

The data regarding the inconclusive patients is presented in Supplementary Table 1 to begin to characterize this difficult to treat group, but it does not provide evidence about the efficacy of our method. Our full & imputed de-identified datasets, as well as a web application to explore our overall models will be published concurrent with this manuscript at `brainmapping.org/MarkCohen/research.html`.

7.4 Results

Our principal finding is that the diverse clinical information collected during a conventional outpatient interview can be combined to accurately distinguish patients with epileptic seizures from those with non-epileptic seizures. At the point of highest overall accuracy (85%, empirical p<0.001), the sensitivity and specificity of both our models were 90% and 70%. As discussed below, these high accuracies were achieved using only a subset of the 91 findings we annotated from the clinical records.

Nine findings were mentioned either positively or negatively in less than 20 patient notes, and therefore were excluded from analysis: transient ischemic attacks, atrial fibrillation, neuro-fibromatosis type 1, neurodegenerative disease, hydrocephalus, psychosomatic disease, hypertensive encephalopathy, metallic taste during seizures, and metastatic neoplasia. There was no significant difference in the probability of missing data for NES versus ES for any MCAR feature (Fisher exact tests, $p>0.068$) except for the duration of seizure disorder and age of onset, which were reported more frequently for patients with ES (95% vs. 90% for NES, Fisher exact tests $p<0.002$). After imputation, the sign and magnitude of the mean difference between the populations on these findings did not change substantially. Cohens Kappa for logical features indicated moderate agreement (κ=0.50), whereas the paired correlation for continuous features showed near perfect agreement (ρ=0.83).

In the summary tables and figures, all reported intervals reflect 95% confidence. Table 7.3 and 7.4 illustrates the significant differences observed between patients with ES and NES. The findings with significant multivariate odds ratios, as determined by MLR with multiple imputation, are illustrated in Figure 7.1. Expanded versions of Table 7.3, Table 7.4, and Figure 7.1, with all studied findings, are available in the online-only supplement. The significant and full odds ratios from the L_1-LR model are available in Supplemental Figures.

The cross-validation performance of the L_1-LR model to predict the vEEG diagnosis is illustrated in Figure 7.2. The cross-validation performance of the MLR model is illustrated in a Supplemental Figure, and was almost identical to the L_1-LR model performance. Performance based on the reason for vEEG admission and diagnostic subclass is listed in Supplemental Figures. No subgroup performed significantly better or worse than would be predicted by the fraction of NES to ES within the group (z-test of proportions, $p>0.1$).

Figure 7.1: Odds ratio of all factors with a significant factors with at least 95% confidence based on multivariate logistic regression (MLR) with multiple imputation. The size of the colored bar indicates the magnitude of the odds ratio, and the color indicates its sign (Green for ES, Blue for NES). The transparent gray overlay indicates the 95% empirical confidence interval of chance. For prevalence, see Supplementary Table 1. All factors were indicators unless the units are specified otherwise. All starred units (*) were normalized across the whole dataset therefore bars reflect the odds ratio of an increase of one standard deviation. Because p–values were not independent across factors, no correction for multiple testing was applied. Abbreviations: number (#,n), seizure disorder (SzD), years old (y.o.), traumatic brain injury (TBI), with (w/), anti-seizure medication (ASM), seizure (Sz), movements (mvmts).

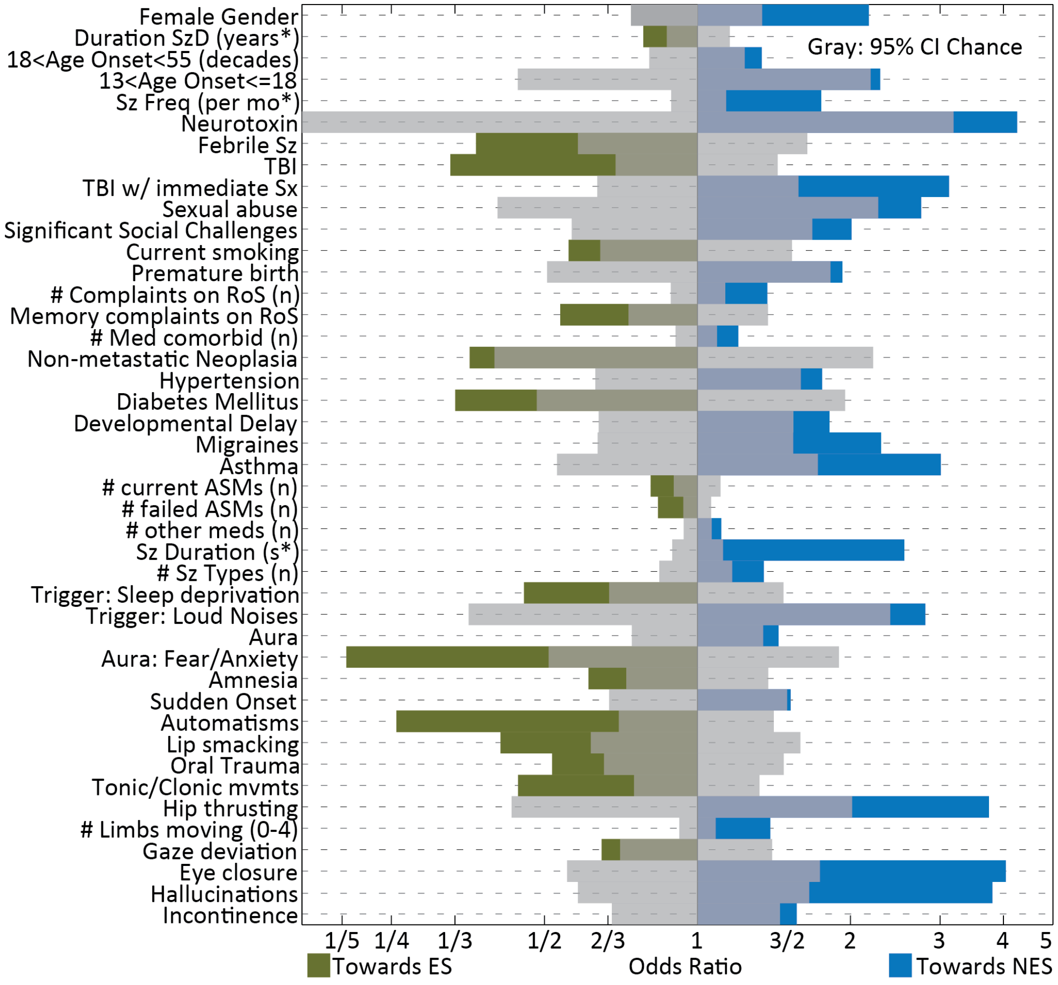


Figure 7.2: Leave-one-out cross-validation performance statistics of the data–driven L1 regularized logistic regression (L_1-LR) to differentiate between Non-Epileptic Seizures (NES) and Epileptic Seizures (ES). Accuracy, sensitivity and specificity (7.2A) describe the point on the receiver operating curve (ROC, 7.2B) that maximizes accuracy. Shading reflects empirical 95% confidence intervals of chance. Abbreviations: area under the ROC (AUC).

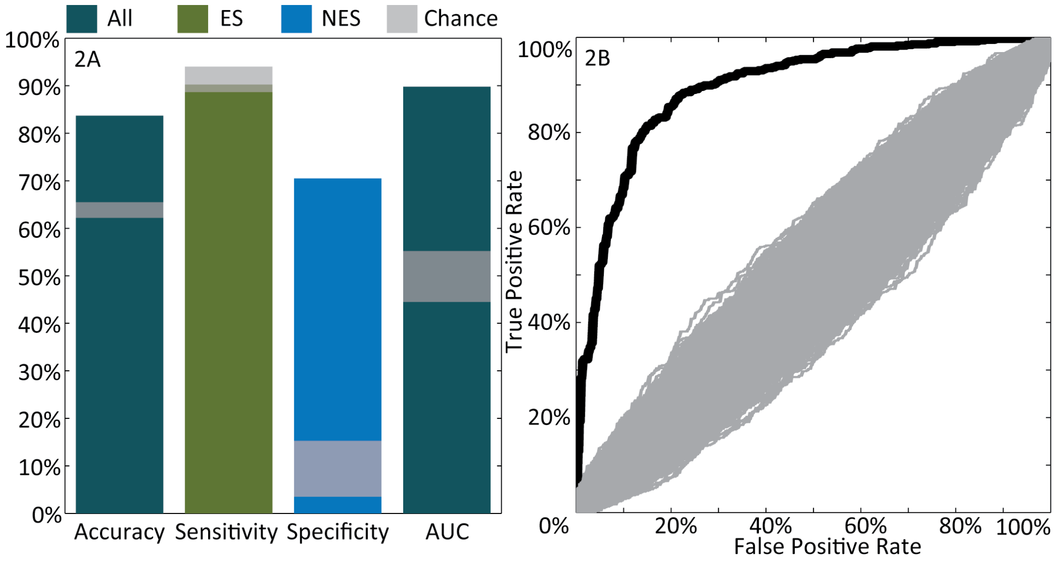


Table 7.3: The significant differences between patients with epileptic seizures (ES), non-epileptic seizures (NES) and inconclusive/mixed diagnosis based on video-electroencephalography. All factors are listed in percentages, unless otherwise specified. *Statistics of seizure duration, seizure frequency, and duration of seizure disorder were calculated in log space, but are displayed in linear values (i.e. displayed value is $10^{[mean[\log_{10} data]]}$). All unlisted factors exhibited no statistically significant differences between ES and NES with greater than 95% confidence (false discovery rate, α=0.05). Dashes in missing data indicate a variable assumes to be biased (see Methods). Abbreviations: number (#), missing data (M), seizure disorder (SzD), standard error (SE), seizures (Sz), anti-seizure medications (ASMs), review of systems (RoS), postictal (PI), deep tendon reflexes (DTRs), comorbidities (comorbid).

Feature Name	%M	Mean	SE	%M	Mean	SE	p
Demographics							
Age (years)	0%	39.8	0.9	0%	35.1	0.6	2.E-05
Female Gender	0%	73%	3%	0%	53%	2%	8.E-09

7.5 Discussion

The high diagnostic value of the clinical interview to differentiate between patients with non-epileptic versus epileptic seizures mirrors the fact that the diagnosis of seizures is inherently clinical. Nevertheless, physicians in the clinic are challenged to integrate a large number of diverse findings in order to make the important single distinction between ES and NES for individual patients. While no substitute for experience and clinical acumen, we believe that providing the clinician with a statistically-derived objective metric has the potential to augment the diagnostic process, reducing time between evaluation and treatment, and improving diagnostic accuracy overall.[62] The results also may be useful to help orient the physician to aspects of the history that are predictive of a diagnosis, even if the physician

Table 7.4: See legend for Table 7.3.

Feature Name	%M	Mean	SE	%M	Mean	SE	*p*
Historical & Physical							
Duration SzD (years*)	12%	2.66	0.30	5%	9.33	0.55	3.E-20
Age Onset $\leq$ 13	10%	15%	2%	5%	47%	2%	4.E-27
18<Age Onset<55 (years)	10%	34.3	0.5	5%	31.1	0.4	6.E-04
Age Onset$\geq$ 55	10%	8%	2%	5%	4%	1%	7.E-03
Sz Freq (per mo*)	26%	13.2	1.4	21%	5.9	0.5	3.E-08
Remote Hx Seizure	-	5%	1%	-	10%	1%	1.E-02
Febrile Sz	-	3%	1%	-	10%	1%	2.E-05
TBI w/ immediate	-	26%	2%	-	16%	2%	8.E-04
Sexual abuse	-	10%	2%	-	1%	0%	2.E-09
Physical abuse	-	9%	2%	-	2%	1%	7.E-06
Social challenges	-	13%	2%	-	5%	1%	2.E-05
# complaints on RoS (n)	42%	8.86	0.55	43%	5.52	0.25	4.E-05
Coord complaints on RoS	-	20%	2%	-	12%	1%	1.E-03
Muscle twitching on RoS	-	10%	2%	-	6%	1%	1.E-02
DTRs (score+)	-	1.99	0.03	-	1.90	0.02	8.E-03
Co-Morbidities							
# medical comorbid (n)	-	2.82	0.16	-	1.18	0.06	1.E-19
# psych comorbid (n)	-	0.68	0.05	-	0.33	0.02	1.E-09
Hypertension	-	17%	2%	-	8%	1%	2.E-04
GERD/Ulcers	-	16%	2%	-	4%	1%	2.E-07
Migraines	-	21%	2%	-	7%	1%	1.E-10
Asthma	-	13%	2%	-	3%	1%	1.E-08
Hypothyroid	-	10%	2%	-	6%	1%	1.E-02
Chronic Pain	-	15%	2%	-	3%	1%	1.E-10
Depression	-	30%	3%	-	16%	1%	4.E-07
Anxiety	-	16%	2%	-	7%	1%	4.E-05
Pharmacologic							
# current AEDs (n)	-	1.65	0.08	-	2.17	0.05	8.E-08
# failed AEDs (n)	-	1.37	0.11	-	2.61	0.10	5.E-17
# psych meds (n)	-	0.54	0.05	-	0.21	0.02	1.E-08
# other meds (n)	-	3.24	0.22	-	1.07	0.08	6.E-19
# non-med suppl (n)	-	0.50	0.06	-	0.28	0.03	3.E-03
Peri-Ictal							
Sz Duration (min*)	35%	3.85	0.40	30%	1.16	0.07	7.E-16
Trigger: sleep dep	-	10%	2%	-	20%	2%	5.E-05
Trigger: menses	-	4%	1%	-	9%	1%	2.E-03
Auras	-	38%	3%	-	46%	2%	3.E-02
Aura: headache	-	6%	1%	-	3%	1%	3.E-02
Aura: metallic taste	-	6%	1%	-	3%	1%	2.E-02
Aura: anxiety/fear	-	1%	1%	-	6%	1%	4.E-04
PI confusion/fatigue	-	29%	3%	-	43%	2%	1.E-05
Directly from sleep	-	13%	2%	-	22%	2%	8.E-04
Amnesia	-	19%	2%	-	27%	2%	8.E-03
Automatisms	-	7%	1%	-	24%	2%	1.E-10
Lip smacking	-	4%	1%	-	11%	1%	2.E-04
Oral Trauma	-	9%	2%	-	17%	1%	2.E-03
Tonic/Clonic mvmts	-	26%	2%	-	48%	2%	1.E-10
Hip Thrusting	-	8%	1%	-	2%	1%	2.E-05
Myoclonus	-	19%	2%	-	11%	1%	1.E-03
Gaze deviation	-	14%	2%	-	20%	2%	1.E-02
Eye closure	-	14%	2%	-	3%	1%	5.E-10
Hallucinations	-	11%	2%	-	6%	1%	2.E-03

does not determine the metric. While prior literature has claimed certain findings as near-perfect classifiers, a clear message of our analysis is that one cannot rely on a single finding to identify NES. Therefore, providers should look for a combination of findings that, in aggregate, raise or lower the suspicion that the patients seizures are non-epileptic. There are two important applications of this work: a resource for illustrating the diagnostic value of a large variety of findings both on their own and in combination, and an objective risk score to assess the likelihood of NES in the outpatient clinic. Patients with high outpatient likelihood for NES should be triaged more quickly towards video-EEG monitoring for definitive diagnosis and appropriate treatment. [36]

7.5.1 Diagnostic Value of Particular Findings

While the number of studied features is too large to provide an interpretation and context for each, we highlight several key findings to help understand our results. This interpretation focuses on psychogenic NES, but it is important to note that 10% of our NES population had physiologic NES. The mechanism for each are very different, but further splitting the classes is outside the scope of this work.

We found that our patients with NES had, by description or diagnosis, more organic and more psychiatric dysfunction. Although psychogenic NES often is considered a conversion disorder, a biological correlate for this increased somatic sensitivity, [277, 278, 279, 280] is apparent in neuroimaging studies.[281, 282, 283, 115, 284, 285] The complaints and common comorbidities were relatively non-specific, with almost all comorbidities and specific complaints being more common in NES. This supports the hypothesis that patients with psychogenic NES may report more disability for the same experience.

Our perspective of looking at features in combination helped us understand how to interpret specific findings within the review of systems (RoS). Patients with

NES reported far more somatic problems during the review of systems, including coordination problems and resting muscle twitches (univariate, MLR & L_1-LR $p<0.05$). When controlling for the number of complaints during the RoS, coordination problems and muscle twitches no longer were suggestive of NES (MLR & L_1-LR $p>0.1$). This suggests that these signs may reflect increased somatic sensitivity but do not have an interpretation outside that context. In contrast, the specific complaint of memory issues indicated ES when controlling for the number of complaints, but not when viewed alone (univariate $p>0.7$, MLR & L_1-LR $p<0.05$). That is, the addition of one more complaint on RoS would usually indicate NES. If, however, the patient reported memory issues, ES was more likely, potentially reflecting hippocampal dysfunction that is characteristic of temporal lobe epilepsy.[30]

Aside from between-seizure features, the ictal features of an NES differed from ES in ways that we did not expect. Classical ictal signs of PNES (i.e. hip thrusting & eye closure) were present much more frequently in NES but also were present in ES (univariate, MLR & L_1-LR $p<0.05$). Even though the shibboleth is that patients with PNES maintain bladder and bowel autonomic tone,[266, 286] this has been disputed by other researchers and our patients with NES who reported ictal injuries and incontinence just as frequently as patients with ES (univariate $p>0.6$).[287, 288] However, ictal incontinence indicated NES in our group, when considered with other factors (MLR & L_1-LR odds ratio for NES 1.5, $p<0.05$). This suggests that when ictal incontinence occurs without findings strongly associated with ES (i.e. a generalized seizure) the incontinence may indicate NES.

Our findings might be used as guidelines to assess specific factors in clinical reports to promote more accurate diagnoses in the future. While our model performed well, an important challenge in the use of clinical notes is that critical details often are not recorded. For instance, patients with ES reportedly were more likely to bite their cheek and the sides of their tongue, whereas patients with

PNES reportedly more frequently bite their lip and the tip of their tongue.[266] We identified oral trauma as a risk factor for ES (univariate, MLR & L_1-LR $p<0.05$); however, the nature of oral trauma was specified in only 26% of notes. Our reported results reflect the diagnostic utility of our defined categories (see online-only appendix), which may or may not be improved by further refinement.

These examples of incontinence and memory dysfunction illustrate the more general observation that multivariate models highlight the importance of considering the patterns of findings to arrive at a more complete picture of the patient, instead of focusing on individual factors on their own. Even factors that were significant on a multivariate level should not be interpreted as diagnostic in isolation.[111] Instead, the pattern associated with changes in this factor result in our ability to differentiate NES from ES.

7.5.2 Developing an Outpatient Assessment Score

In addition to validating and building upon the observations of the extensive prior literature on the differences between patients with NES and ES, our methods could be used to generate an objective, quantitative and individualized assessment of a patients chance of NES using only information acquired in the out-patient setting. While the MLR model identified 43 independent predictive findings to help differentiate NES from ES (Figure 1), asking this many questions of a patient may prove impractical in clinical practice. In comparison, our L_1-LR model balanced the number of contributing findings with the predictive accuracy and achieved similar and impressive out-of-sample performance using just 31 questions (Supplemental Figure 2). These strong retrospective results suggest that the prospective application of one or both of these models may provide valuable clinical insights to identify patients with NES quickly and effectively. Prior to implementation, however, this must be confirmed directly through a prospective study.

It is particularly notable that our methods are based on data from clinical notes that were written for the purpose of documentation and communication, and not for research to assess the diagnostic value of any of the studied factors [89]. Clinical data seldom are pristine, and an analytic method that depends on near-perfect research quality inputs would have limited value in real-world health care settings; we chose instead to develop a method that acknowledges the potential for bias, missing points and other contaminations because these same biases would be present in most outpatient clinics for seizures. For instance, neurologists do not perform a full cardiovascular and psychiatric assessment. Instead, comorbidities come up as part of care for the seizures. Therefore, we expect our estimate of the frequency of other medical and psychiatric comorbidities to be an underestimate of the true burden of disease. As discussed above, patients with NES may or may not have more medical and psychiatric comorbidities than patients with ES, but the perceived impact of these conditions may be higher, leading the patient to be more likely to discuss them with their neurologist. We expect therefore that when domain-specific experts examine our factors with specific, validated data collection and assessment tools, the frequency and magnitude of the differences we saw will change, reflecting a more accurate description. However, we expect our estimates to match the frequencies with which outpatient providers will learn of these conditions. Our inclusion of patients with other neurologic and psychiatric comorbidities, physiologic NES, developmental delay, traumatic brain injury and previous epilepsy surgery prior to being seen at UCLA complicates the analysis and interpretation of results but also ensures that our population best matches the full complexity of patients with seizure-like events at a tertiary epilepsy center. Our lack of a seizure-naive control group both was convenient and valuable for a number of reasons.[81] Our goal was to describe how to differentiate NES from ES. Seizure-naive controls could be differentiated perfectly from our patients, for example, simply by asking if they had experienced repeated seizure-like events.

7.5.3 Limitations

The difference between our results and previous literature from smaller populations is consistent with how the magnitude of statistical effects typically decreases as sample size increases,[86] but also could be explained by differences in our population. UCLA draws patients a large area of the southwestern United States. Our population of patients with ES may have more medically refractory epilepsy and, in particular, more temporal lobe epilepsy and lesional epilepsies than outpatient clinics for epilepsy and other tertiary care centers. The proportion of patients with NES (28%), and the delay to diagnosis (8.7 years) was similar to other centers[42] but the proportion of patients with confirmed or potential mixed NES plus ES (12.6% of all NES) was lower at UCLA than at other centers, although this proportion varies widely by definitions and centers.[289, 290, 269, 291] These populations, however, do not reflect patients with NES that were identified in outpatient clinics and did not require referral to a tertiary care center.

7.5.4 Future Potential and Perspective

If validated prospectively, these results could be used to identify 70% of patients with NES quickly and cost-effectively in a variety of outpatient settings. Because patients with NES are a highly heterogeneous group, our negative predictive value will always be less than 100%. Our data might lead to more timely referral of patients with likely NES for vEEG monitoring; however, we strongly believe all patients with disabling seizures that persist despite trials of two appropriate ASMs deserve a consultation by specialists at a full service tertiary care epilepsy center. A reduction in time to diagnosis of NES has been shown to improve long-term seizure control, while reducing cost and potential for mistreatment.[56, 266] While other risk scores exist,[274, 272, 63] none have achieved such high performance on this large of a population. Clearly, important validation steps like prospective data collection must occur before we can conclude that our score can and should be implemented in outpatient clinics. This work represents a foundational first step in quantitatively assessing the value of combining a wide variety of findings reported in the outpatient clinic for differentiating between NES and ES.

CHAPTER 8

Automated Diagnosis of Epilepsy using EEG Power Spectrum

This work is a reproduction of our work published in Epilepsia. [8] This work was a collaboration with Ariana Anderson, Edward P. Lau, Andrew Y. Cho, Hongjing Xia, Jennifer Bramen, Pamela K. Douglas, Eric S. Braun, John M. Stern and Mark S. Cohen. WTK organized the collaboration, curated the data, wrote the code and the majority of the manuscript. AA, JB & PKD provided guidance for experimental and machine learning design, as well as interpretation of the results. EPL & AYC provided substantial computational support to get the analysis stages to be effective. HX assisted with understanding & preprocessing of the EEG data and extraction of meaningful features. ESB was responsible for Akima interpolation and helped understand the sampling and optimization of the method. JMS provided direction and context with regard to how our work fits within the context of the diagnostic assessment for seizure disorder. MSC assisted with all stages of the experiment and manuscript preparation.

8.1 Abstract

Interictal electroencephalography (EEG) has clinically meaningful limitations in its sensitivity and specificity in the diagnosis of epilepsy because of its dependence on the occurrence of epileptiform discharges. We have developed a computer-aided diagnostic (CAD) tool that operates on the absolute spectral energy of the routine EEG and has both substantially higher sensitivity and negative predictive value than the identification of interictal epileptiform discharges. Our approach used a multilayer perceptron to classify 156 patients admitted for video-EEG monitoring. The patient population was diagnostically diverse with 87 diagnosed with either generalized or focal seizures. The remainder was diagnosed with non-epileptic seizures. The sensitivity was 92% (95% CI: 85-97%) and the negative predictive value was 82% (95% CI: 67%-92%). We discuss how these findings suggest that this CAD can be used to supplement event-based analysis by trained epileptologists.

8.2 Introduction

Epilepsy is common and has a major impact on the global burden of disease. Though epilepsy is defined as an enduring predisposition for seizures, its diagnostic assessment relies on the clinical and/or electrographic description of transient events. Consequentially, the sensitivity of a single outpatient interictal electroencephalography (EEG) is only 50% [56, 57]. If physicians do not observe the hallmark electrical features of interictal epileptiform discharges (IEDs), the assessment is inconclusive. This might help to explain why the average time to the diagnosis of non-epileptic seizures (NES) is 7.2 years [41]. Automated seizure detection algorithms currently help physicians identify these transient events [69], but they do not detect the stable pathology underlying each patients chronic disease. A better understanding of the chronic state of epilepsy has great potential to impact patient care; automated computer methods have the potential to identify this stable abnormality and thereby to increase diagnostic accuracy, saving clinicians valuable time and improving patients quality of care.

Seizure detection and prediction tools in epilepsy have been proposed frequently, yet efficient and effective computer aided diagnostic (CAD) tools have not yet been established. Only three publications address the question of epilepsy diagnosis using interictal scalp EEG alone [72, 73, 74]. All three publications report accuracies in excess of 90%. Other publications using the Freiburg dataset compare scalp EEG from normal controls to interictal intracranial EEG patient with epilepsy, which may limit clinical applicability [70]. Based on their success in the seizure and prediction literature, these tools used largely wavelet-based analysis and time frequency decompositions of short time windows of the signal [66, 67, 69]. However, longer time windows can capture the stable changes baseline dynamics attributable to epilepsy. The previous literature often compares the EEGs of patients with epilepsy to the EEGs from a healthy control population, a question that does not reflect the actual clinical situation. We consider comparing epilepsy to NES mimics the clinical scenario of a patient that needs to be assessed after experiencing a potential seizure event. As we show below, 30 percent of patients admitted for video-EEG monitoring have NES, including some who previously were diagnosed with intractable epilepsy. To develop directly clinically applicable tools, the diagnosis of each patient in the validation set must be certain therefore a careful discussion of the diagnostic assessment of each patient is critical. Similarly, epilepsy is a heterogeneous syndrome. Generally, the CAD literature either studies temporal lobe epilepsy or does not specify diagnostic subclass.

In this report, we outline the success of a novel CAD tool applied to a larger population of patients who have either focal or generalized epilepsies. By comparing to patients with NES and also inspecting time-frequency features of longer time windows of the EEG signal, we harness the stable interictal changes in the EEG that can be used to diagnose epilepsy. Further, we provide a detailed discussion of how such a tool can be used to supplement, not replace, manual analysis.

8.3 Patients and Methods

We studied the diagnostic test results from 514 patients admitted between 2008 and 2011 to the UCLA Seizure Disorder Center video-EEG monitoring unit. A subset of 156 patients was identified for further study because their diagnoses were definitive and they had not experienced previous penetrating head trauma. Within this subset, 87 were diagnosed with epilepsy and 69 were diagnosed with NES (full breakdown in Supplementary Information). Patients with NES and those with epilepsy underwent an identical evaluation. All methods were approved by the UCLA IRB and complied with the Helsinki Declaration.

All scalp EEG recordings were collected in accordance with standardized clinical procedures with a 200 Hz sampling rate using 26 electrodes placed according to the International 10-20 system. During acquisition, an analog 0.5 Hz high pass filter was applied to all recordings. Reviewed data consisted of between 1.5 and 25 hours (mean 9 hours, S.D. 4.5 hours) of archived EEG from either the first or second night of video-EEG monitoring. To assess the diagnostic yield of long term monitoring, we also inspected the records of all 514 reviewed patients admitted to UCLA for video-EEG monitoring.

The mean, standard deviation, minimum and maximum absolute spectral energy for non-overlapping 1 sec, 5 sec, 60 sec, 30 min windows of EEG recordings from all electrodes relative to reference electrode 1, located between Fz and Cz, were calculated using the fast Fourier transform in MATLAB. The absolute value of spectral energy from 1-100 Hz was averaged over 1 Hz spectral bands. Short window lengths measure phenomena analogous to event related spectral perturbations (ERSPs) whereas longer windows capture baseline activity and connectivity. Each input feature corresponds to a separate electrode location, frequency band, statistical parameter and window length. The spectral energy from 58-62 Hz was excluded from all analysis to avoid AC line noise contamination. No other artifacts were removed. Ictal activity and muscle artifact were included in analysis.

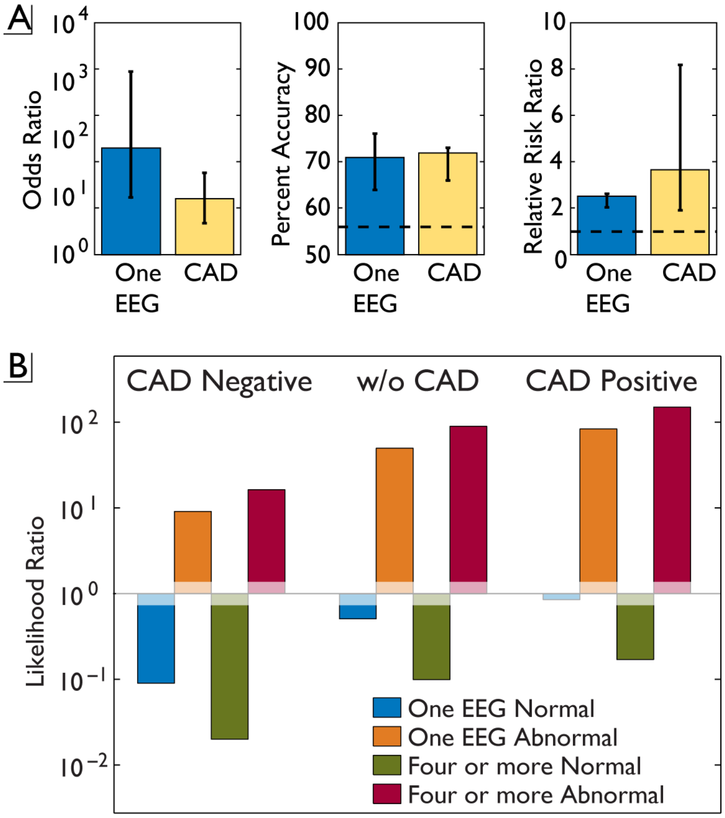
Using a cyclical leave-one-out cross validation technique, a subset of the power spectrum was identified as potentially diagnostic by a highly-efficient minimum redundancy, maximum relevancy (mRMR) feature selection algorithm [5, 6]. This subset was then used as input for the Multilayer Perceptron neural network algorithm as implemented in Weka [292]. For algorithmic details please refer to the supplementary material and Kerr et al. [293].

8.4 Results

The Multilayer Perceptron performance was comparable to manual event-based EEG analysis. Both manual and automated analyses were substantially and significantly better than a chance classifier based on clinical trial statistics (Figure 8.1). The diagnostic accuracy of the CAD tool was 71% (64%-76%, $p < 10^{-4}$) significantly higher than chance: 56%. The risk ratio (the probability that a positive finding occurred in a patient with epilepsy compared to a patient with NES) was 3.68 (1.92-8.19, $p < 10^{-6}$). The odds ratio was 9.32 (3.51-25.73, $p < 10^{-5}$). In the study population, the results of a single outpatient non-video-EEG are not significantly different and have a relative risk ratio, odds ratio and accuracy of 2.52 (2.05-2.64, $p < 10^{-10}$), 99 (8.90-1100, $p < 10^{-3}$) and 72% (66-73%, $p < 10^{-4}$), respectively [56, 57]. All intervals reflect 95% confidence bounds and all p values reflect comparisons to a naive classifier.

In contrast with manual analysis, the performance of our CAD was driven by exceptionally high sensitivity (85%-97%, $p < 10^{-82}$) in comparison to only modest specificity (37%-51%, $p > 0.20$). Consequentially, the negative predictive value (67%-92%, $p < 10^{-24}$) is high compared to the positive predictive value (62%-71%, $p < 10^{-5}$). There was no significant difference in performance for focal and generalized epilepsies (see Suppl. Materials).

Figure 8.1: (A) directly compares the summary statistics of our computer- aided diagnostic (CAD) tool to the same statistics regarding conventional analysis of EEG. (B) assesses the likelihood ratios that can be achieved when our CAD is combined with conventional analysis. Error bars denote 95% CIs and are calculated without normal assumptions. Dashed lines indicate chance or 95% CIs of chance. All effects are significantly different from chance ($p < 0.001$) except when CAD is positive and manual analysis is negative. No comparative effects are significantly different.



These results can be expressed in combination with the results of outpatient non-video EEGs as likelihood ratios, assuming the two tests are independent (Figure 8.1) based on the formula:

$$LR_{-_M+_{CAD}} = \frac{P(-_M|Ep)}{P(-_M|NES)}\frac{P(+_{CAD}|Ep)}{P(+_{CAD}|NES)}.$$

We assume that 99% of neurologically normal patients have negative EEGs and that 50% and 90% of patients with epilepsy have abnormal outpatient EEGs after 1 and 4+ recordings, respectively [56, 57].

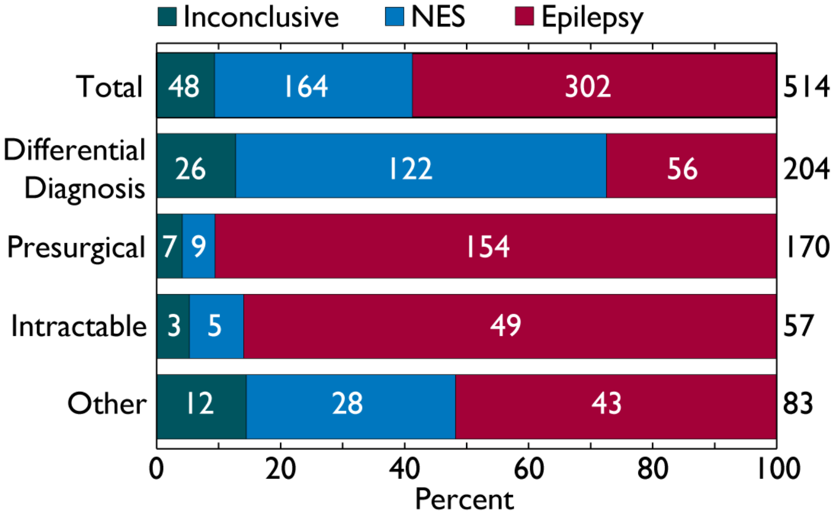
To illustrate the clinical problem further, we addressed the diagnostic yield of long term video-EEG monitoring specifically. As summarized in Figure 8.2, 9 percent of the 514 patients in our sample had inconclusive results upon the completion of monitoring (6%-12%). Six percent of patients admitted for pre-surgical assessment or intractable epilepsy were diagnosed with NES (2-10%).

8.5 Conclusion

Inconclusive EEG results are a significant challenge to the effective treatment of epilepsy. For patients diagnosed with epilepsy, our finding that 6% of patients are later found to have NES is concerning. Further, the most reliable diagnostic test, conventional long term video-EEG monitoring, is inconclusive for roughly 9% of epilepsy patients due to lack of relevant electrophysiological events. To reduce this rate, admission duration must increase. Our technology, however, avoids this problem altogether by focusing on baseline diagnostic features. Successful validation and then implementation of our CAD tool could therefore provide additional information to that could, in time, substantially reduce both of these values. Validation would require a prospective assessment of patients who are later admitted for video-EEG monitoring or retrospective analysis of records from other institution(s).

We hypothesize that our results capitalize both on low frequency trends used

Figure 8.2: Diagnostic yield of long-term video-EEG monitoring. Numbers indicate how many patients are in each class, and the size of the bar denotes percent of total, listed on the right side of the figure, that belong to each class. When the presurgical and intractable classes are combined, 6% of the patients have inconclusive results. NES, nonepileptic seizures.



by previous literature and, potentially, also on high frequency oscillations up to 100 Hz. Most ictal activity is within the 3-25 Hz range [69]. Seizure detection algorithms have achieved impressive results operating on frequency bands less than 40 Hz using much more complex machine learning methods [70]. However, recent evidence in intracranial EEG suggests that patients with epilepsy have increased high frequency oscillations in the 40+ Hz range [66, 67]. Due to the nature of our algorithm, the contribution of each window length, spectral band and electrode location is unclear.

Our entirely automated tool diagnosed patients with performance similar to epileptologists manually reading outpatient EEGs. Our performance was quantitatively less than previous methods. However, ours was designed and tested in the real-world context of an inpatient unit, with its heterogeneous mixture of medications, ages and patient histories. The statistics reveal that our approach has a high negative predictive value whereas manual analysis has, instead, a high positive predictive value. These improvements are based on information not observable without CAD and are independent of rate expertise, suggesting that our methods can be used in combination with manual analysis to improve the diagnostic yield of EEG. This synergistic combination could more efficiently and quickly identify those patients who may require further diagnostic or pre-surgical assessment. Given the broad and growing evidence that early epilepsy surgery – when supported by accurate diagnostics – may be more effective than treatment with AEDs alone [32], we believe that this application offers the potential to meaningfully impact the care of patients with epilepsy.

CHAPTER 9

Computer aided diagnosis and localization of lateralized temporal lobe epilepsy using interictal FDG-PET

This chapter is a reproduction of our work that was published in Frontiers in Neurology. [7] This work was a collaboration with Stefan T. Nguyen, Andrew Y. Cho, Edward P. Lau, Daniel H. Silverman, Pamela K. Douglas, Navya M. Reddy, Ariana Anderson, Jennifer Bramen, Noriko Salamon, John M. Stern and Mark S. Cohen. WTK organized the collaboration, wrote the code and the majority of the manuscript. STN & NMR processed the data through NeuroQ and assisted with manuscript editing. STN also curated the dataset. DHS supervised STN & NMR, and helped with direction and interpretation of PET data. AYC & EPL assisted with parallel processing of the images. PKD, AA & JB assisted with design of the experiment and machine learning algorithms as well as editing of the manuscript. JMS & NS helped understand how this work fits in the context of the pre-surgical and diagnostic assessment for seizure disorder. MSC assisted with all stages of planning and trouble shooting the manuscript.

9.1 Abstract

Interictal FDG-PET (iPET) is a core tool for localizing the epileptogenic focus, potentially before structural MRI, that does not require rare and transient epileptiform discharges or seizures on EEG. The visual interpretation of iPET is challenging and requires years of epilepsy-specific expertise. We have developed an automated computer-aided diagnostic (CAD) tool that has the potential to work both independent of and synergistically with expert analysis. Our tool operates on distributed metabolic changes across the whole brain measured by iPET to both diagnose and lateralize temporal lobe epilepsy. When diagnosing left temporal lobe epilepsy (LTLE) or right TLE (RTLE) versus non-epileptic seizures (NES), our accuracy in reproducing the results of the gold standard long term video-EEG monitoring was 82% (95% confidence interval [CI] 69-90%) or 88% (95% CI 76-94%), respectively. The classifier that both diagnosed and lateralized the disease had overall accuracy of 76% (95% CI 66-84%), where 89% (95% CI 77-96%) of patients correctly identified with epilepsy were correctly lateralized. When identifying LTLE, our CAD tool utilized metabolic changes across the entire brain. By contrast, only temporal regions and the right frontal lobe cortex, were needed to identify RTLE accurately, a finding consistent with clinical observations and indicative of a potential pathophysiological difference between RTLE and LTLE. The goal of CADs is to complement–not replace–expert analysis. In our dataset, the accuracy of manual analysis of iPET (∼80%) was similar to CAD. The square correlation between our CAD tool and manual analysis, however, was only 30%, indicating that our CAD tool does not recreate manual analysis. The addition of clinical information to our CAD, however, did not substantively change performance. These results suggest that automated analysis might provide clinically valuable information to focus treatment more effectively.

9.2 Introduction

It is difficult to differentiate between patients with epilepsy, and those with non-epileptic seizures (NES). The clinical assessment relies on the report of untrained witnesses or the patients themselves. A non-epileptic seizure is defined as the presence of external seizure symptoms and/or signs with no electrographic features characteristic of epilepsy. Long term video-EEG monitoring has shown consistently that roughly one third of patients diagnosed with medication refractory epilepsy in fact suffer from NES [8]. Because they dont suffer from epilepsy, these patients with NES (PWN) are not treated effectively with anti-epileptic drugs (AEDs). For the majority of PWN, the NES are a manifestation of dissociative or conversion disorder in which their psychological challenges manifest themselves physically [294, 295]. A minority of PWN suffers from organic, non-epileptic maladies that can be confused with seizure disorder including, but not limited to, dementia and cardiovascular disease [265]. The gold standard for the differential diagnosis and pre-surgical assessment of epilepsy includes 72 or more hours of video-EEG monitoring [296, 38]. However, 10 percent of patients admitted for this extensive assessment leave with inconclusive results [8]. Considering that one sixth of patients with epilepsy are diagnosed with medication refractory epilepsy [297], improved methods to effectively identify PWN who do not benefit from AEDs effectively could reduce the morbidity and both the financial and social cost of treating epilepsy.

Improved diagnostic tools could also help patients with epilepsy (PWE). The difficulty in ruling out non-epileptic etiologies speaks to the challenge of adequately localizing and characterizing each patients epileptic etiology. The major seizure type discriminations are focal versus generalized; partial versus complex; and lesional versus non-lesional. Each of these key discriminations leads patients down a different treatment path. When medication or other novel treatments like the vagus nerve stimulator fails, as they frequently do, the patient is left to consider resective neurosurgery. Recent reports have shown that surgery is most effective earlier in the course of disease [32]. Improved diagnostic tools could more quickly and effectively diagnose patients with epileptic seizures and therefore speed the progression towards considering the surgical option.

Ultimately, our goal is to establish a general, automated computer-aided diagnostic (CAD) tool that effectively combines clinical information, manual interpretation of EEG and imaging technologies as well as automated analysis of interictal FDG-PET (iPET), EEG, structural MRI (sMRI) and diffusion MRI for all subtypes of epilepsy and NES. To accomplish this, we first must develop effective CAD tools that harness the information from each modality for a limited set of epileptic localizations. We have begun already to address automated analysis of interictal EEG for a wide variety of epilepsy subtypes [8]. Others have described effective CAD tools that diagnose and lateralize temporal lobe epilepsy (TLE) using structural and diffusion MRI [26, 27, 28].

The clinical, metabolic and structural differences between left and right TLE can be subtle. Some theories suggest that TLE is inherently a bilateral disease. Potentially, due to the strong functional link between the hippocampi, the only clinical difference is that in the aura of patient with left TLE (LTLE) more frequently includes language dysfunction. Over time, patients with LTLE more commonly develop verbal memory deficits, compared to non-verbal memory deficits in right TLE (RTLE) [298, 299]. This functional connection between the hippocampi may also lead some patients to be falsely lateralized using scalp EEG because a small seizure onset zone (SOZ) in one hippocampus can induce larger scale ictal activity in the contralateral hippocampus with very little time delay. This can lead neurologists to falsely conclude that the SOZ is either bilateral or in the contralateral hippocampus. Structural and metabolic imaging can reduce these errors by demonstrating that that one temporal lobe is asymmetrically affected, as shown by the previous CAD tools that lateralize TLE [26, 27, 28]. Studies of the functional connectivity of these epileptic networks, however, conclude that there are very few, if any, differences between the two lateralizations [300, 301, 302, 303, 304, 305]. Recently, Pereira et al. suggested that more patterns of functional connectivity change in LTLE compared to RTLE [306]. However, after patients suffer from intractable seizures for 10 or more years, the intrahemispheric hippocampal connectivity linearly increases with the duration of disease, suggesting that over time lateralized disease may become bilateral disease [302]. Because patients with bilateral hippocampal disease are no longer considered surgical candidates, improved methods to distinguish left and right TLE early in the course of disease are needed.

In this manuscript, we discuss the development of an automated computer-aided diagnostic (CAD) tool to diagnose, and lateralize, temporal lobe epilepsy using iPET. We also begin to address how to combine our CAD tool with manual analysis (MA) and incorporate it into clinical practice. Using a mutual information-based feature selection technique, we examine how our methods reveal more about the distributed metabolic abnormalities that are associated with the different anatomical locations of the epileptogenic focus.

The realistic goal of CAD tools is to complement, not to replace, expert analysis. Therefore, we focus on how clinical information and expert analysis can work synergistically with our automated technology. To summarize the major clinical differences, patients with NES are characteristically females in the third decade of life with psychiatric co-morbidities [265]. Patients with epilepsy, however, also have significant psychiatric co-morbidities including potentially reduced financial and social independence due to the suspension of their drivers and, frequently, professional license. Particularly in adult onset epilepsy, age-associated changes in metabolism may confound the interpretation of iPET, possibly leading to an increased diagnostic uncertainty. It is well established that 80 to 90 percent of medication refractory epilepsy is PET positive [22, 23]. The rate of PET positivity in NES has not been studied extensively, therefore the true positive predictive value of iPET is unclear. Although these differences in clinical presentation are salient, their quantitative effect on diagnostic probabilities is unknown. Therefore, we also examined how simple clinical information and expert manual interpretation can be incorporated into our quantitative CAD tool.

The standard of care for the pre-surgical assessment for epilepsy is the manual correlation of iPET with numerous other diagnostic modalities. The goal of this assessment is to simultaneously verify the diagnosis of epilepsy, characterize the seizure etiology and identify the location and extent of the SOZ. Expert radiologists and neurologists can detect metabolic asymmetries indicative of the epileptogenic focus or foci [222]. The exact threshold at which asymmetric metabolism is attributed to pathologic change or seen as a variant of normal is part of the art of neuroradiology [307, 42]. Once non-epileptic etiologies have been ruled out, our previous work demonstrated that the quantitative degree of metabolic asymmetry is correlated with surgical outcome [259]. Surgical outcome is improved further when iPET is co-registered to structural MRI (sMRI) because of improved characterization of the focus or foci [308, 22, 309, 23]. These hypometabolic lesions are thought to be secondary to increased inhibitory neuron cell death, gliosis and abnormal functional connectivity resulting in altered functional metabolism.

The size of the hypometabolic lesion tends to be larger than the SOZ, potentially due to functional changes in nearby tissue secondary to the presence of the epileptogenic lesion [240, 310, 311]. Such reports are major limitations to the wide implementation of iPET in epilepsy practices [312, 58, 313]. In addition to the limitation of counting statistics, that forces the quantitative radioactivity intensity of iPET to be less certain in hypometabolic lesions [233], the biological hypothesis is that the epileptogenic abnormality induces metabolic abnormality at the SOZ and also at closely associated and/or functionally connected regions [314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324]. The epileptogenic lesion commonly is larger and more diffuse in left TLE then right TLE, potentially because of the high degree of functional connectivity between specialized foci within the left temporal lobe associated with language and other functions [325, 79, 326, 327, 328]. These insights parallel the trend in dementia that atrophy starts focally then spreads more quickly to functionally connected regions [329]. The limited sensitivity of iPET unaligned with sMRI to characterize extratemporal lesions may be partly due to the insufficient description of the local functional network of each extratemporal focus and thereby reduced detection of a characteristic pattern of metabolic abnormalities associated with each focus. In general, an improved insight into the clinical interpretation and value of metabolic abnormalities outside the SOZ is needed. To overcome this limitation, the iPET analysis is used in combination with other diagnostic modalities determine which tissue to resect.

Clinical description, EEG, MRI and FDG-PET each describe separate facets of the pathophysiological etiology, and therefore all play critical roles in the diagnosis of epilepsy, and in the identification of the epileptogenic lesion [330]. Each modality, however, also has unique limitations. EEG provides an in-depth description of the seizures and interictal epileptiform spikes. These seizures and spikes, however, are rare events: only 50 percent of PWE exhibit diagnostic interictal epileptiform spikes and/or seizure activity during the first outpatient scalp EEG [57]. The characteristic signs of epilepsy in structural and diffusion MRI may not be measurable until years after the first seizure because these methods require the detection of atrophic tissue and/or subtle regions of cortical dysplasia [77, 331, 332, 333, 334, 335, 336, 337]. Manual analysis (MA) uses the contralateral structure to assess if atrophy is present but a certain degree of asymmetry is expected [26, 28]. It takes years of specific experience in manually analyzing structural MRIs from patients with epilepsy to reliably discriminate between normal variation and pathologic changes. Once these relatively large-scale changes in neural structure have occurred, it is less likely that both invasive and noninvasive treatments will be effective [32]. Interictal FDG-PET can localize the epileptogenic lesion without observing rare events and, potentially, before changes are detectible on sMRI and/or diffusion tensor imaging (DTI)

[338, 339, 340, 341, 308, 342, 343, 344, 239, 345, 346, 241, 347, 348, 22, 235, 349, 309, 350, 23, 336, 351, 352, 353]. As discussed above, the presence of metabolic abnormalities outside the SOZ, however, complicates the effective localization of the SOZ using iPET alone [240]. An improved description of these induced changes outside the SOZ may help spare healthy tissue from resective surgery. Given the recent report that resective neurosurgery for epilepsy is more effective earlier in disease [32]; we believe that iPET may play a critical role in characterizing patients with unremarkable MRIs and inconclusive EEGs earlier in the course of their disease.

9.3 Materials and Methods

9.3.1 Patient data

All of the 105 patients that were included in our analysis were admitted to the University of California, Los Angeles (UCLA) Seizure Disorder Centers video-EEG Epilepsy Monitoring Unit (EMU) between 2005 and 2012. Each patients diagnosis was based on a consensus panel review of their clinical history, physical and neurological exam, neuropsychiatric testing, video-EEG, interictal FDG-PET, ictal FDG-PET, structural and diffusion MRI and/or CT scan. This multimodal assessment is the gold standard for epilepsy diagnosis and localization of the epileptic focus [296, 38]. The patients included in this analysis were chosen because they had an FDG-PET after 2005; had no history of penetrative neurotrauma, including neurosurgery; were determined by consensus diagnosis to have a single, lateralized epileptogenic focus; and had no suspicion of mixed non-epileptic and epileptic seizure disorder. These patients were diagnosed either with left temporal lobe epilepsy (LTLE, n=39), right temporal lobe epilepsy (RTLE, n=34) or non-epileptic seizures (NES, n=32). PET images were determined to be interictal by clinical findings and concurrent scalp EEG.

PET and MRI images were acquired according to the best clinical practices at the time of acquisition. PET/CT studies were acquired using a Siemens Biograph scanner. After a minimum fasting period of 6 hours, patients received 0.14 mCi/kg of 18F-FDG-PET intravenously. During the ensuing 40 minute uptake period with concomitant EEG monitoring to confirm interictal status, the patients waited in a quiet, dimply lit room with their eyes open. PET images were reconstructed with an iterative algorithm (OSEM: 2 iterations, 8 subsets). CT images were reconstructed using filtered back projection at 3.4 mm axial intervals to match the slice separation of the PET data, and used for attenuation correction.

9.3.2 Computer aided diagnostic tool training and validation

Automated analysis of the interictal FDG-PET records was performed in four stages. (1) First, each image was screened for gross structural and/or metabolic abnormalities by S.T.N., N.M.R., and/or W.T.K. (n=21). These excluded subjects are not reflected in the sample sizes quoted above. (2) NeuroQ (Syntermed, GA) was used to segment each brain into 47 regions of interest (ROIs) and then to calculate the average radioactivity in each ROI, normalized by the whole brain radioactivity (Supplemental Table 1). (3) The minimum redundancy-maximum relevancy (mRMR) toolbox for MATLAB (Mathworks, MA) was used to generate a ranked list of the ROI metabolisms (features) within each training set that were maximally relevant to the diagnosis of epilepsy and minimally redundant with all higher ranked features [5, 6]. The representative number of features to exclude and quantal levels was selected based on our method discussed previously [8, 293] (see below). In each of the training sets, the feature ranking was determined exclusive of the test patients data. We expect the ranked lists to be similar, but not identical, across training sets. For purely illustrative purposes, the full dataset was used to create the ranked list in Table 9.2. (4) Weka was used to implement leave-one-out cross-validation of a cost-sensitive Multilayer Perceptron

(MLP) that was weighted to maximize balanced accuracy, defined by the mean of sensitivity and specificity [292]. Using this method, we examined our ability to diagnose either LTLE or RTLE from NES and assessed our ability to diagnose and lateralize disease simultaneously. For the remainder of this manuscript, the latter tool that discriminates LTLE versus RTLE versus NES is called the trinary classifier. Similarly, the binary CAD tools are referred to by the laterality of epilepsy that is being detected. The comparison to NES is not stated, but can be assumed. We then compared our CAD tools performance to the results of MA alone.

9.3.3 Machine learning algorithmic details

The Multilayer Perceptron (MLP) was implemented with default parameters in Weka [292]. All input features were normalized to values between negative and positive 1. No limit was set on the number of hidden layers or nodes within each hidden layer. These parameters were optimized within each training set independently. The learning rate and momentum were set to 0.3 and 0.2, respectively. Five hundred epochs were used for training. During training, models with more than 20 consecutive errors were excluded. The trinary classifier was created by decomposing the three class problem into three 1-against-1 problems that were combined using majority voting. No three-way ties occurred during training or testing.

Balanced accuracy was optimized using a cost-sensitive classifier in which a false positive was given a cost of $n+$ and a false negative was given the cost of $n-$, where $n+$ and $n-$ represent the number of patients with epilepsy and non-epileptic seizures in the full sample, respectively. In the trinary classifier, the cost was set as the sum of the number of patients in the other two diagnostic classes.

Cyclical leave one out cross validation (CL1OCV) was used to assess the performance of the MLP. In this paradigm, all but one patient was used to determine the features selected and train the algorithm. The single remaining patient is tested using the model built upon the other patients. The identity of the test patient is permuted until all patients have been the test case once and only once. To determine the number and identity of the input features, the mRMR algorithm requires the number of input features, F, and quantal levels, Q, be set a priori. For the calculation of mutual information, the features were smoothed into Q quantal bins akin to the bins in a histogram. Classification, however, utilizes unsmoothed features. The choice of input features smoothed into quantal levels was determined to be most representative of the performance of the algorithm across a wide variety of choices of F and Q [293]. This choice was made by selecting a point within a region of F-Q parameter space that performed significantly better than the naive classifier with 95% confidence based on random field theory correction where the spatial smoothness is estimated directly from the data (for more details, see [354, 355]). The naive classifier classifies all test exemplars as the most common class in the training set. Under the CL1OCV procedure, these input features were determined independently for each of the training samples. The illustrated rank order of features was calculated based on the full dataset, and does not necessarily match the rank list of any individual training sample.

When clinical information was incorporated into the algorithm, the same methodology was applied as above, except that all exemplars with missing data were excluded from analysis. In these additional analyses, we did not re-sample the parameter space of F and Q. We simply used the selections determined in the previous analysis.

9.3.4 Manual analysis of PET and MRI records

Manual analyses of the iPET and sMRI records were performed based on the review of clinical records primarily written by Dr. Noriko Salamon. Dr. Salamon has 10 years of experience in the pre-surgical assessment of epilepsy using FDG-PET and MRI. All manual interpretation was conducted for the clinical assessment of each patient when it occurred, prior to the CAD tool development. Therefore, Dr. Salamon was blinded to the automated results. Due to the unclear relationship between structural and metabolic abnormalities, asymmetries and epilepsy, all abnormal results were interpreted to be consistent with some form of epilepsy. Not all patients had sMRI (n=6) and iPET (n=1) reports available; therefore all analysis regarding MA of neuroimaging includes only patients with available records. These patients had raw iPET data available; they therefore were included in the automated analysis.

9.3.5 Combination of clinical information with computer aided diagnostic information

To examine the combined power of clinical knowledge, MA and our automated analysis, we assessed the linear correlation of detecting epilepsy with CAD compared to MA, and also incorporated clinical information and MA into our algorithm in two ways. First, the clinical literature suggests that patients with NES are more likely to be female, begin having seizures in the third decade of life, have a decreased duration of disease and have increased seizure frequency (Table 9.1). Although we did not see a significant difference seizure frequency within our dataset, we included this features to better match clinical practice. These clinical features were then added to the input and leave-one-out cross validation was repeated. Secondly, to explore how our computational methods can complement clinical wisdom, we included the results of MA of the iPET and sMRI as two additional input features and re-evaludated CAD performance. For the trinary classifier only, we split each of the features describing the iPET and sMRI MA to indicate if a left and/or right sided abnormality was reported.

To assess the applicability of our CAD as a separate modality that could

Table 9.1: This table reflects the clinical information known before the application of the CAD tool. All times are listed in years (y) unless otherwise specified by days (d), weeks (w), or months (m). Manual analysis of all patients' iPET and sMRI were not done, therefore we list the number with available manual results. *, §, or ¶ indicate that the value for NES vs. LTLE, NES vs. RTLE, or LTLE vs. RTLE, respectively, is statistically significant from both the LTLE and RTLE groups with at least 95% confidence using a two-sample z-test of proportions or MannWhitney U test, where appropriate. No other differences are statistically significant ($p > 0.10$).

		NES	LTLE	RTLE
Age	Mean ± SD	37 ±14*	38±12	36±13
	Min-Max (Median)	16-76 (38)	18-54 (40)	17-67 (35)
	N	32	39	34
Sex	% Female±SE	78.1±7.3*§	53.8±8.0	35.3±8.2
Duration of disease	Mean ± SD	12±12*§	22±15	20±13
	Min-Max (Median)	10d-40y (7)	6m-53y (21)	2y-48y(19)
Seizure frequency	Mean ± SD	3.2/d ± 5.9/d	1.2/d ± 2.4/d	1.5/w ± 1.7/w
	Min-Max (Median)	0.3/m-25/d (3/d)	0.2/m-11/d (1/w)	0.1/m-1/d (0.8/w)
iPET manual	% Positive±SE	18.8±6.9*§	76.9±6.7	87.9±5.7
	N	32	39	33
sMRI manual	% Positive±SE	34.5±8.8*§	73.7±7.1	87.5±5.8
	N	29	38	32

be considered as part of the clinical assessment of epilepsy, we calculated the likelihood ratios (LRs) of each of the combinations of our CAD with MA of iPET and/or sMRI. This was done only for the binary classifiers, because likelihood ratios have a clear formulation only for binary outcomes. The likelihood ratio is defined by the likelihood that a patient with a certain combination of diagnostic outcomes has epilepsy, divided by the likelihood that the same patient has NES. Intuitively, a likelihood ratio of two implies that the patient is twice as likely to have epilepsy. The 95% confidence intervals of chance were calculated using exact binomial intervals by considering the likelihood ratio of a classifier that diagnosed patients according to their prior likelihood alone, conditioned upon the assumption that the same total number of patients would have the diagnostic outcome of interest. For example, 39 of 71 patients had LTLE when we discriminated between

LTLE and NES, therefore the median LR is 1.2. Thirty-five patients from the NES versus LTLE group had negative MA of their iPET. Therefore, we use a binomial distribution with 35 trials and success probability of 39 over 71 to yield a 95% confidence interval of 0.94-3.38.

9.4 Results

All of our results are compared to the gold standard diagnosis from the consensus panel. The clinical trial statistics of each of our automated diagnostic tool matched, but were not redundant with, expert manual analysis of both interictal PET and sMRI (Figure 9.1). All intervals reflect 95% confidence intervals and all p-values correspond to differences from a naive classifier. The binary CAD tool for RTLE had accuracy of 88% (69-90%), compared to the accuracy of MA of iPET (85%, [72-92%]) and sMRI (77%, [63-85]). The binary tool for LTLE had accuracy of 83% (69-90%), compared to the accuracy of MA of iPET (79%, [66-88%]) and sMRI (70%, [56-81%]). The pattern in sensitivities, specificities and odds ratios all parallel this trend where our automated diagnostic tools are non-statistically superior to MA or iPET, which, in turn, are non-statistically superior to MA of sMRI (Figure 9.1). The accuracy of our trinary CAD tool that simultaneously diagnoses epilepsy and lateralize disease was 76% (66-84%), where 89% (77-96%) of patients correctly identified with epilepsy were also lateralized correctly. MA to diagnose and lateralize was 78% (69-86%) accurate with 89% (76-94%) correctly lateralized using iPET and 71% (61-80%) accurate with 91% (78-97%) correctly lateralized using sMRI.

The rank order of the features used in our algorithm parallel the clinical observation that the epileptogenic networks in LTLE are broader than in RTLE. The LTLE vs NES classifier achieved its performance by utilizing trends across almost the entire brain by including 42 of the 47 features in the final algorithm. In

Figure 9.1: CAD tool performance matches manual analysis. These figures indicate the accuracy, sensitivity and specificity of the LTLE (A), RTLE (B) and trinary (C) classifiers. The performance of our CAD tools matched that of MA and was superior to just using gender alone. The error bars indicate standard error of the mean performance for each measure. The translucent region indicates the performance of a naive classifier. *Indicates significant differences from the naive classifier with a confidence level of 95% or more.

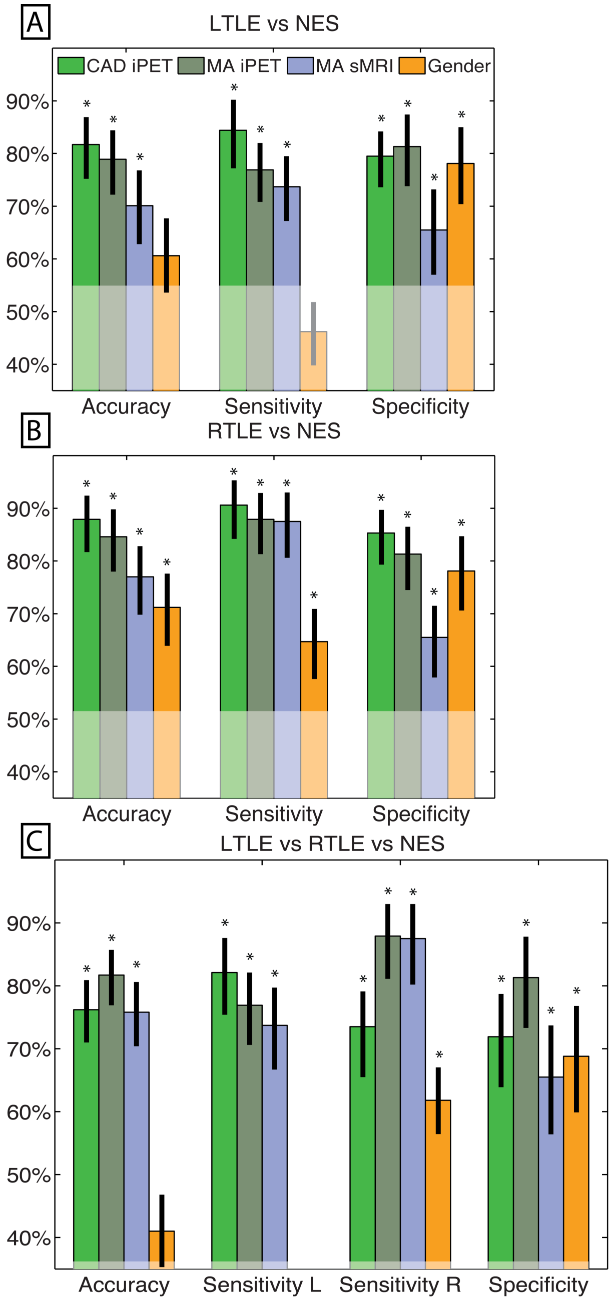


Table 9.2: This table illustrates the top six informative and non-redundant regions of interest (ROIs) that may contribute to each of the CAD tools, as determined by the minimum redundancy-maximum relevancy criteria (mRMR)[5, 6]. The illustrated rank order of features was calculated based on the full dataset and does not necessarily match the rank list of any individual training sample. The leading L or R indicates left or right. The lowercase letters indicate inferior (i), lateral (l), median (m), anterior (a), and posterior (p). The lagging C signifies cortex. Note that the LTLE vs. NES and trinary classifiers include information from 42 and 30 ROIs, respectively. To better understand the benefit of mRMR, this list can be directly compared to the list of ROIs ranked by t-statistics in Table A1 in Appendix.

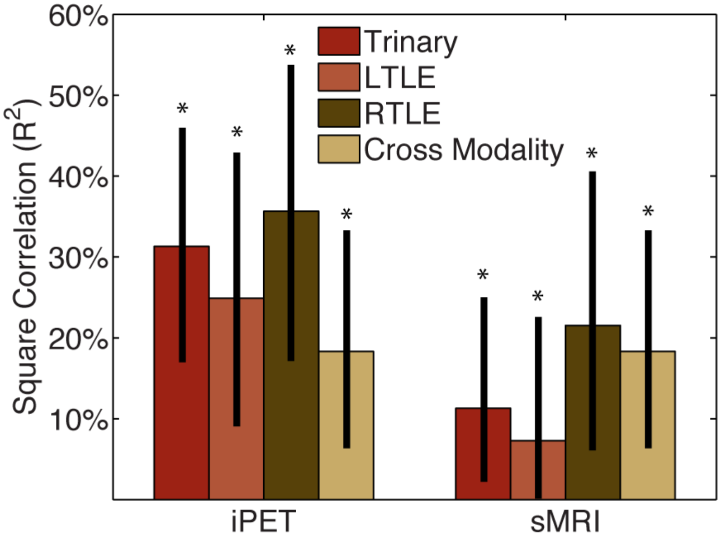
		Region of Interest	
mRMR rank	LTLE vs NES	RTLE vs NES	Trinary
1	Midbrain	R ila temporal C	R ila temporal C
2	L ilp temporal C	R ilp temporal C	L ilp temporal C
3	R ilp temporal C	L sensorimotor C	L sensorimotor C
4	L associative visual C	L sl temporal C	R ilp temporal C
5	L Broca's Region	R thalamus	R sl temporal C
6	L s frontal C	R i frontal C	R pm temporal C

contrast, the RTLE vs NES classifier only needed to measure the metabolism in 6 regions–bilateral temporal cortex and two associated regions of cortexto achieve its impressive performance (Table 9.2). As expected, the trinary classifier utilized an intermediate number of features to achieve its accuracy (30 of 47). The rank list of these features matches the biological intuition based on knowledge about the potential connectivity of epileptogenic networks (Table 9.2).

We then considered how this CAD information could be used in combination with clinical information or expert analysis. The squared correlation of our CAD tool with manually interpreted iPET was 0.25 (0.09-0.43), 0.32 (0.17-0.54) and

0.34 (0.17-0.46) for the LTLE, RTLE and trinary classifiers, respectively (Figure 9.2). The squared correlation of our tool with manually interpreted sMRI was 0.07 (0.001-0.23), 0.21 (0.06-0.40) and 0.11 (0.02-0.25) for the LTLE, RTLE and trinary classifiers respectively. For comparison, the squared correlation between manually interpreted iPET and sMRI was 0.17 (0.06-0.33).

Figure 9.2: CAD tool is not redundant with manual analysis. The squared correlation of our CAD tools' results with those of MA of the iPET or sMRI from the same patients was below 50%. This indicates that while some information is shared, the majority of information provided by our CAD tools is not captured by MA. The correlation between MA of iPET and sMRI is similar in magnitude to the correlation of CAD with MA, therefore the CAD could potentially be seen as similar to another informative modality. *Indicates significant differences of the correlation from zero with a confidence level of 95% or more.



When the same automated analysis was used to combine clinical findings with our iPET data, performance did not change significantly. After the four clinical factors were added to the input of our tools, the accuracy changed to 79% (66-88%), 68% (56-79%) and 64% (54-73%) for the LTLE, RTLE and trinary classifiers, respectively (Figure 9.3). These accuracies do not substantively change when only sex and duration of disease were considered (results not shown). Adding the results of MA of both iPET and sMRI to our iPET data changed the accuracy to 82% (73-91%), 77% (67-88%) and 68% (59-77%) for the LTLE, RTLE and trinary classifiers, respectively. When all information sources contribute to the algorithm, the accuracy changed to 77% (68-88%), 74% (64-85%) and 76% (68-84%) for the LTLE, RTLE and trinary classifiers, respectively.

We combined the results of MA were combined with our CAD tool manually using likelihood ratios. After doing so, the likelihood was generally only significant if all considered modalities agreed. Viewed alone, MA and our CAD increased the likelihood of the predicted outcome between two and nine-fold ($p < 0.02$; Figure 9.4A). When two analysis streams were combined, if both analyses agreed, the likelihood of the predicted outcome was increased between eight and 27 fold ($p < 3 \times 10^{-4}$; Figure 9.4B and 9.4C). If all three analyses agreed, the likelihood of the predicted outcome increased more than 15 fold ($p < 1.3 \times 10^{-5}$; Figure 9.4D). However, in most cases, if there was any disagreement, the likelihood did not change significantly, most probably due to the small numbers of patients with each potential outcome. There are two key exceptions: (1) Given iPET results indicating NES over RTLE using either MA or CAD, the sMRI could be largely ignored ($p < 1.1 \times 10^{-2}$). (2) If both MA and CAD of iPET agreed that a patient suffered from LTLE and not NES, the sMRI results could be similarly ignored ($p < 3.3 \times 10^{-2}$).

9.5 Discussion

These results demonstrate how our CAD tool has the potential for clinically application, while also confirming and elucidating the distributed effects of epilepsy on the entire brain. Our CAD tools diagnostic performance of TLE matches, but is

Figure 9.3: Automated combination of clinical information with automated analysis of iPET images. The automated combination of clinical information and/or MA with our analysis produced no significant change in performance for the LTLE (A), RTLE (B) or trinary (C) classifiers, relative to the CAD operating on automated values alone. The unshaded bars indicate the performance of similarly constructed CAD tools using clinical information or the results of MA alone. The shaded bars indicate the modified performance when information from NeuroQ is added. The horizontal line indicates the mean accuracy of each CAD tool without clinical information. The translucent region indicates the performance of a naive classifier.

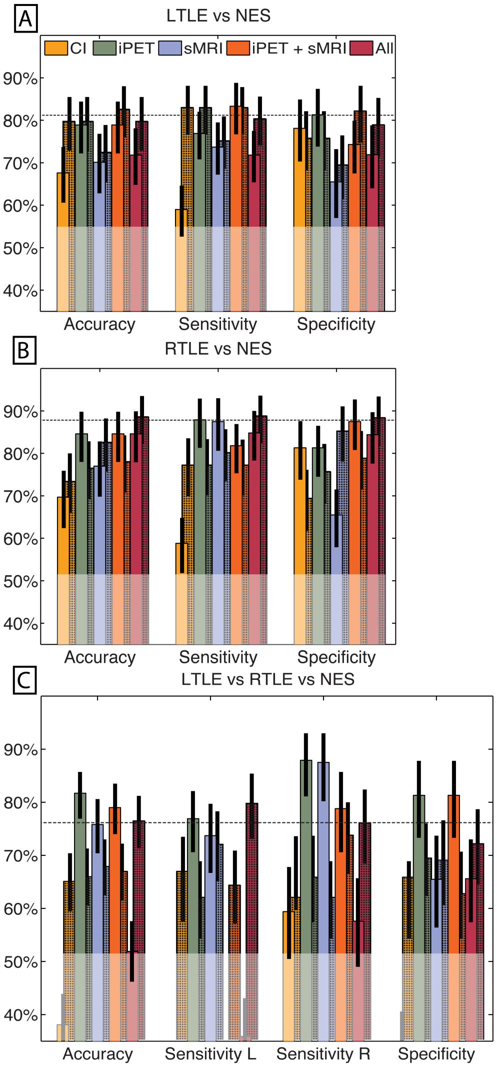
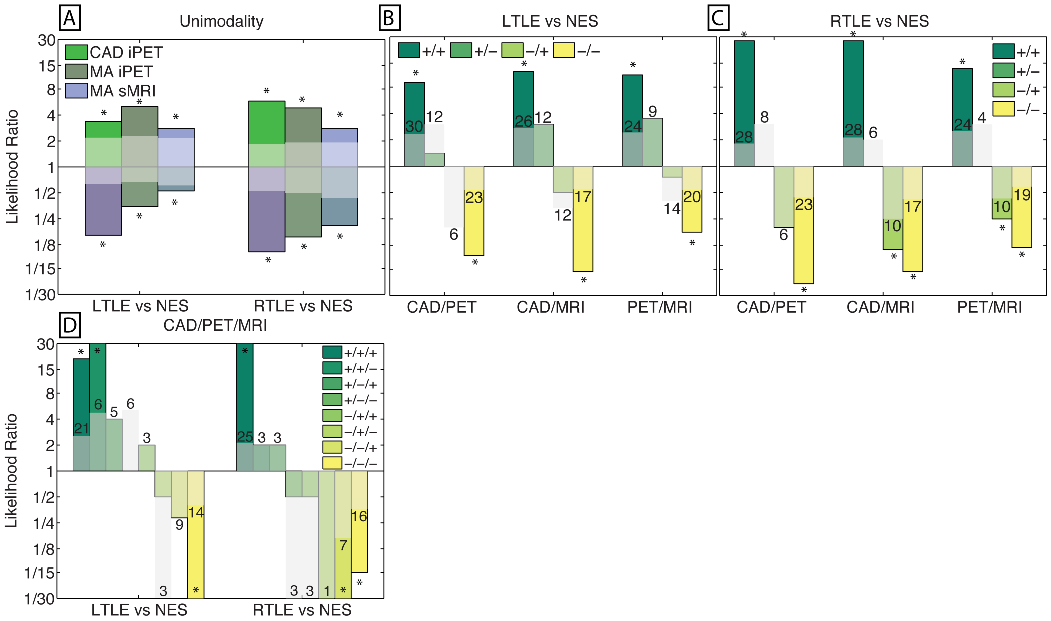


Figure 9.4: Combination of clinical information and CAD results using likelihoods. Columns in this log plot above 1 indicate that the seizures are more likely to be epileptic whereas the columns below 1 indicate a non-epileptic etiology is more probable. (A) Illustrates the positive and negative likelihood ratio of each analysis method considered individually. (B,C) Illustrate the likelihood ratios of each possible outcome when two analysis methods are combined. (D) Indicates the likelihood ratios of each possible outcome when all analysis methods are combined. If all modalities agree, the likelihood non-significantly increases with the addition of each modality. However, if there is disagreement, the likelihood ratio is generally not significantly different from chance. The translucent bars indicate the 95% confidence interval for chance with the relevant sign (see Materials and Methods).The numbers above the translucent bars indicate the total number of patients with each outcome. The bars that go off the scale of the graph diverge toward zero or infinity because no patients of a certain class had that outcome. *Indicates significant differences of the correlation from zero with a confidence level of 95% or more.



not redundant with, expert manual analysis of iPET and sMRI. When considered in the context of recent reports of CAD tools for epilepsy based on structural MRI and interictal EEG data [26, 27, 28, 8], CAD is proving especially applicable to epilepsy. Further, if more work confirms the hypothesis that metabolic changes in iPET are observable before the structural changes in sMRI, our iPET tool may have better clinical utility than these existing sMRI tools. In contrast to MA, this and other CAD tools can be quickly and efficiently applied by minimally trained technicians, emergency physicians and primary care providers as preliminary analysis of the iPET images [89, 62]. The performance of MA can vary with experience and fatigue of the observer; automated tools are consistent over time. Upon further validation, these CAD results could also be incorporated into the consensus diagnoses with minimal cost if iPET already has been obtained.

9.5.1 Clinical Impact

Our CAD tools could provide valuable clinical information that may help readily identify which treatments may be effective in patients who present with uncharacterized, and/or medication refractory seizures [8, 89]. In particular, 15 of our 105 patients were admitted twice to achieve definitive characterization or localization of their seizures. The appropriate binary classifier correctly diagnosed 12 (80%) of these challenging patients. This valuable information might reduce the need for multiple video-EEG admissions. Additionally, 28% (9/32) of our PWN were admitted for improved characterization of their previously-diagnosed "epilepsy," and 16% (12/73) of our PWE were admitted for the differential diagnosis of epilepsy, indicating that non-epileptic etiologies were not ruled out sufficiently. The trinary CAD effectively diagnosed 67% (14) of these particularly challenging patients. Despite this impressive performance, the ultimate goal of CAD, however, is to complementnot replacemanual analysis.

9.5.2 Combination of automated analysis with clinical wisdom

Our finding that performance almost uniformly, but non-statistically, decreased when the automated algorithm incorporated clinical information indicates that automated analysis cannot and should not replace manual interpretation across information modalities. We suspect that this performance decreased due to ineffective modeling of the contribution of the clinical information and over-fitting. The statistical distribution of the clinical factors was very different from the metabolic data therefore the same model likely cannot effectively utilize both modalities. The efficient incorporation of multimodality information into machine learning is an active area of theoretical research, and well-validated methods are not yet available. Now that CAD tools using interictal EEG [8], structural MRI [26, 27, 28] and iPET have been published, we believe it will be extremely exciting to assess how these various tools can be combined.

We expected that the best performance would be achieved when our CAD is used synergistically with MA. The low correlations between the CAD results and MA suggest that our CAD tool provides information that is not evident on visual inspection. These results emphasize that PET is not redundant with MRI [356]. Physicians could learn to view CAD as analogous to another imaging modality that provides valuable, but not perfectly diagnostic, clinical insight. This synergistic application of computer aided diagnosis after manual interpretation already has proven beneficial in the detection of lung nodules by the FDA and is an active area of translational research [89, 156]. The key differences between MA and automated analysis are the ability to entirely ignore certain pieces of data, and to rule that the results are inconclusive.

The results summarized above, and the likelihood ratios for each analysis stream individually, show that both MA and CAD are useful clinically. If the analysis streams agree, the diagnostic certainty increases substantially, but at a cost: as more analyses are added, more patients have inconclusive results because the analyses did not agree, and the likelihood ratios are not significant. Even though our sample size is large compared to other studies of this type, there were not enough patients in our dataset with each diagnostic outcome to explain the clinical implication of disagreeing analyses adequately. This matter of inconclusive results is a common challenge faced in clinical practice. Physicians struggle regularly with those types of decisions. When MA of iPET and sMRI are combined, they need to agree to yield meaningful results. However, our analysis shows that in some specific cases, if both the MA and CAD of iPET agree, the sMRI is not needed. This parallels the finding we suggested above: iPET may be more clinically useful than sMRI to diagnose and lateralized epilepsy. The hypometabolic abnormality may be present earlier in disease [338, 339, 340, 341, 308, 342, 343, 344, 239, 345, 346, 241, 347, 348, 22, 235, 349, 309, 350, 23, 336, 351, 352, 353], and it may provide slightly more accurate disease characterization, as seen in our dataset. In settings where the PET scanner is not combined with the MRI scanner, and/or when the cost of imaging is a limiting factor (both common occurrences) the effective application of our CAD could result in substantial cost savings.

9.5.3 Pathophysiological Insights

Our methods also reveal a potential difference in the pathophysiology of left versus right TLE. This may help explain why CAD tools perform slightly better when diagnosing RTLE compared to LTLE [26, 27, 28]. The finding that mostly bilateral temporal ROIs, the right inferior frontal cortex and left sensorimotor cortex provide non-redundant diagnostic information for RTLE is consistent with the clinical wisdom that the epileptogenic network in RTLE is more focal than in LTLE. The inclusion of temporal regions echoes the conventional wisdom that focal hypometabolism and asymmetry reflect characteristic changes due to epilepsy.

This suggests that conservative resection of the temporal lobe may result in increased rates of seizure freedom in RTLE compared to LTLE due to complete resection of the SOZ. Further, seizures that originate in the left temporal lobe may secondarily generalize more frequently in LTLE. These differences have not yet been studied clinically.

The trends in the extratemporal regions included in the algorithms suggest that the primary lesion may induce metabolic changes in functionally or anatomically associated regions. This is substantiated further by the finding that almost all regions of the brain provide informative diagnostic information in LTLE. This in turn mirrors the increased stereotypic connectivity of the left temporal lobe. Even though the interconnectivity of the right hemisphere is higher than the left hemisphere, the left hemisphere has strong connections between specialized foci [325, 326, 327]. We hypothesize that the SOZ may induce abnormal metabolism along these strong, stereotyped connections. This change cannot be attributed to language specifically in our dataset because we did not identify the laterality of language dominance in our patients. Compared to our t-statistics ranking, it may seem surprising that the metabolism of the midbrain was ranked first by mRMR for LTLE versus NES. This rank may indicate a nonlinear change in the metabolism within the dorsal midbrain anticonvulsant zone, which has itself been identified in animals to be part of the network that modulates seizure threshold [357]. The exact relationship between epilepsy and midbrain metabolism is unclear, however. The lack of distributed atrophy in LTLE measured by sMRI suggests that these changes are not associated with distributed cell death or gliosis [26, 27, 28]. Instead, we hypothesize that this change instead reflects abnormal metabolism in these regions due to altered neural connectivity and/or activity secondary to the epileptogenic lesion. This is supported by the finding that LTLE was associated with more changes in functional connectivity than RTLE was [306]. This also explains why we observed metabolic changes in the right thalamus in

RTLE: recent work demonstrates that the connectivity of the right thalamus with the right hippocampus is reduced in RTLE [303]. The presence of such distributed changes also supports the finding that the size of the hypometabolic lesion visualized on PET may be larger than the SOZ [240, 310, 311]. It is particularly interesting to note that the extent of these distributed changes is underappreciated by t-statistics comparing LTLE to NES. This indicates that there is a complex, likely nonlinear, relationship between the metabolism of the hypometabolic lesion and its associated tissue that may be better understood by mutual information.

The inclusion of the contralateral hippocampus in both of the binary classifiers lends itself to multiple interpretations that are all supported by biologically sound hypotheses. Firstly, a salient feature of LTLE or RLTE could be asymmetric metabolism, as suggested clinically; therefore the metabolism of the contralateral hippocampus was compared to the observed metabolism in the ipsilateral hippocampus. Alternatively, the interhemispheric connectivity between the hippocampi is high, therefore under our hypothesis that changes in metabolism spread according to functional connections, the metabolism in the contralateral hippocampus may be one of the first induced changes due to the epileptic lesion. Lastly, if LTLE and RTLE are inherently bilateral diseases then the metabolism in the contralateral hippocampus may also be abnormal. This also provides an explanation for why LTLE and RTLE were not perfectly distinguished.

In addition to diagnosing epilepsy, our algorithm lateralized disease efficiently with an accuracy of approximately 90 percent when epilepsy was diagnosed correctly. This impressive accuracy could be clinically useful for pre-surgical planning, when used in combination with other clinical and radiological information. Although our current sample size is too small to fully assess this potential fully, our results suggest that similar methodology could be applied to a larger dataset with more diverse and specific SOZ localizations to yield an objective and reliable tool to assist in pre-surgical SOZ localization. Our data suggest that this approach likely would identify and utilize distributed metabolic findings associated with each epileptic lesion to improve performance. Instead of blurring the boundary of the SOZ by detecting affected tissue outside the SOZ, the improved understanding of these distributed effects may lead to more refined characterization of this clinically vital SOZ. However, the spatial resolutions of our outcome classes were insufficient to assess the utility of this method directly to identify candidate lesions for resective surgery.

While our lateralization accuracy is exciting, there is also a potential clinical interpretation of the patients who were falsely lateralized. Functional connectivity between the temporal lobes is particularly strong. In a minority of patients, this connectivity allows epileptogenic activity to spread quickly from the seizure onset zone to the contralateral temporal lobe on EEG, resulting in the appearance of either bilateral or falsely-lateralized disease. Similarly to the distributed networks discussed above, this high degree of functional connectivity also may induce metabolic abnormalities in the contralateral temporal lobe that may be indistinguishable from the primary lesion. This hypothesis can be tested by comparing these falsely-lateralized patients to patients with bilateral temporal lobe epilepsy. This comparison requires a detailed methodological treatment of non-mutually exclusive classes in machine learning and therefore lies outside the scope of the current manuscript.

To characterize these and other pathophysiological insights, most studies utilize healthy neurologically normal controls. In contrast, we prefer the use of PWN as our control group. In brief, when constructing a control group, one aims to match the patients in the pathologic group in all aspects other than the pathology. In contrast to neurologically normal controls when compared to PWE, PWNs have been exposed similarly to AEDs and other medications, have increased prevalence of TBI and some other risk factors for epilepsy [265], have regular and frequent meetings with health care providers, and have much more strict inclusion criteria.

Lastly, and perhaps most importantly, physicians do not consider whether all of their patients have epilepsy; they assess only the patients with seizures. Therefore, in our opinion, the use of PWN as the control group is a benefit in of our study because it maximizes the clinical relevance of our results while simultaneously improving its statistical selectivity.

9.5.4 Limitations and future directions

Because our retrospective dataset was collected as part of clinical care, our approach has a few important limitations. The accuracy of MA reported in our patients is worse than the rates quoted in previous literature [22, 309, 23]. Given UCLAs status as a tertiary referral center, the decrease in manual accuracy likely indicates that our patients had more heterogeneous etiologies and/or were more complex and difficult to diagnose than other centers. This suggests that our CAD tool may perform better on other datasets. Our iPETs and MRIs were collected on varying cameras with varying resolutions. This demonstrates the flexibility of our automated analysis using NeuroQ. The efficacy of the MA of older and limited resolution data may not be comparable to that of more current and higher resolution data. After establishing the efficacy of our method, we plan to both validate our tool prospectively on data from other centers, and to incorporate multi-center data into our algorithm to further improve its performance. Additionally, we only discuss the combination of CAD results with independently derived MA. Future work will examine the efficacy of CAD tools informed by MA and vice versa.

Critics of our approach might claim that the significant gender and age difference of the patients with NES compared to patients with epilepsy may lead to our CAD simply detecting the age and/or gender of the patients. While we do not expect this to be the case for RTLE, the utilization of language areas by the LTLE classifier might reflect differences in gender, and not epileptogenic pathology. However, the performance of our CAD was significantly higher than when clinical information was used directly, therefore the algorithm utilized more information than just clinical data to achieve its strong performance. These significant differences in clinical factors largely mirror the observed differences in clinic; therefore our dataset better matches the population for which our CAD tool would be applied. The only noteable exception is the significant age difference between LTLE and RTLE, which was unexpected. Due to the naturalistic nature of our data collection scheme, we did not correct for this difference. However, we note similarly to the NES group, the use of age alone was significantly worse than our tool and the addition of age to the iPET data to control for its effect did not significantly change performance.

Another key caveat to the direct clinical application of our tool to clinical practice is the fact that epilepsy is an extremely heterogeneous disease. The generalization of our method to bilateral temporal lobe epilepsy (TLE), extratemporal foci and multifocal epilepsy will be critical before it can be incorporated into clinical practice. In particular, even though non-epileptic seizures mimic all types of seizures, it is uncommon for TLE to be mistaken for NES. Instead, it is more common that non-epileptic seizures appear to have a focus in frontal cortex [358]. Therefore, the literature suggests that the highest impact CAD tool would discriminate between frontal lobe epilepsy and NES and another, separate tool could be used to lateralize TLE. Based on our results above (see section 4.1), we believe that our TLE-specific tool may be clinically applicable. For the first publication demonstrating the applicability of computer aided diagnosis based on iPET data, we chose to focus on the diagnosis and lateralization of TLE, based, based on prior findings that the sensitivity of iPET is highest for TLE. Our future work then can address generalizing our methods to the other epilepsies, including bilateral TLE and frontal lobe epilepsy.

9.6 Conclusion

Despite a few caveats, and upon further validation with data from other centers, our automated methods could provide unique information for the effective and efficient characterization of epilepsy, with the potential to decrease the fraction of patients with non-epileptic seizures that are being treated (inappropriately) with AEDs, and to more quickly triage patients with medication refractory epilepsy towards surgical intervention. This may help achieve the ultimate goal: a global reduction in seizures [32].

CHAPTER 10

Multimodal diagnosis of epilepsy using conditional dependence and multiple imputation

This chapter is a reproduction of our work that appeared in the proceedings of the International Workshop in Pattern Recognition in Neuroimaging. [?] This work was a collaboration with Eric S. Hwang, Kaavya R. Raman, Sarah E. Barritt, Akash B. Patel, Justine M. Le, Jessica M. Hori, Emily C. Davis, Chelsea T. Braesch, Emily A. Janio, Edward P. Lau, Andrew Y. Cho, Ariana Anderson, Daniel H.S. Silverman, Noriko Salamon, Jerome Engel, Jr., John M. Stern, and Mark S. Cohen. ESH, KRR, SEB, ABP, JML, JMH, ECD, CTB, and EAJ read all of the clinical & radiological notes and annotated the meaningful data in them. AYC assisted with parallelizing code and performing permutation tests. AA helped design the machine learning portion and interpret the multimodal results. DHSS, NS, JE and JMS helped fit our work in context of the diagnostic assessment of seizure disorder. MSC helped with all stages of these processing and manuscript.

10.1 Abstract

The definitive diagnosis of the type of epilepsy, if it exists, in medication-resistant seizure disorder is based on the efficient combination of clinical information, long-term video-electroencephalography (EEG) and neuroimaging. Diagnoses are reached by a consensus panel that combines these diverse modalities using clinical wisdom and experience. Here we compare two methods of multimodal computer-aided diagnosis, vector concatenation (VC) and conditional dependence (CD), using clinical archive data from 645 patients with medication-resistant seizure disorder, confirmed by video-EEG. CD models the clinical decision process, whereas VC allows for statistical modeling of cross-modality interactions. Due to the nature of clinical data, not all information was available in all patients. To overcome this, we multiply-imputed the missing data. Using a C4.5 decision tree, single modality classifiers achieved 53.1%, 51.5% and 51.1% average accuracy for MRI, clinical information and FDG-PET, respectively, for the discrimination between non-epileptic seizures, temporal lobe epilepsy, other focal epilepsies and generalized-onset epilepsy (vs. chance, $p<0.01$). Using VC, the average accuracy was significantly lower (39.2%). In contrast, the CD classifier that classified with MRI then clinical information achieved an average accuracy of 58.7% (vs. VC, $p<0.01$). The decrease in accuracy of VC compared to the MRI classifier illustrates how the addition of more informative features does not improve performance monotonically. The superiority of conditional dependence over vector concatenation suggests that the structure imposed by conditional dependence improved our ability to model the underlying diagnostic trends in the multimodality data.

10.2 Introduction

The diagnosis of seizure disorder is challenging, and relies on the effective integration of multiple streams of information, or modalities. Clinicians must combine clinical information, obtained through the clinical interview, with various technological modalities including, but not limited to, scalp electroencephalography (EEG), structural and diffusion magnetic resonance imaging (MRI), and fluorodeoxyglucose positron emission tomography (PET). Each modality provides in-complete but complementary information upon which a diagnosis can be built, and each modality has its own limitations. Clinical information depends typically upon accurate reporting from patients and/or caregivers who are untrained observers, and some work has shown that their reports are no more accurate than random guessing [63]. Neuroimaging relies on the development of observable structural and/or metabolic abnormalities that are associated, but not necessarily by cause or effect, with epileptogenic regions. Based on analysis of these factors, clinicians are able to provide effective treatment for two-thirds of patients with seizure disorder.

When a patient has failed two or more antiepileptic drugs (AEDs), or the etiology of the seizures is unclear, they are admitted for long-term video-EEG monitoring. During these admissions, 20 to 30% of patients with medication-resistant seizure disorder are found to have non-epileptic seizures [8]. For those patients with epilepsy, the goal of long-term monitoring is to determine if the seizures have focal or generalized onset and, if the seizures have focal onset, determine where the focus is and if it is surgically resectable [359]. Each of these determinations leads to changes in the treatment plan to target the cause of the seizures more effectively.

Our objective in designing computer-aided diagnostic tools (CADTs) is to improve diagnostic accuracy and certainty by providing information complementary to clinicians' judgment. This has the potential to decrease the cost of and time to diagnosis by providing clinicians' information that they would not otherwise have access to. Due to the inherently multimodal nature of the diagnosis of epilepsy, we focus on how to develop effective multimodal CADTs using the information available to clinicians.

In this manuscript, we assess the efficacy of two methods of multimodal learning: *vector concatenation* (VC) and *conditional dependence* (CD), with simplified

Table 10.1: Summary of the most prevalent features in each diagnostic group, prior to multiple imputation. Abbreviations: Temporal Lobe Epilepsy (TLE), Other Focal Epilepsy (OFE), Generalized-onset epilepsy (Gen), Unspecified-onset Epileptic Seizures (UES), Non-Epileptic Seizures (NES), Magnetic Resonance Imaging (MRI), fluoro-deoxyglucose Positron Emission Tomography (PET).

mean (standard error of the mean)	TLE	OFE	Gen	UES	NES
Female (%)	51 (3)	60 (5)	53 (7)	49 (6)	71 (3)
Age (years)	38.1 (0.8)	33.5 (1.4)	32.3 (2.0)	34.5 (1.7)	38.4 (1.2)
Duration Seizure Disorder (log_{10}year)	1.074 (0.033)	1.065 (0.044)	1.002 (0.072)	0.959 (0.082)	0.464 (0.066)
Seizure Frequency (log_{10}Seizures/month)	0.787 (0.049)	0.988 (0.083)	0.789 (0.132)	0.807 (0.100)	1.148 (0.069)
History of Stroke (%)	3 (1)	5 (2)	6 (3)	8 (3)	9 (2)
History of Febrile Seizures (%)	16 (3)	18 (4)	12 (6)	13 (4)	9 (3)
History of Neurotrauma (%)	35 (3)	31 (4)	24 (6)	25 (5)	36 (4)
History of Neuroinfection (%)	16 (3)	8 (3)	3 (3)	8 (4)	16 (3)
Abnormal MRI (%)	68 (3)	56 (5)	41 (7)	49 (6)	24 (3)
Abnormal PET (%)	71 (3)	48 (5)	36 (8)	43 (6)	26 (6)
Mesial Temporal Sclerosis (%)	68 (3)	34 (5)	27 (6)	35 (5)	10 (2)
Other MRI Findings (%)	45 (3)	48 (5)	37 (7)	41 (6)	22 (3)
Temporal Hypometabolism (%)	60 (3)	29 (3)	33 (8)	37 (6)	18 (5)
Other PET Findings (%)	27 (3)	27 (5)	13 (5)	13 (4)	11 (4)

data from clinical information (CI), MRI and PET. Vector concatenation represents a purely information theory perspective that relies on algorithms to discover the relationships between modalities. For other applications, VC has resulted in decreased performance relative to single modality models, likely due to overfitting and the "curse of dimensionality." CD attempts to overcome these limitations by considering each modality sequentially [360]. CD also models clinical practice, where clinicians make a preliminary diagnosis based on the clinical interview, then look to technological data to modify that initial impression.

10.3 Methods

All 645 selected patients with medication-intractable seizures were admitted to the University of California, Los Angeles adult (age 13-88) video-EEG epilepsy monitoring unit (UCLA EMU) between the years of 2006 and 2013. Patients were split according to their definitively diagnosed etiology: temporal lobe epilepsy (TLE, n=235), other focal-onset epilepsy (OFE, n=109), generalized-onset epilepsy (Gen, n=50), unspecified epilepsy (UES, n=81) and non-epileptic seizures (NES, n=170). Patients diagnosed with unspecified epilepsy had confirmed epilepsy, but the seizure onset zone was not determined. Definitive diagnosis was based on consensus panel review of long term scalp video-EEG, MRI, FDG-PET, clinical history, physical and neurologic exam, and/or neuropsychiatric testing. Not all patients underwent all studies. Patients with prior neurosurgery, those with inconclusive video-EEG results, and events suspicious for mixed NES and epilepsy seizure disorder were excluded from analysis (n=219). This work was approved by the UCLA Institutional Review Board and was consistent with the Helsinki declaration. Written informed consent was obtained from all patients (or guardians of patients).

Our analysis focused on three modalities: CI, MRI and PET. All data were acquired as part of the patients' clinical care according to the resources available at the time of care. Simple clinical information was extracted, including age, gender, duration of seizure disorder prior to neuroimaging, seizure frequency and a history of clinically suspected stroke, febrile seizures, focal or generalized neurotrauma, and neuroinfection. For patients with multiple neuroimages, only the most recent, pre-operative scan of each modality was included. Neuroimaging results were based on review of clinical records written primarily, but not exclusively, by Dr. Noriko Salamon, who is an expert in the interpretation of neuroimaging for the diagnosis and pre-surgical assessment of epilepsy. The MRI findings were simplified into binary indicator variables for extratemporal FLAIR or T2 hyperintensities, evidence of mesial temporal sclerosis, mass/space occupying lesion, encephalomalacia, cavernoma/hemangioma/angioma, cortical dysplasia, ischemic changes, gliosis, grey or white matter heterotopia, diffuse atrophy, focal extratemporal atrophy, meningioma, encephalocele, non-specific tumor, edema, vascular abnormality, cortical thickening, tuberosclerosis, unspecified lesion, cerebellar tonsil ectopia, abnormal gyration/sulcus structure, neurocystocercosis, hydrocephalus, and other MRI finding. The PET findings were simplified into indicators for hypo- or hyper-metabolism in the temporal lobe, frontal lobe, occipital lobe, parietal lobe, insula, diffuse cerebral cortex, cerebellum or whole brain diffuse hypometabolism, as well as foci of abnormal metabolism (i.e. high metabolism in white matter). Both neuroimaging modalities also included an aggregate indicator of abnormal findings.

Our data were extracted entirely from real-world clinical archives; not all data values were available for all patients. For the purposes of data imputation, we split the missing data into two groups. Duration of seizure disorder (0.5% missing) and seizure frequency (7% missing) were considered to be missing completely at random (MCAR), because these variables clearly are defined for every patient, and there was no trend in percent missing in any diagnostic subgroup. In contrast, if the clinical notes did not mention a historical factor (i.e., neurotrauma), we assumed that the patient had no history of this factor because the clinician is biased to report a historical factor if it exists. Overall 624 (97%) and 486 (75%) patients had MRI and PET records, respectively. The presence or absence of neuroimaging was not a significant predictor of diagnosis, when other clinical factors were taken into account (data not shown). Therefore, we assumed that this data was MCAR. We multiply imputed the data 20 independent times using the mi package in R [361]. Based on their theoretical and observed distribution, duration and seizure frequency were log transformed to maintain linearity. For the neuroimaging, there was insufficient information to impute each individual abnormality, therefore only the aggregate abnormality indicator for each modality was imputed. Separate analysis was conducted on each imputed dataset and results were aggregated with respect to the within and between imputation variance [362].

All classifications were based on C4.5 decision trees in Hall2009 [228] with leave-one-out cross-validation (LOOCV), and performance was compared to chance distributions determined by permutation tests. Briefly, at each node, the C4.5 finds the feature and threshold that maximizes the normalized information gain. In LOOCV, one patient is excluded from all training. Once the decision tree is built, its performance is assessed on this “unseen” patient. For each method, we evaluated the overall accuracy, sensitivity for each diagnostic class (TLE, OFE, Gen, UES, NES). UES patients were considered correctly classified if they were predicted to have any type of epilepsy, but not NES. All other patients were considered misclassified if they were predicted to have UES. This penalty was reflected in the cost matrix of the C4.5 classifier. To compare multiple classifiers head-to-head we calculated the paired performance change, where the difference in accuracy is paired within patient, then averaged across patient because the performance on each patient cannot be assumed to be independent across classifiers. The null distribution for all performance measures was calculated by conducting the same analysis (imputation, training, LOOCV and aggregating results across imputed datasets as in [362]) on data with permuted diagnostic labels, without replacement. At least 100 permutations were done on each imputed dataset. The rank order of performance measures from the permutations were used as as empirical markers for the 1% quantile bins of each chance, or null, distribution used to determine significance, because the permuted labels had no relation to the underlying diagnostic class.

We compared VC and CD. VC ignores the modality structure and treats all features as components of one large model. CD, otherwise known as “stacking” [360], classifies each patient into discrete, multivariate classes based on each modality individually in a specified order. Intuitively, for each test case the classifier gives a preliminary diagnosis based on the first modality. Then, a second layer classifier is learned from all training samples that also were classified as that same preliminary diagnosis, either correctly or incorrectly. To frame this theoretically, Bayes theorem states that:

$$P(Dx|Data_{M1,M2}) \propto P(Data_{M1,M2}|Dx)P(Dx) \tag{10.1}$$

where Dx and $Data$ indicate the diagnosis and data, respectively. In CD, we factor $P(Dx|Data_{M1,M2})$ by each modality to get:

$$\begin{aligned} P(Dx|Data_{M1,M2}) \propto & P(Data_{M2}|Dx, Data_{M1}) \\ & \cdot P(Data_{M1}|Dx)P(Dx) \end{aligned} \tag{10.2}$$

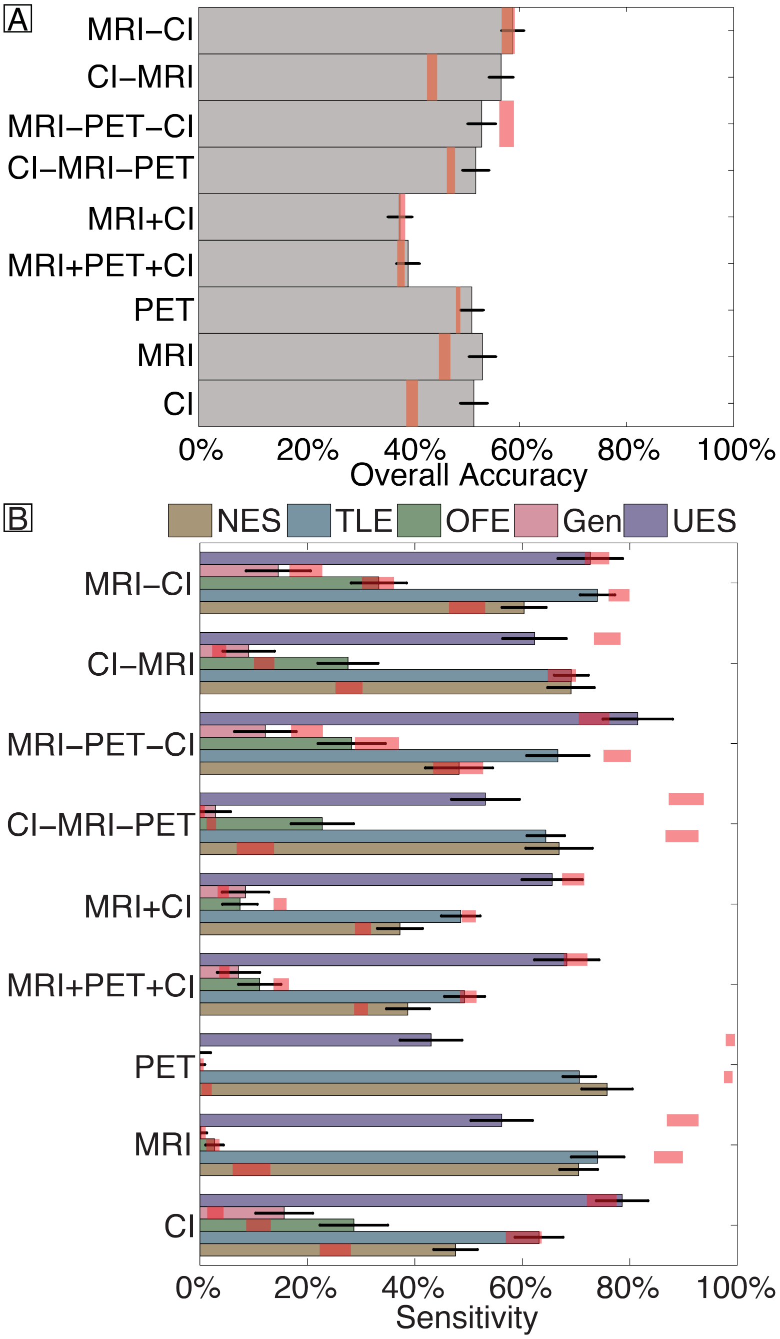
where $M1$ and $M2$ indicate two modalities, in order. Therefore, $P(Data_{M2}|Dx)$ is conditionally dependent on $Data_{M2}$. Although we have described two-modality CD, this reasoning can be extended to apply to m modalities for any positive integer m. The final predicted diagnosis is the diagnosis that maximizes this likelihood, given the data and the classification model used to estimate the probabilities.

10.4 Results

The LOOCV accuracy and per-class sensitivity, taking into account the multiple imputations [362], of the single and multimodality classifiers is illustrated in Figure 10.1. The accuracy of the single modality classifiers was 53.1%, 51.5%, and 51.1% fo MRI, CI, and PET, respectively. The accuracy of VC was 39.2% and 37.7% using MRI+PET+CI and just MRI+CI, respectively. The accuracy of CD was 58.7%, 56.6%, 52.9%, and 51.8% when modalities were considered in the order MRI→CI, CI→MRI, MRI→PET→CI, and CI→MRI→PET, respectively. All accuracies were significantly better than chance ($p<0.01$) except the MRI+CI, MRI→PET→CI, and MRI→CI ($p>0.1$). All pairwise comparisons revealed that all classifiers were superior to vector concatenation ($p<0.01$), but no other pairwise comparisons were significant ($p>0.08$).

Table 10.1 illustrates the distribution of the considered diagnostic features,

Figure 10.1: Overall accuracy (A) and per-class sensitivity (B) of each classifier. Error bars reflect binomial theoretical standard error bars, with multiple imputation. Red shading reflects the 95% quantile bounds from permutation tests. Vector concatenation and conditional dependence are indicated by + and -, respectively. For conditional dependence, the order of modalities is read from left to right. Abbreviations: Clinical information (CI).



except for the long list of neuroimaging indicators, by diagnostic class. All trees were more than 10 nodes deep and were too large for display.

10.5 Discussion

In real-world applications, combining information from multiple modalities does not always improve accuracy; this combination must consider the statistical and practical limitations inherent in modeling high dimensional data. Conditional dependence (CD) was superior to vector concatenation (VC) in overcoming these limitations, but did not result in a significant improvement over the single best modality classifier: the MRI.

The efficacy of CD relies on efficiently splitting the patients into more homogenous subgroups. The curse of dimensionality states that as the number of dimensions increases the number of samples needed to achieve the same sampling density increases exponentially. This curse can be overcome if the data truly exist in a lower dimensional subspace. This can occur when there are subgroups of patients within each diagnostic class that are more similar to each other, and therefore are distributed over a relatively limited region of feature space. These subgroups can be discovered using hypothesis-driven methods like CD, or through data-driven "committee-of-experts" methods that we will examine in the future. We hypothesize that, when applied in the most efficient order (neuroimaging first), CD identifies subgroups of patients with similar etiology. The relatively simple clinical variables then can identify if the clinical presentation of this etiology matches with the expected presentation of patients with similar etiologies. In particular, this order is interesting because it is the opposite of how clinicians diagnose patients. This illustrates how the ideal structure of automated computer analysis may differ from how clinicians' diagnose, due to the relative strengths of each analysis method. This reflects our belief that CADTs cannot, and should not, replace clinicians' expertise.

Even though neuroimaging-first produced higher accuracies than CI-first, this was not significantly higher than the accuracy on permuted diagnostic labels. Variation of chance between 36% ($n_{\text{TLE}}/n_{\text{total}}$) and 49% ($n_{\text{TLE}} + n_{\text{UES}}/n_{\text{total}}$) was expected due to the latent structure of the data and classifiers naively diagnosing all patients as the most common class (TLE), which also was considered correct for patients with UES. However, chance accuracies of 58% for the neuroimaging-first CD classifiers seem inflated, for a number of reasons that can and should be explored. For instance, latent structure of the data could have been used to identify coherent subclasses that the randomly permuted diagnostic labels did not break up. This exploration is outside the scope of this short manuscript.

While most of our diagnostic accuracies were significantly above chance, they were too low to be readily applicable to clinical medicine. We expect that CADT performance would improve by including more detailed clinical information, including ictal semiology and co-morbidity profile; as well as integrating in automated MRI- and/or PET-based CADTs that utilize features not appreciated by radiologists (i.e. [27, 26, 28, 7]). However, the addition of these other diagnostic features could magnify the problem of the curse of dimensionality. We, therefore, chose to focus first on simplified, high-salience features to assess multimodal classification methods.

To develop this CADT, we relied solely on archived clinical data from a tertiary epilepsy center, which has its benefits and limitations. The primary benefit is that the information we used reflects the information that would be available in clinic. This ensures that the CADT performance on this data is more similar to how the CADT would perform when applied in a similar setting, at the cost of accurately describing the underlying pathology [81]. As discussed above, the clinical information may be misreported, and radiologists cannot determine the epileptogenic region in all patients. Therefore, even though our CADTs may be clinically applicable, these observed trends may or may not reflect the true pathologic process of disease.

Archived clinical data often are limited because some data are missing. In this case, we multiply-imputed the missing durations, seizure frequency and neuroimaging results based on multilinear trends in all of the other included variables. This allowed the imputed missing data points to contribute to the MRI- and PET-based classifiers. While we expect the variance and, therefore the uncertainty, of each diagnosis to increase with the amount of missing data, in the case of our CADT, multiple imputation has the additional benefit of allowing us to apply one unified model to all patients, irrespective of what data has been collected.

10.6 Conclusion

Conditional dependence resulted in a more clinically-applicable CADT compared to vector concatenation. The imposed structure of conditional dependence improved performance. The opposite order of modalities in our analysis suggests that computers view the data differently from clinicians and could provide a non-redundant, complementary perspective on the data that could improve diagnostic accuracy and certainty, when combined with clinicians' expertise.

CHAPTER 11

Parameter Selection in Mutual Information-Based Feature Selection in Automated Diagnosis of Multiple Epilepsies Using Scalp EEG

This chapter is a reproduction of our work that appeared in the proceedings of the International Workshop on Pattern Recognition in Neuroimaging.[293] This work is a collaboration with Ariana Anderson, Hongjing Xia, Eric S. Braun, Edward P. Lau, Andrew Y. Cho, and Mark S. Cohen. EPL and AYC assisted with computational processing and parrallelization of code. AA assisted with machine learning design and interpretation of results. HX assisted with implementing the mRMR feature selection, as well as understanding the structure of the data. MSC assisted with all stages of planning, implementation and manuscript preparation.

11.1 Abstract

Developing EEG-based computer aided diagnostic (CAD) tools would allow identification of epilepsy in individuals who have experienced possible seizures, yet such an algorithm requires efficient identification of meaningful features out of potentially more than 35,000 features of EEG activity. Mutual information can be used to identify a subset of minimally-redundant and maximally relevant (mRMR) features but requires a priori selection of two parameters: the number of features of

interest and the number of quantization levels into which the continuous features are binned. Here we characterize the variance of cross-validation accuracy with respect to changes in these parameters for four classes of machine learning (ML) algorithms. This assesses the efficiency of combining mRMR with each of these algorithms by assessing when the variance of cross-validation accuracy is minimized and demonstrates how naive parameter selection may artificially depress accuracy. Our results can be used to improve the understanding of how feature selection interacts with four classes of ML algorithms and provide guidance for better a priori parameter selection in situations where an overwhelming number of redundant, noisy features are available for classification.

11.2 Introduction

The accuracy of machine learning (ML) relies on the identification of salient features that reflect, at least partially, the discrimination in question. Ideally that feature space is sparse and, in clinical classification, it is based on biological features with prior likelihood of involvement in the medical condition. In complex, heterogeneous clinical syndromes, such as epilepsy, there are large numbers of computational features with sound biological support. ML methods can be used to identify features that discriminate the patients from controls. This can elucidate the pathology underlying complex disorders but there are an overwhelming number of possible features, many of which are redundant. A key challenge to this approach is: how does one select the number of features to include? ML methods often use a hypothesis-driven approach to select a small subset of features and explain their discriminative efficacy without reference to excluded features. The salience of these hypothesized features can be confirmed using principled data-driven feature selection algorithms that leverage features against each other and results in a better characterization and therefore classification of disease. Mutual information (MI) is particularly capable of considering the interactions of thousands of features simultaneously, using model-free methods to identify those that are minimally-redundant and maximally-relevant (mRMR) to the classification [6]. This depends on the choice of two parameters: the number of quantal levels, Q, in which to bin the continuous features, and the number of features, F, selected to classify.

In this work we use resting state EEG data to distinguish whether a given patient suffers from epilepsy or instead has experienced non-epileptic seizures (NES). Conventional methods initially miss greater than 50% of patients with epilepsy and further assessment inadequately diagnoses up to 30% of patients. Starting from roughly 40,000 different summaries (features) per subject of EEG behavior, we use mRMR to select features and create an ML classifier that discriminates between epilepsy and NES. We demonstrate how parameter choices of (Q, F) affect the mean and variability of the cross-validation accuracy; arbitrary parameter selection can lead to models that systematically classify worse than chance while naive attempts to optimize parameters within the model can lead to bias. We demonstrate the varying effect of parameter selection on four classes of machine learning algorithms: Support Vector Machines (SVM), Multilayer Perceptrons (MLP), Bayesian Logistic Regression (BLR) and Alternative Decision Trees (ADT). Accurate discrimination will translate directly to reduced morbidity, and the results of our sampling of the parameter space can be used to guide others in the selection of these critical parameters when utilizing mRMR.

In mRMR, all continuous features must be smoothed into Q a priori selected discrete bins. Redundancy between features is computed by calculating the MI between features:

$$MI(X_i, X_j) = \sum_{i=1}^{Q} \sum_{j=1}^{Q} \frac{n_{ij}}{N} \log_2 \frac{N n_{ij}}{n_{i\cdot} n_{\cdot j}}, \tag{11.1}$$

where n_{ij} is the number of elements in bin (i, j) from the joint time series (X_i, X_j).

For features in which the classes are separable, a clear choice for Q is 2. Real data, however, is rarely separable. If the continuous scale of a feature is meaningful, then discretizing the data results in a loss of information that increases with the log of the chosen bin size [363]. If bin size is minimized with one exemplar per bin, then n_{ij} is uniformly one, resulting in an inaccurate MI calculation. Therefore the optimal value of Q is likely intermediate. We hypothesize that near the optimal number of quantal levels, the variance of the accuracy is decreased due to effective calculation of mutual information.

A limitation common to mRMR and many other feature selection algorithms is that the number of features, F, must be selected prior to testing. Selection of too few features omits valuable discriminative information, whereas selection of too many features risks over fitting. Because the magnitudes of these accuracy decreasing forces are both minimized at the optimum, we expect the variance of cross-validation accuracy to decrease around the optimal number of features. When too many features are specified, the accuracy in the training set is relatively stable, whereas the test accuracy varies artifactually. Similarly, when the number of features is inadequate to explain the test set variation the accuracy is highly affected by the addition or subtraction of salient features. This optimum number of features might not be conserved across different ML algorithms, as algorithms vary substantially in the degree to which features with low signal to noise ratios contribute to the classification. SVM and MLP perform remarkably well using extremely high dimensional neuroimaging data but fail when considering small numbers of highly salient features, relative to BLR or decision trees. Algorithms that omit information distributed across many features may not capture highly discriminatory information; however algorithms that integrate many features may be incorporating redundant information. Reducing redundancy using mRMR ensures that the utilized subset closely represents the full dataset thereby minimizing the computational burden of operating on non-contributory information and reducing the effect of redundant, low salience features. The relevancy criteria used in MI may screen out noise features, but it is not guaranteed that this translates to higher classification accuracy. The selection of support vectors in SVM suppresses data points far from the decision boundary, whereas MI incorporates all data points. Therefore, multiple ML classifiers must be tested with different numbers of input features to generalize the effect of parameter selection on classification accuracy.

11.3 Methods

11.3.1 Patient & EEG Processing Information

Our subjects include 156 patients admitted to the UCLA Seizure Disorder Center Epilepsy video-EEG Monitoring Unit (EMU) from 2009-2011. Upon the completion of monitoring, 87 were diagnosed with a diverse set of epilepsies and the remaining 69 were diagnosed with non-epileptic seizures by clinical criteria. All scalp EEG recordings were collected in accordance with standardized clinical procedures with a 200 Hz sampling rate using 26 electrodes placed according to the International 10-20 system. During acquisition, an analog 0.5 Hz high pass filter was applied to all recordings. Reviewed data consisted of between 1.5 and 25 hours (mean 9 hours, S.D. 4.5 hours) of archived EEG from either the first or second night of video-EEG monitoring. This work is compliant with the UCLA IRB (IRB#11-000916, IRB#11-002243).

The mean, standard deviation, minimum and maximum power spectra for non-overlapping 1 sec, 5 sec, 60 sec, 30 min windows of EEG recordings from all electrodes relative to reference electrode 1 were calculated in MATLAB. The absolute value of spectral energy from 1-100 Hz was averaged over 1 Hz spectral bands for each of the 26 electrodes. Short window lengths measure phenomena analogous to event related spectral perturbations (ERSPs) whereas longer windows capture baseline activity and connectivity. The power spectra from 58-62 Hz were excluded from all analysis to avoid AC line noise, leading to 39,174 features per subject describing EEG activity. No other artifacts were removed. Ictal activity, muscle artifact and bad channels were included in analysis.

11.3.2 Sampling, Feature Selection and Classification

The most relevant and least redundant of the 39,174 features were selected for specific (Q, F) using the highly efficient mRMR feature selection algorithm optimized and released for MATLAB and C++ by Ding & Peng [6]. All machine learning algorithms were implemented using default parameters in Weka 3.6.4[292] using the full continuous range of each selected features. Accuracy is based on cyclical leave-one-out cross validation that left one subject out of both the feature selection and ML training.

We sampled the cross-validation response surface of (Q, F) using a series of grids with highly parallel computing. The computational burden of each sample is $O(F^3)$ therefore the space was more densely sampled for low F. Sampling points with more than 2,400 features took over 156 days and is therefore infeasible. Sampling 17,677 of the more than 365,000 possible parameter combinations took more than 144 cpu-years therefore the use nested cross-validation, and permutations are infeasible. The possible discontinuity and non-convexity of the space violates the assumptions of most joint optimization procedures.

We then examined the local variation in the 3D space and also the trends of accuracy and variance across each parameter individually. The visualizations of the 3D space utilize Akima bivariate interpolation to fill in unsampled points [4].

When modeling variation along an individual parameter ignoring the other, we interpolated the value of unsampled parameters using a Loess smoother [364]. Because higher numbers of features were sampled less densely, the smoother was trained on log-features in order to maintain a consistent sampling density across the domain.

11.4 Results

Figure 11.1 illustrates the cross-validation accuracy for each of the four algorithms. To illustrate the full space, the F dimension is shown in log steps. On average, the SVM outperformed the other algorithms that otherwise seemed relatively indistinguishable. For extreme quantal levels, the accuracy of the MLP was comparable to the SVM. The maximum accuracy achieved was 86, 70, 69 and 71% for BLR, ADT, SVM and MLP, respectively. The minimum accuracy was 8, 32, 43 and 29% for BLR, ADT, SVM and MLP, respectively. The accuracy for a naive classifier in this setting is 56% (95% CI: 48-64%). Falsely assuming that the sampled points were randomly selected without replacement, the 95 percent confidence intervals for the mean cross validation accuracy for BLR, ADT, SVM and MLP were: 54.7-54.9; 54.0-54.2; 57.8-57.9 and 55.5-55.7%.

11.4.1 Variance with respect to Feature Number

As illustrated in Figure 11.2, the variance of accuracy of most algorithms decreases with increasing feature number. The notable exception is SVM, which had higher accuracy and lower variance across almost all of the space. The minimum for each algorithm was reached at 2,400; 2,400; 1 and 234 feature(s) for BLR, ADT, SVM and MLP, respectively.

11.4.2 Variance with respect to Quantal Level

As illustrated in Figure 11.3, the variance of accuracy is relatively constant except for high Q. The variance of SVM is lower than all algorithms across all selections.

Figure 11.1: The cross validation accuracy of all four classifiers. The unsampled points are filled using Akima bivariate interpolation [4]. The bottom right corner is set to 0 due to lack of support. All values less than 40% are rounded up to 40% to maintain contrast. Without multiple testing correction, individual yellow to red points are significantly more accurate than a naive classifier whereas deep blue points are significantly worse.

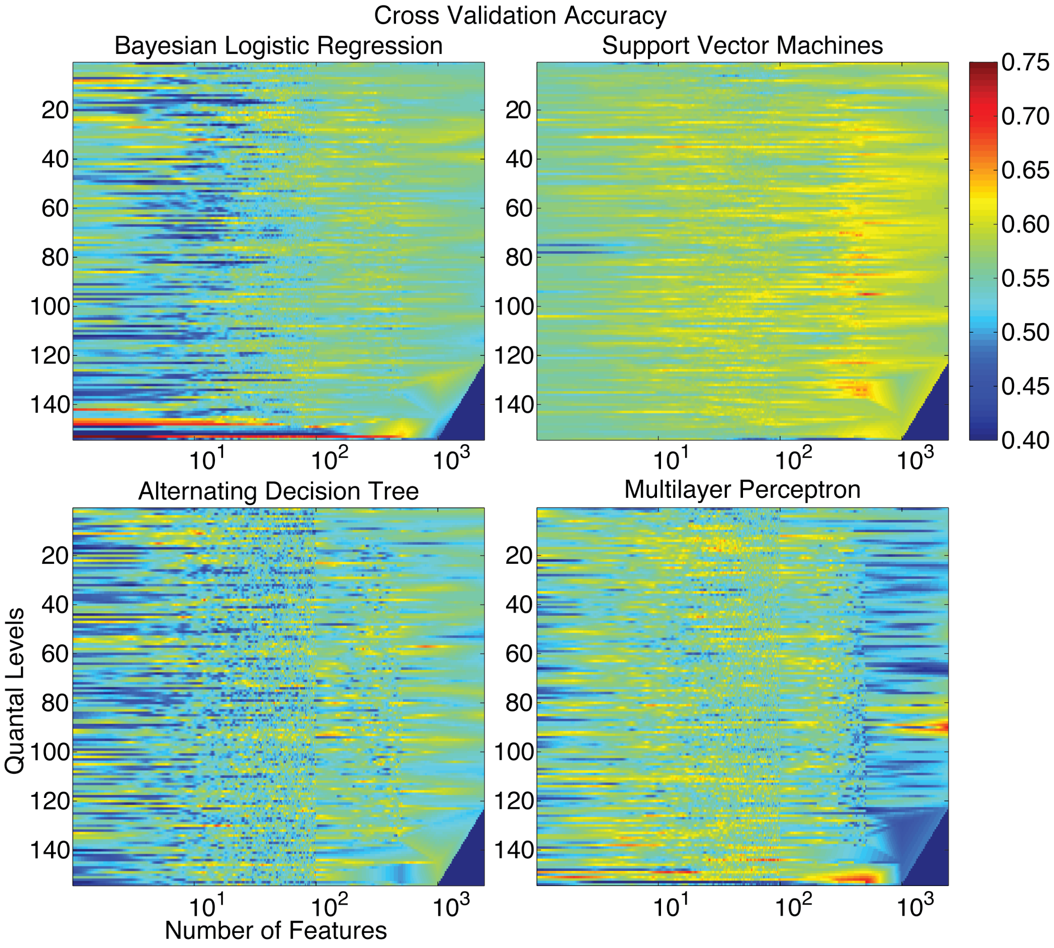


Figure 11.2: Variance of cross validation accuracy with respect to number of input features. Thickness represents standard error.

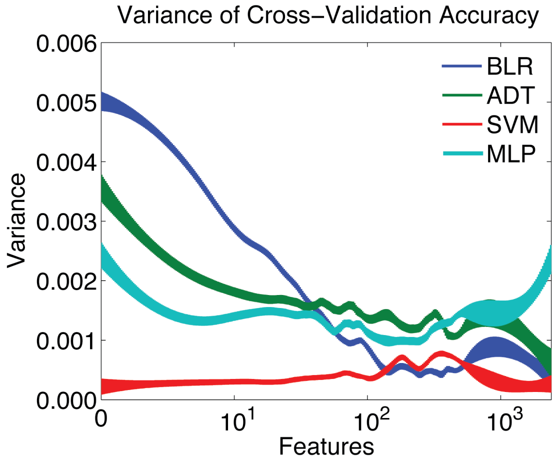
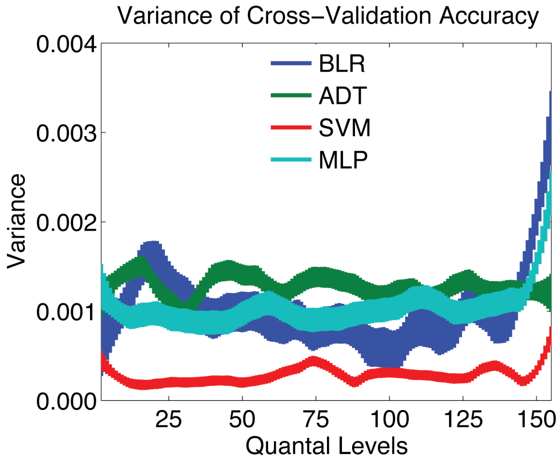


Figure 11.3: Variance of cross validation accuracy with respect to number of quantal levels. Thickness represents standard error.



11.5 Discussion

We note a few key observations regarding the behavior of the cross-validation accuracy. (1) The distribution of all algorithms has substantial negative skew. (2) The number of features is responsible for most of the variation in accuracy. (3) The variance of cross-validation accuracy largely is independent of the choice of quantal level. (4) SVM has decreased variance and increased accuracy within this system compared to the other algorithms.

When visualizing the overall cross-validation accuracy (Fig. 1), the majority of points seem to be less than chance, 56%. This, however, is not the case. Negative bias occurs because the majority of points are sampled for low feature number relative to the optimum, causing these less accurate points to be over represented. This skew means that a naive or random choice of parameters could lead to a conclusion that no discriminatory signal exists when a signal indeed exists. This illustrates the need to better understand the effect of parameter selection.

It is apparent that as long as an intermediate number of quantal levels are chosen, the variance of accuracy is relatively constant. This suggests that the mRMR algorithm is resistant to small variations in the selection of this parameter and confirming our hypothesis that variance is decreased around the optimum Q. Even as the variance of accuracy is constant across quantal levels (Fig. 1), the accuracy is even more consistent within each quantal level.

Similarly, the variance with respect to number of features had very similar trends across the four algorithms. All of the algorithms achieved a local minimum of variance at around 200 to 500 features. This suggests that across all algorithms, this may represent the number of non-redundant features in the data that hold diagnostic information. As expected from the bias-variance tradeoff, the accuracy grows according to a roughly sigmoidal function that peaks around 500 features. After this optimum, the accuracy then falls, possibly due to over fitting, as discussed in the introduction. Due to the decreasing trend in variance for all but the MLP, it is not guaranteed that this represents a global optimum.

The magnitude of each of the variances suggests that on a large scale, the number of features is much more important than the choice of quantal level. Around the region with decreased variation in F, however, the variance from Q is of similar magnitude. This suggests that this space may be explored efficiently using coordinate descent.

It is particularly interesting to note the large difference in variance for low feature numbers. The most salient example is the BLR that has between 1.5 and 5 fold more variance for low F than the other algorithms. BLR achieves both the maximum and minimum global accuracy with low F and high Q, therefore this performance may be due to noise instead of (in)effective modeling of the underlying pathology or MI. This is confirmed by our hypothesis that with high Q the probability distribution of each feature approaches the uniform; therefore the MI calculation may be ineffective.

On the other hand, the SVM was impacted the least by the parameter selection. The variance with respect to both parameters was less for the SVM than for most of the other algorithms across all parameters. As discussed in the introduction, this may be because the underlying weighting of relevancy by mRMR is substantially different from the SVM. The minimum accuracy observed for the SVM was substantially higher than the minimum for all other algorithms. The optimum accuracy was also achieved across intermediate quantal levels, suggesting that it reflects an effective modeling of the underlying data to discriminate epilepsy from non-epileptic seizures. We caution against interpretation of the extrema because their significance can only be assessed using random field theory [365] and/or bias correction [366], both of which are out of scope for this article.

Based on these and other results, we believe that the power spectrum of EEG holds valuable diagnostic information for epilepsy but that this diagnostic information is hidden among a large degree of noise [74]. A deeper understanding of parameter selection may lead to the efficient implementation of power spectrum information on an automated diagnostic tool for epilepsy.

In general, a priori selection of the number of input features and quantal levels in mRMR has the same challenges as other feature selection algorithms even though it involves a joint optimization because accuracy is generally invariant of quantal level. The selection of the optimum number of features requires sampling to determine the region that maximizes both classification accuracy and minimizes the variance with respect to changes in number of features. This ensures that the accuracy is not inflated artifactually but also correctly reports the best accuracy that can be achieved in practice.

CHAPTER 12

Hyperparameter Optimization with Random Field Theory without Nested Cross-Validation

This is an early and partial draft of our manuscript in progress. This is a collaboration with Ryan M. McCarthy, Andrew Y. Cho, Edward P. Lau, Marc A. Suchard and Mark S. Cohen. WTK came up with the idea, organized the collaboration, wrote most of the code and the manuscript. RMC assisted with initial explorations of the idea and wrote some simulations. AYC and EPL assisted with the computational aspects of this work. MAS helped verify and develop the analytical foundation for this work, as well as assisted in the other aspects of the work. MSC assisted with framing, design of experiments, manuscript preparation and interpretation of results.

12.1 Abstract

Machine learning models effectively estimate the optimal value of parameters inherent to the model based on training data, but determining the optimal value of hyperparameters comparatively less well defined. We define hyperparameters as selected values that are not optimized jointly with the parameters inherent in the log-likelihood of the data, given the model. Common examples of hyperparameters are the soft margin parameter, C, in support vector machines (SVM); the regularization parameter, λ, in regularized machine learning models; and the number of features selected from a high dimensional dataset using a filter-based feature selec-

tion. Learning of these hyperparameters is critical to developing highly accurate and applicable models, but current methods to learn these values typically rely on *a priori* selection or nested cross-validation. In this manuscript, we theoretically pose and empirically validate a novel method for assessing the significance of cross-validation accuracy and optimizing hyperparameters by understanding the inherent dependency between the cross-validation accuracy for similar hyperparameter choices using the ideas of random field theory. This method has the potential to reduce the computation cost of fitting highly accurate models, especially in applications with limited training data, while reducing the complexity of interpreting the final solution. It also gives quantifiable evidence about the stability of the learned solution, with respect to changes in hyperparameter selection.

12.2 Introduction

Training effective machine learning models relies on learning optimum parameters, ψ, and hyperparameters, θ, to describe the problem at hand accurately. Each of the many machine learning models is based on maximizing some form of log-likelihood of the data given ψ and the chosen model. We define a hyperparameter as a value or choice made outside of the context of maximizing the log-likelihood that affects the structure and effectiveness of the model. In this manuscript, we theoretically pose and empirically validate a novel method based on random field theory (RFT) for assessing the significance of observed cross-validation (CV) accuracies achieved over a range of θ that maximizes the data available for training and validation, while reducing computational cost.

This addresses a key challenge with the application of machine learning models that include hyperparameters. In many manuscripts, θ seems to be chosen arbitrarily. Hopefully, this arbitrary choice was made prior to observing the CV accuracy that was achieved by a given θ. If it wasn't and the researchers instead chose the θ that resulted in the best performance, their performance would be inflated and the estimates of significance would be inaccurate. For even the most responsible researchers, there is little guidance provided in the choice of many types of θ. For example, when applying a filter feature selection, one must choose the number of features, F, to allow through the filter, out of the total number of features, m. The best evidence is that F should be chosen based on biological prior knowledge, or simply prior knowledge with the classification scheme[3].

This challenge of selecting θ is that without observing the CV performance, it is difficult to predict performance or the sensitivity of performance on the choice of θ. The methods we outline here, while applied to hyperparameters in machine learning, can be applied to other statistical problems where an arbitrary parameter(s) must be chosen to assess how well a model fits, and the sensitivity of model fitting to the choice of parameter is unknown.

The applicability of a machine-learning model to unseen data is assessed commonly through CV or testing its performance on an external or left out dataset. This requires splitting the data into at least two groups: training and validation. The training group is used to optimize ψ. However, separate data is needed to optimize θ to limit overfitting. This can be done by further splitting the training set into a smaller training set to optimize ψ and a testing set to optimize θ. Frequently, this is done within the setting of cyclical K-fold nested cross-validation, where each exemplar is used successively for training, testing and validation. The method of splitting the full dataset into subsections ensures that the reported generalization accuracy is not overinflated by overfitting the data. Overfitting is where the learned ψ and θ capture trends present in the available data that do not translate to unseen data.

However, this practice of successively splitting the data results in fewer exemplars to train ψ and θ, increases computational cost, and decreases the interpretability of the model. If K-fold nested CV is employed with 10-fold outer and inner loops, then we estimate 100 different pairs of $\hat{\psi}$ and $\hat{\theta}$. For the duration of this manuscript, we denote estimates of values with a hat. Ten-fold nested CV, however, results in only 80% of the available data being used to learn ψ, arguably the most important part of a model. One generalizable finding in machine learning is that the more training data that is used to learn the model, the better the model reflects the underlying problem at hand. Consequentially, one could conduct leave-one-out CV in both the inner and outer folds, resulting in n^2 different pairs, where n is the number of available exemplars. This moderate increase in the number of training exemplars may result in only moderately improved accuracies; the reported accuracy is what determined the impact or applicability of machine learning models. However, even as leave-one-out nested CV may improve generalization accuracy, it does so at the expense of computational cost, it also decreases the interpretability of the learned ψ and θ. For some models, averaging ψ and θ is a valid procedure to result in a single, interpretable summary statistic, but this ignores the potential strong dependency between ψ and θ. While the variability around these averages can be quantified, there is insufficient data to assess how this variability influences CV accuracy.

To address these challenges, we propose to apply the ideas of random field theory to understand how cross-validation accuracy varies “spatially” with respect to changes in θ. This directly studies the sensitivity of the reported CV accuracy with respect to θ, thereby improving the interpretability of $\hat{\theta}$. In addition, by using statistical theory to describe the “spatial” relationship, we remove the need for the nested CV loop without compromising the validity of the reported CV accuracy. Lastly, and most importantly, by understanding how to correct for testing multiple θ, we develop a statistical procedure for determining the significance of an observed CV accuracy, as compared to chance.

To our knowledge, there are only two methods that address this specific problem. First, one could consider each assessment of the CV error for different θ a statistically independent test. Therefore, one could apply a Bonferroni correction to assess the significance of the highest achieved CV error. This, however, ignores the dependence that we will make clear below. Secondly, Tibshirani & Tibshirani proposed that when testing multiple θ, the bias for choosing the best observed CV error is the same as the bias for choosing the best observed training error. Therefore, this bias can be corrected by the following expression:

$$Error_{\text{Corrected}} = Error_{\text{CV}}(\hat{\theta}) + Bias = 2 \cdot Error_{\text{CV}}(\hat{\theta}) - \frac{1}{K}\sum_{k=1}^{K} e_k(\hat{\theta}_k) \qquad (12.1)$$

where $e_k(\hat{\theta}_k)$ is the prediction error on the *training set* of the optimum choice of θ within cross-validation fold k. To demonstrate the superiority of random field theory to these two methods, we will compare the ability of these methods to achieve an unbiased estimate of the true CV accuracy in each application.

This manuscript can be split into five building sections: (1) theoretical framing of random field theory for this application, (2) analytical exploration of the shape of the random field, (3) empirical estimation of the random field in null data, (4) application of this theory to real world data from patients with epilepsy, and (5) comparison to alternate correction methods. Subsequently, we discuss how this theory can be applied for planning sampling schemes of θ and for correcting observed cross-validation accuracies when multiple θ are sampled regularly.

12.3 Theoretical Framing of Random Field Theory Application

In random field theory, one considers a stationary random variable that varies across space such that nearby values are dependent based on a given smoothing kernel. By describing this kernel, one can derive analytical expressions for the probability that the random variable will achieve a certain value, or more extreme, over a region with a given size. Through the Statistical Parametric Mapping software, Gaussian random field theory, in particular, has become a popular multiple testing correction in neuroimaging data. For the duration of this manuscript, we use quotations around the phrase "spatial" to denote that not all θ represent spatial dimensions, but there is an analogy between our theory and the theory of Gaussian random fields for neuroimaging data, where θ are spatial dimensions.

For our application, consider CV accuracy, ϵ, a random variable that varies as a function of θ. In that way, we think of θ as describing a position in "space". For each CV exemplar, there are two versions of ϵ that we can consider: ϵ as an indicator for correct prediction, therefore $\epsilon \in \{0, 1\}$; or ϵ as the amount of predictive error, with $\epsilon \in \mathbb{R}$. In both cases, it is clear for an arbitrary θ_1 that as $|\theta_1 - \theta_2|$, the expected values of $E\left[\epsilon(\theta_1)\right]$ and $E\left[\epsilon(\theta_2)\right]$ become more similar. It follows that the correlation, $Corr\left[\epsilon(\theta_1), \epsilon(\theta_2)\right]$, is non-zero for θ_1 and θ_2 that are sufficiently close together. Therefore, $\epsilon(\theta_1)$ and $\epsilon(\theta_2)$ are not statistically independent. To derive a null probability distribution for ϵ in this random field, one must determine the structure and range of the dependence of ϵ with respect to $|\theta_1 - \theta_2|$. This null probability distribution can be used to accomplish our ultimate goal: determine the significance of an observed ϵ achieved over a range of θ.

12.3.1 Chosen Examples

There are a great variety of hyperparameters. We will address a few major categories of hyperparameters including regularization costs, λ; the soft margin parameter, C, in support vector machines (SVM); the number of selected input features, F, from a filter-based feature selection; and the number of quantal levels, Q, chosen for to calculate the mutual information for minimum redundancy, maximum relevancy (mRMR) feature selection [**?**]. While these do not comprise the whole variety of hyperparameters, these choices allow the reader to assess the wide applicability of our proposed perspective. For simplicity, we discuss hyperparameter selection and its relevance to binary classification problems, although these ideas could translate to multiclass and other problems.

First, regularization parameters are used to apply the prior hypothesis that the optimal solution is sparse in the number of features, X, that contribute to the solution. This is accomplishing by incorporating a regularization penalty into the log-likelihood in the following form:

$$L(Y, X|\psi, \theta) = \lambda\|\beta\| + \ell(Y, X|\psi), \tag{12.2}$$

where Y is the vector of class membership, X is the matrix input data, ψ is the set of parameters including β, the weights placed on the input data, θ is the set of hyperparametrs including λ, the regularization parameter and $\ell(Y, X|\psi)$ is the chosen log-likelihood function. In our applications, we will explore when ℓ corresponds to logistic regression. Conventionally, λ is taken as unity, or may be scaled dependent on the length of β or size of X.

Secondly, we study the effect of the soft margin parameter, C, in a SVM model. In SVM, we optimize the following Lagrange equation:

$$\begin{aligned} L(Y, X|\psi, \theta) =& \lambda\|\beta\|_2^2 + \sum_{i=1}^{n} \frac{\|w\|_2}{2} + \sum_{i=1}^{n} \alpha_i \left[y_i \left(w \cdot x_i - b \right) - 1 \right] \\ & + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \beta_i \xi_i, \end{aligned} \tag{12.3}$$

where α is the vector of weights on exemplars to indicate the support vectors, b is the bias due to imbalanced classes or non-centered input data, ξ is the vector of distances from the separating hyperplane on the wrong side. When C is zero, SVM simplifies to linear discriminant analysis, depending on the kernel used. When C is ∞, the soft margin SVM is equivalent to a hard-margin SVM. Simultaneously, the value of C is related to the amount of regularization applied, therefore λ and C are not statistically independent. Conventionally, C is taken as unity or is optimized using nested-cross validation.

Lastly, we study two hyperparameters related to feature selection, because hyperparameters occur more frequently when multiple optimization methods are applied, but the additional methods are not incorporated into the log-likelihood optimization. The simplest filter-based feature selection only utilizes the F features with the highest t-statistic difference between the two classes. Equivalently, one could select a particular t-statistic cutoff with which features needed to pass prior to incorporation in the classification model.

Next, we address hyperparameter seletion the mRMR feature selection [**?**, **?**], which requires a priori selection of the number of selected features, F, and the number of quantal levels, Q, used in the calculation of mutual information. As discussed in [**?**], the calculation of mutual information requires discretizing the data or selecting a Gaussian smoothing kernel with a chosen σ. Both of these feature selection hyperparameters result in the following log-likelihood:

$$L(Y, X|\psi, \theta) = \ell(Y, X^{(1..F)}|\psi) \tag{12.4}$$

where $X^{(i)}$ is ith ranked feature based on either t-statistic ranking or maximizing the mutual information between the included features and the predicted class, minus the average redundant information across input features. Mutual information is defined for discrete input data, therefore continuous data must be quantized, with a certain number of quantal bins, Q. Ideally, F and/or Q are chosen prior to assessing the cross-validation accuracy of the model, but when preparing a manuscript, it is very tempting to simply report the optimal value without discussing the process with which that optimum was reached. Conventionally, if one seeks to optimize F and/or Q, it must be done through nested-cross-validation.

12.4 Analytical Exploration of 'Spatial' Dependence of Hyperparameter Choice

To apply random field theory to each of these hyperparameters, we must determine the structure of the dependence of ϵ on θ around the optimal value of θ. For sake of comparison, in Gaussian random field theory, the correlation of nearby ϵ resembles a Gaussian random variable in $|\theta_1 - \theta_2|$. Equivalently, this suggests that $\epsilon(\theta_1)$ can be approximated locally using a first order Taylor approximation, i.e. $\frac{\partial^m \epsilon(\theta_1)}{\partial \theta^m}$ is zero for $m > 1$. To explore the structure of the dependence, we will examine the analytical form of the derivatives of the likelihood function with respect to θ in each of our chosen examples. This assumes that the log-likelihood is a valid proxy for ϵ and that the unseen validation data comes from the same distribution as the training data.

12.4.1 Regularization Hyperparameter, λ

As an example of a regularized machine learning model, we consider a canonical L_2 regularized logistic regression model, due to it's firm analytical motivation and amenability to differentiating. For the purposes of this analytical exploration, we consider $\epsilon_i(\theta) = y_i - \pi_i$ where where $y_i \in 0, 1$ is the known class membership for exemplar i, π_i is the estimated probability exemplar i is in class $y_i = 1$. We denote $\epsilon(\theta) = \sum_{i=1}^{n} \epsilon_i(\theta)$. In logistic regression, the log-likelihood function we seek to maximize using the training data is:

$$\ell(\beta) = \sum_{i=1}^{n} y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) \text{ where } \pi_i = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}} \tag{12.5}$$

where β is a to-be-estimated vector of weights on the input data and X_i is a vector of known input data, including an intercept term. We add regularization into the log-likelihood function by adding an L_2 penalty, with the set of hyperparameters $\theta = \{\lambda\}$, resulting in a regularized log-likelihood of:

$$\ell_R(\beta, \lambda) = \lambda \|\beta\|_2^2 + \sum_{i=1}^{n} y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) \tag{12.6}$$

To learn about $\frac{\partial \epsilon(\lambda)}{\partial \lambda}$, we use the log-likelihood, $\ell_R(\beta, \lambda)$ as a proxy for $\epsilon(\theta)$. We simplify the problem at hand using the chain rule, which states that

$$\frac{\partial \epsilon}{\partial \lambda} = \frac{\partial \epsilon}{\partial \beta} \frac{\partial \beta}{\partial \lambda} \tag{12.7}$$

First, we optimize β by setting $\frac{\partial \ell(\beta,\lambda)}{\partial \beta}$ equal to zero. This result in the expression (for derivation, see Appendix A):

$$\beta = \frac{1}{\lambda} \sum_{i=1}^{n} [y_i - \pi_i] \, X_i \tag{12.8}$$

We note the similarity of this expression to the transcendental equation that has no closed-form solution. In unregularized logistic regression, the transcendental equation is optimized frequently using Newton's method, suggesting that a second order Taylor assumption for $\frac{\partial \ell(\beta,\lambda)}{\partial \beta}$ is valid locally. Using this expression, we also can determine $\frac{\partial \beta}{\partial \lambda}$ by differentiating the modified transcendental equation, resulting in the following expression (for derivation, see Appendix A):

$$\frac{\partial \beta}{\partial \lambda} = \frac{\sum_{i=1}^{n} y_i X_i}{\lambda - H(\beta)} - \frac{\beta}{\lambda - H(\beta)} \tag{12.9}$$

where $H(\beta)$ is the Hessian of β that is used within Newton's method to optimize $\ell(\beta)$ with respect to β. Since we are interested in developing a method to determine significance, we are most interested in the structure of the random field under the null hypothesis that there is no predictive relationship between X, the matrix of input data, and Y, the vector of class memberships. In this case, $E(\beta) = 0$, which simplifies the above expression to:

$$\frac{\partial \beta}{\partial \lambda} = \frac{\sum_{i=1}^{n} y_i X_i}{\lambda - H(\beta)} \tag{12.10}$$

From this expression, it is apparent that (1) the random field may not be stationary across λ and (2) subsequent derivatives of ϵ with respect to λ clearly will be non-zero. Therefore, while the surface of ϵ among changing λ may be a random field, we believe that it may be non-stationary and non-Gaussian.

12.4.2 Soft Margin Hyperparameter, C, in SVM

Similarly to above, we begin by differentiating the objective function of the SVM above (eq. 1) with respect to C. However, in contrast to using $\ell(\psi, \theta)$ as a proxy for $\epsilon(\theta)$, we note that ξ_i is the continuous $\epsilon_i(\psi,\theta)$ that we are interested in studying, therefore we rearrange the expression in terms of these terms to yield:

$$\begin{aligned}\frac{\partial f(\phi,\theta=\{C\})}{\partial C}=0=&w\frac{\partial w}{\partial C}+\frac{\partial}{\partial C}\sum_{i=1}^{n}\alpha_i\left[y_i\left(w\cdot x_i-b\right)-1\right]+\|\xi\|_1+\frac{\partial}{\partial C}\sum_{i=1}^{n}\beta_i\xi_i\\ -\|\xi\|_1=&w\frac{\partial w}{\partial C}+\sum_{i=1}^{n}\left[\alpha_i y_i\left(\frac{\partial w}{\partial C}\cdot x_i-\frac{\partial b}{\partial C}\right)+\left[y_i(w\cdot x_i-b)-1\right]\frac{\partial \alpha_i}{\partial C}\right.\\ &\left.+\beta_i\frac{\partial \xi_i}{\partial C}+\xi_i\frac{\partial \beta_i}{\partial C}\right]\end{aligned}$$

Unfortunately, the soft-margin SVM has no closed form solution and is, instead, solved through quadratic programming. Therefore, to understand the structure of the partial derivative with respect to C, we must utilize our knowledge of the inter-relationship between the estimated variables. First, we consider the Lagrange variables, α and β, which are critical to the difference between a soft and hard-margin SVM. In a hard-margin SVM, C is effectively ∞ because no misclassification errors are allowed. In terms of the objective function, this is equivalent to stating that $\xi_i=0\forall i$ or $\beta_i=\infty\forall i$. When C is finite, misclassification errors are allowed, but up to a point. For a constant C, it can be shown trivially that $C=\beta_i+\alpha_i$ (see Appendix B). This linear equation suggests that the derivative of β_i and α_i with respect to C is linear, for all i. To phrase the linear relationship between β_i and C in words as compared to variables, the weight placed on misclassifying exemplar i (β_i) by a distance ξ_i changes linearly with changes in C. Up to a reasonable approximation, this suggests that the optimized ξ_i changes linearly with C. Looking back to our original equation, this suggests that:

$$-\|\xi\|_1\approx w\frac{\partial w}{\partial C}+\sum_{i=1}^{n}\left[\alpha_i y_i\left(\frac{\partial w}{\partial C}\cdot x_i-\frac{\partial b}{\partial C}\right)+\text{ constant terms}\right]\tag{12.11}$$

The exception to the linearity of α and β with respect to ∂C is that α_i and β_i are $\geq 0\forall i$, therefore non-linearities could occur on the boundary (i.e. $\alpha_i=0$ for $C=C_1$ but $\alpha_i>0$ for $C=C_2$ for arbitrary i). This represents when the identity of the support vectors and/or the misclassified training exemplars changes.

This leaves the terms that define the maximum margin hyperplane. Given constant support vectors, α, SVM mathematically is equivalent to weighted and regularized linear discriminant analysis. Within the region of C that the set of non-zero α does not change, it is intuitive to hypothesize that this would induce a linear change in w and b, but difficult to prove analytically (and outside the scope of this manuscript). Outside this region, where the set of non-zero α is not constant, there is no analytical prediction for the structure of the derivative with respect to C.

Therefore, there is some theoretical evidence that for the predictive error, $\|\xi\|_1$, changes approximately linearly with C within local regions for a soft-margin SVM. However, longer range changes may induce highly non-linear changes in $\|\xi\|_1$ due to changes in the identity of support vectors. From a purely theoretical standpoint, it is difficult to predict the structure of these changes.

12.4.3 Feature Selection Hyperparameters

Unlike the previous two examples, the feature selection hyperparameters are not trivially incorporated into the log-likelihood functions, and they are discrete as compared to continuous values. Therefore, we consider the discrete derivatives of ϵ with respect to changes in the number of input features, F, and the number of quantal levels used to calculate mutual information, Q.

First we discuss F. In both feature selection methods, the goal is to order the input data based on decreasing utility in the discrimination of classes. This suggests that higher ranked features will hold more information than lower ranked features. However, since we are developing a theory to determine significance, we are interested in behavior under the null hypothesis when no features hold discriminative information. Therefore, the ranking of features should be meaningless and each feature is equally likely to change the learned model as much as any other feature. (Due to overfitting and the curse of dimensioanlity, this is not true in practice.)

To begin, consider the effect of increasing F by one feature on the discrete accuracy, ϵ. For each exemplar, there is a probability, q, that the exemplar was predicted correctly to start with. Further, for each exemplar there is a probability, p, that the predicted solution will change with the addition of a new feature. This results in a discrete derivative of:

$$\epsilon(F) - \epsilon(F+1) = n(-qp + (1-q)p) = np(1-2q) \tag{12.12}$$

where n is the number of exemplars.

Next, consider increasing F by two features. This result in the following expression:

$$\begin{aligned} \epsilon(F) - \epsilon(F+2) =& n(-qp(1-p) + (1-q)p(1-p) - qp^2 + (1-q)(1-p)^2) \\ =& n(1-p-q-4pq). \end{aligned}$$

This considers all possibilities of being correct, changing once or changing twice, assuming that the probability of change is constant across F. What becomes clear is that these derivatives do not seem to decrease in magnitude as step size increases, therefore a local low order Taylor expansion likely will not provide a good approximation. Therefore, just as in the cases of the regularization parameter, λ; and the soft margin parameter, C; this random field seems to have non-Gaussian long-range dependence.

Lastly, consider the effect of Q on cross-validation error. We note that there was nothing we assumed about F that was specific to the structure of F. Similarly, under the null hypothesis there is no relationship between the input data and class membership. Therefore, there is no optimal Q that could be used to describe the mutual information between the input data and class membership. In that way, we expect changing Q to result in random changes in the model, and thereby the predicted class, just as we assumed with F. The same structure of derivatives is expected to be seen in Q as it was in F. The one exception is that the value of p may vary depending on the number of training exemplars used to calculate the mutual information, and the dimensionality of the input data.

12.4.4 Summary of Analytical Exploration of Dependence

In each of these examples, the analytical prediction of the shape of the dependence of each hyperparameter, θ, was different. This suggests that there may be no universal theory that predicts the "spatial" dependence of CV accuracy and/or goodness of fit on hyperparameter choice. In each of these examples, there also were sections where analytical approximations or formula-free intuition needed to be used to yield interpretable results. This suggests that there may not be a closed form solution that predicts the structure of the dependence. In the next section, we use simulated null data to produce empirical probability distributions that could be used for determining significance.

12.5 Empirical validation of "Spatial" Dependence in Null Data

Given the long-range and non-Gaussian structure suggested by the analytical exploration, we generated null data to examine the structure of the dependence of $\epsilon(\theta)$ for each θ we explored above. This null data included 150 exemplars with a dimensionality, m, of 47, 150 and 1000. At random, half of these exemplars were denoted to be class 1. Each input feature included data sampled independently from a standard Gaussian distribution. Each machine-learning model was trained and validated with leave-one-out cross-validation. This was conducted for 10,000 independent datasets. For each of our chosen examples, we compare the observed long-range structure as compared to a Gaussian random field. The full-width half max of the Gaussian random field was matched with the observed first derivative cross-validation accuracy with respect to changes in θ, assuming a stationary random field. Additionally, we used these results to estimate an empirical probability distribution of CV accuracy achieved across each of these random fields, both with respect to maximum achieved value and the size of the contiguous region within the field that achieved this value.

The ultimate goal is to determine the statistical dependence of CV accuracy, $\epsilon(\theta)$, on θ. Therefore, for each type of θ we illustrate the spatial relationship by (1) addressing stationarity by showing how the expected null CV accuracy varies across θ, (2) addressing structure by calculating the correlation by exemplars of $\epsilon(\theta_1)$ with $\epsilon(\theta_2)$ for varying $|\theta_1 - \theta_2|$ and (3) empirically calculating the probability distribution function (PDF) of the maximum $\epsilon(\hat{\theta})$ achieved over a contiguous region. For comparison, each of these illustrations are plotted against what they would be if the random field was stationary and Gaussian with a matching mean $\epsilon(\hat{\theta})$ and first derivative, $\frac{\partial \epsilon(\hat{\theta})}{\partial \theta}$.

12.5.1 Regularization Hyperparameter, λ

To sample a wide range of feasible λ, we tested $1/\lambda$ between values of 0.05 and 100 with a step size of 2.5. We set $1/\lambda$ because this is the units typically used in studies that select a non-unitary λ. This wide range was chosen because in pilot studies, very high dependence was seen for limited ranges. For each of the dimensionalities examined, the CV accuracy was stationary in the mean across all studied values of λ (student's t-tests, p>0.05).

12.5.2 Soft Margin Hyperparameter, C, in SVM

We sampled a wide range of C from 1/300 to 2 in steps of 1/150. This represents a wide range of feasible choices for C. For each of the dimensionalities examined, the CV accuracy was stationary in the mean across the range of C (student's t-tests, p>0.05).

12.5.3 Feature Selection Hyperparameters

For the number of features, F, and quantal levels, Q, we tested all possible choices for each dataset. First, we examine the use of a t-statistic filter feature selection prior to regularized logistic regression with a λ of 1. In this setting, we only modify the choice of F. Similar to all the previous simulations, the CV accuracy was stationary in the mean across all values of F (student's t-test, p>0.05).

12.5.4 Summary of Dependence in Null Datasets

These null simulations reveal that in some cases (λ and C), there exist long range dependencies in the cross-validation accuracy on the choice of θ, therefore the choice of θ does not effect the overall solution. In other cases (F and Q), the magnitude of the dependence fell off like a decaying exponential. This behavior was difficult to predict based on the analytical derivations above, and suggest that empirical simulations were necessary to understand the nature of the dependency.

Even though the analytical structure of $\frac{\partial \epsilon(\lambda)}{\partial \lambda}$ was shown to be an inverse function, whereas $\frac{\partial \epsilon(C)}{\partial C}$ was expected to be linear, the shape of their resulting correlation and probability distribution function of the field were extremely similar. This may be due to a floor effect in the sensitivity of $\epsilon(\theta)$ with respect to changing θ. Our analytical exploration showed that long range correlation was possible, but did not suggest that the magnitude of correlation would be so high.

The similarity of the dependency of $\epsilon(\theta)$ to F and Q was predicted by our analytical reasoning, even though these hyperparameters were not encorporated explicitly into the objective functions. Our ability to generate an consistent empirical model of the correlation of ϵ with varying θ that did not depend on m

suggests that, in some circumstances, null permutations may not be necessary to assess the significance of observed CV accuracies.

12.6 Empirical Validation of “Spatial Dependence” in Epilepsy Data

The clean results that were achieved using null data confirm that (1) spatial dependence exists and (2) reporting the best CV accuracy achieved after testing multiple θ results in an over-inflated CV accuracy. To explore how the nature of this spatial dependence and bias changes when real world data is used, we examined the same parameters with real data. The difference between simulated data and real data is that only one dataset is available and there is latent correlation structure within X.

In the null data, we generated 10,000 independent datasets. For real data, however, we do not have the luxuary of having 10,000 independent datasets. Instead, we approximate 10,000 independent datasets by permuting the class labels, Y, without replacement. This preserves the latent structure within X but breaks the relationship of Y with X. Thereby, we can generate the same empirical correlation plots and probability distribution of CV accuracy that we did with the null data.

For each type of θ, we discuss the difference between the correlation and probability distribution of CV accuracy for null data and real data, as well as the comparative significance of the observed CV accuracies when the relationship between Y and X was preserved.

Two separate datasets from patients with seizure disorder were used to validate our method. To assess structure corresponding to the regularization parameter, λ; and the soft margin parameter, C, in SVM, and the number of features, F, to include after t-stastic filtering, we utilized a dataset with 951 patients built to discriminate patients with epilepsy (ES) or non-epileptic seizures (NES) based on 84 features reported in outpatient clinical reports. In the original manuscript that described a machine learning model trained using this data, missing data was multiply inputed. For sake of clarity in this manuscript, we treat the first imputation as the true data. Due to the reliance of primarily binary features in clinical reports, we use a different dataset with continuous input data to inspect the effect of feature selection hyperparameter Q. The continuous dataset includes 150 patients and was built to discriminate patients with left versus right temporal lobe epilepsy (LTLE versus RTLE) and NES based on 47 z-transformed FDG-PET based features. In this application, a multilayer perceptron neural network classifier was applied after feature selection to learn the most discriminative model, as was done for the null permutations above. For more information regarding these datasets, please refer to the Chapters 7 and 9.

12.6.1 Regularization Hyperparameter, λ

The goal of regularization in this application is not to make an underdetermined model better determined. In this case, the number of exemplars far exceeds the number of features; therefore the model probably is well determined. Instead, regularization is used to limit the size of the final solution to maximize ease of application of the model.

12.6.2 Soft Margin Hyperparameter, C, in SVM

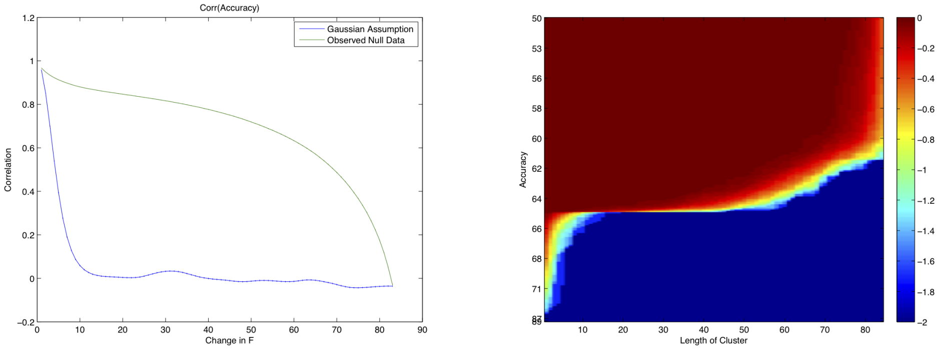
The benefit of optimizing C in this application is in determining if the insight that the model should be most sensitive to difficult to classify exemplars leads to improved classification accuracies over the entire dataset.

12.6.3 Feature Selection Hyperparameters

Our two datasets provide examples of two important cases: the clinical report dataset has many more exemplars than total features, whereas the PET dataset has slightly fewer exemplars per class than total features. Feature selection is useful in both cases because it limits the size of the final model. For unregularized models, however, feature selection is critical in the PET dataset to have well determined and stable estimates of ψ, θ, and overall performance.

First we address optimizing the number of features, $\theta = F$, using t-statistic filtering prior to modeling with an ℓ_2 regularized logistic regression model. Similar to the null datasets above, we tested all 84 possible values for F. Just as in the null datasets, the overall mean accuracy for each value of F was stationary approximately across the whole range of F. The correlation of $\epsilon(\theta)$ and the resulting empirical PDF, however, differed substantially compared to the null permutations.

Figure 12.1: (Left) The correlation of cross-validataion accuracy by subject for the clinical dataset at one value of F_1 compared to F_2 where the distance between F_1 and F_2 is modified. (Right) The log PDF of maximum cross-validataion accuracy achieved over a range of F.



12.6.4 Summary of Empirical Validation in Epilepsy Data

12.7 Comparison to Alternative Correction Methods

As discussed above, there are two alternatives to the random field theory method described here that do not require nested-cross validation: Bonferroni correction and the training-set bias correction proposed by Tibshirani & Tibshirani. Both of these corrections do not explicitly take into account the spatial extent of the maxima achieved, although a simple extension of the Tibshirani & Tibshirani theory could do this.

First, we illustrate the high degree of overcorrection present in the Bonferroni correction. The 95% confidence threshold of the maximum CV accuracy for our null data achieved for a single value of θ is much less conservative than the 95% confidence threshold suggested by the Bonferroni correction where the number of θ tested, n_θ, was the correction factor. Similarly, if we consider the significance of the maximum CV accuracy achieved by each classification method for the real data, we observe that the Bonferroni correction is extremely overconservative.

This was expected because our theory and empirical results suggest that $\epsilon(\theta_1)$ is not independent of $\epsilon(\theta_2)$ for similar θ_1 and θ_2. Therefore, the n_θ is a vast overestimate of the number of independent statistical comparisons that were performed. If we use the $\alpha_{\text{Uncorrected}} = 0.05$ cutoff of the probability distribution estimated with the null data, we can estimate the number of independent statistical comparisons through the following reasoning:

$$\text{For } P\left[|\epsilon(\hat{\theta})| \geq \xi | \epsilon(\hat{\theta}) \sim \text{empirical PDF}\right] = \alpha_{\text{Uncorrected}},$$

$$\text{Calculate } P\left[|e| \geq \xi | e \sim Binomial(n, 50\%)\right] = \alpha_{\text{Bonf}}$$

$$\frac{\alpha_{\text{Uncorrected}}}{n_{\text{Indep Tests}}} = \alpha_{\text{Bonf}} \Rightarrow n_{\text{Indep Tests}} = \frac{\alpha_{\text{Uncorrected}}}{\alpha_{\text{Bonf}}}.$$

In lay language, this states that we find the threshold, ξ, that results in a false positive error of $\alpha_{\text{Uncorrected}}$ based on our empirical probability distribution function for $\epsilon(\hat{\theta})$. The probability that a binomial random variable, e, with n trials and a success probability of 50% achieves a value as extreme or more extreme than ξ is an estimate of α_{Bonf}. We can rearrange the original Bonferroni formula to solve for the number of independent tests, $n_{\text{Indep Tests}}$.

The theory of Tibshirani & Tibshirani states that we can use the average prediction error on the *training* set to approximate the bias we incur by choosing the best cross-validation accuracy (see Formula 1). Their method is a correction to the estimated $\epsilon(\hat{\theta})$ instead of an estimate of significance. We translate our estimate of significance into a correction method through the following reasoning:

$$\text{For } P\left[|e| \geq \epsilon(\hat{\theta})_{\text{Uncorrected}} | X, Y, e \sim \text{empirical PDF}\right] = p_{\text{RFT}},$$
$$\text{Calculate } P\left[|e| \geq \epsilon(\hat{\theta})_{\text{Corrected}} | X, Y, e \sim \mathit{Binomial}(n, 50\%)\right] = p_{\text{RFT}}.$$

In lay language, we estimate the probability, p_{RFT}, that our observed $\epsilon(\hat{\theta})_{\text{Uncorrected}}$, or a value more extreme, was achieved based on our empirical probability distribution function for $\epsilon(\hat{\theta})$. We use the inverse binomial distribution with n samples and a 50% probability of success to estimate what cross-validation error, $\epsilon(\hat{\theta})_{\text{Corrected}}$, that would result in p_{RFT}, if we did not test multiple θ.

In their formulation of this theory, they considered the bias in selecting the single best $\epsilon(\theta)$. We extend this theory, we consider the bias in selecting the best $\epsilon(\theta)$ achieved over a contiguous set of θ. This makes the comparison with random field theory more apt, because it incorporates both the magnitude and extant of the reported $\epsilon(\theta)$.

12.8 Verification of Asymptotic Approach of the Optimal Value of θ

While the previous sections have shown that for a given application, we can correct accurately for reporting the best, or a cluster of the best, $\epsilon(\theta)$, these sections do not verify that our method converges on the true correct value for θ, as sample size increases. To verify this, we create artificial datasets where the true optimal value for θ is known.

We construct these artificial datasets with a known solution similar to our null simulations studying the number of features selected in a t-statistic filtering prior to application of an ℓ_2 regularized logistic regression model. Instead of each input dimension holding no information, we randomly add independent Gaussian signal to 75 of the 150 input features with a univariate signal to noise ratio of 0.1. This ensures that a large amount of data is necessary to accurately model each of these informative dimensions is large, so that we can observe how our method converges on the correct $\hat{\theta} = 75$ as $n \to \infty$.

12.9 Discussion

The selection of hyperparameters, θ, can have either a minor or huge impact on the cross-validation error, $\epsilon(\theta)$, that is reported. If we assess $\epsilon(\theta)$ for multiple θ, then report the optimal $\epsilon(\hat{\theta})$, our cross-validation accuracies and their corresponding estimates of significance will be inflated. Using the ideas behind random field theory, we can estimate accurately both the significance of $\epsilon(\hat{\theta})$ and the sensitivity of the solution to that particular choice of θ.

The net result of applying this theory is two fold: (1) better cross-validation performance can be achieved through improved sampling of θ and mildly increasing the size of the training set and (2) our understanding of the sensitivity of the solution with respect to changes in θ improves substantially. The benefit of improved performance is clear, because the applicability of our methods critically depends on how well our models perform on out-of-sample data. By correcting for multiple testing, we maintain the out-of-sample nature of cross-validation exemplars. Additionally, instead of needing to choose the default θ for a given application, we have established a valid and conservative method for improving the fit of our models by determining the optimal θ for our specific dataset.

Prior to explicitly measuring the sensitivity to θ, there was little rigorous statistical understanding for the impact of hyperparameters, which made it difficult to compare the optimal value of θ across applications. If a significant cluster of $\epsilon(\theta)$ is identified in one application, this could help limit the search space of θ for other similar applications, thereby reducing the number of computationally expensive permutation tests.

Additionally, our theory provides a method for comparing the value of $\hat{\theta}$ for different applications. For example, the optimal number of metabolic regions needed to differentiate of non-epileptic seizures (NES) from left temporal lobe epilepsy (LTLE) was high than the number needed to differentiate NES from right temporal lobe epilepsy (RTLE). As discussed in our previous work, this has a clinically and biologically salient interpretation: the metabolic changes associated with LTLE involve a larger network of regions than in RTLE. This suggests that RTLE is a more focal disease, and could be more amenable to surgical resection. Alternatively, even if there are not different rates of post-operative seizure-freedom, these results suggest that LTLE could be associated with more dysfunction outside the epileptogenic region. This suggests that resection of the epileptogenic region, but not associated regions, could provide seizure control and improvements in function in the associated regions. Without a statistically rigorous theory for comparing these two values, these observations were speculative. With our increased understanding, we can state confidently that the difference in the number of regions necessary to differentiate LTLE from NES and RTLE from NES was different statistically with an appropriate amount of confidence.

We clearly illustrated the major limitations in the Tibshirani method and in Bonferroni correction. Both of these methods ignore the inherent dependence of ϵ on θ. By explicitly understanding this dependence, we were able to make substantial improvements in these methods.

The substantial difference between the structure of the random field between the null simulations and the permutation tests showed that this structure may depend on X. Therefore, null simulations prior to measuring $\epsilon(\theta)$ can assist in planning the sampling of θ and to give researchers a general idea of the structure of the field. However, permutation tests may be necessary to estimate the significance of the observed results accurately.

12.9.1 Recommendations for Determining Significance of Observed Maximum Cross-Validation Accuracy

We showed here that our perspective of using random field theory to correct for sampling θ outside the context of nested cross-validation does not lead to overinflated p-values. If a researcher as already determined that permutation tests are necessary to estimate empirical probability distributions of chosen summary parameters, then no additional processing has to occur to apply our methods. Therefore, it is not necessary to perform both permutation tests and nested cross-validation. Instead, as we showed here, we recommend pre-determining a regular grid sampling across the range of feasible values for θ, then performing permutation tests to correct for the multiple testing that occurred by assessing the cross-validation performance at multiple θ.

If permutation tests are not necessary inherently to the chosen approach of the researcher, then we recognize that the computational cost of permutations frequently is much higher than nested cross-validation. Therefore, if the goal of the classification is solely to produce a single, highly accurate predictive tool, then our method is unnecessary. If, however, one seeks to develop an accurate and interpretable model at all stages, then we recommend sampling the cross-validation performance for a selection of potential θ and correcting for multiple testing using our method. Even if the primary interest of the work is not to learn about θ, the interpretation of the parameters inherent to the model, ψ, are inexorably linked to θ. Therefore, improving our understanding of θ also improves our understanding of ψ.

12.9.2 Planning Sampling Schemes for Hyperparameters

In addition to our primary goal, performing null simulations on a proposed set of θ prior to analysis of real data also allows researchers to make more principled choices. To define notation, we designate the set of θ that one wants to assess the performance of as $\{\theta\}$. In the above work, we assumed that researchers had predetermined $\{\theta\}$. However, in addition to determining significance, null simulations can be used to determine $\{\theta\}$. For the assumptions of stationarity of the mean and correlation structure to hold, one must determine a sampling scheme for θ where the $\{\theta\}$ are evenly spaced. By performing a small set of null simulations for a proposed $\{\theta\}$, these assumptions can and should be tested directly.

In addition to checking stationarity, prior null simulations can determine the required density of $\{\theta\}$ over the search space, and the minimum performance necessary to achieve a significant result. As the above examples showed, testing multiple values of the regularization parameter, λ, did not result in substantial changes to the performance. Therefore, only a couple of representative values needed to be tested to characterize the performance over a wide range of λ. The number of features selected using t-statistic filtering, however, resulted in major changes to the performance. Especially when the dimensionality of X was 1,000, it was not necessary to test every possible value of F. Our results suggest that if one tested every 100 F, then adjacent samples would be less than 50% correlated. Further, if every possible value for F was chosen, then the observed performance at one specific value would need to be greater than 90% to be significant, statistically.

If we believed, due to issues in data quality or consistency, that the maximum achievable performance was less than 90%, then we should limit the size of $\{\theta\}$ so that the magnitude of the potential bias from choosing the best performance for a specific $\hat{\theta}$ is reduced. Therefore, prior null simulations can help researchers effectively plan their experiments based on the performance they seek to achieve, and the known correlation structure of performance on θ.

Our method also can be extended trivially to calculate the statistical power of a given $\{\theta\}$. Just as the null simulations estimate an empirical probability distribution for performance, one could use simulations to estimate the probability distribution for performance if a difference of a given magnitude did indeed exist. Due to the major differences between the null simulations and the permutation tests, we expect that these power calculations would be approximate, as are most power calculations.

12.9.3 Limitations and Future Directions

Our theory relies on the strong assumption and verification that cross-validation performance is a stationary random field across the range of θ. As we did here, it is critical to check this assumption prior to the application of our method. In that way, our perspective of random field theory is not a cure all. If these strong assumptions are met, then the validation data can be used both to assess out-of-sample performance and optimize θ. If not, then over-sampling θ and reporting the best performance remains a dubious practice, at best.

An important limitation to our method, as currently framed, is that $\{\theta\}$ must be defined independent of the observed performance. If $\{\theta\}$ is selected adaptively (i.e. Newton's method), then simply testing the same $\{\theta\}$ would result in an over-estimate of significance. The effective search area of adaptive methods is much larger than the actual search area because low-performance regions are left untested. Therefore, even though it is tempting, researchers should not test many, seemingly random values for θ until a desired result is achieved. Even if they use our method to correct for the $\{\theta\}$ they tested, our estimate of significance would be inflated. Further work must be done to estimate the significance of results when adaptive sampling schemes for θ are implemented.

12.10 Conclusion

Our method of applying the ideas of random field theory to determining the significance of observed cross-validation accuracies for a range of hyperparameters allows for improved reported accuracies and understanding of the sensitivity of these accuracies to changes in hyperparameters. This method vastly out-performed the two alternate methods of correction.

CHAPTER 13

Discussion & Conclusions

In the above chapters, we described the initial development of CADTs designed to be applicable to clinical care of seizure disorder, as well as an important section of novel statistical methods to train those algorithms better. These manuscripts demonstrate the wide potential and success of these methods, given just a few years of development.

In part 2 of this work, we addressed a basic challenge to training and optimizing machine learning models. Even though the methods for learning parameters inherent to these models are efficient and effective, the effect of hyperparameters previously were not rigorously studied or understood. Our methods both provide an effective, if not efficient, method for optimizing hyperparameters as well as understanding the sensitivity of the observed solution to the choice of hyperparameters. Previous methods to address this basic issue in machine learning did not have as strong backing in statistical rigor [367, 368, 366]. While we did not find succinct analytical expressions that could be predicted a priori based on the structure of the log-likelihood functions, future work could help provide a mechanistic and statistically rigorous explanation for the structure we observed with our empirical methods. This could reduce the need for computationally expensive empirical simulations. These simulations are the primary limitation to the application of our method.

We illustrated our theory for a regularly sampled, predetermined range of hyperparameters. Of course, this is not the only method for sampling and optimizing a hyperparameter of interest. Other adaptive sampling schemes, like Newtons method described in Didactic Background, are governed by the same theory of a random field. The methods for determining significance under such a sampling scheme are undetermined and need more study. Our methods and results provide a necessary base from which these follow up studies can be conducted.

The motivation for the development of this novel statistical perspective was difficulties that were encountered when developing the CADTs. When seeking to train effective CADTs, we were unsatisfied with the lack of statistical rigor in the previous practices in selecting hyperparameters a priori and the inability of previous methods to incorporate the spatial dependence of performance on the hyperparameter. While we do not claim that our method obviates the challenge we encountered, it does begin to provide an approach to address it.

Prior to Kloppels landmark papers in the analysis of MRIs from patients with Alzheimers disease with a SVM in 2008 [168, 369], machine-learning tools had not been applied extensively to high dimensional clinical data with the goal of assisting in the diagnosis of patients. In subsequent years, the wide applicability of this perspective has been demonstrated for many diseases including dementia, neuropsychiatric disorders and radiology [62]. In particular, recent work demonstrated how a machine-learning tool to detect cancerous lung nodules was as accurate as expert radiologists, and had a slightly lower false positive rate [156]. Despite the numerous academic manuscripts describing the success of machine learning, there are few FDA-approved CADTs because there are a number of practical and theoretical barriers to implementation [62].

In this work, we demonstrated that this machine-learning perspective could assist at multiple key decision points in the diagnosis of patients with seizure disorder. Prior to this work, the epilepsy literature focused primarily on utilizing conventional statistics to study the difference between populations of patients with and without particular seizure subtypes. The plethora of papers describing the difference between healthy controls and patients with NES or ES using clinical assessments and neuroimaging data provide evidence for this focus on population statistics [18]. This helps understand the population of interest, but does not provide as much evidence about the individuals that comprise that population. Concurrent with our work, there has begun to be a slow shift towards personalized predictive algorithms that seek to use these differences in populations to diagnose individual patients.

These personalized predictive algorithms can be applied to many clinically relevant challenges other than diagnostics. In epilepsy and other disorders, the same statistical methods could be used to predict which patients may respond more favorably to particular medical or surgical treatments.

In particular, reliably identifying patients that will become seizure free with minimal function loss after resective surgery for epilepsy would have a high clinical impact. For patients with mesial temporal lobe epilepsy, two thirds of patients have a favorable outcome after surgery. The rate of success in extratemporal epilepsy is lower. If we could identify patients who would not benefit from surgery before we operate, that would reduce morbidity, mortality through SUDEP, and save appreciable cost in treating the epilepsy. The challenge in predicting post-operative outcome is that relatively few patients undergo surgery, and maintaining contact with patients to follow their progress 5 or 10 years after surgery is time-intensive and expensive.

Another high impact application of these types of statistical methods is in predicting which patients may respond more favorably to which anti-seizure medications. One particularly salient treatment choice occurs when determining if a patient with infantile spasms should be treated with vigabatrin or high dose adrenocorticotropic hormone (ACTH). The cost of ACTH is high in both absolute dollars and side effects, but without effective treatment these patients will have devestating neural damage. However, not all patients respond to ACTH.

If we could identify which patients could be treated effectively with vigabatrin or ACTH prior to treatment, then we could save substantial time and money. The challenge to developing these treatment predictive tools is that there is not a rich literature describing biomarkers for which patients will respond. Therefore, we must generate, evaluate and validate any potential predictive biomarker using high quality data.

However, in order for these personalized predictive methods to be applied directly in the epilepsy clinic, a number of practical and theoretical barriers must be addressed. We highlighted the importance of choosing the appropriate control group to reflect the clinical question the CADT seeks to address [81]. Unfortunately, previous methods focus on differentiating patients with ES or NES from seizure-naive controls, which does not reflect the clinical challenge [72, 26, 27, 73, 28, 74]. Previous work also focused on collecting research quality measures on recruited patient populations to validate their methods. We extended this work by operating on clinical quality data from an unselected patient population [37]. This extension addresses how well these CADTs could perform when applied to real world data, as clinicians see them. Our use of archived records from an EHR highlights that the methods we developed here could be applied to other similar datasets from other tertiary care centers for seizure disorder. In addition to increasing sample size, and thereby performance (see bias-variance trade-off in Didactic Background Material) the application of these methods to data from other centers would allow us to determine how consistent the diagnostic trends are across practice locations. This would help us know if separate CADTs must be trained at each location, or if a single, generalizable CADT could be applied nationally or internationally. Despite these novel and important extensions of previous work, there are a number of necessary next steps that must be taken prior to applying CADTs directly within clinical care.

The most salient next step is a prospective assessment of our CADTs, which were all trained based on retrospective patient samples. It is incorrect statistically to state that cross-validation artificially inflates performance and, consequentially, the significance of results. Prospective assessment tests how well the trends seen in retrospective data translate to new patients. Especially with psychosomatic disorders like psychogenic NES, the character of the seizures may change over time depending on many factors including but not limited to cultural beliefs and patient education initiatives [370, 371, 372]. For all diseases, the pattern of comorbidities and historical factors may change over time due to more effective acute and chronic treatment of these events. In particular, we hope the current focus on the treatment and prevention of head injuries will change how many patients develop either epileptic or non-epileptic seizure disorder after these events [373, 374, 375]. Additionally, by identifying the common comorbidities in psychogenic NES, we can identify where prevention and education resources can be targeted to reduce the number of patients that experience their first psychogenic seizure or the burden of psychogenic seizures in general. Lastly and most importantly, when CADTs are utilized in clinic, the patient serves as a prospective application of the method. Therefore, prospective validation of CADTs is critically needed to determine how they will perform in real world clinical application.

Additionally, we note that substantial further development is needed to integrate multiple modalities into a single CADT. The majority of the work above addressed the initial development of single modality CADTs that are pre-requisites for the multimodal work. At each stage of the diagnostic process, newly acquired information must be used to update the prevailing assessment of the patient. By developing multimodal, update-able algorithms, future work could hope to mirror the diagnostic process. However, it is important to stress that these tools do not aim to replace clinical reasoning. Instead, they aim to provide information to the clinician that would not otherwise be appreciated.

One important question about all of these CADTs is how will they be integrated into the clinical process, once fully validated? To assess this, we need to observe directly how clinicians utilize the information as we provide it to them. This means conducting randomized validation studies where clinicians are provided information from CADTs, or they conduct the standard of care. This allows for direct comparison of the diagnostic performance and follow up outcomes in patients treated with and without the assistance of CADTs. In this way, we could measure if CADTs have an impact on the public health problems that we described in the introduction. This level of validation is outside the scope of this graduate thesis. However, the work described here provides a strong foundation for future CADTs to address those dismal statistics we outlined in the introduction.

CHAPTER 14

Conflicts of Interests

The primary author and all collaborators declare no conflicts of interest in this work. Dr. Silverman was a co-inventor of the NeuroQ software that was utilized to extract features from the FDG-PET data. This software is now licensed to UCLA. Drs. Stern, Engel, Silverman and Salamon were involved in the clinical care of the patients at UCLA, but no clinical decisions or changes in clinical practice were made based on this work or because of their knowledge that this work was being done.

CHAPTER 15

Abbreviations, Notation & Glossary

15.1 Abbreviations

- 2D: Two dimensional
- 3D: Three dimensional
- AD: Alzheimers disease
- AED: Anti-epileptic medication
- ACTH: Adrenocorticotropic hormone
- ADNI: Alzheimer's disease neuroimaging intiative
- ADT: Altenating decision tree
- ANOVA: Analysis of variance
- ARRA: American Recovery and Reinvestment Act of 2009
- ASM: Anti-seizure medication
- AUC: Area under the ROC
- AVP: atrial ventricular premature contraction
- BLR: Bayesian Logistic Regression
- BTLE: Bilateral temporal lobe epilepsy

- CAD: Computer aided diagnostic
- CADT: Computer-aided diagnostic tool
- Caltech: California institute of technology
- CD: Conditional dependence
- CI: Confidence interval
- CL10CV: Cyclical leave-one-out cross validation
- CoD: Curse of dimensionality
- CT: Computed tomography
- DGSOM: David Geffen school of medicine
- DD-FS: Data-driven feature selection
- DDx: Differential diagnosis
- DTI: Diffusion tensor imaging
- DTR: Deep tendon reflexes
- DOPA: Dihydroxyphenylanine
- EDE: European database on epilepsy
- EEG: Electroencephalography
- EHR: Electronic health record
- EKG: Electrocardiogram (acronym from the German)
- EM: Expectation maximization
- ERSD: Event-related spectral perturbation

- ES: Epileptic seizure
- FBP: Filtered back projection
- FDG-PET: flourodeoxyglucose positron emission tomography
- FLE: Frontal lobe epilepsy
- FS: Feature selection
- FTLD: Fronto-temporal lobar dementia
- GERD: Gasto-esophageal reflux disorder
- GPRD: General practice reserach database
- GUI: Graphical user interface
- IC: Independent component
- ICA: Independent component analysis
- IRB: Institutional Review Board
- L_1-LR: L_1 regularized logistic regression
- LBBB: left bundle branch block
- LONI: Laboratory of Neuroimaging
- LOOCV: Leave-one-out cross validation
- LTLE: left temporal lobe epilepsy
- LDA: Fisher linear discriminant analysis
- LLE: Local linear embedding
- MA: Manual analysis

- MCI: Mild cognitive impairment
- MCMC: Markov-chain Monte Carlo
- MCAR: Missing completely at random
- MDS: Multidimensional scaling
- MEG: Magnetoencephalography
- MI: Mutual information or myocardial infarction
- ML: Machine learning
- MLR: Multivariate logistic regression
- MLP: Multilayer perceptron
- MSE: The variance of the data explained by the model.
- MSTP: Medical scientist training program
- MRI: Magnetic Resonance Imaging
- mRMR: Minimum redundancy, maximum relevancy
- NES: Non-epileptic seizure
- NF1: Neurofibromatosis type 1
- NNMF: Non-negative matrix factorization
- NNP: Neurologicall normal patients
- NOS: Not otherwise specified
- NIH: National Institute of Health
- OLE: Occipital lobe epilepsy

- OSEM: Ordered subset expectation maximization
- PC: Principal component
- PCA: Principal component analysis
- PDF: Probability distribution function
- iPET: interictal PET
- PET: Positron emission tomography
- PGP: Personal genome project
- PI: Post-ictal
- PLE: Parietal lobe epilepsy
- PNES: Psychogenic non-epileptic seizure
- PVC: Ventricular premature contraction
- PWE: Patients with epilepsy
- PWN: Patients with non-epileptic seizures
- QDA: Fisher quadratic discriminant analysis
- RBBB: right bundle branch block
- RFI: Request for information
- RFE: Recursive feature seleimnation
- ROC: Receiver operating curve
- ROI: Region of interest
- ROS: Review of systems

- sMRI: Structural magnetic resonance imaging
- SD: Standard deviation
- SE: Standard error
- SIBTP: Systems and integrative biology training program
- SNR: Signal to noise ratio
- SOZ: Seizure-onset zone
- SM: Sensorimotor cortex
- SPECT: Single-photon emission computed tomography
- SVM: Support vector machine
- SSE: Sum of the squared error of a model
- TBI: Traumatic brain injury
- TIA: Temporary ischemic attack
- TLE: Temporal lobe epilepsy
- TV: Total variation
- UCLA: University of California, Los Angeles
- US: United States
- VC: Vector concatenation
- vEEG: video-electroencephalography

15.2 Variable & Notation Definition

- $[\cdot,\cdot]$ and $(\cdot,\cdot)$: This is interval notation where square brackets indicate that the interval includes the boundary ($\geq$ or $\leq$), whereas soft brackets indicate exclusion of the boundary ($>$ or $<$).
- $\hat{\cdot}$: Any variable with a hat over it is estimated from the data.
- $\|\cdot\|_r$: The r norm of a variable, for $r \in \mathbb{R}_{\geq 0}$.
- $\bigtriangledown$: The gradient or first derivative operator
- $\sim$: The variable preceeding this symbol is distributed according to the distribution described after this symbol.
- $< \cdot, \cdot > = \cdot^T \cdot$: The inner product of two vectors or matrices
- $\left.\frac{\partial^k f(\beta)}{\partial \beta^k}\right|_{\beta=\beta_0}$: the k^{th} partial derivative of the generic function, f, with respect to β calculated at β_0
- α: The false positive rate OR, in the context of an SVM, the Lagrange vector defining the support vectors
- β: linear weights of the input data to predict the independent predicted variable OR, equivalently, the vector with which multidimensional data is projected onto to make a classification
- χ^2_ν: A chi-squared distributed variable with ν degrees of freedom, defined as the sum of the square of ν standard Gaussian random variables.
- D: a generic diagonal matrix
- ϵ: binary or continuous error of the predicted model
- $\in$: The variable preceding this symbol is within the set listed after this symbol.

- i: subject or datapoint-specific index
- j: the covariate or dimensionality-specific index
- $E(\cdot)$: the expectation (or average) of a variable
- F_{ν_1,ν_2}: An F-statistic with ν_1 and ν_2 degrees of freedom, defined by the ratio of two χ^2 variables.
- $g(\cdot)$ or $f(\cdot)$: a generic function of the variable
- $G(j,k)$: the Gini impurity for a given covariate j and threshold k (see Machine Learning Classifiers).
- $H(\cdot) = \bigtriangledown^2\cdot$: The Hessian or second derivative of a vector valued variable
- ℓ: log-likelihood function of the data given the model
- L: likelihood function of the data given the model
- L_r: the distance between two points, as defined by the r-Minkowski metric where $r \in \mathbb{R}_{\geq 0}$.
- $logit(\cdot) = \frac{\log\cdot}{\log(1-\cdot)}$: the logit transform function
- $MI(\cdot,\cdot)$: The mutual information between two variables
- m: the dimensionality of X prior to any feature selection (width of X)
- n: the total number of exemplars in a dataset (height of X or length of Y)
- π: The transcendental number that begins with 3.14 or the estimate of the probability of logistic data
- ψ: The set of parameters that are estimated by optimizing a log-likelihood or objective function
- $P(a|u)$: The probability of a, given or conditional on u.

- $\mathbb{R}_{\geq 0}$: The set of all non-negative real numbers.

- σ: standard deviation of error of a predictive model

- $tr(\cdot)$: the trace of a matrix

- θ: The set of hyperparameters that define the structure of the model, but are not optimized jointly with ψ.

- w: A separating hyperplane that is perpendicular to β.

- Variance

  - $Var(\cdot) = \sigma^2 I_n$: The variance matrix of a vector.

  - $Var_S(\cdot) = \sigma^2$: The scalar variance of a variable or vector. If $\cdot$ is a vector, then each element of the vector is considered an independent sample that can be used to estimate the sample variance.

- X: matrix of input data

  - X_i: vector of input data from the i^{th} datapoint

  - X_j: vector of the j^{th} covariate of data from all exemplars

- Y: vector of independent predicted or output variable

  - y_i: value of independent predicted or output variable from the i^{th} datapoint. Note that computer scientists and those that use SVM think of $y_i \in \{\pm 1\}$ whereas statisticians tend to think of $y_i \in \{0, 1\}$. Depending on the context, we switch between these definitions to maintain consistency with the prevailing literature.

15.3 Glossary of Terminology

- Exemplar: Individual point of data, with an associated outcome and input data
- Expectation: Average or mean of a variable
- Free parameter: see Parameter
- Gaussian Distribution: The standard Gaussian or normal distribution: $P(Y = y|E(Y) = \mu, Var(Y) = \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$
- Parameter: a variable that is estimated through a statistical optimization process
- Hyperparameter: a variable that is critical to a model, but is not jointly optimized with parameters
- Ictal: Another word for seizure.
  - Interictal: Between seizures
  - Peri-ictal: Around seizures
  - Pre-ictal: Before seizures
  - Post-ictal: After seizures
- Likelihood function: an explicit mathematical expression of the likelihood of the data, given the model
- Manifold: A non-linear lower dimensional surface that exists within a higher dimensional space
- Normal distribution: see Gaussian distribution
- Objective function: an explicit mathematical formulation that is minimized and/or maximized to estimate free parameters of the mode

- Loss or penalty function: The function that defines how well your model fits the training datal

- Overfitting: Using non-generalizable trends in the training data to improve the log-likelihood or objective function

- Sensitivity: the accuracy of the predicted classification on the diseased patients or the $y_i = 1$ patients

- Space: the concept of visualizing data as points or vectors with respect to dimensional axes. Dimensional axes can be thought of as a reference frame for how to compare data.

  - Observational space: dimensional axes defined by the raw measurements made

  - Transformed space: dimensional axes are defined by the transformed measurements

- Specificity: the accuracy of the predicted classification on the not diseased patients or the $y_i = 0$ patients

- Standard deviation: the square root of variance

- Standard error: an estimation of the unbiased population standard deviation around an estimate of the mean, defined by $SE(X) = \sqrt{\frac{Var(X)}{n}}$.

- Variance: A quantification of the spread of a variable, X, around its mean, $E(X)$ defined by $Var(X) = E(XX^T) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - E(X))^2$

- Supervised algorithm: An algorithm that uses the class information, Y, from the training set to inform the solution.

- Unsupervised algorithm: An algorithm that does *not* use class information, Y, from the training set to find a solution.

- Semi-supervised algorithm: An algorithm that includes input data both with known and unknown class information to find a solution.

- Voxel: Volumetric pixel

References

[1] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–6, 2000.

[2] United States Department of Health Prevention and Human Services. Centers for Disease Control and. National ambulatory medical care survey., 2009.

[3] Carlton Chu, Ai-Ling Hsu, Kun-Hsien Chou, Peter Bandettini, and Ching-Po Lin. Does feature selection improve classification accuracy? impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage*, 2011.

[4] H. Akima. Algorithm 761: scattered-data surface fitting that has the accuracy of a cubic polynomial. *ACM Transactions on Mathematical Software*, 22:362–371, 1996.

[5] Peng H.C. Ding, C. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*, 3(2):185–205, 2005.

[6] Long F. Ding C. Peng, H. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[7] W. T. Kerr, S. T. Nguyen, A. Y. Cho, E. P. Lau, D. H. Silverman, P. K. Douglas, N. M. Reddy, A. Anderson, J. Bramen, N. Salamon, J. M. Stern, and M. S. Cohen. Computer-aided diagnosis and localization of lateralized temporal lobe epilepsy using interictal fdg-pet. *Front Neurol*, 4:31, 2013.

[8] W. T. Kerr, A. Anderson, E. P. Lau, A. Y. Cho, H. Xia, J. Bramen, P. K. Douglas, E. S. Braun, J. M. Stern, and M. S. Cohen. Automated diagnosis of epilepsy using eeg power spectrum. *Epilepsia*, 53(11):e189–92, 2012.

[9] M. Reuber and C. E. Elger. Psychogenic nonepileptic seizures: review and update. *Epilepsy Behav*, 4(3):205–16, 2003.

[10] A. T. Berg, S. F. Berkovic, M. J. Brodie, J. Buchhalter, J. H. Cross, W. van Emde Boas, J. Engel, J. French, T. A. Glauser, G. W. Mathern, S. L. Moshe, D. Nordli, P. Plouin, and I. E. Scheffer. Revised terminology and concepts for organization of seizures and epilepsies: report of the ilae commission on classification and terminology, 2005-2009. *Epilepsia*, 51(4):676–85, 2010.

[11] . Center for Disease Control. Epilepsy, 2011.

[12] S. Arroyo, M. J. Brodie, G. Avanzini, C. Baumgartner, C. Chiron, O. Dulac, J. A. French, and J. M. Serratosa. Is refractory epilepsy preventable? *Epilepsia*, 43(4):437–44, 2002.

[13] P. R. Camfield and C. S. Camfield. Antiepileptic drug therapy: when is epilepsy truly intractable? *Epilepsia*, 37 Suppl 1:S60–5, 1996.

[14] J. A. French. Refractory epilepsy: clinical overview. *Epilepsia*, 48 Suppl 1:3–7, 2007.

[15] P. Kwan and M. J. Brodie. Early identification of refractory epilepsy. *N Engl J Med*, 342(5):314–9, 2000.

[16] F. Semah, M. C. Picot, C. Adam, D. Broglin, A. Arzimanoglou, B. Bazin, D. Cavalcanti, and M. Baulac. Is the underlying cause of epilepsy a major prognostic factor for recurrence? *Neurology*, 51(5):1256–62, 1998.

[17] M. J. Jackson. Concise guidance: diagnosis and management of the epilepsies in adults. *Clin Med*, 14(4):422–7, 2014.

[18] P. Dickinson and K. J. Looper. Psychogenic nonepileptic seizures: a current overview. *Epilepsia*, 53(10):1679–89, 2012.

[19] S. R. Benbadis, E. O'Neill, W. O. Tatum, and L. Heriaud. Outcome of prolonged video-eeg monitoring at a typical referral epilepsy center. *Epilepsia*, 45(9):1150–3, 2004.

[20] D. F. Ghougassian, W. d'Souza, M. J. Cook, and T. J. O'Brien. Evaluating the utility of inpatient video-eeg monitoring. *Epilepsia*, 45(8):928–32, 2004.

[21] C. M. Michel, G. Lantz, L. Spinelli, R. G. De Peralta, T. Landis, and M. Seeck. 128-channel eeg source imaging in epilepsy: clinical yield and localization precision. *J Clin Neurophysiol*, 21(2):71–83, 2004.

[22] K.K. Lee and N. Salamon. [18f] fluorodeoxyglucosepositron-emission tomography and mr imaging coregistration for presurgical evaluation of medically refractory epilepsy. *AJNR Am J Neuroradiol*, 30:1811–1816, 2009.

[23] N. Salamon, J. Kung, S. J. Shaw, J. Koo, S. Koh, J. Y. Wu, J. T. Lerner, R. Sankar, W. D. Shields, Jr. Engel, J., I. Fried, H. Miyata, W. H. Yong, H. V. Vinters, and G. W. Mathern. Fdg-pet/mri coregistration improves detection of cortical dysplasia in patients with epilepsy. *Neurology*, 71(20):1594–601, 2008.

[24] C. Tonini, E. Beghi, A. T. Berg, G. Bogliun, L. Giordano, R. W. Newton, A. Tetto, E. Vitelli, D. Vitezic, and S. Wiebe. Predictors of epilepsy surgery outcome: a meta-analysis. *Epilepsy Res*, 62(1):75–87, 2004.

[25] S. Wiebe, W. T. Blume, J. P. Girvin, and M. Eliasziw. A randomized, controlled trial of surgery for temporal-lobe epilepsy. *N Engl J Med*, 345(5):311–8, 2001.

[26] N. Farid, H. M. Girard, N. Kemmotsu, M. E. Smith, S. W. Magda, W. Y. Lim, R. R. Lee, and C. R. McDonald. Temporal lobe epilepsy: quantitative mr volumetry in detection of hippocampal atrophy. *Radiology*, 264(2):542–550, 2012.

[27] N. K. Focke, M. Yogarajah, M. R. Symms, O. Gruber, W. Paulus, and J. S. Duncan. Automated mr image classification in temporal lobe epilepsy. *Neuroimage*, 59(1):356–62, 2012.

[28] S. Keihaninejad, R. A. Heckemann, I. S. Gousias, J. V. Hajnal, J. S. Duncan, P. Aljabar, D. Rueckert, and A. Hammers. Classification and lateralization of temporal lobe epilepsies with and without hippocampal atrophy based on whole-brain automatic mri segmentation. *PLoS One*, 7(4):e33096, 2012.

[29] S. Corkin, D. G. Amaral, R. G. Gonzalez, K. A. Johnson, and B. T. Hyman. H. m.'s medial temporal lobe lesion: findings from magnetic resonance imaging. *J Neurosci*, 17(10):3964–79, 1997.

[30] R. Mayeux, J. Brandt, J. Rosen, and D. F. Benson. Interictal memory and language impairment in temporal lobe epilepsy. *Neurology*, 30(2):120–5, 1980.

[31] W. B. Scoville and B. Milner. Loss of recent memory after bilateral hippocampal lesions. *J Neurol Neurosurg Psychiatry*, 20(1):11–21, 1957.

[32] Jr. Engel, J., M. P. McDermott, S. Wiebe, J. T. Langfitt, J. M. Stern, S. Dewar, M. R. Sperling, I. Gardiner, G. Erba, I. Fried, M. Jacobs, H. V. Vinters, S. Mintzer, and K. Kieburtz. Early surgical therapy for drug-resistant temporal lobe epilepsy: a randomized trial. *JAMA*, 307(9):922–30, 2012.

[33] G. W. Mathern. Challenges in the surgical treatment of epilepsy patients with cortical dysplasia. *Epilepsia*, 50 Suppl 9:45–50, 2009.

[34] G. W. Mathern, C. C. Giza, S. Yudovin, H. V. Vinters, W. J. Peacock, D. A. Shewmon, and W. D. Shields. Postoperative seizure control and antiepileptic drug use in pediatric epilepsy surgery patients: the ucla experience, 1986-1997. *Epilepsia*, 40(12):1740–9, 1999.

[35] D. J. Dlugos. The early identification of candidates for epilepsy surgery. *Arch Neurol*, 58(10):1543–6, 2001.

[36] Jr. LaFrance, W. C., G. A. Baker, R. Duncan, L. H. Goldstein, and M. Reuber. Minimum requirements for the diagnosis of psychogenic nonepileptic seizures: a staged approach: a report from the international league against epilepsy nonepileptic seizures task force. *Epilepsia*, 54(11):2005–18, 2013.

[37] W.T. Kerr, C.T. Braesch, E.J. Janio, J.M. Le, J.M. Hori, A.B. Patel, N.L. Gallardo, J. Bauirjan, A.M. Chau, S.E. Barritt, E. S. Hwang, E.C. Davis, A.Y. Cho, J. Gordon, D. Torres-Barba, J. Jr. Engel, M.S. Cohen, and J.M. Stern. Accurate differentiation of epileptic and non-epileptic seizures through quantitative combination of findings in the clinical history. *Epilepsy & Behavior*, page [submitted], 2015.

[38] Jr. LaFrance, W. C. and O. Devinsky. The treatment of nonepileptic seizures: historical perspectives and future directions. *Epilepsia*, 45 Suppl 2:15–21, 2004.

[39] B. F. Shneker and J. O. Elliott. Primary care and emergency physician attitudes and beliefs related to patients with psychogenic nonepileptic spells. *Epilepsy Behav*, 13(1):243–7, 2008.

[40] T. S. Walczak, S. Papacostas, D. T. Williams, M. L. Scheuer, N. Lebowitz, and A. Notarfrancesco. Outcome after diagnosis of psychogenic nonepileptic seizures. *Epilepsia*, 36(11):1131–7, 1995.

[41] N. M. Bodde, J. L. Brooks, G. A. Baker, P. A. Boon, J. G. Hendriksen, and A. P. Aldenkamp. Psychogenic non-epileptic seizures–diagnostic issues: a critical review. *Clin Neurol Neurosurg*, 111(1):1–9, 2009.

[42] M. Reuber, G. Fernandez, J. Bauer, C. Helmstaedter, and C. E. Elger. Diagnostic delay in psychogenic nonepileptic seizures. *Neurology*, 58(3):493–5, 2002.

[43] R. C. Martin, F. G. Gilliam, M. Kilgore, E. Faught, and R. Kuzniecky. Improved health care resource utilization following video-eeg-confirmed diagnosis of nonepileptic psychogenic seizures. *Seizure*, 7(5):385–90, 1998.

[44] Jr. Engel, J. Surgery for seizures. *N Engl J Med*, 334(10):647–52, 1996.

[45] Jr. King, J. T., M. R. Sperling, A. C. Justice, and M. J. O'Connor. A cost-effectiveness analysis of anterior temporal lobectomy for intractable temporal lobe epilepsy. *J Neurosurg*, 87(1):20–8, 1997.

[46] J. T. Langfitt, R. G. Holloway, M. P. McDermott, S. Messing, K. Sarosky, A. T. Berg, S. S. Spencer, B. G. Vickrey, M. R. Sperling, C. W. Bazil, and S. Shinnar. Health care costs decline after successful epilepsy surgery. *Neurology*, 68(16):1290–8, 2007.

[47] C. E. Begley, M. Famulari, J. F. Annegers, D. R. Lairson, T. F. Reynolds, S. Coan, S. Dubinsky, M. E. Newmark, C. Leibson, E. L. So, and W. A. Rocca. The cost of epilepsy in the united states: an estimate from population-based clinical and survey data. *Epilepsia*, 41(3):342–51, 2000.

[48] J. A. Cramer, Z. J. Wang, E. Chang, A. Powers, R. Copher, D. Cherepanov, and M. S. Broder. Healthcare utilization and costs in adults with stable and uncontrolled epilepsy. *Epilepsy Behav*, 31:356–62, 2014.

[49] L. Gao, L. Xia, S. Q. Pan, T. Xiong, and S. C. Li. Burden of epilepsy: A prevalence-based cost of illness study of direct, indirect and intangible costs for epilepsy. *Epilepsy Res*, 110:146–56, 2015.

[50] A. N. Wilner, B. K. Sharma, A. Thompson, A. Soucy, and A. Krueger. Diagnoses, procedures, drug utilization, comorbidities, and cost of health care for people with epilepsy in 2012. *Epilepsy Behav*, 41:83–90, 2014.

[51] G. A. Baker, A. Jacoby, D. Buck, C. Stalgis, and D. Monnet. Quality of life of people with epilepsy: a european study. *Epilepsia*, 38(3):353–62, 1997.

[52] L. S. Boylan, L. A. Flint, D. L. Labovitz, S. C. Jackson, K. Starner, and O. Devinsky. Depression but not seizure frequency predicts quality of life in treatment-resistant epilepsy. *Neurology*, 62(2):258–61, 2004.

[53] J. P. Szaflarski, C. Hughes, M. Szaflarski, D. M. Ficker, W. T. Cahill, M. Li, and M. D. Privitera. Quality of life in psychogenic nonepileptic seizures. *Epilepsia*, 44(2):236–42, 2003.

[54] J. P. Szaflarski, M. Szaflarski, C. Hughes, D. M. Ficker, W. T. Cahill, and M. D. Privitera. Psychopathology and quality of life: psychogenic non-epileptic seizures versus epilepsy. *Med Sci Monit*, 9(4):CR113–8, 2003.

[55] J. M. Travaline, R. Ruchinskas, and Jr. D'Alonzo, G. E. Patient-physician communication: why and how. *J Am Osteopath Assoc*, 105(1):13–8, 2005.

[56] F. Brigo. An evidence-based approach to proper diagnostic use of the electroencephalogram for suspected seizures. *Epilepsy Behav*, 21(3):219–22, 2011.

[57] Sethuraman G. Kotagal U. Buncher R. Gilbert, D.L. Meta-analysis of eeg test performance shows wide variation among studies. *Neurology*, 60:564–570, 2003.

[58] T. R. Henry, M. Chupin, S. Lehericy, J.P. Strupp, M.A. Sikora, Z.Y. Sha, K. Ugurbil, and P-F. Van de Moortele. Hippocampal sclerosis in temporal lobe epilepsy: Findings at 7 t. *Radiology*, 261(1):199–209, 2011.

[59] T. R. Henry, T. L. Babb, Jr. Engel, J., J. C. Mazziotta, M. E. Phelps, and P. H. Crandall. Hippocampal neuronal loss and regional hypometabolism in temporal lobe epilepsy. *Ann Neurol*, 36(6):925–7, 1994.

[60] T. R. Henry, H. T. Chugani, B. W. Abou-Khalil, W. H. Theodore, and B. E. Swartz. *Positron emission tomography in presurgical evaluation of epilepsy.* Surgical treatment of the epilepsies. Raven Press, New York, 2nd ed. edition, 1993.

[61] R. S. Fisher, W. van Emde Boas, W. Blume, C. Elger, P. Genton, P. Lee, and Jr. Engel, J. Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ilae) and the international bureau for epilepsy (ibe). *Epilepsia*, 46(4):470–2, 2005.

[62] Shaefer-Prokop C.M. Prokop M. van Ginneken, B. Computer-aided diagnosis: How to move from the laboratory to the clinic. *Radiology*, 261(3):719–732, 2011.

[63] T. U. Syed, Jr. LaFrance, W. C., E. S. Kahriman, S. N. Hasan, V. Rajasekaran, D. Gulati, S. Borad, A. Shahid, G. Fernandez-Baca, N. Garcia, M. Pawlowski, T. Loddenkemper, S. Amina, and M. Z. Koubeissi. Can semiology predict psychogenic nonepileptic seizures? a prospective study. *Ann Neurol*, 69(6):997–1004, 2011.

[64] R. Duncan, C. D. Graham, and M. Oto. Neurologist assessment of reactions to the diagnosis of psychogenic nonepileptic seizures: relationship to short- and long-term outcomes. *Epilepsy Behav*, 41:79–82, 2014.

[65] K. K. McMillan, M. J. Pugh, H. Hamid, M. Salinsky, J. Pugh, P. H. Noel, E. P. Finley, L. K. Leykum, H. J. Lanham, and Jr. LaFrance, W. C. Providers' perspectives on treating psychogenic nonepileptic seizures: frustration and hope. *Epilepsy Behav*, 37:276–81, 2014.

[66] M. Ayala, M. Cabrerizo, P. Jayakar, and M. Adjouadi. Subdural eeg classification into seizure and nonseizure files using neural networks in the gamma frequency band. *J Clin Neurophysiol*, 28:20–29, 2011.

[67] J. Jacobs, K. Kobayashi, and J. Gotman. High-frequency changes during interictal spikes detected by time-frequency analysis. *Clin Neurophysiol*, 122(1):32–42, 2011.

[68] L. Kuhlmann, A. N. Burkitt, M. J. Cook, K. Fuller, D. B. Grayden, L. Seiderer, and I. M. Mareels. Seizure detection using seizure probability estimation: comparison of features used to detect seizures. *Ann Biomed Eng*, 37(10):2129–45, 2009.

[69] M. E. Saab and J. Gotman. A system to detect the onset of epileptic seizures in scalp eeg. *Clin Neurophysiol*, 116(2):427–42, 2005.

[70] A.T. Tzallas, M.G. Tsipouras, and D.I. Fotiadis. Epileptic seizure detection in eegs using time-frequency analysis. *IEEE Transactions on Information Technology in Biomedicine*, 13(5):703–710, 2009.

[71] J. Jacobs, R. Zelmann, J. Jirsch, R. Chander, C. E. Dubeau, and J. Gotman. High frequency oscillations (80-500 hz) in the preictal period in patients with focal seizures. *Epilepsia*, 50(7):1780–92, 2009.

[72] Gao J-M. Lie D.Y.C. Zhang Y. Oommen K.J. Bao, F.S. Automated epilepsy diagnosis using interictal scalp eeg. In *31st Annual International Conference of the IEEE EMBS*, pages 6603–6607.

[73] H. Isik and E. Sezer. Diagnosis of epilepsy from electroencephalography signals using multilayer perceptron and elman artificial neural networks and wavelet transform. *J Med Syst*, 36(1):1–13, 2010.

[74] E. Sezer, H. Isik, and E. Saracoglu. Employment and comparison of different artificial neural networks for epilepsy diagnosis from eeg signals. *J Med Syst*, 36(1):347–62, 2010.

[75] F. G. Woermann and C. Vollmar. Clinical mri in children and adults with focal epilepsy: a critical review. *Epilepsy Behav*, 15(1):40–9, 2009.

[76] A. Bernasconi and N. Bernasconi. Unveiling epileptogenic lesions: the contribution of image processing. *Epilepsia*, 52 Suppl 4:20–4, 2011.

[77] A. Bernasconi, N. Bernasconi, B. C. Bernhardt, and D. Schrader. Advances in mri for 'cryptogenic' epilepsies. *Nat Rev Neurol*, 7(2):99–108, 2011.

[78] F. G. Woermann, S. M. Sisodiya, S. L. Free, and J. S. Duncan. Quantitative mri in patients with idiopathic generalized epilepsy. evidence of widespread cerebral structural changes. *Brain*, 121 (Pt 9):1661–7, 1998.

[79] Z. Haneef, A. Lenartowicz, H. J. Yeh, Jr. Engel, J., and J. M. Stern. Effect of lateralized temporal lobe epilepsy on the default mode network. *Epilepsy Behav*, 25(3):350–7, 2012.

[80] H. O. Luders, I. Najm, D. Nair, P. Widdess-Walsh, and W. Bingman. The epileptogenic zone: general principles. *Epileptic Disord*, 8 Suppl 2:S1–9, 2006.

[81] W.T. Kerr, A.Y. Cho, A. Anderson, P.K. Douglas, E. P. Lau, E. S. Hwang, K. R. Raman, A. Trefler, S. T. Nguyen, N. M. Reddy, D. H. Silverman, and M. S. Cohen. Balancing clinical and pathologic relevence in the machine

learning diagnosis of epilepsy. In *International Workshop Pattern Recognition in Neuroimaging*. IEEE.

[82] F. T. Sun, M. J. Morrell, and Jr. Wharen, R. E. Responsive cortical stimulation for the treatment of epilepsy. *Neurotherapeutics*, 5(1):68–74, 2008.

[83] C. la Fougere, A. Rominger, S. Forster, J. Geisler, and P. Bartenstein. Pet and spect in epilepsy: a critical review. *Epilepsy Behav*, 15(1):50–5, 2009.

[84] F. Mauguiere and P. Ryvlin. The role of pet in presurgical assessment of partial epilepsies. *Epileptic Disord*, 6(3):193–215, 2004.

[85] T. J. O'Brien, K. Miles, R. Ware, M. J. Cook, D. S. Binns, and R. J. Hicks. The cost-effective use of 18f-fdg pet in the presurgical evaluation of medically refractory focal epilepsy. *J Nucl Med*, 49(6):931–7, 2008.

[86] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafo. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*, 14(5):365–76, 2013.

[87] K. Friston. Ten ironic rules for non-statistical reviewers. *Neuroimage*, 61(4):1300–10, 2012.

[88] M. Ingre. Why small low-powered studies are worse than large high-powered studies and how to protect against "trivial" findings in research: comment on friston (2012). *Neuroimage*, 81:496–8, 2013.

[89] W. T. Kerr, E. P. Lau, G. E. Owens, and A. Trefler. The future of medical diagnostics: large digitized databases. *Yale J Biol Med*, 85(3):363–77, 2012.

[90] M. Balish, P. S. Albert, and W. H. Theodore. Seizure frequency in intractable partial epilepsy: a statistical analysis. *Epilepsia*, 32(5):642–9, 1991.

[91] M. J. Hayat and M. Higgins. Understanding poisson regression. *J Nurs Educ*, 53(4):207–15, 2014.

[92] F. Kianifard and P. P. Gallo. Poisson regression analysis in clinical research. *J Biopharm Stat*, 5(1):115–29, 1995.

[93] E. Briscoe and J. Feldman. Conceptual complexity and the bias/variance tradeoff. *Cognition*, 118(1):2–16, 2011.

[94] Q. Noirhomme, D. Lesenfants, F. Gomez, A. Soddu, J. Schrouff, G. Garraux, A. Luxen, C. Phillips, and S. Laureys. Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. *Neuroimage Clin*, 4:687–94, 2014.

[95] M. Kottas, O. Kuss, and A. Zapf. A modified wald interval for the area under the roc curve (auc) in diagnostic case-control studies. *BMC Med Res Methodol*, 14:26, 2014.

[96] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):272–297, 1995.

[97] G.F. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, 1968.

[98] B. Mwangi, T. S. Tian, and J. C. Soares. A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, 12(2):229–44, 2014.

[99] W. T. Kerr, P. K. Douglas, A. Anderson, and M. S. Cohen. The utility of data-driven feature selection: Re: Chu et al. 2012. *Neuroimage*, page [in press], 2013.

[100] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *Ieee Transactions on Neural Networks*, 10(3):626–634, 1999.

[101] A. Anderson, I. D. Dinov, J. E. Sherin, J. Quintana, A. L. Yuille, and M. S. Cohen. Classification of spatially unaligned fmri scans. *Neuroimage*, 49(3):2509–19, 2010.

[102] A. Anderson, J. S. Labus, E. P. Vianna, E. A. Mayer, and M. S. Cohen. Common component classification: what can we learn from machine learning? *Neuroimage*, 56(2):517–24, 2011.

[103] P. Comon. Independent component analysis, a new concept. *Signal Processing*, 36(3):287–314, 1994.

[104] C. Jutten and J. Herault. Blind separation of sources .1. an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

[105] S. A. Meda, B. Narayanan, J. Liu, N. I. Perrone-Bizzozero, M. C. Stevens, V. D. Calhoun, D. C. Glahn, L. Shen, S. L. Risacher, A. J. Saykin, and G. D. Pearlson. A large scale multivariate parallel ica method reveals novel imaging-genetic relationships for alzheimer's disease in the adni cohort. *Neuroimage*, 60(3):1608–21, 2012.

[106] R. C. Thornton, R. Rodionov, H. Laufs, S. Vulliemoz, A. Vaudano, D. Carmichael, S. Cannadathu, M. Guye, A. McEvoy, S. Lhatoo, F. Bartolomei, P. Chauvel, B. Diehl, F. De Martino, R. D. Elwes, M. C. Walker, J. S. Duncan, and L. Lemieux. Imaging haemodynamic changes related to

seizures: comparison of eeg-based general linear model, independent component analysis of fmri and intracranial eeg. *Neuroimage*, 53(1):196–205, 2010.

[107] M. Chagoyen, P. Carmona-Saez, H. Shatkay, J. M. Carazo, and A. Pascual-Montano. Discovering semantic features in the literature: a foundation for building functional associations. *BMC Bioinformatics*, 7:41, 2006.

[108] A. Anderson, P. K. Douglas, W. T. Kerr, V. S. Haynes, A. L. Yuille, J. Xie, Y. N. Wu, J. A. Brown, and M. S. Cohen. Non-negative matrix factorization of multimodal mri, fmri and phenotypic data reveals differential changes in default mode subnetworks in adhd. *Neuroimage*, 102 Pt 1:207–19, 2014.

[109] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J Mach Learn Res*, 3:1157–1182, 2003.

[110] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[111] S. Haufe, F. Meinecke, K. Gorgen, S. Dahne, J. D. Haynes, B. Blankertz, and F. Biessmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014.

[112] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: a library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[113] E.D. Dohmatob, A. Gramfort, B. Thirion, and G. Varoquaux. Benchmarking solvers for tv-1 least-squares and logistic regression in brain imaging, 2014.

[114] M. Dubois, F. Hadj-Selem, T. Lofstedt, M. Perrot, C. Fischer, V. Frouin, and E. Duchesnay. Predictive support recovery with tv-elastic net penalty and logistic regression: An application to structural mri, 2014.

[115] A. Labate, A. Cerasa, M. Mula, L. Mumoli, M. C. Gioia, U. Aguglia, A. Quattrone, and A. Gambardella. Neuroanatomic correlates of psychogenic nonepileptic seizures: a cortical thickness and vbm study. *Epilepsia*, 53(2):377–85, 2012.

[116] European database on epilepsy, 2007.

[117] P.S. Aisen. Adni 2 study, 2008.

[118] Y. C. Chen, J. C. Wu, I. Haschler, A. Majeed, T. J. Chen, and T. Wetter. Academic impact of a public electronic health database: bibliometric

analysis of studies using the general practice research database. *PLoS One*, 6(6):e21404, 2011.

[119] G.M. Church. Personal genome project mission, 2012.

[120] M. Hunter, R. L. Smith, W. Hyslop, O. A. Rosso, R. Gerlach, J. A. Rostas, D. B. Williams, and F. Henskens. The australian eeg database. *Clin EEG Neurosci*, 36(2):76–81, 2005.

[121] J. E. Lunshof, R. Chadwick, D. B. Vorhaus, and G. M. Church. From genetic privacy to open consent. *Nat Rev Genet*, 9(5):406–11, 2008.

[122] A. Provost. Australian eeg database, 2011.

[123] D. Schrader, R. Shukla, R. Gatrill, K. Farrell, and M. Connolly. Epilepsy with occipital features in children: factors predicting seizure outcome and neuroimaging abnormalities. *Eur J Paediatr Neurol*, 15(1):15–20, 2011.

[124] M.W. Weiner. Letter of welcome from the adni principal investigator, 2009.

[125] United States Congress. American recovery and reinvestment act, 2009.

[126] NIH. Final nih statement on sharing research data, 2003.

[127] NIH. Expansion of sharing and standardization of nih-funded human brain imaing data, 2011.

[128] B. Brockstein, T. Hensing, G. W. Carro, J. Obel, J. Khandekar, L. Kaminer, C. Van De Wege, and R. de Wilton Marsh. Effect of an electronic health record on the culture of an outpatient medical oncology practice in a four-hospital integrated health care system: 5-year experience. *J Oncol Pract*, 7(4):e20–4, 2011.

[129] GPRD. General practice research database, 2012.

[130] Gutthann S.P. Rodrigues, L.A.G. Use of the uk general practice research database for pharmacoepidemiology. *Br J Clin Pharmacol*, 45:419–425, 1998.

[131] Z. Wang, A. D. Shah, A. R. Tate, S. Denaxas, J. Shawe-Taylor, and H. Hemingway. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS One*, 7(1):e30412, 2012.

[132] E. R. Weitzman, L. Kaci, and K. D. Mandl. Sharing medical data for health research: the early personal health record experience. *J Med Internet Res*, 12(2):e14, 2010.

[133] P. A. Teixeira, P. Gordon, E. Camhi, and S. Bakken. Hiv patients' willingness to share personal health information electronically. *Patient Educ Couns*, 84(2):e9–12, 2011.

[134] R. Whiddett, I. Hunter, J. Engelbrecht, and J. Handy. Patients attitudes towards sharing their health information. *International Journal of Medical Informatics*, 75:530–541, 2005.

[135] B. Malin, K. Benitez, and D. Masys. Never too old for anonymity: a statistical standard for demographic data sharing via the hipaa privacy rule. *J Am Med Inform Assoc*, 18(1):3–10, 2011.

[136] Rebecca Skloot. *The immortal life of Henrietta Lacks*. Crown Publishers, New York, 2010.

[137] Congress. Health insurance portability and accountability act, 1996.

[138] WMA General Assembly. World medical association declaration of helsinki: Ethical principles for medical research involving human subjects, 2008 2008.

[139] Crawford K. Neu, S.C. Loni deidentification debablet, 2005.

[140] R.R. Schaller. Moore's law: Past, present and future. *IEEE Spectrum*, 34(6):52–59, 1997.

[141] K. Rupp. The economic limit to moore's law. *IEEE Trans Semiconductor Manufacturing*, 24(1):1–4, 2011.

[142] P. Coupe, S. F. Eskildsen, J. V. Manjon, V. Fonov, and D. L. Collins. Simultaneous segmentation and grading of anatomical structures for patient's classification: Application to alzheimer's disease. *Neuroimage*, 2011.

[143] M. Liu, D. Zhang, and D. Shen. Ensemble sparse classification of alzheimer's disease. *Neuroimage*, 60(2):1106–1116, 2012.

[144] Dorizzi B.-Boudy J. Andreao, R.V. Ecg signal analysis through hidden markov models. *IEEE Trans. Biomed. Eng*, 53:1541–1549, 2006.

[145] O'Dwyer M.-Reilly R.B. Chazal, R. Automated classification of heartbeats using ecg morphology and heartbeat interval features. *IEEE Trans. Biomed. Eng*, 51:1196–1206, 2004.

[146] F. M. Cuthill and C. A. Espie. Sensitivity and specificity of procedures for the differential diagnosis of epileptic and non-epileptic seizures: a systematic review. *Seizure*, 14(5):293–303, 2005.

[147] Reilly-R.B. de Chazal, F. A patient adapting heart beat classifier using ecg morphology and heartbeat interval features. *IEEE Trans. Biomed. Eng*, 53:2535–2543, 2006.

[148] Khazaee A.-Ranaee V. Ebrahimzadeh, A. Classification of electrocardiogram signals using supervised classifiers and efficient features. *Comput. Methods Programs Biomed.*, 99:179–194, 2010.

[149] F. Hoeft, B. D. McCandliss, J. M. Black, A. Gantman, N. Zakerani, C. Hulme, H. Lyytinen, S. Whitfield-Gabrieli, G. H. Glover, A. L. Reiss, and J. D. Gabrieli. Neural systems predicting long-term outcome in dyslexia. *Proc Natl Acad Sci U S A*, 108(1):361–6, 2011.

[150] Kiranyaz S.-Gabbouj M. Ince, T. A generic and robust system for automated patient-specific classification of electrocardiogram signals. *IEEE Trans. Biomed. Eng*, 56:1415–1526, 2009.

[151] Chakraborty C.-Ray A.K. JoyMartis, R. A two-stage mechanism for registration and classification of ecg using gaussian mixture model. *Pattern Recognit*, 42:2979–2988, 2009.

[152] Peterson C.-Braccini G. Edenbrandt L. Sornmo L. Langerholm, M. Clustering ecg complexes using hermite functions and self-organizing maps. *IEEE Trans. Biomed. Eng*, 47:839–847, 2000.

[153] Mitra M.-Chaudhuri B.B. Mitra, S. A rough set-based inference enginer for ecg classification. *IEEE Trans. Instrum. Meas.*, 55:2198–2206, 2006.

[154] D. San-juan, A. T. Claudia, G. A. Maricarmen, M. M. Adriana, J. S. Richard, and A. V. Mario. The prognostic role of electrocorticography in tailored temporal lobe surgery. *Seizure*, 20(7):564–9, 2011.

[155] Wu Y.H.-Hu W.C. Shyu, L.Y. Using wavelet transform and fuzzy neural network for vpc detection from the holter ecg. *IEEE Trans. Biomed. Eng*, 51:1269–1273, 2004.

[156] van Klaveren-R.J. de Bock G.H. Zhao Y. Vernhout R. Leusveld A. Scholten E. Verschakelen J. Mali W. de Koning H. Oudkerk M. Wang, Y. No benefit for consensus double reading at baseline screening for lunch cancer with the use of semiautomated volumetry software. *Radiology*, 262(1):320–326, 2012.

[157] Chou-K.T. Yu, S. N. Selection of significant for ecg beat classification. *Expert Syst Appl*, 36:2088–2096, 2009.

[158] A. E. Zadeh and A. Khazaee. High efficient system for automatic classification of the electrocardiogram beats. *Ann Biomed Eng*, 39(3):996–1011, 2011.

[159] N. Cowan. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav Brain Sci*, 24(1):87–114; discussion 114–85, 2001.

[160] G. Veneri, E. Pretegiani, P. Federighi, F. Rosini, A. Federico, and A. Rufa. Evaluating human visual search performance by monte carlo methods and heuristic model. In *10th IEEE International Conf Information and Applications in Biomedicine*.

[161] Y. Harrison and J. A. Horne. The impact of sleep deprivation on decision making: a review. *J Exp Psychol Appl*, 6(3):236–49, 2000.

[162] P. Kohl, D. Noble, L.R. Winslow, and P.J. Hunter. Computational modeling of biological systems: Tools and visions. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 358(1766):579–610, 2000.

[163] S. J. Wang, B. Middleton, L. A. Prosser, C. G. Bardon, C. D. Spurr, P. J. Carchidi, A. F. Kittler, R. C. Goldszer, D. G. Fairchild, A. J. Sussman, G. J. Kuperman, and D. W. Bates. A cost-benefit analysis of electronic medical records in primary care. *Am J Med*, 114(5):397–403, 2003.

[164] T. Balli and R. Palaniappan. Classification of biological signals using linear and nonlinear features. *Physiol Meas*, 31(7):903–20, 2010.

[165] J. N. Gelinas, A. W. Battison, S. Smith, M. B. Connolly, and P. Steinbok. Electrocorticography and seizure outcomes in children with lesional epilepsy. *Childs Nerv Syst*, 27(3):381–90, 2011.

[166] R. Rodrigues Tda, E. B. Sternick, and C. Moreira Mda. Epilepsy or syncope? an analysis of 55 consecutive patients with loss of consciousness, convulsions, falls, and no eeg abnormalities. *Pacing Clin Electrophysiol*, 33(7):804–13, 2010.

[167] I.G. Fita, A. Enciu, and B.P. Stanoiu. New insights on alzheimer's disease diagnostic. *Rom J Morphol Embryol*, 52(3 Suppl):975–979, 2011.

[168] S. Kloppel, C. M. Stonnington, J. Barnes, F. Chen, C. Chu, C. D. Good, I. Mader, L. A. Mitchell, A. C. Patel, C. C. Roberts, N. C. Fox, Jr. Jack, C. R., J. Ashburner, and R. S. Frackowiak. Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method. *Brain*, 131(Pt 11):2969–74, 2008.

[169] Z. Dai, C. Yan, Z. Wang, J. Wang, M. Xia, K. Li, and Y. He. Discriminative analysis of early alzheimer's disease using multi-modal imaging and multilevel characterization with multi-classifier (m3). *Neuroimage*, 2011.

[170] Su-S-C. Huang C-H. Wang J.J. Xu W-C. Wei Y-Y. Lee S.T. Lee, J-D. Combination of multiple features in support vector machine with principle component analysis in application for alzheimer's disease diagnosis. *Lecture notes in computer science*, 5864:512–519, 2009.

[171] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc*, 18(5):601–6, 2011.

[172] Shah N.-Hanson-P. Balasubramaniam S.-Smith S.A. Pakhomov, S.V. Automatic quality of life prediction using electronic medical records. In *American Medical Informatics Association Symposium*, pages 545–549.

[173] S.B. McGrayne. *The theory that would not die: How Bayes' rule cracked the Enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy.* Yale University Press, Devon, Pennsylvania, 2011.

[174] M. A. Oquendo, E. Baca-Garcia, A. Artes-Rodriguez, F. Perez-Cruz, H. C. Galfalvy, H. Blasco-Fontecilla, D. Madigan, and N. Duan. Machine learning and data mining: strategies for hypothesis generation. *Mol Psychiatry*, 2012.

[175] Y. Cho, J. K. Seong, Y. Jeong, and S. Y. Shin. Individual subject classification for alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. *Neuroimage*, 2011.

[176] Suchard M.A. Huelsenbeck, J.P. A nonparametric method for accommodating and testing across-site rate variation. *Systematic Biology*, 56:975–987, 2007.

[177] Agile diagnosis, 2012.

[178] B. Goldman. Doctors make mistakes: Can we talk about that? In *TED*. TED.

[179] E. Tuv, A. Borisov, G. Runger, and K. Torkkola. Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research*, 10:1341–1366, 2009.

[180] M. Esterman, Y. C. Chiu, B. J. Tamber-Rosenau, and S. Yantis. Decoding cognitive control in human parietal cortex. *Proc Natl Acad Sci U S A*, 106(42):17974–9, 2009.

[181] J. D. Johnson, S. G. McDuff, M. D. Rugg, and K. A. Norman. Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. *Neuron*, 63(5):697–708, 2009.

[182] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns. *Neuroimage*, 43(1):44–58, 2008.

[183] C. Ecker, V. Rocha-Rego, P. Johnston, J. Mourao-Miranda, A. Marquand, E. M. Daly, M. J. Brammer, C. Murphy, and D. G. Murphy. Investigating the predictive value of whole-brain structural mr scans in autism: a pattern classification approach. *Neuroimage*, 49(1):44–56, 2010.

[184] D. Dai, J. Wang, J. Hua, and H. He. Classification of adhd children through multimodal magnetic resonance imaging. *Front Syst Neurosci*, 6:63, 2012.

[185] S. Song, Z. Zhan, Z. Long, J. Zhang, and L. Yao. Comparative study of svm methods combined with voxel selection for object category classification on fmri data. *PLoS One*, 6(2):e17191, 2011.

[186] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–30, 2001.

[187] N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional brain mapping. *Proc Natl Acad Sci U S A*, 103(10):3863–8, 2006.

[188] M. Bjornsdotter, K. Rylander, and J. Wessberg. A monte carlo method for locally multivariate brain mapping. *Neuroimage*, 56(2):508–16, 2011.

[189] Y. Liu, H.H. Zhang, C. Park, and J. Ahn. The lq support vector machine. *Contemporary Mathematics*, 443:35–48, 2007.

[190] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987.

[191] M.J. KcKeown, S. Makeig, G.G. Brown, Jung T.-P., S.S. Kindermann, A.J. Bell, and T.J. Sejnowski. Analysis of fmri data by blind separation into independent spatial components, 1997.

[192] C. Hinrichs, V. Singh, L. Mukherjee, G. F. Xu, M. K. Chung, S. C. Johnson, and ADNI. Spatially augmented lpboosting for ad classification with evaluations on the adni dataset. *Neuroimage*, 48(1):138–149, 2009.

[193] P.K. Douglas, S. Harris, A. Yuille, and M.S. Cohen. Performance comparison of machine learning algorithms and number of independent components used in fmri decoding of belief vs disbelief. *Neuroimage*, 56(2):544–553, 2010.

[194] K. Franke, G. Ziegler, S. Kloppel, and C. Gaser. Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: exploring the influence of various parameters. *Neuroimage*, 50(3):883–92, 2010.

[195] K. Franke, E. Luders, A. May, M. Wilke, and C. Gaser. Brain maturation: predicting individual brainage in children and adolescents using structural mri. *Neuroimage*, 63(3):1305–12, 2012.

[196] J. H. Yang and V. Honavar. Feature subset selection using a genetic algorithm. *Ieee Intelligent Systems & Their Applications*, 13(2):44–49, 1998.

[197] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[198] T. G. Dietterich. Ensemble methods in machine learning. *Multiple Classifier Systems*, 1857:1–15, 2000.

[199] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[200] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28(2):337–374, 2000.

[201] N. Kwak and C. H. Choi. Input feature selection for classification problems. *Ieee Transactions on Neural Networks*, 13(1):143–159, 2002.

[202] J. M. Leiva-Murillo and A. Artes-Rodriguez. Maximization of mutual information for supervised linear feature extraction. *Ieee Transactions on Neural Networks*, 18(5):1433–1441, 2007.

[203] H. Liu and R. Setiono. Feature selection via discretization. *Ieee Transactions on Knowledge and Data Engineering*, 9(4):642–645, 1997.

[204] R. Setiono and H. Liu. Neural-network feature selector. *Ieee Transactions on Neural Networks*, 8(3):654–662, 1997.

[205] H. B. Zhang and G. Y. Sun. Feature selection using tabu search method. *Pattern Recognition*, 35(3):701–711, 2002.

[206] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[207] K. H. Brodersen, T. M. Schofield, A. P. Leff, C. S. Ong, E. I. Lomakina, J. M. Buhmann, and K. E. Stephan. Generative embedding for model-based classification of fmri data. *PLoS Comput Biol*, 7(6):e1002079, 2011.

[208] A. Anderson, J. Bramen, P. K. Douglas, A. Lenartowicz, A. Cho, C. Culbertson, A. L. Brody, A. L. Yuille, and M. S. Cohen. Large sample group independent component analysis of functional magnetic resonance imaging using anatomical atlas-based reduction and bootstrapped clustering. *Int J Imaging Syst Technol*, 21(2):223–231, 2011.

[209] J.B. Colby, J.D. Rudie, J.A. Brown, P.K. Douglas, M. S. Cohen, and Z. Shehzad. Insights into multimodal imaging classification of adhd. *Frontiers in Systems Neuroscience*, In press, 2012.

[210] M. D. Greicius, G. Srivastava, A. L. Reiss, and V. Menon. Default-mode network activity distinguishes alzheimer's disease from healthy aging: evidence from functional mri. *Proc Natl Acad Sci U S A*, 101(13):4637–42, 2004.

[211] M. Zibulevsky and B. A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13(4):863–882, 2001.

[212] J. Sui, T. Adali, G. D. Pearlson, and V. D. Calhoun. An ica-based method for the identification of optimal fmri features and components using combined group-discriminative techniques. *Neuroimage*, 46(1):73–86, 2009.

[213] A. R. Franco, M. V. Mannell, V. D. Calhoun, and A. R. Mayer. Impact of analysis methods on the reproducibility and reliability of resting-state networks. *Brain Connect*, 3(4):363–74, 2013.

[214] I. Osorio, A. Lyubushin, and D. Sornette. Automated seizure detection: unrecognized challenges, unexpected insights. *Epilepsy Behav*, 22 Suppl 1:S7–17, 2011.

[215] J. C. Sackellares. Seizure prediction. *Epilepsy Currents*, 8(3):55–59, 2008.

[216] O. A. Rosso, A. Mendes, R. Berretta, J. A. Rostas, M. Hunter, and P. Moscato. Distinguishing childhood absence epilepsy patients from controls by the analysis of their background brain electrical activity (ii): a combinatorial optimization approach for electrode selection. *J Neurosci Methods*, 181(2):257–67, 2009.

[217] E. Santiago-Rodriguez, T. Harmony, L. Cardenas-Morales, A. Hernandez, and A. Fernandez-Bouzas. Analysis of background eeg activity in patients with juvenile myoclonic epilepsy. *Seizure*, 17(5):437–45, 2008.

[218] L. M. Binder and M. C. Salinsky. Psychogenic nonepileptic seizures. *Neuropsychol Rev*, 17(4):405–12, 2007.

[219] H. Patel, E. Scott, D. Dunn, and B. Garg. Nonepileptic seizures in children. *Epilepsia*, 48(11):2086–92, 2007.

[220] M. A. Rogawski and W. Loscher. The neurobiology of antiepileptic drugs. *Nat Rev Neurosci*, 5(7):553–64, 2004.

[221] C.M. Schneider-Mizell, J.M. Parent, E. Ben-Jacob, M.R. Zochowski, and L.M. Sander. From network structure to network reorganization: implications for adult neurogenesis. *Physical Biology*, 7:1–11, 2010.

[222] C. Person, L. Koessler, V. Louis-Dorr, D. Wolf, L. Maillard, and P.Y. Marie. Analysis of the relationship between interictal electrical source imaging and pet hypometabolism. In *IEEE Eng Med Biol Soc*, pages 3723–3726.

[223] C. Hoppe and C. E. Elger. Depression in epilepsy: a critical review from a clinical perspective. *Nat Rev Neurol*, 7(8):462–72, 2011.

[224] A. B. Ettinger. Psychotropic effects of antiepileptic drugs. *Neurology*, 67(11):1916–25, 2006.

[225] S. Vincentiis, K. D. Valente, S. Thome-Souza, E. Kuczinsky, L. A. Fiore, and N. Negrao. Risk factors for psychogenic nonepileptic seizures in children and adolescents with epilepsy. *Epilepsy Behav*, 8(1):294–8, 2006.

[226] G. L. Holmes, J. C. Sackellares, J. McKiernan, M. Ragland, and F. E. Dreifuss. Evaluation of childhood pseudoseizures using eeg telemetry and video tape monitoring. *J Pediatr*, 97(4):554–8, 1980.

[227] A. M. Kanner, S. C. Schachter, J. J. Barry, D. C. Hersdorffer, M. Mula, M. Trimble, B. Hermann, A. E. Ettinger, D. Dunn, R. Caplan, P. Ryvlin, and F. Gilliam. Depression and epilepsy, pain and psychogenic non-epileptic seizures: clinical and therapeutic perspectives. *Epilepsy Behav*, 24(2):169–81, 2012.

[228] M. Hall, E. Frank, G. L. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[229] J. D. Bremner, M. Narayan, E. R. Anderson, L. H. Staib, H. L. Miller, and D. S. Charney. Hippocampal volume reduction in major depression. *Am J Psychiatry*, 157(1):115–8, 2000.

[230] T. Frodl, E. M. Meisenzahl, T. Zetzsche, C. Born, C. Groll, M. Jager, G. Leinsinger, R. Bottlender, K. Hahn, and H. J. Moller. Hippocampal changes in patients with a first episode of major depression. *Am J Psychiatry*, 159(7):1112–8, 2002.

[231] M. A. Mikati. *Kliegman: Nelson Textbook of Pediatrics*. Saunders, Philadelphia, 19th edition, 2011.

[232] A. M. Isom, G. A. Gudelsky, S. C. Benoit, and N. M. Richtand. Antipsychotic medications, glutamate, and cell death: A hidden, but common medication side effect? *Med Hypotheses*, 80(3):252–8, 2013.

[233] W. T. Kerr and E. P. Lau. Poisson noise obscures hypometabolic lesions in pet. *Yale J Biol Med*, 85(4):541–9, 2012.

[234] J. W. Pan, A. Williamson, I. Cavus, H. P. Hetherington, H. Zaveri, O. A. Petroff, and D. D. Spencer. Neurometabolism in human epilepsy. *Epilepsia*, 49 Suppl 3:31–41, 2008.

[235] J. T. Lerner, N. Salamon, J. S. Hauptman, T. R. Velasco, M. Hemb, J. Y. Wu, R. Sankar, W. Donald Shields, Jr. Engel, J., I. Fried, C. Cepeda, V. M. Andre, M. S. Levine, H. Miyata, W. H. Yong, H. V. Vinters, and G. W. Mathern. Assessment and surgical outcomes for mild type i and severe type ii cortical dysplasia: a critical review and the ucla experience. *Epilepsia*, 50(6):1310–35, 2009.

[236] J. Anderson and K. Hamandi. Understanding juvenile myoclonic epilepsy: Contributions from neuroimaging. *Epilepsy Res*, 2011.

[237] T. Butler, M. Ichise, A. F. Teich, E. Gerard, J. Osborne, J. French, O. Devinsky, R. Kuzniecky, F. Gilliam, F. Pervez, F. Provenzano, S. Goldsmith, S. Vallabhajosula, E. Stern, and D. Silbersweig. Imaging inflammation in a patient with epilepsy due to focal cortical dysplasia. *J Neuroimaging*, 2011.

[238] F. Chassoux, S. Rodrigo, F. Semah, F. Beuvon, E. Landre, B. Devaux, B. Turak, C. Mellerio, J.-F. Meder, F.-X. Roux, C. Daumas-Dupont, P. Merlet, O. Dulac, and C. Chiron. Fdg-pet improves surgical outcome in negative mri taylor-type focal cortical dysplasias. *Neurology*, 75:2168–2175, 2010.

[239] J. Duncan. The current status of neuroimaging for epilepsy. *Curr Opin Neurol*, 22(2):179–84, 2009.

[240] T. R. Henry and D. D. Roman. Presurgical epilepsy localization with interictal cerebral dysfunction. *Epilepsy Behav*, 20(2):194–208, 2011.

[241] Y. H. Kim, H. C. Kang, D. S. Kim, S. H. Kim, K. W. Shim, H. D. Kim, and J. S. Lee. Neuroimaging in identifying focal cortical dysplasia and prognostic factors in pediatric and adolescent epilepsy surgery. *Epilepsia*, 52(4):722–7, 2011.

[242] A. Kumar, C. Juhasz, E. Asano, S. Sood, O. Muzik, and H. T. Chugani. Objective detection of epileptic foci by 18f-fdg pet in children undergoing epilepsy surgery. *J Nucl Med*, 51(12):1901–7, 2010.

[243] N. Madan and P. E. Grant. New directions in clinical imaging of cortical dysplasias. *Epilepsia*, 50 Suppl 9:9–18, 2009.

[244] A. A. Cohen-Gadol, B. G. Wilhelmi, F. Collignon, J. B. White, J. W. Britton, D. M. Cambier, T. J. Christianson, W. R. Marsh, F. B. Meyer, and G. D. Cascino. Long-term outcome of epilepsy surgery among 399 patients with nonlesional seizure foci including mesial temporal lobe sclerosis. *J Neurosurg*, 104(4):513–24, 2006.

[245] A. E. Elsharkawy, F. Behne, F. Oppel, H. Pannek, R. Schulz, M. Hoppe, G. Pahs, C. Gyimesi, M. Nayel, A. Issa, and A. Ebner. Long-term outcome of extratemporal epilepsy surgery among 154 adult patients. *J Neurosurg*, 108(4):676–86, 2008.

[246] M. A. Murphy, T. J. O'Brien, K. Morris, and M. J. Cook. Multimodality image-guided surgery for the treatment of medically refractory epilepsy. *J Neurosurg*, 100(3):452–62, 2004.

[247] M. Defrise, D.W. Townsend, and F. Deconinck. Statistical noise in three-dimensional positron tomography. *Phys Med Biol*, 35(1):131–138, 1990.

[248] L. D. Nickerson, S. Narayana, J. L. Lancaster, P. T. Fox, and J. H. Gao. Estimation of the local statistical noise in positron emission tomography revisited: practical implementation. *Neuroimage*, 19(2 Pt 1):442–56, 2003.

[249] K. M. Hanson. On the optimality of the filtered backprojection algorithm. *J Comput Assist Tomogr*, 4(3):361–3, 1980.

[250] H. M. Hudson and R. S. Larkin. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Trans Med Imaging*, 13(4):601–9, 1994.

[251] T. F. Budinger. Pet instrumentation: what are the limits? *Semin Nucl Med*, 28(3):247–67, 1998.

[252] D. W. Townsend. Physical principles and technology of clinical pet imaging. *Ann Acad Med Singapore*, 33(2):133–45, 2004.

[253] K. Lange and R. Carson. Em reconstruction algorithms for emission and transmission tomography. *J Comput. Assist. Tomog.*, 8:306–316, 1984.

[254] Kenneth Lange. *Optimization*, volume 2. Springer, New York, USA, 2004.

[255] W. Chen, D. H. Silverman, S. Delaloye, J. Czernin, N. Kamdar, W. Pope, N. Satyamurthy, C. Schiepers, and T. Cloughesy. 18f-fdopa pet imaging of brain tumors: comparison study with 18f-fdg pet and evaluation of diagnostic accuracy. *J Nucl Med*, 47(6):904–11, 2006.

[256] C. S. Lee, A. Samii, V. Sossi, T. J. Ruth, M. Schulzer, J. E. Holden, J. Wudel, P. K. Pal, R. de la Fuente-Fernandez, D. B. Calne, and A. J. Stoessl. In vivo positron emission tomographic evidence for compensatory changes in presynaptic dopaminergic nerve terminals in parkinson's disease. *Ann Neurol*, 47(4):493–503, 2000.

[257] J. Booij, G. Tissingh, A. Winogrodzka, and E. A. van Royen. Imaging of the dopaminergic neurotransmission system using single-photon emission tomography and positron emission tomography in patients with parkinsonism. *Eur J Nucl Med*, 26(2):171–82, 1999.

[258] J. Booij and H. W. Berendse. Monitoring therapeutic effects in parkinson's disease by serial imaging of the nigrostriatal dopaminergic pathway. *J Neurol Sci*, 310(1-2):40–3, 2011.

[259] T. W. Lin, M. A. de Aburto, M. Dahlbom, L. L. Huang, M. M. Marvi, M. Tang, J. Czernin, M. E. Phelps, and D. H. Silverman. Predicting seizure-free status for temporal lobe epilepsy patients undergoing surgery: prognostic value of quantifying maximal metabolic asymmetry extending over a specified proportion of the temporal lobe. *J Nucl Med*, 48(5):776–82, 2007.

[260] D. H. Silverman, C. L. Geist, H. A. Kenna, K. Williams, T. Wroolie, B. Powers, J. Brooks, and N. L. Rasgon. Differences in regional brain metabolism associated with specific formulations of hormone therapy in postmenopausal women at risk for ad. *Psychoneuroendocrinology*, 36(4):502–13, 2011.

[261] J. Fox. *Applied Regression Analysis, Linear Models, and Related Methods.* Sage Publications, 1 edition, 1997.

[262] J. Fox. *Applied regression analysis and general linear models.* Sage Publications, Inc, 2nd edition, 2008.

[263] G.K. Robinson. That blup is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32, 1991.

[264] S. Kim, L. Shen, A. J. Saykin, and J. D. West. Visual exploration of genetic association with voxel-based imaging phenotypes in an mci/ad study. *Conf Proc IEEE Eng Med Biol Soc*, 2009:3849–52, 2009.

[265] K. Sahaya, S. A. Dholakia, and P. K. Sahota. Psychogenic non-epileptic seizures: a challenging entity. *J Clin Neurosci*, 18(12):1602–7, 2011.

[266] F. Brigo, M. Storti, P. Lochner, F. Tezzon, A. Fiaschi, L. G. Bongiovanni, and R. Nardone. Tongue biting in epileptic seizures and psychogenic events: an evidence-based perspective. *Epilepsy Behav*, 25(2):251–5, 2012.

[267] R. J. Wilkus, C. B. Dodrill, and P. M. Thompson. Intensive eeg monitoring and psychological studies of patients with pseudoepileptic seizures. *Epilepsia*, 25(1):100–7, 1984.

[268] U. Seneviratne, D. Rajendran, M. Brusco, and T. G. Phan. How good are we at diagnosing seizures based on semiology? *Epilepsia*, 2012.

[269] U. Seneviratne, D. Reutens, and W. D'Souza. Stereotypy of psychogenic nonepileptic seizures: insights from video-eeg monitoring. *Epilepsia*, 51(7):1159–68, 2010.

[270] J. O. Elliott and C. Charyton. Biopsychosocial predictors of psychogenic non-epileptic seizures. *Epilepsy Res*, 108(9):1543–53, 2014.

[271] H. Luders, J. Acharya, C. Baumgartner, S. Benbadis, A. Bleasel, R. Burgess, D. S. Dinner, A. Ebner, N. Foldvary, E. Geller, H. Hamer, H. Holthausen, P. Kotagal, H. Morris, H. J. Meencke, S. Noachtar, F. Rosenow, A. Sakamoto, B. J. Steinhoff, I. Tuxhorn, and E. Wyllie. Semiological seizure classification. *Epilepsia*, 39(9):1006–13, 1998.

[272] T. U. Syed, A. M. Arozullah, K. L. Loparo, R. Jamasebi, G. P. Suciu, C. Griffin, R. Mani, I. Syed, T. Loddenkemper, and A. V. Alexopoulos. A self-administered screening instrument for psychogenic nonepileptic seizures. *Neurology*, 72(19):1646–52, 2009.

[273] L. Szabo, Z. Siegler, L. Zubek, Z. Liptai, I. Korhegyi, B. Bansagi, and A. Fogarasi. A detailed semiologic analysis of childhood psychogenic nonepileptic seizures. *Epilepsia*, 53(3):565–70, 2012.

[274] V. Patterson, P. Pant, N. Gautam, and A. Bhandari. A bayesian tool for epilepsy diagnosis in the resource-poor world: development and early validation. *Seizure*, 23(7):567–9, 2014.

[275] D.B. Rubin. Inference and missing data (with discussion). *Biometrika*, 63:581–592, 1976.

[276] D.B. Rubin. *Multiple imputation for non-response in surveys.* John Wiley & Sons, New York, 1987.

[277] S. R. Benbadis. A spell in the epilepsy clinic and a history of ”chronic pain” or ”fibromyalgia” independently predict a diagnosis of psychogenic seizures. *Epilepsy Behav*, 6(2):264–5, 2005.

[278] R. Dixit, A. Popeschu, A. Bagic, G. Ghearing, and R. Henrdrickson. Medical comorbidities in patients with psychogenic nonepileptic spells (pnes) referred for video-eeg monitoring. *Epilepsy & Behavior*, 28:137–140, 2013.

[279] B. A. Dworetzky, A. Strahonja-Packard, C. W. Shanahan, J. Paz, B. Schauble, and E. B. Bromfield. Characteristics of male veterans with psychogenic nonepileptic seizures. *Epilepsia*, 46(9):1418–22, 2005.

[280] K. R. Kaufman, A. Mohebati, and A. Sotolongo. Pseudoseizures and hysterical stridor. *Epilepsy Behav*, 5(2):269–72, 2004.

[281] M. Arthuis, J. A. Micoulaud-Franchi, F. Bartolomei, A. McGonigal, and E. Guedj. Resting cortical pet metabolic changes in psychogenic non-epileptic seizures (pnes). *J Neurol Neurosurg Psychiatry*, 2014.

[282] J. R. Ding, D. An, W. Liao, J. Li, G. R. Wu, Q. Xu, Z. Long, Q. Gong, D. Zhou, O. Sporns, and H. Chen. Altered functional and structural connectivity networks in psychogenic non-epileptic seizures. *PLoS One*, 8(5):e63850, 2013.

[283] J. Hovorka, T. Nezadal, E. Herman, I. Nemcova, and M. Bajacek. Psychogenic non-epileptic seizures, prospective clinical experience: diagnosis, clinical features, risk factors, psychiatric comorbidity, treatment outcome. *Epileptic Disord*, 9 Suppl 1:S52–8, 2007.

[284] M. Reuber, G. Fernandez, C. Helmstaedter, A. Qurishi, and C. E. Elger. Evidence of brain abnormality in patients with psychogenic nonepileptic seizures. *Epilepsy Behav*, 3(3):249–254, 2002.

[285] S. J. van der Kruijs, N. M. Bodde, M. J. Vaessen, R. H. Lazeron, K. Vonck, P. Boon, P. A. Hofman, W. H. Backes, A. P. Aldenkamp, and J. F. Jansen. Functional connectivity of dissociation in patients with psychogenic nonepileptic seizures. *J Neurol Neurosurg Psychiatry*, 83(3):239–47, 2012.

[286] P. de Timary, P. Fouchet, M. Sylin, J. P. Indriets, T. de Barsy, A. Lefebvre, and K. van Rijckevorsel. Non-epileptic seizures: delayed diagnosis in patients presenting with electroencephalographic (eeg) or clinical signs of epileptic seizures. *Seizure*, 11(3):193–7, 2002.

[287] A. A. Asadi-Pooya, M. Emami, and Y. Emami. Ictal injury in psychogenic non-epileptic seizures. *Seizure*, 23(5):363–6, 2014.

[288] E. Peguero, B. Abou-Khalil, T. Fakhoury, and G. Mathews. Self-injury and incontinence in psychogenic seizures. *Epilepsia*, 36(6):586–91, 1995.

[289] S. R. Benbadis, V. Agrawal, and W. O. th Tatum. How many patients with psychogenic nonepileptic seizures also have epilepsy? *Neurology*, 57(5):915–7, 2001.

[290] R. P. Lesser. Psychogenic seizures. *Neurology*, 46(6):1499–507, 1996.

[291] K. R. Sigurdardottir and E. Olafsson. Incidence of psychogenic seizures in adults: a population-based study in iceland. *Epilepsia*, 39(7):749–52, 1998.

[292] Frank E. Hall-M.A. Holmes G.-Pfahringer B. Ruetemann P. Witten-I.H. Bouckaert, R.R. Weka-experiences with a java open-source project. *J Mach Learn Res*, 11:2533–2541, 2010.

[293] W.T. Kerr, A. Anderson, H. Xia, E.S. Braun, E.P Lau, A.Y. Cho, and M. S. Cohen. Parameter selection in mutual information-based feature selection in automated diagnosis of multiple epilepsies using scalp eeg, 2012.

[294] R. L. Marchetti, D. Kurcgant, J. Gallucci Neto, M. A. Von Bismark, and L. A. Fiore. Evaluating patients with suspected nonepileptic psychogenic seizures. *J Neuropsychiatry Clin Neurosci*, 21(3):292–8, 2009.

[295] R. L. Marchetti, D. Kurcgant, J. G. Neto, M. A. von Bismark, L. B. Marchetti, and L. A. Fiore. Psychiatric diagnoses of patients with psychogenic non-epileptic seizures. *Seizure*, 17(3):247–53, 2008.

[296] D. E. Cragar, D. T. Berry, T. A. Fakhoury, J. E. Cibula, and F. A. Schmitt. A review of diagnostic techniques in the differential diagnosis of epileptic and nonepileptic seizures. *Neuropsychol Rev*, 12(1):31–64, 2002.

[297] M. Privitera. Current challenges in the management of epilepsy. *Am J Manag Care*, 17 Suppl 7:S195–203, 2011.

[298] R. C. Delaney, A. J. Rosen, R. H. Mattson, and R. A. Novelly. Memory function in focal epilepsy: a comparison of non-surgical, unilateral temporal lobe and frontal lobe samples. *Cortex*, 16(1):103–17, 1980.

[299] H. Kim, S. Yi, E. I. Son, and J. Kim. Differential effects of left versus right mesial temporal lobe epilepsy on wechsler intelligence factors. *Neuropsychology*, 17(4):556–65, 2003.

[300] W. Liao, Z. Zhang, Z. Pan, D. Mantini, J. Ding, X. Duan, C. Luo, Z. Wang, Q. Tan, G. Lu, and H. Chen. Default mode network abnormalities in mesial temporal lobe epilepsy: a study combining fmri and dti. *Hum Brain Mapp*, 32(6):883–95, 2011.

[301] C. McCormick, M. Quraan, M. Cohn, T. A. Valiante, and M. P. McAndrews. Default mode network connectivity indicates episodic memory capacity in mesial temporal lobe epilepsy. *Epilepsia*, 2013.

[302] V. L. Morgan, B. P. Rogers, H. H. Sonmezturk, J. C. Gore, and B. Abou-Khalil. Cross hippocampal influence in mesial temporal lobe epilepsy measured with high temporal resolution functional magnetic resonance imaging. *Epilepsia*, 52(9):1741–9, 2011.

[303] V. L. Morgan, H. H. Sonmezturk, J. C. Gore, and B. Abou-Khalil. Lateralization of temporal lobe epilepsy using resting functional magnetic resonance imaging connectivity of hippocampal networks. *Epilepsia*, 2012.

[304] F. Pittau, C. Grova, F. Moeller, F. Dubeau, and J. Gotman. Patterns of altered functional connectivity in mesial temporal lobe epilepsy. *Epilepsia*, 53(6):1013–23, 2012.

[305] Z. Zhang, G. Lu, Y. Zhong, Q. Tan, W. Liao, Z. Wang, K. Li, H. Chen, and Y. Liu. Altered spontaneous neuronal activity of the default-mode network in mesial temporal lobe epilepsy. *Brain Res*, 1323:152–60, 2010.

[306] F. R. Pereira, A. Alessio, M. S. Sercheli, T. Pedro, E. Bilevicius, J. M. Rondina, H. F. Ozelo, G. Castellano, R. J. Covolan, B. P. Damasceno, and

F. Cendes. Asymmetrical hippocampal connectivity in mesial temporal lobe epilepsy: evidence from resting state fmri. *BMC Neurosci*, 11:66, 2010.

[307] S. R. Benbadis, W. O. th Tatum, F. R. Murtagh, and F. L. Vale. Mri evidence of mesial temporal sclerosis in patients with psychogenic nonepileptic seizures. *Neurology*, 55(7):1061–2, 2000.

[308] P. S. Chandra, N. Salamon, J. Huang, J. Y. Wu, S. Koh, H. V. Vinters, and G. W. Mathern. Fdg-pet/mri coregistration and diffusion-tensor imaging distinguish epileptogenic tubers and cortex in patients with tuberous sclerosis complex: a preliminary report. *Epilepsia*, 47(9):1543–9, 2006.

[309] S. Rastogi, C. Lee, and N. Salamon. Neuroimaging in pediatric epilepsy: a multimodality approach. *RadioGraphics*, 28:1079–1095, 2008.

[310] C. Juhasz, F. Nagy, C. Watson, E. A. da Silva, O. Muzik, D. C. Chugani, J. Shah, and H. T. Chugani. Glucose and [11c]flumazenil positron emission tomography abnormalities of thalamic nuclei in temporal lobe epilepsy. *Neurology*, 53(9):2037–45, 1999.

[311] P. Matheja, T. Kuwert, P. Ludemann, M. Weckesser, C. Kellinghaus, G. Schuierer, B. Diehl, E. B. Ringelstein, and O. Schober. Temporal hypometabolism at the onset of cryptogenic temporal lobe epilepsy. *Eur J Nucl Med*, 28(5):625–32, 2001.

[312] S. F. Barrington, M. Koutroumanidis, A. Agathonikou, P. K. Marsden, C. D. Binnie, C. E. Polkey, M. N. Maisey, and C. P. Panayiotopoulos. Clinical value of ”ictal” fdg-positron emission tomography and the routine use of simultaneous scalp eeg studies in patients with intractable partial epilepsies. *Epilepsia*, 39(7):753–66, 1998.

[313] E. L. So, T. J. O’Brien, B. H. Brinkmann, and B. P. Mullan. The eeg evaluation of single photon emission computed tomography abnormalities in epilepsy. *J Clin Neurophysiol*, 17(1):10–28, 2000.

[314] S. Arnold, G. Schlaug, H. Niemann, A. Ebner, H. Luders, O. W. Witte, and R. J. Seitz. Topography of interictal glucose hypometabolism in unilateral mesiotemporal epilepsy. *Neurology*, 46(5):1422–30, 1996.

[315] V. Bouilleret, S. Dupont, L. Spelle, M. Baulac, Y. Samson, and F. Semah. Insular cortex involvement in mesiotemporal lobe epilepsy: a positron emission tomography study. *Ann Neurol*, 51(2):202–8, 2002.

[316] D. J. Dlugos, J. Jaggi, W. M. O’Connor, X. S. Ding, M. Reivich, M. J. O’Connor, and M. R. Sperling. Hippocampal cell density and subcortical metabolism in temporal lobe epilepsy. *Epilepsia*, 40(4):408–13, 1999.

[317] T. R. Henry, J. C. Mazziotta, and Jr. Engel, J. Interictal metabolic anatomy of mesial temporal lobe epilepsy. *Arch Neurol*, 50(6):582–9, 1993.

[318] T. R. Henry, J. C. Mazziotta, Jr. Engel, J., P. D. Christenson, J. X. Zhang, M. E. Phelps, and D. E. Kuhl. Quantifying interictal metabolic activity in human temporal lobe epilepsy. *J Cereb Blood Flow Metab*, 10(5):748–57, 1990.

[319] E. M. Lee, K. C. Im, J. H. Kim, J. K. Lee, S. H. Hong, Y. J. No, S. A. Lee, J. S. Kim, and J. K. Kang. Relationship between hypometabolic patterns and ictal scalp eeg patterns in patients with unilateral hippocampal sclerosis: An fdg-pet study. *Epilepsy Res*, 84(2-3):187–93, 2009.

[320] N. Nelissen, W. Van Paesschen, K. Baete, K. Van Laere, A. Palmini, H. Van Billoen, and P. Dupont. Correlations of interictal fdg-pet metabolism and ictal spect perfusion changes in human temporal lobe epilepsy with hippocampal sclerosis. *Neuroimage*, 32(2):684–95, 2006.

[321] V. Rusu, F. Chassoux, E. Landre, V. Bouilleret, F. Nataf, B. C. Devaux, B. Turak, and F. Semah. Dystonic posturing in seizures of mesial temporal origin: electroclinical and metabolic patterns. *Neurology*, 65(10):1612–9, 2005.

[322] B. Sadzot, R. M. Debets, P. Maquet, C. W. van Veelen, E. Salmon, W. van Emde Boas, D. N. Velis, A. C. van Huffelen, and G. Franck. Regional brain glucose metabolism in patients with complex partial seizures investigated by intracranial eeg. *Epilepsy Res*, 12(2):121–9, 1992.

[323] M. R. Sperling, R. C. Gur, A. Alavi, R. E. Gur, S. Resnick, M. J. O'Connor, and M. Reivich. Subcortical metabolic alterations in partial epilepsy. *Epilepsia*, 31(2):145–55, 1990.

[324] S. Takaya, T. Hanakawa, K. Hashikawa, A. Ikeda, N. Sawamoto, T. Nagamine, K. Ishizu, and H. Fukuyama. Prefrontal hypofunction in patients with intractable mesial temporal lobe epilepsy. *Neurology*, 67(9):1674–6, 2006.

[325] T. R. Barrick, C. E. Mackay, S. Prima, F. Maes, D. Vandermeulen, T. J. Crow, and N. Roberts. Automatic analysis of cerebral asymmetry: an exploratory study of the relationship between brain torque and planum temporale asymmetry. *Neuroimage*, 24(3):678–91, 2005.

[326] Y. Iturria-Medina, A. Perez Fernandez, D. M. Morris, E. J. Canales-Rodriguez, H. A. Haroon, L. Garcia Penton, M. Augath, L. Galan Garcia, N. Logothetis, G. J. Parker, and L. Melie-Garcia. Brain hemispheric structural efficiency and interconnectivity rightward asymmetry in human and nonhuman primates. *Cereb Cortex*, 21(1):56–67, 2011.

[327] A. Kucyi, M. Moayedi, I. Weissman-Fogel, M. Hodaie, and K. D. Davis. Hemispheric asymmetry in white matter connectivity of the temporoparietal junction with the insula and prefrontal cortex. *PLoS One*, 7(4):e35589, 2012.

[328] A. W. Toga and P. M. Thompson. Mapping brain asymmetry. *Nat Rev Neurosci*, 4(1):37–48, 2003.

[329] J. Zhou, E. D. Gennatas, J. H. Kramer, B. L. Miller, and W. W. Seeley. Predicting regional neurodegeneration from the healthy brain functional connectome. *Neuron*, 73(6):1216–27, 2012.

[330] A. F. Struck, L. T. Hall, J. M. Floberg, S. B. Perlman, and D. A. Dulli. Surgical decision making in temporal lobe epilepsy: a comparison of [(18)f]fdg-pet, mri, and eeg. *Epilepsy Behav*, 22(2):293–7, 2011.

[331] K. Dabbs, T. Becker, J. Jones, P. Rutecki, M. Seidenberg, and B. Hermann. Brain structure and aging in chronic temporal lobe epilepsy. *Epilepsia*, 53(6):1033–43, 2012.

[332] E. Jung da and J. S. Lee. Multimodal neuroimaging in presurgical evaluation of childhood epilepsy. *Korean J Pediatr*, 53(8):779–85, 2010.

[333] R. S. Liu, L. Lemieux, G. S. Bell, S. M. Sisodiya, P. A. Bartlett, S. D. Shorvon, J. W. Sander, and J. S. Duncan. The structural consequences of newly diagnosed seizures. *Ann Neurol*, 52(5):573–80, 2002.

[334] D. C. Reutens, J. M. Stevens, D. Kingsley, B. Kendall, I. Moseley, M. J. Cook, S. Free, D. R. Fish, and S. D. Shorvon. Reliability of visual inspection for detection of volumetric hippocampal asymmetry. *Neuroradiology*, 38(3):221–5, 1996.

[335] M. H. Schmidt and B. Pohlmann-Eden. Neuroimaging in epilepsy: the state of the art. *Epilepsia*, 52 Suppl 4:49–51, 2011.

[336] B. E. Swartz, U. Tomiyasu, A. V. Delgado-Escueta, M. Mandelkern, and A. Khonsari. Neuroimaging in temporal lobe epilepsy: test sensitivity and relationships to pathology and postoperative outcome. *Epilepsia*, 33(4):624–34, 1992.

[337] W. Van Paesschen, J. S. Duncan, J. M. Stevens, and A. Connelly. Longitudinal quantitative hippocampal magnetic resonance imaging study of adults with newly diagnosed partial seizures: one-year follow-up results. *Epilepsia*, 39(6):633–9, 1998.

[338] K. Benedek, C. Juhasz, O. Muzik, D. C. Chugani, and H. T. Chugani. Metabolic changes of subcortical structures in intractable focal epilepsy. *Epilepsia*, 45(9):1100–5, 2004.

[339] D. E. Blum, T. Ehsan, D. Dungan, J. P. Karis, and R. S. Fisher. Bilateral temporal hypometabolism in epilepsy. *Epilepsia*, 39(6):651–9, 1998.

[340] V. Brodbeck, L. Spinelli, A. M. Lascano, C. Pollo, K. Schaller, M. I. Vargas, M. Wissmeyer, C. M. Michel, and M. Seeck. Electrical source imaging for presurgical focus localization in epilepsy patients with normal mri. *Epilepsia*, 51(4):583–91, 2010.

[341] R. P. Carne, T. J. O'Brien, C. J. Kilpatrick, L. R. MacGregor, R. J. Hicks, M. A. Murphy, S. C. Bowden, A. H. Kaye, and M. J. Cook. Mri-negative pet-positive temporal lobe epilepsy: a distinct surgically remediable syndrome. *Brain*, 127(Pt 10):2276–85, 2004.

[342] S. Chinchure, C. Kesavadas, and B. Thomas. Structural and functional neuroimaging in intractable epilepsy. *Neurol India*, 58(3):361–70, 2010.

[343] R. M. Debets, B. Sadzot, J. W. van Isselt, G. J. Brekelmans, L. C. Meiners, A. O. van Huffelen, G. Franck, and C. W. van Veelen. Is 11c-flumazenil pet superior to 18fdg pet and 123i-iomazenil spect in presurgical evaluation of temporal lobe epilepsy? *J Neurol Neurosurg Psychiatry*, 62(2):141–50, 1997.

[344] A. Drzezga, S. Arnold, S. Minoshima, S. Noachtar, J. Szecsi, P. Winkler, W. Romer, K. Tatsch, W. Weber, and P. Bartenstein. 18f-fdg pet studies in patients with extratemporal and temporal epilepsy: evaluation of an observer-independent analysis. *J Nucl Med*, 40(5):737–46, 1999.

[345] W. D. Gaillard, S. Bhatia, S. Y. Bookheimer, S. Fazilat, S. Sato, and W. H. Theodore. Fdg-pet and volumetric mri in the evaluation of patients with partial epilepsy. *Neurology*, 45(1):123–6, 1995.

[346] B. Jupp, J. Williams, D. Binns, R. J. Hicks, L. Cardamone, N. Jones, S. Rees, and T. J. O'Brien. Hypometabolism precedes limbic atrophy and spontaneous recurrent seizures in a rat model of tle. *Epilepsia*, 53(7):1233–44, 2012.

[347] R. C. Knowlton, R. A. Elgavish, A. Bartolucci, B. Ojha, N. Limdi, J. Blount, J. G. Burneo, L. Ver Hoef, L. Paige, E. Faught, P. Kankirawatana, K. Riley, and R. Kuzniecky. Functional imaging: Ii. prediction of epilepsy surgery outcome. *Ann Neurol*, 64(1):35–41, 2008.

[348] R. C. Knowlton, K. D. Laxer, G. Ende, R. A. Hawkins, S. T. Wong, G. B. Matson, H. A. Rowley, G. Fein, and M. W. Weiner. Presurgical multimodality neuroimaging in electroencephalographic lateralized temporal lobe epilepsy. *Ann Neurol*, 42(6):829–37, 1997.

[349] C. J. Liew, Y. M. Lim, R. Bonwetsch, S. Shamim, S. Sato, P. Reeves-Tyer, P. Herscovitch, I. Dustin, A. Bagic, G. Giovacchini, and W. H. Theodore. 18f-fcway and 18f-fdg pet in mri-negative temporal lobe epilepsy. *Epilepsia*, 50(2):234–9, 2009.

[350] P. Ryvlin, L. Cinotti, J. C. Froment, D. Le Bars, P. Landais, M. Chaze, G. Galy, F. Lavenne, J. P. Serra, and F. Mauguiere. Metabolic patterns associated with non-specific magnetic resonance imaging abnormalities in temporal lobe epilepsy. *Brain*, 114 (Pt 6):2363–83, 1991.

[351] W. H. Theodore, D. Katz, C. Kufta, S. Sato, N. Patronas, P. Smothers, and E. Bromfield. Pathology of temporal lobe foci: correlation with ct, mri, and pet. *Neurology*, 40(5):797–803, 1990.

[352] S. G. Uijl, F. S. Leijten, J. B. Arends, J. Parra, A. C. van Huffelen, and K. G. Moons. The added value of [18f]-fluoro-d-deoxyglucose positron emission tomography in screening for temporal lobe epilepsy surgery. *Epilepsia*, 48(11):2121–9, 2007.

[353] C. H. Yun, S. K. Lee, S. Y. Lee, K. K. Kim, S. W. Jeong, and C. K. Chung. Prognostic factors in neocortical epilepsy surgery: multivariate analysis. *Epilepsia*, 47(3):574–9, 2006.

[354] A. Chauvin, K. J. Worsley, P. G. Schyns, M. Arguin, and F. Gosselin. Accurate statistical tests for smooth classification images. *J Vis*, 5(9):659–67, 2005.

[355] K. J. Worsley, A. C. Evans, S. Marrett, and P. Neelin. A three-dimensional statistical analysis for cbf activation studies in human brain. *J Cereb Blood Flow Metab*, 12:900–918, 1992.

[356] T. R. Henry, D. A. Ross, L. A. Schuh, and I. Drury. Indications and outcome of ictal recording with intracerebral and subdural electrodes in refractory complex partial seizures. *J Clin Neurophysiol*, 16(5):426–38, 1999.

[357] S. Shehab, M. Simkins, P. Dean, and P. Redgrave. The dorsal midbrain anticonvulsant zone–i. effects of locally administered excitatory amino acids or bicuculline on maximal electroshock seizures. *Neuroscience*, 65(3):671–9, 1995.

[358] Jr. LaFrance, W. C. and S. R. Benbadis. Differentiating frontal lobe epilepsy from psychogenic nonepileptic seizures. *Neurol Clin*, 29(1):149–62, ix, 2011.

[359] K. M. Sauro, S. Macrodimitris, C. Krassman, S. Wiebe, N. Pillay, P. Federico, W. Murphy, and N. Jette. Quality indicators in an epilepsy monitoring unit. *Epilepsy Behav*, 33:7–11, 2014.

[360] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[361] S. Yu-Sung, A. Gelman, J. Hill, and M. Yajima. Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *J Stat. Softw.*, 45(2):1–31, 2011.

[362] D.B. Rubin. Multiple imputation after 18+ years (with discussion). *JASA*, 91:473–489, 1996.

[363] Morton S.C. Hall, P. On the estimation of entropy. *Ann. Inst. Statist. Math*, 45(1):69–88, 1993.

[364] Grosse E. Shyu-W.M. Cleveland, W.S. *Chapter 8:Local regression models.* Wadsworth & Brooks/Cole, 1992.

[365] K. J. Worsley, J. E. Taylor, F. Tomaiuolo, and J. Lerch. Unified univariate and multivariate random field theory. *Neuroimage*, 23 Suppl 1:S189–95, 2004.

[366] Tibshirani R. Tibshirani, R.J. A bias corection for the minimum error rate in cross-validation. *Ann. Appl. Stat*, 3(2):822–829, 2009.

[367] C.E. Bonferroni. *Il calcolo delle assicurazioni su gruppi di teste.* Studi in Onore del Professore Salvatore Ortu Carboni. Rome: Italy, 1935.

[368] C.E. Bonferroni. Teoria statistica delle classi e calcolo della probabilita. *Pubblicazioni del R Instituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[369] S. Kloppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, Jr. Jack, C. R., J. Ashburner, and R. S. Frackowiak. Automatic classification of mr scans in alzheimer's disease. *Brain*, 131(Pt 3):681–9, 2008.

[370] P. Agarwal, M. M. Mehndiratta, A. R. Antony, N. Kumar, R. N. Dwivedi, P. Sharma, and S. Kumar. Epilepsy in india: nuptiality behaviour and fertility. *Seizure*, 15(6):409–15, 2006.

[371] D. M. An, X. T. Wu, B. Yan, J. Mu, and D. Zhou. Clinical features of psychogenic nonepileptic seizures: a study of 64 cases in southwest china. *Epilepsy Behav*, 17(3):408–11, 2010.

[372] W. Silva, B. Giagante, R. Saizar, L. D'Alessio, S. Oddo, D. Consalvo, P. Saidon, and S. Kochen. Clinical features and prognosis of nonepileptic seizures in a developing country. *Epilepsia*, 42(3):398–401, 2001.

[373] L. C. Frey. Epidemiology of posttraumatic epilepsy: a critical review. *Epilepsia*, 44 Suppl 10:11–7, 2003.

[374] A. Pakalnis and J. Paolicchi. Psychogenic seizures after head injury in children. *J Child Neurol*, 15(2):78–80, 2000.

[375] L. E. Westbrook, O. Devinsky, and R. Geocadin. Nonepileptic seizures after head injury. *Epilepsia*, 39(9):978–82, 1998.